Asking the Right Questions: Increasing Fairness and Accuracy of Personality Assessments with Computerised Adaptive Testing

Yin Lin

Supervisor: Dr Anna Brown

School of Psychology, University of Kent

A thesis submitted for the Degree of

Doctor of Philosophy

July 2020

Word Count: 48,504

ABSTRACT

Personality assessments are frequently used in real-life applications to predict important outcomes. For such assessments, the forced choice (FC) response format has been shown to reduce response biases and distortions, and computerised adaptive testing (CAT) has been shown to improve measurement efficiency. This research developed FC CAT methodologies under the framework of the Thurstonian item response theory (TIRT) model. It is structured into a logical sequence of three areas of investigation, where the findings from each area inform key decisions in the next one. First, the feasibility of FC CAT is tested empirically. Analysis of large historical samples provides support for item parameter invariance when an item appears in different FC blocks, with person score estimation remaining very stable despite minor violations. Remedies for minimising the risk of assumption violations are also developed. Second, the design of the FC CAT algorithm is optimised. Current CAT methodologies are reviewed and adapted for TIRT-based FC assessments, and intensive simulation studies condense the design options to a small number of practical recommendations. Third, the practicality and usefulness of FC CAT is examined. An adaptive FC assessment measuring the HEXACO model of personality is developed and trialled empirically. In conclusion, this research mapped out a blueprint for developing FC CAT that use the TIRT model, highlighting the benefits, limitations, and key directions for further research.

Keywords: Forced choice, computerised adaptive testing, multidimensional item response theory, Thurstonian IRT model.

PUBLICATIONS

Journal Publication

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement, 77*(3), 389-414. doi:10.1177/0013164416646162

Conference Presentation

Lin, Y., & Brown, A. (2015). Item parameters in forced-choice personality assessments: Does context in which items appear matter? Paper presented at the *80th Annual Meeting of the Psychometric Society,* Beijing, China.

Lin, Y., & Brown, A. (2015). Response formats and trait estimation efficiency in computerized adaptive testing. Paper presented at the *5th Conference of the International Association for Computerized Adaptive Testing*, Cambridge, UK.

Lin, Y., Brown, A., & Affourtit, M. (2017). Comparing scoring methods for IRT-based multidimensional forced choice assessments. Paper presented at the *82nd Annual Meeting of the Psychometric Society,* Zurich, Switzerland.

ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council and SHL (ESRC CASE studentship, grant reference ES/J500148/1).

First and foremost, I would like to thank my supervisor, Dr Anna Brown, for her guidance and encouragement throughout this journey. Without her inspiration, I wouldn't have imagined starting this long-term project. As a researcher, I have learned so much from Anna – from research methods, to academic writing, to navigating publications and academia. As a person, I'm very grateful to have Anna as a mentor and role model, helping me balance work and studies, and believing in me even when I doubt myself. It is thanks to Anna that I managed to get through the highs and lows of the last few years and get to where I am today. I have no doubt that her influence will continue to shape me going into the future, whether as a researcher or as a person.

I would also like to thank my industrial supervisors, Dr Tracy Kantrowitz and Dr Sara Gutierrez at SHL, for providing research feedback and organisational support to make this thesis possible. Their guidance had been instrumental in accessing adequate data sources for Studies 1, 4 and 6, and in making the research relevant to field applications as well as academic research.

My thanks also go to Paul Williams for designing and building the data collection website for Study 6. As a web developer and software engineer, he provided invaluable insights into user interface, user experience and network performance, helping to ensure a standardised experience across randomised conditions to enable fair comparisons during analysis.

I must also thank all my colleagues at the University of Kent and at SHL for expanding the horizons of my knowledge and skills through countless formal or

informal discussions; thank the editors and anonymous reviewers from the journals we've submitted to for providing constructive feedback to this research; and thank all the anonymous participants who generously provided their response data to help the advancement of scientific knowledge.

Last but not least, I would like to thank my family and friends for their unwavering patience and support. The last few years have transformed me for the better, and I'm forever indebted to the people in my life for making it happen.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: PERSONALITY AND PERSONALITY ASSESSMENT

As an integral part of an individual's psychology, personality permeates many aspects of one's life. Ozer & Benet-Martinez (2006) summarised research that related personality to many important outcomes at individual, interpersonal and institutional levels – from happiness and physical wellbeing to relationship quality and political attitudes. Moreover, for many outcomes, the influence of personality was on par with or even greater than the influence of social economic status or cognitive ability (Almlund, Duckworth, Heckman, & Kautz, 2011; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). The breadth and depth of the impact of personality thus makes it a crucial tool in understanding and predicting individual life choices and outcomes across many fields of psychological research and practice. In educational psychology, personality has been shown to shape subject choices and academic performance (e.g., Mendolia & Walker, 2014; Trapmann, Hell, Hirn, & Schuler, 2007; O'Connor & Paunonen, 2007; Poropat, 2009). In health psychology, personality has been shown to correlate with physical and mental wellbeing (e.g., Caspi, Roberts, & Shiner, 2005; Miller, Smith, Turner, Guijarro, & Hallet, 1996; Trull & Sher, 1994). In work psychology, personality has been shown to predict job performance and occupational outcomes across many roles and industries, making it a useful tool for employee recruitment, development and appraisal (e.g., Barrick & Mount, 1991; Hurtz & Donovan, 2000; Ones, Dilchert, Viswesvaran, & Judge, 2007; Salgado, 1997, 2002, 2003; Tett, Jackson, & Rothstein, 1991). Regardless of the field of application, personality research and practice involve the quantification of individuals' latent personality traits, which are typically measured through the administration of personality assessments.

This chapter summarises the current status of personality assessment practices. The limitations of traditional personality assessments using fixed questionnaires and

rating scales are described, followed by the review of two increasingly-popular measurement techniques – forced choice (FC) and computerised adaptive testing (CAT). The chapter concludes with the potential benefits arising from combining these two measurement techniques using the Thurstonian Item Response Theory (TIRT) model (Brown & Maydeu-Olivares, 2011), finishing with an outline of the research questions that this thesis addresses.

Traditional Personality Assessments

As a result of their utility, personality assessments (also referred to as "inventories", "instruments" or "tests") are popular in many real-life applications. For example, in a survey of 3,135 human resources professionals from around the world, 60% of the respondents' organisations were using personality assessments pre-hire, with a further 15% planning to do so in the near future (Kantrowitz, Tuzinski, & Raines, 2018).

Due to the lack of better alternatives in many practical settings, the measurement of personality is almost always conducted using a self-report questionnaire. Traditional personality questionnaires typically share two features: first, single-stimulus (SS) response formats are adopted, asking respondents to describe themselves in relation to a series of items, one at a time, using rating scales (usually ordered categories); second, multiple traits representing different factors and facets of personality are assessed, each measured by a small, static set of items. For example, the main NEO[1] Personality Inventories, including NEO-PI-R and NEO-PI-3, assess the Five-Factor Model of personality (FFM; Digman, 1990) using 240 items administered in a SS response format with a five-point rating scale from "strongly disagree" to "strongly agree" (Costa &

[1] "NEO" was originally an acronym for the "Neuroticism-Extraversion-Openness" inventory. However, with the subsequent additions of the Agreeableness (A) and Conscientiousness (C) factors, "NEO" is now merely the instrument brand name and no longer an acronym.

McCrae, 1992; McCrae, Costa, & Martin, 2005). As another example, the main inventory for the six-factor HEXACO[2] model of personality (Ashton et al., 2004), the HEXACO-PI-R, has 60-, 100- and 200-item versions all administered in a SS response format with a reversed five-point rating scale from "strongly agree" to "strongly disagree" (Ashton & Lee, 2009; Lee & Ashton, 2004, 2006, 2018). However, despite the prevalence of this traditional format in numerous historical and current personality assessments, the practical challenges they face have been widely acknowledged and well documented.

With regards to the response format, the SS response format is susceptible to various response biases and distortions. Response biases and distortions arise due to: 1) differences in interpretation of the rating scale (Friedman & Amoo, 1999); 2) individual response styles such as central/extreme tendency, acquiescence and socially desirable responding (Paulhus, 1991; Paulhus & Vazire, 2007); and 3) intentional manipulations of responses to manage impression, also known as faking (e.g., Donovan, Dwight, & Hurtz, 2003; Griffith, Chmielowski, & Yoshita, 2007; Viswesvaran & Ones, 1999). Unintentional response styles are inherent in the measurement methodology and affect all individuals, albeit to varying degrees (e.g., van Herk, Poortinga, & Verhallen, 2004). Intentional response distortions are some individuals' attempt at influencing assessment scores through response manipulation, which is easily achievable with a SS response format (e.g., Martin, Bowen, & Hunt, 2002). Although not all individuals would engage in faking, it is especially prevalent in high-stakes settings, such as pre-employment assessments (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Donovan et al., 2003; Donovan, Dwight, & Schneider, 2014; Griffith et al., 2007; Landers, Sackett, &

[2] The "HEXACO" model of personality comprises of six dimensions, namely Honesty-Humility (H), Emotionality (E), eXtraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). This model is described further in Chapter 4.

Tuzinski, 2011; O'Connell, Kung, & Tristan, 2011; Rosse, Stecher, Miller, & Levin, 1998) and college entrance exams (e.g., Griffin & Wilson, 2012; Lönnqvist, 2014; Yusoff, 2013). Whether unintentional or intentional, response biases introduce systematic nuisance variances that not only impair measurement equivalence between individuals and groups, but also weaken construct and criterion-related validity of personality instruments (e.g., Christiansen, Goffin, Johnston, & Rothstein, 1994; Donovan et al., 2014; Ellingson, Sackett, & Hough, 1999; Hirsh & Peterson, 2008; Mueller-Hanson, Heggestad, & Thornton, 2003; Paulhus, Bruce, & Trapnell, 1995; Peterson, Griffith, O'Connell, & Isaacson, 2008; Rosse et al., 1998; Schmit & Ryan, 1993; Topping & O'Gorman, 1997; van Herk et al., 2004).

With regards to the complex multi-faceted nature of personality, personality inventories often require many items for comprehensive measurement, leading to long assessment times (Kantrowitz, Grelle, & Lin, 2019). More specifically, psychological models of personality often have a small number of broad factors that further subdivide into narrower facets. For example, the FFM is further divided into 30 facets (Costa & McCrae, 1992; McCrae et al., 2005), while the HEXACO model is further divided into 24 facets plus an interstitial scale (Ashton & Lee, 2007; Lee & Ashton, 2008). As a result of this multidimensional structure, reliable measurement of all facets naturally leads to long assessments with many items and increased risk of test fatigue (Kantrowitz et al., 2019). Ackerman and Kanfer (2009) classified the cognitive fatigue arising from prolonged testing into two types: 1) objective cognitive fatigue, i.e., a decrease in actual cognitive functioning, and; 2) subjective cognitive fatigue, i.e., a shift in motivational and attitudinal standings. With cognitive fatigue being a relatively under-researched topic in psychology (Matthews, 2011), there has been very limited empirical research on cognitive fatigue in personality assessments specifically. Even in the related field of cognitive ability testing, findings remain inconsistent with regards to whether prolonged

assessment time lead to reduced assessment scores (Ackerman and Kanfer, 2006). However, Ackerman and Kanfer (2009) highlighted some individual differences in subjective cognitive fatigue during cognitive ability tests that also have relevance to personality assessments. In an empirical study, they found that subjective cognitive fatigue arising from prolonged testing was more severe for individuals with higher levels of neuroticism and anxiety. Moreover, subjective cognitive fatigue was related to reduced effort, which then lead to a significant reduction in test performance. Although Ackerman and Kanfer's (2009) study focused on ability tests only, and there appears to be no literature investigating whether long personality assessments induce subjective cognitive fatigue akin to those induced by long cognitive ability tests, there is no shortage of test takers expressing aversion to long tests. It follows that individuals who are high in neuroticism may be more adversely affected by a long personality assessment: they would feel fatigue earlier than others, which leads to less concentrated efforts in responding, which leads to greater measurement error. In addition to the considerations around measurement accuracy and candidate experience, longer assessments also bear economic consequences – even a small increase in time requirement per candidate multiply into many hours of human costs for test takers and administrators in large-scale assessment programs.

For decades, researchers and practitioners have striven to make comprehensive and reliable personality assessments that are bias-free, fake-resistant, and time-efficient. Various measurement techniques have been applied to combat the shortcomings of the traditional assessment format, two of which are the focus of this thesis. On one hand, in order to address response biases associated with the SS response format, researchers have turned their attention towards an alternative, forced-choice (FC) response format, where respondents indicate their preference among several items at a time by ranking them instead of rating each individually. On the other hand, in order to minimise

assessment time, there is increasing interest in computerised adaptive testing (CAT), where the questions are tailored to each individual to maximise measurement efficiency.

The Forced-Choice (FC) Response Format

Forcing choice between personality items has emerged as an approach to prevent biases and distortions (Nederhof, 1985; Zavala, 1965). Questionnaires using the FC format place items into blocks and ask respondents to rank the items within the block according to the extent they describe their personality. An example of a FC block is given in Table 1. Each FC block contains two or more items. Blocks with two, three, and four items are referred to as pairs/dyads, triplets/triads, and quads/tetrads respectively. Each item in the block is an indicator for an underlying trait of interest. When items within the same block are indicators for different traits, the format is said to be multidimensional forced choice (MFC).

Table 1. Example of a FC block with three items

Please select one statement that is most true or typical of you, and another statement that is least true of you:	Most	Least
I am lively in conversation		
I persevere with tasks		
I avoid taking criticism personally		

For decades, assessments using the FC format faced issues with ipsative scores (Cornwell & Dunlap, 1994; Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988). An assessment's scores are "ipsative" or "purely ipsative" if their total is a constant for all response sets, or "quasi-ipsative" or "partially ipsative" if the total score is not a constant but there are still trade-offs between scores across different traits (Hicks, 1970; Meade, 2004). FC assessments often give rise to ipsative scores if classical test theory (CTT) scoring is applied, i.e., each FC question is given a fixed number of total points,

which are distributed to different scales based on the comparative responses. Ipsativity leads to unnatural constraints in scale variance-covariance matrices (Clemans, 1966), thus distorting the scales' factor structures and reliabilities (Meade, 2004), as well as compromising the scores' interpersonal comparability (Johnson et al., 1988). Ipsativity is therefore a significant measurement issue. However, with the development of Item Response Theory (IRT) modelling of FC responses, scores from FC assessments are no longer ipsative (Brown, 2016; Brown & Maydeu-Olivares, 2011, 2013; Chernyshenko et al., 2009; Stark, Chernyshenko, & Drasgow, 2005).

Research on the FC format has demonstrated that it removes all uniform response biases including central/extreme tendency and acquiescence (Cheung & Chan, 2002), provides greater resistance to motivated distortions (e.g., Cao & Drasgow, 2019; Christiansen, Burns, & Montgomery, 2005; Hirsh & Peterson, 2008; Jackson, Wroblewski, & Ashton, 2000; Lee, Joo, & Lee, 2019; Martin et al., 2002; O'Neill et al., 2016; Pavlov, Maydeu-Olivares, & Fairchild, 2019; Usami, Sakamoto, Naito, & Abe, 2016), and increases differentiations between the constructs measured (e.g., Brown, Inceoglu & Lin, 2017). The practical benefits of FC thus made it an attractive option for improving assessment fairness and accuracy when biases and distortions are of concern, for example in cross-cultural studies affected by culturally-specific response styles (van de Vijver & Leung, 1997; van Herk et al., 2004), and in high-stakes assessments affected by faking (e.g., Viswesvaran & Ones, 1999). For a full discussion, see Brown and Maydeu-Olivares (2011, 2013) for an in-depth summary of the advantages of the FC response format over the SS response format.

Many operational personality assessments already adopt the FC response format, e.g., the Edwards Personal Preference Schedule (EPPS; Edwards, 1973), the Gordon Personal Profile Inventory (Gordon, 1993), the Myers-Briggs Type Indicator (MBTI;

Myers, McCaulley, Quenk, & Hammer, 1998), the Employee Screening Questionnaire (ESQ; Jackson, 2002), and the Occupational Personality Questionnaire (OPQ; Bartram, Brown, Fleck, Inceoglu, & Ward, 2006). Salgado and Táuriz (2014) conducted a very thorough meta-analysis of criterion-related validity of FC personality assessments in occupational and educational applications. Collating findings from 122 independent samples reported up until September 2011, they found FC measures to produce similar or even higher criterion-related validity coefficients than those reported in previous meta-analyses covering mostly SS personality inventories (e.g., Barrick & Mount, 1991; Salgado, 1997). Salgado and Táuriz thus concluded that "FC inventories can be a good alternative to SS questionnaires for making academic and personnel decisions."

It is worth noting that all the FC instruments included in Salgado and Táuriz's (2014) meta-analysis were classically scored, and thus open for further measurement optimisation using an appropriate IRT model (for options, see Brown, 2016). For example, Brown and Bartram (2009) refined a classically-scored FC personality assessment using IRT methodologies, successfully reducing assessment time by 40-50% while maintaining similar levels of score reliability. The application of modern IRT methodologies to FC thus help to minimise assessment time through more efficient extraction of information from comparative data. Furthermore, the availability of IRT models in conjunction with computer-based testing technology opens up the possibility of shortening assessments even further through computerised adaptive testing (CAT).

Computerised Adaptive Testing (CAT)

Computerised adaptive testing (CAT) tailors an assessment to each and every individual in real time – the most informative questions for a candidate are presented, based on existing intelligence about them (e.g., their response to previous questions in the assessment, their results from previous assessment sessions). In order to conduct

CAT, an IRT model is needed to conceptualise, model and quantify information collection (see Chapter 2 for details), and a computer algorithm is needed to drive the assessment assembly (see Chapter 3 for details).

CAT has demonstrated great utility in the field of cognitive ability testing, with studies showing 50% reduction in test lengths compared to static paper-and-pencil versions (Embretson & Reise, 2000, p. 268). Comparatively, adaptive personality assessments have been somewhat sparse. Nevertheless, existing findings from a series of real-data simulation studies for SS personality assessments replicated similar levels of adaptive efficiency gains as those reported for cognitive ability tests. For example, Waller and Reise (1989) demonstrated that IRT-scored adaptive personality scales could be 50%-75% shorter than their classically-scored paper-and-pencil counterparts, although they did not quantify what proportion of that reduction was attributable to IRT and CAT respectively. Waller (1999) further demonstrated that CAT reduced the number of items needed for the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & Mckinley, 1940). Similarly, Hol, Vorst and Mellenbergh (2008) studied the adjective checklist (ACL; Gough & Heilbrun, 1980), and found that an adaptive version only required as few as 33% of the original items to reach similar levels of measurement accuracy as the IRT-scored full length test. Independently, two studies (Makransky, Mortensen, & Glas, 2013; Reise & Henson, 2000) involving the NEO-PI-R (Costa & McCrea, 1992) found that CAT achieved similar measurement accuracy with merely 50% of the original test length. Furthermore, Nieto et al. (2017) developed a new item pool for the FFM and found that CAT administrations required as few as 4 items per facet. These comparable findings reported by multiple researchers across different item pools and independent samples clearly demonstrated the power of CAT in SS personality assessments. It is therefore not surprising to see adaptive SS personality assessments in real-life applications, for example, the computerised adaptive assessment

of personality disorders (CAT-PD; Simms et al., 2011), and the adaptive schizotypal personality questionnaire (Moore, Calkins, Reise, Gur, & Gur, 2018).

Following the development of IRT models for FC responses, applications of adaptive FC personality assessments have also gained popularity, but mainly in the field of occupational psychology. Houston, Borman, Farmer, and Bearden (2006) developed the Navy Computer Adaptive Personality Scales (NCAPS), a successful proof-of-concept for an operational FC CAT of personality traits for US military. The Global Personality Inventory – Adaptive (GPI-A; SHL, 2009-2014), a measure for general workplace personality traits, was then developed following the same methodological setup as NCAPS. Both NCAPS and GPI-A use unidimensional FC blocks, and they employ unidimensional IRT models for item selection and scoring of each personality trait. In other words, they consist of instances of independent, unidimensional FC CATs presented in parallel. Such a unidimensional setup was methodologically simpler, but removed the possibility of further measurement efficiency gains from correlated traits within a multidimensional CAT[3], nor did it take advantage of the potential increase in resistance to socially desirable responding and faking that the multidimensional FC format can provide. Multidimensional FC CATs (i.e., administering multidimensional FC questions and employing multidimensional IRT models for item selection and scoring) only emerged in the last decade, including the Tailored Adaptive Personality Assessment System (TAPAS; Drasgow et al., 2012) and the Adaptive Employee Personality Test (ADEPT-15; Boyce, Conway, & Caputo, 2014). A series of studies

[3] Segall (1996) demonstrated that "multidimensional adaptive testing can provide equal or higher reliabilities with about one-third fewer items than are required by one-dimensional adaptive testing". Wang and Chen (2004) further demonstrated via simulations that the comparative advantage of multidimensional CAT over unidimensional CAT was greater with higher trait correlations and larger trait counts – both features are typical for personality assessments.

comparing adaptive and static FC assessments confirmed similar levels of measurement length reductions as those seen when comparing adaptive and static SS assessments, typically reaching the same level of true score correlation at about half the test length (Stark & Chernyshenko, 2007, 2011; Stark, Chernyshenko, Drasgow, & White, 2012). FC CATs therefore appear to improve measurement efficiency while also ensure good resistance to biases and distortions, effectively combating the two practical challenges of traditional personality assessments.

Research Questions

Despite recent advancements in FC CAT research, it remains a relatively new and under-explored topic. On one hand, there is next to no empirical evidence on the influence of context on item functioning within FC blocks. Ortner (2008) showed in an empirical study that item order within a SS personality assessment could have a significant impact on measurement, thereby raising caution on the standard assumption in CAT that an item's properties stay the same regardless of the items surrounding it. As the influence of the context around an item is a concern even for SS response formats with no explicit item interactions, it ought to be even more important for FC response formats where the responding process requires items to be directly compared. Yet it seems illogical that this fundamental assumption of context-invariance of item properties, one that can call the feasibility of FC CAT into question, has never[4] been tested empirically. In order to address this concern, Chapter 2 reports an empirical investigation into the robustness of this fundamental assumption.

[4] A new study (Morillo et al., 2019) on this topic has since been published following the publication of Study 1 of this thesis (Lin & Brown, 2017).

On the other hand, there is also very limited knowledge of the functioning of FC CAT with dominance items. A dominance item is characterised by a monotonic relationship between the probability of endorsement of the item and the underlying personality trait[5] it indicates. In other words, as the personality trait value increases, the probability of agreeing with the item monotonically increases if the item is positively keyed, or monotonically decreases if the item is negatively keyed. For example, "I am organised" is a dominance item for Conscientiousness. Existing personality measures tend to employ dominance items by default. However, most published research on FC CAT, as well as all four operational FC CATs (i.e., NCAPS, GPI-A, TAPAS and ADEPT), adopt ideal-point items (Coombs, 1964). An ideal-point item is characterised by a curvilinear relationship between the probability of endorsement of the item and the underlying personality trait it indicates. In other words, there is a particular trait value at which point the probability of agreeing with the item peaks, and deviations from this ideal point in either direction on the personality trait lowers the probability of endorsement. For example, "I am sometimes organised and sometimes forgetful" is an ideal-point item for Conscientiousness.

Dominance and ideal point items exhibit different item characteristics, have different response processes, and demand different IRT models (Brown, 2015). It follows that the techniques for and the findings from ideal-point FC CATs cannot be generalised to dominance FC CATs. While one very recent study (Chen, Wang, Chiu, & Ro, 2019) did explore FC CAT with dominance items, it adopted the Rasch ipsative model that produces scores "with the constraint of zero sum across dimensions for every person" (Wang, Qiu, Chen, Ro, & Jin, 2017), thus focusing on within-person profiling

[5] It should be noted that, although it is theoretically possible for an item to indicate multiple traits, such a setup tends to be impractical for personality measurement. This thesis therefore focuses on the situations where each item indicates one and only one trait.

rather than cross-person comparisons of assessment results. In order to bridge this gap, Chapter 3 formulates and optimises algorithm components for FC CATs using the dominance Thurstonian IRT model (TIRT; Brown & Maydeu-Olivares, 2011, 2013). Since most existing personality items are dominance items, advancing research on dominance FC CAT methodologies enables the utilisation of validated historical content in the creation of new FC CATs, as opposed to needing to develop and validate new ideal-point items from scratch. Then, Chapter 4 tests the methodology empirically through the development of a dominance FC CAT for personality measurement.

This thesis mapped out a rough blueprint for the development of dominance FC CATs. However, constrained by its scope, there are still many open questions requiring further research. Chapter 5 summarises the findings of this thesis, considers its implications for research and practice, and outlines important areas for further investigation.

CHAPTER 2: FOUNDATIONS FOR FC CAT

The most natural way of formulating a FC CAT is through the utilisation of an item response theory (IRT) model. An IRT model serves two purposes in a FC CAT. First, it enables the establishment of interpersonally comparable person scores from relative-to-self (or ipsative) responses resulting from the FC format. Second, it enables adaptive assessment tailoring through parameterisation of the psychometric properties of items.

A number of IRT models have been developed for the FC response format, e.g., the probabilistic, multidimensional unfolding model (Zinnes & Griggs, 1974), the hyperbolic cosine unfolding model for pairwise preferences (Andrich, 1995), the multi-unidimensional pairwise preference (MUPP) model (Stark, 2002; Stark, Chernyshenko, & Drasgow, 2005), and the Thurstonian IRT (TIRT) model (Brown & Maydeu-Olivares, 2011, 2013). Brown (2016) discussed the similarities and differences between such models and how they can be organised in a unified framework. For this thesis, the TIRT model is chosen. The TIRT model is able to handle multidimensionality, is flexible when modelling FC blocks of any size, and is compatible with the most commonly used dominance items. Moreover, the TIRT model has demonstrated great usability and utility in empirical applications (e.g., Brown & Bartram, 2009, 2009-2011; Brown, Inceoglu & Lin, 2017).

This chapter is structured as follows. First, the mathematical formulation of TIRT is described in detail. Then, the essential assumption of item parameter invariance (regardless of the place in a test where that item appears) for FC CAT is discussed, followed by an empirical study examining this assumption (Study 1). Finally, the chapter concludes with a summary of findings and implications for further research.

The Thurstonian Item Response Theory (TIRT) Model

Response Modelling

In TIRT, the full or partial ranking response to a FC block of size n is decomposed into $n(n-1)/2$ pairwise comparisons, as shown in Table 2 (Brown & Maydeu-Olivares, 2012).

Table 2. Decomposing FC blocks into pairwise comparisons

Block size (n)	Items/stimuli	Binary Outcomes
2 ("pairs")	i, k	$\{i, k\}$
3 ("triplets")	i, k, l	$\{i, k\}$, $\{i, l\}$, $\{k, l\}$
4 ("quads")	i, k, l, o	$\{i, k\}$, $\{i, l\}$, $\{i, o\}$, $\{k, l\}$, $\{k, o\}$, $\{l, o\}$

Then, from the ranking response to the FC block, the binary outcome for any constituting pairwise comparison $\{i, k\}$ can be deduced. The binary outcome variable $Y_{\{i,k\}}$ is coded as described in Equation 1 (Maydeu-Olivares & Böckenholt, 2005).

$$Y_{\{i,k\}} \equiv \begin{cases} 1 & if\ item\ i\ is\ preferred\ over\ item\ k \\ 0 & if\ item\ k\ is\ preferred\ over\ item\ i \\ missing & if\ the\ outcome\ of\ comparison\ is\ unknown \end{cases} \quad (1)$$

TIRT models the responding process behind this binary outcome by Thurstone's Law of Comparative Judgement (Thurstone, 1927), which states that the two items' psychological utility values within a respondent (denoted t_i and t_k, the person index is omitted in the notations) determine the outcome of the comparative judgement (Equation 2, Brown & Maydeu-Olivares, 2011).

$$Y_{\{i,k\}} = \begin{cases} 1 & if\ t_i \geq t_k \\ 0 & if\ t_i < t_k \end{cases} \quad (2)$$

A respondent's psychological utility value for an item is modelled as a function of their psychological profile and the characteristics of the item (Equation 3, Brown & Maydeu-Olivares, 2011). The respondent's psychological profile is modelled as a latent trait column vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_S)^T$ with S dimensions. The characteristics of item i are modelled through TIRT item parameters: μ_i is the mean utility of the item; $\boldsymbol{\lambda_i} = \left(\lambda_{i_1}, \dots, \lambda_{i_S}\right)^T$ is a column vector of S factor loadings; ε_i is a normally distributed error term with mean 0 and unique variance ψ_i^2.

$$t_i = \mu_i + \boldsymbol{\lambda_i}^T \boldsymbol{\eta} + \varepsilon_i \tag{3}$$

Based on this set-up, the response probabilities of the binary outcome $Y_{\{i,k\}}$ can be deduced, giving the Item Response Function (IRF) of the TIRT model (Equation 4, Brown & Maydeu-Olivares, 2011). In this expression, $\gamma_{\{i,k\}} \equiv \mu_k - \mu_i$ is the threshold parameter for the pairwise comparison, and Φ represents the standard normal cumulative distribution function.

$$p_{\{i,k\}}(\boldsymbol{\eta}) \equiv P\left(Y_{\{i,k\}} = 1 | \boldsymbol{\eta}\right) = \Phi\left(\frac{-\gamma_{\{i,k\}} + (\boldsymbol{\lambda_i} - \boldsymbol{\lambda_k})^T \boldsymbol{\eta}}{\sqrt{\psi_i^2 + \psi_k^2}}\right) \equiv \Phi\left(z_{\{i,k\}}\right) \tag{4}$$

Most practical FC assessments fall into a special case where items are factorially simple, i.e., each item indicates one and only one latent trait. In other words, for each item i, the factor loading vector $\boldsymbol{\lambda_i}$ contains one and only one non-zero entry $\lambda_{i_{s_i}}$ corresponding to the latent trait η_{s_i} indicated by the item. In such cases, Equation 4 simplifies to Equation 5 (Brown & Maydeu-Olivares, 2011).

$$p_{\{i,k\}}(\boldsymbol{\eta}) = \Phi\left(\frac{-\gamma_{\{i,k\}} + \lambda_{i_{s_i}} \eta_{s_i} - \lambda_{k_{s_k}} \eta_{s_k}}{\sqrt{\psi_i^2 + \psi_k^2}}\right) \tag{5}$$

The likelihood for an observed binary response $Y_{\{i,k\}}$ or for an entire response string of binary responses $\boldsymbol{Y}$ can then be expressed by Equation 6 and Equation 7 respectively (Brown & Maydeu-Olivares, 2011). Note that Equation 7 assumes that the pairwise comparisons are conditionally independent (i.e., the errors for pairwise comparisons are independent from each other), which is true in the case of tests using only FC blocks with two items. Larger FC blocks with three or more items violate this assumption because multiple pairwise comparisons in such a block will involve the same item, leading to correlated errors even after controlling for the latent traits (Brown & Maydeu-Olivares, 2011). For example, ranking responses to a triplet $\{i, k, l\}$ are decomposed into pairs $\{i, k\}$, $\{i, l\}$ and $\{k, l\}$. Then, pairs $\{i, k\}$ and $\{i, l\}$ have correlated errors due to the common item i; pairs $\{i, k\}$ and $\{k, l\}$ have correlated errors due to the common item k; and pairs $\{i, l\}$ and $\{k, l\}$ have correlated errors due to the common item l. Therefore, for FC blocks with three or more items, Equation 7 is an approximation with the simplifying assumption of local independence across pairwise comparisons within the same FC block.

$$L\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right) = p_{\{i,k\}}(\boldsymbol{\eta})^{Y_{\{i,k\}}}\left(1 - p_{\{i,k\}}(\boldsymbol{\eta})\right)^{1-Y_{\{i,k\}}} \tag{6}$$

$$L(\boldsymbol{Y}|\boldsymbol{\eta}) = \prod_{\{i,k\}} L\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right) \tag{7}$$

It is worth noting that the TIRT model is a variant of the multidimensional 2-parameter normal-ogive (M2PNO) model (Bock & Schilling, 2003; McDonald 1999; Samejima, 1974), which has an IRF as described in Equation 8. Clearly, Equation 8 and Equation 4 are equivalent with assignments as detailed in Equation 9.

$$p_i(\boldsymbol{\theta}) \equiv P(U_i = 1|\boldsymbol{\theta}) = \Phi(\boldsymbol{a_i}^T\boldsymbol{\theta} + d_i) \quad (8)$$

$$\boldsymbol{\theta} = \boldsymbol{\eta}\ ;\ \boldsymbol{a_i} = \frac{\boldsymbol{\lambda_i - \lambda_k}}{\sqrt{\psi_i^2 + \psi_k^2}}\ ;\ d_i = \frac{-\gamma_{\{i,k\}}}{\sqrt{\psi_i^2 + \psi_k^2}} \quad (9)$$

The TIRT model, however, has some special features compared to the M2PNO model, leading to distinctions in assumptions and intended applications. Firstly, the research and applications of the M2PNO model had focused mainly on ability measurement, and hence typically assumed the elements of $\boldsymbol{a_i}$ to always have non-negative values. However, TIRT focuses on measuring preferences, using non-cognitive statements that sometimes result in negative loading values in $\boldsymbol{\lambda_i}$. Furthermore, even if no items are negative indicators of their intended traits, combining them into FC blocks will inevitably result in negative values for some $\boldsymbol{a_i} = \frac{\boldsymbol{\lambda_i - \lambda_k}}{\sqrt{\psi_i^2 + \psi_k^2}}$. The two models thus differ in terms of their assumption regarding the possible signs of the loading/ slope/ discrimination parameters.

Secondly, the M2PNO model assumes local independence between any two item responses. However, for responses collected using a FC format, unless the block size n is 2, there will be multiple pairwise comparisons resulting from each FC block. In order to account for item overlaps between pairs from the same FC block, the TIRT model adopts additional structures and constraints, including: 1) equal factor loadings when an item contributes to multiple pairs within the same block, and 2) correlated error structures between pairs involving the same item. The formulation of correlated error structures is described in more details in Brown and Maydeu-Olivares (2011), and results in separate identification of unique variance parameters (whereas in the M2PNO model, error variances are all fixed to 1 for model identification).

Despite these differences, the similarities between TIRT and M2PNO models meant that much research, methods and practices from the relatively-mature M2PNO model and other related models (e.g., the Multidimensional 2-Parameter Logistic model, McKinley & Reckase, 1983) are relevant and likely extendable to the TIRT model, even though the response formats look very different on the surface level.

Information and Standard Error of Measurement (SEM)

IRT models use information functions to describe the measurement gain provided by each item. Equation 10 shows the item information function (IIF) for a general multidimensional IRT model (Reckase, 2009). In this expression, $\boldsymbol{\alpha}$ is a vector of angles with the coordinate axes, indicating a direction in the multidimensional space; $\nabla_{\boldsymbol{\alpha}}$ is the gradient (i.e., directional derivative) in the direction of $\boldsymbol{\alpha}$ (Equation 11, Reckase, 2009); and $I_i^{\alpha}(\boldsymbol{\theta})$ is the information from item i in direction $\boldsymbol{\alpha}$ for an individual with trait profile $\boldsymbol{\theta}$.

$$I_i^{\alpha}(\boldsymbol{\theta}) = \frac{[\nabla_{\boldsymbol{\alpha}} p_i(\boldsymbol{\theta})]^2}{p_i(\boldsymbol{\theta})\big(1 - p_i(\boldsymbol{\theta})\big)} \tag{10}$$

$$\nabla_{\boldsymbol{\alpha}} p_i(\boldsymbol{\theta}) = \sum_{s=1}^{S} \frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_s} cos\alpha_s \tag{11}$$

The same concept can be applied to the TIRT model, leading to a similar expression (Equation 12) for the information gain from a pairwise comparison for a general direction $\boldsymbol{\alpha}$ in the multidimensional space (Brown & Maydeu-Olivares, 2011). The gradient term for TIRT can be deduced by combining Equation 4 and Equation 11, giving Equation 13, where ϕ represents the standard normal density function.

$$I^{\alpha}_{\{i,k\}}(\boldsymbol{\eta}) = \frac{\left[\nabla_{\boldsymbol{\alpha}} p_{\{i,k\}}(\boldsymbol{\eta})\right]^2}{p_{\{i,k\}}(\boldsymbol{\eta})\left(1 - p_{\{i,k\}}(\boldsymbol{\eta})\right)} \tag{12}$$

$$\nabla_{\boldsymbol{\alpha}} p_{\{i,k\}}(\boldsymbol{\eta}) = \sum_{s=1}^{S} \frac{cos\alpha_s\left(\lambda_{i_s} - \lambda_{k_s}\right)}{\sqrt{\psi_i^2 + \psi_k^2}} \phi\left(z_{\{i,k\}}\right) \tag{13}$$

For the special case where items are factorially simple (i.e., as described in Equation 5), there are only two non-zero factor loadings $\lambda_{i_{s_i}}$ and $\lambda_{k_{s_k}}$, and Equation 13 further reduces to Equation 14.

$$\nabla_{\boldsymbol{\alpha}} p_{\{i,k\}}(\boldsymbol{\eta}) = \left(\frac{cos\alpha_{s_i}\lambda_{i_{s_i}} - cos\alpha_{s_k}\lambda_{k_{s_k}}}{\sqrt{\psi_i^2 + \psi_k^2}}\right) \phi \left(\frac{-\gamma_{\{i,k\}} + \lambda_{i_{s_i}}\eta_{s_i} - \lambda_{k_{s_k}}\eta_{s_k}}{\sqrt{\psi_i^2 + \psi_k^2}}\right) \tag{14}$$

When the direction $\boldsymbol{\alpha}$ in the multidimensional space aligns with the direction of a latent trait axis (denoted $\boldsymbol{\alpha}^s$ for the s^{th} trait), the cosine term is equivalent to the Pearson correlation between latent traits (Bock ,1975), giving rise to Equation 15, where $\boldsymbol{cor}$ denotes the $S \times S$ correlation matrix between latent traits.

$$cos\alpha^s_{s_i} = cor_{s,s_i}\ ;\ cos\alpha^s_{s_k} = cor_{s,s_k} \tag{15}$$

Combining Equations 12, 14 and 15, the information contributions from a pairwise comparison of factorially simple items for measuring the s^{th} trait can be deduced and take the form of Equation 16.

$$I^{\alpha^s}_{\{i,k\}}(\boldsymbol{\eta}) = \frac{\left[\left(\frac{cor_{s,s_i}\lambda_{i_{s_i}} - cor_{s,s_k}\lambda_{k_{s_k}}}{\sqrt{\psi_i^2 + \psi_k^2}}\right) \phi \left(\frac{-\gamma_{\{i,k\}} + \lambda_{i_{s_i}}\eta_{s_i} - \lambda_{k_{s_k}}\eta_{s_k}}{\sqrt{\psi_i^2 + \psi_k^2}}\right)\right]^2}{p_{\{i,k\}}(\boldsymbol{\eta})\left(1 - p_{\{i,k\}}(\boldsymbol{\eta})\right)} \tag{16}$$

When one of the items indicates the s^{th} trait (i.e., $s = s_i$ or $s = s_k$, see Brown & Maydeu-Olivares, 2011 for full formulae), the information gain for the scale is direct and forms the core of measurement. When neither of the items indicates the s^{th} trait (i.e., $s \neq s_i$ and $s \neq s_k$), there is still peripheral information gain for the scale if the latent traits are correlated. It is sometimes helpful to distinguish core information gain (Equation 17) from peripheral information gain, in order to focus on core information as the main basis of measurement.

$$CI_{\{i,k\}}^{\boldsymbol{\alpha}^s}(\boldsymbol{\eta}) = \begin{cases} I_{\{i,k\}}^{\boldsymbol{\alpha}^s}(\boldsymbol{\eta}) & if\ s \in \{s_i, s_k\} \\ 0 & if\ s \notin \{s_i, s_k\} \end{cases} \tag{17}$$

Information at the test level is then calculated as the sum of information from all constituting pairwise comparisons across all FC blocks (Equation 18, Brown & Maydeu-Olivares, 2011). As in the case of total response likelihood (Equation 7), Equation 18 assumes local independence between pairwise comparisons, which is only true in the case of tests using FC pairs. For larger FC blocks, Equation 18 is an approximation with the simplifying assumption of local independence.

$$I^{\boldsymbol{\alpha}}(\boldsymbol{\eta}) = \sum_{\{i,k\}} I_{\{i,k\}}^{\boldsymbol{\alpha}}(\boldsymbol{\eta}) \tag{18}$$

In addition to the information gain from assessment responses, prior information also contributes to measurement when Bayesian trait estimators are adopted. TIRT often assumes a multivariate standard normal distribution for the latent traits. The multivariate normal prior leads to prior information for each trait equalling the "diagonal element of the inverted trait covariance matrix" (Brown & Maydeu-Olivares, 2011). The posterior information in the direction of the s^{th} trait can thus be deduced (Equation 19, Brown & Maydeu-Olivares, 2011).

$$I_{Pos}^{\boldsymbol{\alpha}^s}(\boldsymbol{\eta}) = I^{\boldsymbol{\alpha}^s}(\boldsymbol{\eta}) + (\boldsymbol{cov}^{-1})_{s,s} \tag{19}$$

Following standard IRT methodology, the standard errors of measurement (SEMs) associated with elements of the latent trait estimate vector $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_S)^T$ are then calculated as the inverse of the square root of information values in the directions of the latent trait axes (Equation 20 if the trait estimator is not Bayesian, Equation 21 if the trait estimator is Bayesian).

$$SEM(\hat{\eta}_s) = \frac{1}{\sqrt{I^{\boldsymbol{\alpha}^s}(\hat{\boldsymbol{\eta}})}} \tag{20}$$

$$SEM(\hat{\eta}_s) = \frac{1}{\sqrt{I_{Pos}^{\boldsymbol{\alpha}^s}(\hat{\boldsymbol{\eta}})}} \tag{21}$$

Fisher Information Matrix (FIM)

In addition to the IRT information functions, the Fisher Information Matrix (FIM) is often useful in CAT research. The FIM for TIRT is deduced here. Brown & Maydeu-Olivares (2017; expression B.3) provided the FIM for a graded preference response to pair $\{i, k\}$ where no intransitive preferences are possible. Graded preference is a comparative judgement expressed in C ordered response categories between two items. For example, item i can be preferred "much more" – "a little more" – "a little less" – "much less" to item k (here, the number of ordered categories $C = 4$). The FC format modelled by TIRT is a special case of the graded preference format with $C = 2$ and no intransitivities, thus leading to a FIM as shown in Equation 22. In this expression, the function $p_{\{i,k\}}$ is as defined in Equations 4. The block-diagonal design matrix of contrasts $\boldsymbol{A}$ captures the assignment of items (columns) to blocks (with rows corresponding to pairs within blocks). The matrix $\boldsymbol{\Lambda}$ captures factor loadings of items (rows) on latent traits (columns). The matrix $\boldsymbol{A\Lambda}$ therefore details the factor loadings of

each pair (rows) on each latent trait (columns). The $\{i,k\}$ subscript for the matrix $\boldsymbol{A\Lambda}$ denotes the row in the matrix associated with pair $\{i,k\}$. It follows that $(\boldsymbol{A\Lambda})_{\{i,k\}} = \boldsymbol{\lambda}_i - \boldsymbol{\lambda}_k$, giving Equation 23.

$$\boldsymbol{F}_{\{i,k\}}(\boldsymbol{\eta}) = \frac{(\boldsymbol{A\Lambda})_{\{i,k\}}(\boldsymbol{A\Lambda})_{\{i,k\}}^{T}}{\psi_i^2 + \psi_k^2}\left(\frac{\left[\phi\left(z_{\{i,k\}}\right)\right]^2}{1 - p_{\{i,k\}}} + \frac{\left[\phi\left(z_{\{i,k\}}\right)\right]^2}{p_{\{i,k\}}}\right) \tag{22}$$

$$= \frac{\left[\phi\left(z_{\{i,k\}}\right)\right]^2(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_k)^{\boldsymbol{T}}(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_k)}{p_{\{i,k\}}\left(1 - p_{\{i,k\}}\right)(\psi_i^2 + \psi_k^2)} \tag{23}$$

In the case where items are factorially simple, i.e., $\lambda_{i_{s_i}}$ and $\lambda_{k_{s_k}}$ being the only non-zero entries of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_k$ respectively, the element on the s^{th} row and v^{th} column in the FIM can be simplified further, giving Equation 24 if $s_i = s_k$, or Equation 25 if $s_i \neq s_k$. The FIM of a FC pair thus only has one non-zero entry for unidimensional comparisons, or four non-zero entries for multidimensional comparisons. Similar to the FIM of multidimensional items in regular IRT models, the FIM of a multidimensional pairwise comparison has rank one and is singular (Mulder & van der Linden, 2009).

$$\left[\boldsymbol{F}_{\{i,k\}}(\boldsymbol{\eta})\right]_{sv} = \begin{cases} \dfrac{\left[\phi\left(z_{\{i,k\}}\right)\right]^2\left(\lambda_{i_{s_i}} - \lambda_{k_{s_i}}\right)^2}{p_{\{i,k\}}\left(1 - p_{\{i,k\}}\right)(\psi_i^2 + \psi_k^2)} & if\ s = v = s_i = s_k \\ 0 & otherwise \end{cases} \tag{24}$$

$$\left[\boldsymbol{F}_{\{i,k\}}(\boldsymbol{\eta})\right]_{sv} = \begin{cases} \dfrac{\left[\phi\left(z_{\{i,k\}}\right)\right]^2\left(\lambda_{i_{s_i}}\right)^2}{p_{\{i,k\}}\left(1 - p_{\{i,k\}}\right)(\psi_i^2 + \psi_k^2)} & if\ s = v = s_i \neq s_k \\ \dfrac{\left[\phi\left(z_{\{i,k\}}\right)\right]^2\left(\lambda_{k_{s_k}}\right)^2}{p_{\{i,k\}}\left(1 - p_{\{i,k\}}\right)(\psi_i^2 + \psi_k^2)} & if\ s = v = s_k \neq s_i \\ \dfrac{\left[\phi\left(z_{\{i,k\}}\right)\right]^2\left(-\lambda_{i_{s_i}}\lambda_{k_{s_k}}\right)}{p_{\{i,k\}}\left(1 - p_{\{i,k\}}\right)(\psi_i^2 + \psi_k^2)} & if\ \{s,v\} = \{s_i, s_k\} \\ 0 & otherwise \end{cases} \tag{25}$$

As per standard IRT models, FIM at the test level is then calculated as the sum of FIM from all constituting pairwise comparisons across all FC blocks (Equation 26, Segall, 1996). Similar to the case of total response likelihood (Equation 7) and total IRT information (Equation 18), Equation 26 assumes local independence between pairwise comparisons, which is only true in the case of tests using FC pairs. For larger FC blocks, Equation 26 is an approximation with the simplifying assumption of local independence.

$$\boldsymbol{F}(\boldsymbol{\eta}) = \sum_{\{i,k\}} \boldsymbol{F}_{\{i,k\}}(\boldsymbol{\eta}) \tag{26}$$

Parameter Invariance Foundation for FC CAT

In the most unconstrained form of FC CAT, items are adaptively assembled into FC blocks, and the properties of FC blocks are derived from the properties of the constituting items. This simple process requires an item to function in exactly the same way regardless of what other items appear in the same FC block. In IRT terms, this is equivalent to making the assumption that the item parameters are invariant with respect to context – the items surrounding the target item in the FC block.

However, the way items are combined into and explicitly compared within FC blocks can potentially introduce contextual changes, leading to respondents viewing the items in a different light. The impact of context on item functioning is neither new nor unique to forced choice. For example, Strack, Martin and Schwarz (1988) showed that by simply swapping the order of two satisfaction items, their correlational relationship changed, producing the item-order effect. At the same period, Knowles (1988) demonstrated that the constructs being measured by a personality assessment become clearer to the respondents as they consider more items, leading to more “polarized, consistent, and reliable” responses in items appearing later in the assessment, producing the serial-order effect. More recently, Steinberg (2001) showed that presenting two SS

items on anger experience and anger expression next to each other lead to more extreme responses than when they were presented on their own. Phenomenon as such can lead to change in item properties and thus item parameter shifts.

While item parameter shifts due to change in context are relevant for both linear and CAT assessments, in practice this problem can be fully addressed for linear FC assessments. With a fixed FC form, estimation of item parameters can be done using this particular linear form. In this case, the context (i.e., surrounding items in the same block) remains constant between calibration and application of the assessment. In the more complex case when multiple, parallel linear FC forms with overlapping items are employed, the forms can be calibrated independently and subsequently equated at the form level, without necessarily imposing the parameter invariance assumption on the common items. It is only when the items move blocks from one form to the next, for example in FC CAT or any non-adaptive but dynamic FC assessments, that context differences between calibration and application become inevitable, and thus item parameter invariance becomes a paramount assumption. In other words, there is no guarantee that people will interpret each and every item in a consistent way (leading to invariant item parameters), when other items around it change as in the case of FC CAT.

Empirical studies are needed to examine the effect of context on item functioning. While recent findings have provided some reassurance on the stability of person parameter estimation when FC block compositions vary (Lin, Inceoglu, & Bartram, 2013), and found item parameter estimates from SS and FC response data to be reasonably comparable (Morillo et al., 2019), examination of the item parameter stability assumption across different FC blocks had been largely ignored by most researchers. As an important pre-requisite assumption of FC CAT, the research question is whether varying contexts have negligible impact on people's FC responding

behaviours and thus on the subsequently deduced item parameters. More specifically, research should quantify the level of item parameter stability when the context around one item is altered due to the presence of other items.

Empirical Examination of Parameter Invariance Assumption (Study 1)

This study explored the effect of context on item functioning in FC blocks, by examining empirically estimated item parameters across two instruments. The first instrument was compiled of FC blocks of three items, whereas in the second, the context was manipulated by adding one item to each block, resulting in FC blocks of four. The robustness of the parameter invariance assumption required for CAT was examined, and situations where this assumption was violated were identified. Practical strategies to avoid such violations were suggested to inform future FC CAT designs.

Method

Instruments

The Occupational Personality Questionnaire (OPQ32) is an assessment of people's behavioural preference or style in the workplace, providing measurement for 32 traits (Bartram et al., 2006). The present study utilises two versions of this assessment that employ a multidimensional forced-choice (MFC) format (i.e., a FC format where items in the same block indicate different traits): the OPQ32i and the OPQ32r. Both versions request respondents to choose the statement that is "most" and "least" like them within each of the 104 FC blocks. However, the OPQ32i blocks consist of four items (so-called "quads") and OPQ32r blocks consist of three items ("triplets"). The OPQ32r triplets were developed through removing one item per quad from OPQ32i (Brown & Bartram, 2009-2011). Except wording improvements for 5 items, all other remaining items were exactly the same across versions. This nested

design allows studying the effect on responding behaviour of contextual change caused by an additional, distractor item in the same FC block.

Samples

Table 3. Study 1 sample composition

Sample characteristics		Quad instrument (OPQ32i)	Triplet instrument (OPQ32r)
Time of data collection		2004-2009	2009-2011
Gender	Male	62%	61%
	Female	38%	39%
	Missing	<1%	0%
Age	Below 20	1%	4%
	20-29	23%	33%
	30-39	32%	24%
	40-49	30%	21%
	50-59	12%	8%
	60 or above	1%	<1%
	Missing	1%	10%
Ethnicity	White	82%	56%
	Other	8%	8%
	Missing	10%	36%
N		62,639	22,610

Data from prior live administrations of the OPQ32 in UK English in the United Kingdom was used in this study after anonymisation. The samples were collected through a large number of assessment projects, which were typically for employee selection or development purposes. Respondents in the first sample (N=62,639) completed the older, quad instrument between 2004 and 2009. Respondents in the second sample (N=22,610) completed the newer, triplet instrument between 2009 and

2011. As shown in Table 3, the two samples had very similar gender compositions – each had just over 60% males and just under 40% females. In each sample, all working ages were represented, and the majority of respondents were white.

Analysis strategy

Analysis was structured in four main steps. Firstly, to create the foundation for all subsequent analyses, item parameters for the quad and triplet instruments were estimated independently using their respective samples, and equated to the same scales in order to remove sample-specific metric differences in the resulting model parameters. Secondly, to examine the impact of instrument design change on people's responding behaviour at item level, item parameters for the quad and triplet instruments were compared directly. Thirdly, to identify underlying reasons of item parameter differences, qualitative contextual analysis of item content was conducted. Finally, to examine the robustness of measurement at trait level, trait score estimates based on different item parameter sets were compared.

Item parameter estimation

The two samples were analysed using the TIRT model. Firstly, the "most" and "least" responses to FC blocks were converted to binary outcomes associated with pairwise comparisons within blocks. Each block of four items was coded as six pairwise comparisons. The quad instrument thus had $104 \times 6 = 624$ binary outcomes. Each block of three items was coded as three pairwise comparisons, and the triplet instrument had $104 \times 3 = 312$ binary outcomes.

Secondly, a TIRT model (Brown & Maydeu-Olivares, 2011) with 32 correlated latent traits indicated by their respective observed binary outcomes was fitted to each sample independently using the Unweighted Least Squares estimator in Mplus software

(Muthén & Muthén, 1998-2012). The conditional probability for a positive outcome of pairwise comparison was modelled as described in Equation 5 (each item in the OPQ32 instruments indicates one and only one latent trait, and items within the same FC block indicate different latent traits).

To enhance parallelism in the comparison of model parameters later, the models only considered binary outcomes shared by both instruments – that is, the 312 binary outcomes as in the triplet version. The outcomes unique to the quad instrument were ignored for two reasons. First, they were not relevant for answering the question of how people's responding behaviour changed when a fourth item was added into the same block. The fourth item acted merely as a distractor (context) in the present study's design. It existed only in the quad version, and therefore the parameters relevant to this distractor item could not be estimated for the triplet version, and therefore provided no basis for any parameter comparison. Second, the inclusion of the additional outcome variables when estimating the model parameters for the quad instrument would make the two models non-equivalent, thus introducing an extra source of difference into the comparison of model parameters. The only type of difference of interest to this study was the differences caused by empirical behaviour change between the two versions.

The OPQ32 instruments employed a well-established model of workplace personality (Bartram et al., 2006). Many studies had replicated OPQ32 scale correlations, and found them to be very stable across contexts and even language versions (for example, see SHL, 2014, Table 15). For the present study, both samples were collected from the same country (United Kingdom), in the original English language version. The IRT scoring protocol applied to UK English OPQ32 data in operational settings uses Bayesian maximum-a-posteriori estimation, informed by the prior distribution of the 32 traits with the correlation matrix established on "a

representative sample of the British population collected by the Office of National Statistics in parallel to their Labour Force Survey", and contained 2028 individuals (Bartram et al., 2006, Table 1). Therefore, the trait correlations in our models were fixed to these same correlations in order to define the factorial space. Furthermore, the origin and unit for each latent trait was set so that the sample's latent trait mean was 0 and standard deviation was 1. For model identification, the unique variance of one item per FC block was fixed arbitrarily to 0.5 (see Brown & Maydeu-Olivares, 2012). To ensure comparability of parameter estimates across instruments, for each corresponding FC block in quad and triplet instruments, the same item was chosen for fixing the unique variance.

However, the partial ranking design of the quad instrument resulted in some missing outcomes that needed additional treatment before item parameters could be estimated. Missing data arose because the "most" and "least" response format did not provide full rank ordering information for blocks of four items – the rank order of the two unselected items was not collected by design. The mechanism was missing at random (MAR), but not missing completely at random (MCAR), since the pattern of missingness was fully determined by the observed responses (Brown & Maydeu-Olivares, 2012). The TIRT models use limited information estimators (i.e. ULS) based on tetrachoric correlations of the observed binary dummy variables. Previous research by Asparouhov and Muthén (2010) showed that limited information estimators such as the ULS used in the present study result in biased parameter estimates when data were missing at random (MAR) but not completely at random (MCAR). Because the focus of the present study is on the item parameters, any systematic parameter estimation bias is unacceptable. However, the above bias can be eliminated almost completely using multiple imputation with as few as five replications (Asparouhov & Muthén, 2010). Following the guidance developed specifically for FC data by Brown and Maydeu-

Olivares (2012), multiple imputation with 10 replications was applied to handle the MAR data in the quad instrument, in order to prevent any bias in parameter estimation.

Table 4. Stability of quad instrument item parameter estimates across 10 imputations

Item parameter	Standard deviation for item parameter estimates across imputations			
	Mean across all items	SD across all items	Min across all items	Max across all items
Threshold $\gamma_{\{i,k\}}$	0.007	0.009	0.001	0.079
Loading $\lambda_{i_{s_i}}$	0.008	0.007	0.001	0.051
Uniqueness ψ_i^2	0.013	0.024	0.000	0.206

Due to the very large size of the quad instrument (416 items, resulting in 624 dummy observed variables), it was not possible to run multiple imputation on the entire instrument all at once. Instead, the quad instrument was divided into 12 similarly-sized subsections covering all 104 FC blocks. Multiple imputation was then conducted using all available data for each of the subsections. Even with this sub-sectioning, due to very large samples used in this study, Bayesian estimation of the unrestricted model required for multiple imputation for each subsection still took up to one day to complete. A total of 10 samples were imputed for each subsection and the resulting data subsequently merged across subsections to reconstruct the complete instrument. The TIRT model was then fitted to each of the 10 imputed samples. All 10 models converged and gave expected parameter estimates, which were stable across imputations (see the imputation statistics in Table 4). The estimates from the 10 models were then averaged to give the final IRT parameter estimates for the quad instrument.

Item parameter equating

The parameters of a multidimensional IRT model have a degree of arbitrariness – they are indeterminant until the trait directions, origins and units have been fixed

(Reckase, 2009, p. 233-234). In the present study, the IRT models for the triplet and quad instruments were constructed using two different samples. To identify trait directions, both models were estimated while fixing the correlations between latent traits, thus ensuring identical factorial space. To identify latent trait metrics for each model, the latent trait origins and units were fixed to reflect the means and standard deviations of the individual samples. However, the two samples were far from randomly-equivalent, and thus it was fully expected that the resulting latent trait metrics of the two models would be different. As a result, the item parameters of the two models were not directly comparable. Therefore, equating was required to place the item parameters on the same scale before subsequent analyses and comparisons.

As described in the TIRT Model section, the TIRT model is a variant of the M2PNO model with some special features. Metric transformation equations for the M2PNO model have long been published (e.g., Davey, Oshima & Lee, 1996). For the TIRT model, however, additional attention is needed to handle the unique variance parameters, thus demanding the deduction of new metric transformation equations, as detailed below.

With latent trait directions fixed to be equivalent across models, transforming of origins and units could be captured by a linear transformation as per unidimensional equating (Equation 27; Kolen & Brennan, 2004, p. 162). In the present study, the aim of equating was to find optimal coefficients x_s and y_s to transform the metric of the quad instrument model (η_s) to the metric of the triplet instrument model (η_s^*).

$$\eta_s^* = x_s \eta_s + y_s \tag{27}$$

Transforming the metric of latent traits has implications on item parameter values. For the IRT model to be invariant after transformation, the conditional probability of responses needs to remain unchanged (Reckase, 2009, p. 235), leading to

Equation 28. Therefore, the conversions of the threshold and the two factor loadings between the old and new metrics are as shown in Equations 29, 30 and 31.

$$p_{\{i,k\}}(\boldsymbol{\eta}) = \Phi\left(\frac{-\gamma_{\{i,k\}} + \lambda_{i_{s_i}}\eta_{s_i} - \lambda_{k_{s_k}}\eta_{s_k}}{\sqrt{\psi_i^2 + \psi_k^2}}\right)$$

$$= \Phi\left(\frac{-\gamma^*_{\{i.k\}} + \lambda^*_{i_{s_i}}\eta^*_{s_i} - \lambda^*_{k_{s_k}}\eta^*_{s_k}}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}}\right)$$

$$= \Phi\left(\frac{-\gamma^*_{\{i,k\}} + \lambda^*_{i_{s_i}}y_{s_i} - \lambda^*_{k_{s_k}}y_{s_k} + \lambda^*_{i_{s_i}}x_{s_i}\eta_{s_i} - \lambda^*_{k_{s_k}}x_{s_k}\eta_{s_k}}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}}\right) \tag{28}$$

$$\frac{-\gamma_{\{i,k\}}}{\sqrt{\psi_i^2 + \psi_k^2}} = \frac{-\gamma^*_{\{i,k\}} + \lambda^*_{i_{s_i}}y_{s_i} - \lambda^*_{k_{s_k}}y_{s_k}}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}} \tag{29}$$

$$\frac{\lambda_{i_{s_i}}}{\sqrt{\psi_i^2 + \psi_k^2}} = \frac{\lambda^*_{i_{s_i}}x_{s_i}}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}} \tag{30}$$

$$\frac{\lambda_{k_{s_k}}}{\sqrt{\psi_i^2 + \psi_k^2}} = \frac{\lambda^*_{k_{s_k}}x_{s_k}}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}} \tag{31}$$

Note that the unique variances provide essential scaling for thresholds and loadings pre- and post-transformation, but their own units are arbitrary. Because the models for the two instruments were fitted using identical unique variance identification constraints, the units for unique variances in the quad instrument model and the triplet instrument model are the same (i.e., $\psi_i^{*2} = \psi_i^2$). With this, Equations 29-31 simplify to Equations 32-34.

$$-\gamma_{\{i,k\}} = -\gamma^*_{\{i,k\}} + \lambda^*_{i_{s_i}} y_{s_i} - \lambda^*_{k_{s_k}} y_{s_k} \quad (32)$$

$$\lambda_{i_{s_i}} = \lambda^*_{i_{s_i}} x_{s_i} \quad (33)$$

$$\lambda_{k_{s_k}} = \lambda^*_{k_{s_k}} x_{s_k} \quad (34)$$

With the transformation method determined, the next step was finding the equating coefficients x_s and y_s for each latent trait. The data structure called for a common-item non-equivalent group linking design (Kolen & Brennan, 2004, p. 19). Given the nested structure of the two instruments, all but five items with wording change could be used as common items, thus giving a high proportion of common items far exceeding the essential requirements. When equating, however, the common items are assumed to function in exactly the same way across instruments (Kolen & Brennan, 2004, p. 19). This assumption may not always hold in the present study, where contextual change across instruments takes place. However, the impact on the results due to possible violation of this assumption was expected to be small if the vast majority of items functioned in the same way across instruments. With this, the coefficients x_s and y_s were subsequently estimated by linear equating (Equation 35; Kolen & Brennan, 2004, p. 31). In Equation 35, η_s denotes the latent trait in the default metric of the quad instrument model, thus $mean(\eta_s) = 0$ and $SD(\eta_s) = 1$ for the quad sample; η^*_s denotes the latent trait in a new metric, estimated by fitting a new model to the quad instrument sample, with all common item parameters fixed to values from the triplet instrument model, and $mean(\eta^*_s)$ and $SD(\eta^*_s)$ freely estimated. For the current study, Equation 35 further simplifies to Equation 36, thus giving linking coefficients x_s and y_s as shown in Equations 37 and 38.

$$\frac{\eta_s^* - mean(\eta_s^*)}{SD(\eta_s^*)} = \frac{\eta_s - mean(\eta_s)}{SD(\eta_s)} \quad (35)$$

$$\eta_s^* = SD(\eta_s^*)\eta_s + mean(\eta_s^*) \quad (36)$$

$$x_s = SD(\eta_s^*) \quad (37)$$

$$y_s = mean(\eta_s^*) \quad (38)$$

The linking coefficients x_s and y_s for each of the 32 latent traits were thus obtained by extracting the latent $mean(\eta_s^*)$ and $SD(\eta_s^*)$ estimates from Mplus outputs. Given the large sample sizes and similar sample characteristics across instruments, the latent trait distributions were expected to be similar and it was therefore not surprising that most x_s coefficients were close to 1 and most y_s coefficients were close to zero (Table 5), with the deviations from the expected values reflecting differences between the two samples. The x_s parameters ranged from 0.782 to 1.016, indicating that the latent trait standard deviations of the quad sample were between 78% and 102% (i.e. generally smaller) of the triplet sample. One tentative explanation of such differences might be population change over time – perhaps the UK population from which operational assessment data were collected had become more diverse, thus explaining the variance increase from the older quad sample to the newer triplet sample. Another potential explanation might be demographic composition differences between the two samples. For example, there were a larger proportion of younger respondents in the triplet sample, which might explain why the "Rule Following" trait showed the largest variance increase. The item parameters for the quad instrument model were then equated using these coefficients as shown in Equations 32-34 before subsequent analysis.

Table 5. Equating coefficients for linear transformations between latent trait metrics

Latent trait (η_s)	x_s	y_s
1 Persuasive	0.929	−0.168
2 Controlling	0.908	−0.151
3 Outspoken	0.896	−0.111
4 Independent Minded	0.861	0.049
5 Outgoing	0.897	−0.043
6 Affiliative	0.852	−0.091
7 Socially Confident	0.831	−0.147
8 Modest	0.828	0.079
9 Democratic	1.016	−0.082
10 Caring	0.835	−0.233
11 Data Rational	0.819	−0.179
12 Evaluative	0.861	−0.257
13 Behavioural	0.910	−0.146
14 Conventional	0.901	−0.337
15 Conceptual	0.890	−0.196
16 Innovative	0.905	−0.258
17 Variety Seeking	0.830	0.033
18 Adaptable	0.841	0.031
19 Forward Thinking	0.884	−0.144
20 Detail Conscious	0.865	−0.298
21 Conscientious	0.864	−0.373
22 Rule Following	0.782	−0.358
23 Relaxed	0.921	−0.090
24 Worrying	0.809	0.085
25 Tough Minded	0.897	−0.147
26 Optimistic	0.885	−0.117
27 Trusting	0.807	−0.051
28 Emotionally Controlled	0.825	−0.057
29 Vigorous	0.785	−0.324
30 Competitive	0.952	−0.058
31 Achieving	0.886	−0.318
32 Decisive	0.896	0.030
Mean	**0.871**	**−0.138**

Stability of item parameters

After equating, the item parameter sets were compared directly to establish their level of stability across the two instruments. The means and standard deviations of the

differences and absolute differences were calculated. Note that the loading, threshold and unique variance parameters were scaled arbitrarily in accordance with the unique variance model identification constraints, and thus the size of the differences must be interpreted in line with the scaling of the parameters.

The relationships between parameter sets were also examined graphically using scatter plots. Multivariate outliers away from the equating line, which had standardized residuals of magnitude above 3, were identified and studied in the qualitative phase of the analysis.

Qualitative analysis of item context

Qualitative analysis of items was conducted for FC blocks containing outliers as identified by the previous step of the analysis. To avoid confirmation bias, analysis was conducted purely though qualitative review of item text, without referring to their item parameter estimates. For each block concerned, analysis explored contextual changes across the triplet and quad versions of the block. Potential causes of parameter shifts were formulated, and predictions were made as to what the shifts may be. For a particular pairwise comparison of two items, contextual changes can cause parameter shifts in the following ways:

- When the context caused the likelihood of endorsement for one item over the other to change for the average person, the threshold is expected to shift;
- When the context moderated the relationships between items and their underlying traits, the loadings are expected to shift;
- When the context changed the amount of variation in the responses that cannot be explained by the underlying traits, the unique variances are expected to shift;

- When the context introduced sources of biases into the responding process, the existing model is insufficient for describing the full responding process, and all parameters can shift in unpredictable ways.

Themes emerging from qualitative analyses are reported in the Results section. Some general hypotheses of how the identified themes may influence item parameters in FC CAT are proposed in the Discussion.

Stability of trait score estimation

The ultimate goal of studying item parameter shift was to ensure stability of measurement at the trait level for each respondent. To assess this, respondents' scores based on parameter sets estimated from the two different instruments were compared. The sample taking the triplet instrument was selected for this analysis, because the binary outcomes of all pairwise comparisons were known in this sample. This sample was first scored using the parameters estimated from the triplet instrument, and then, separately, scored again using the before-equating parameters estimated from the quad instrument. Responses associated with the five items with wording change across instruments were not scored. At the end of this scoring process, each respondent in the sample had two sets of scores – one based on triplet instrument parameters, and the other based on quad instrument parameters. The trait scores estimated using the quad instrument parameters were then transformed using Equations 27 to align the metrics. The resulting two sets of trait score estimates were then compared as follows:

- Stability of rank ordering of individuals on a particular trait – correlations of the trait score estimates;
- Stability of rank ordering of individuals' personality profiles as a whole – correlations of profile locations (defined as the average score across all traits for each individual);

- Stability of rank ordering of traits for a particular individual – profile similarities (defined as the correlation between the two score profiles for the same individual based on two different parameter sets);
- Size of the differences between trait score estimates – relative and absolute differences between trait score estimates from different parameter sets.

Results

Stability of item parameters

Analysis was conducted on item parameter estimates that were neither associated with the 5 items with wording change, nor fixed in the model estimations. For example, there were 312 uniqueness terms in the models, one for each of the 312 items. However, 104 of them were fixed for model identification purposes and 5 were associated with items with wording change, thus reducing the total number of parameter estimates for analysis to 312 – 104 – 5 = 203.

The parameter estimates were aligned across the instruments, giving mean differences close to zero for all – thresholds, factor loadings and unique variances (Table 6). The parameters estimates also demonstrated strong linear relationships, as can be seen in the scatter plots of equated quad instrument parameters against triplet instrument parameters (Figures 1-3) and their very high correlations (Table 6). Estimates of item thresholds (see Figure 1) were mostly stable, giving a correlation of .975. Estimates of factor loadings (Figure 2) were less stable, giving a correlation of .878. Unique variance parameters turned out to be the most volatile to estimate across instruments (Figure 3), but still produced a high correlation of .841. Regarding the spread of the estimates, while Figure 1 shows a uniform spread around the equating line for thresholds, Figure 2 shows clear heterogeneity in the spread of the factor loadings. Specifically, larger slopes varied much more between the instruments than smaller

slopes did. The same was true for the uniquenesses (Figure 3). The greater fluctuations seen in loading and unique variance parameters were not surprising. Simulation studies by Brown and Maydeu-Olivares (2012, Tables 3 and 4) showed that loading parameters were typically recovered less accurately than threshold parameters, with larger loading values providing greater space for fluctuations than smaller loading values. The uniqueness parameters were estimated with even less precision.

Table 6. Comparing item parameter sets estimated from quad and triplet instruments

Parameter set comparison		Item parameter		
		Threshold $\gamma_{\{i,k\}}$	Loading $\lambda_{i_{s_i}}$	Uniqueness ψ_i^2
No. of free estimates		302	307	203
Quad (equated)	Mean	–0.009	0.731	0.484
	SD	0.735	0.290	0.459
Triplet	Mean	–0.028	0.726	0.497
	SD	0.751	0.330	0.737
Difference	Mean	0.019	0.005	–0.013
	SD	0.167	0.158	0.430
Absolute difference	Mean	0.121	0.104	0.201
	SD	0.116	0.118	0.381
Correlation		.975	.878	.841

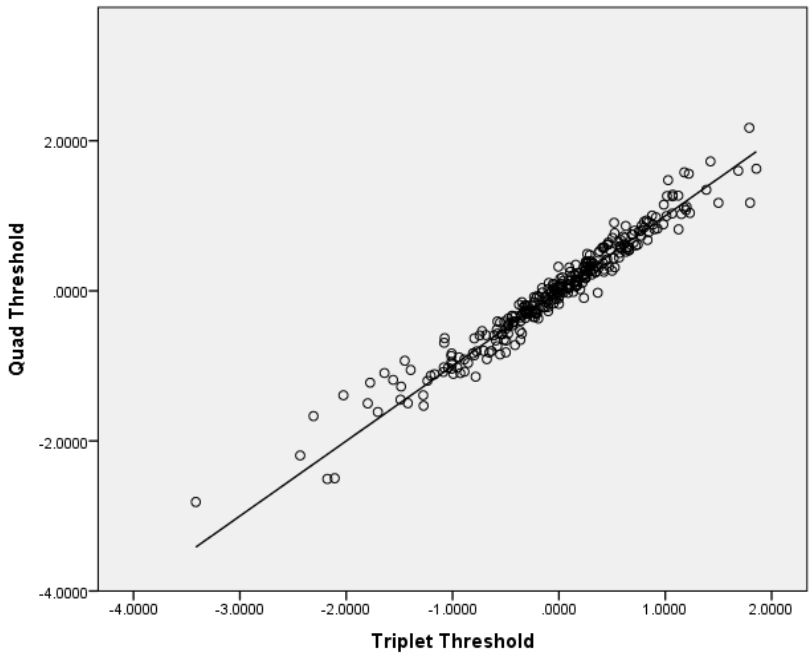


Figure 1. Scatter plot of estimated threshold parameters from quad and triplet instruments

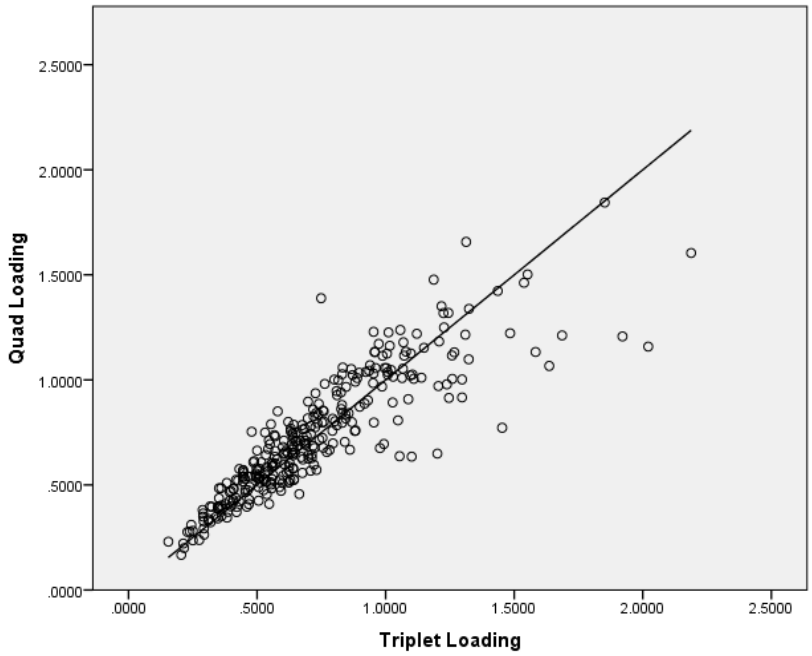


Figure 2. Scatter plot of estimated loading parameters from quad and triplet instruments

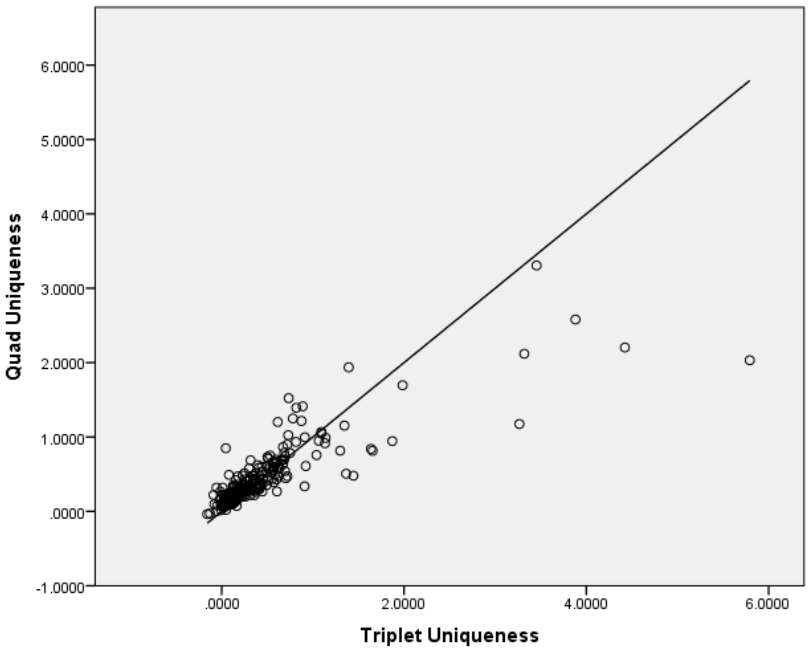


Figure 3. Scatter plot of estimated uniqueness parameters from quad and triplet instruments

Outliers

Table 7. Outliers with respect to parameter invariance from quad and triplet instruments

	Parameters			Affected items			Affected blocks		
	Total	Outlier	%	Total	Outlier	%	Total	Outlier	%
$\gamma_{\{i,k\}}$	302	7	2.3%	307	12	3.9%	104	5	4.8%
$\lambda_{i_{s_i}}$	307	8	2.6%	307	8	2.6%	104	5	4.8%
ψ_i^2	203	4	2.0%	307	4	1.3%	104	4	3.8%

Between 2.0% and 2.6% outliers were identified for each type of parameter (Table 7). Note that each threshold outlier affected two items, while each loading or uniqueness outlier affected only one item. In total, 17 (5.5%) of the 307 common items (i.e., 312 items in the triplet version minus five items with wording change) were

marked as outliers in at least one of the parameters. These outlier items were found in eight (7.7%) of the 104 blocks.

Qualitative analysis of item context

Items within the eight FC blocks containing outliers were studied to identify contextual changes across the instruments. This analysis identified a number of recurring themes, which are outlined below and illustrated by examples.

Theme 1: change in relative item endorsement levels

Change in relative item endorsement level was observed in three blocks. The block containing items 189-192 gives a good example.

Item		Quad	Triplet	Scale
189	I consider what motivates people	√	√	Behavioural
190	I am easily bored by repetitive work	√	√	Variety Seeking
191	I worry before an interview	√	√	Worrying
192	I finish things on time	√		Conscientious

The triplet version contains items 189, 190 and 191. In the workplace, item 189 is likely to be perceived as most desirable, so the relative endorsement levels of item 189 against items 190 and 191 are likely to be high. In the quad version, the desirability of item 192 is likely to be high. As a result, item 189 is no longer the obvious “best answer” in the quad, as it may be in the triplet. So the endorsement level of item 189 against items 190 and 191 is likely to be lower in the quad version. To put this in terms of parameters, the pairs {i189, i190} and {i189, i191} in the triplet version are likely to have lower threshold parameters (i.e., easier to endorse the first item) than in the quad version, which is what was observed.

Theme 2: change in item's discrimination levels

Change in item discrimination levels was observed in five blocks. The block containing items 141-144 gives a good illustration.

Item		Quad	Triplet	Scale
141	I am lively in conversation	√		Outgoing
142	I follow rules and regulations	√	√	Rule Following
143	I persevere with tasks	√	√	Conscientious
144	I avoid talking about my successes	√	√	Modest

The triplet version contains items 142, 143 and 144, and it is clear to the respondent that they all refer to distinct attributes. The additional item 141 in the quad version, however, is very similar to item 144 in content – both items have an element of talking to people. This "talking" emphasis in the same block creates an unintended contrast between items 141 and 144. As a result, item 144 may shift from being a positive indicator of Modest to being a negative indicator of Outgoing. Thus, the factor loading for item 144 on the Modest trait were expected to be lower in the quad version – exactly what was observed in the IRT parameter estimates. Predictions of shifts of other parameters in this block, however, were not as successful. It was hypothesised that item 142 would be unaffected by the shared "talking" element, and therefore the parameters for item 142 should not change. This prediction was not accurate and the loading for item 142 was actually lower in the quad version, suggesting that some additional factors were at play.

The qualitative study of change in context was unfortunately not always as simple as the examples given here. Often, multiple themes were present in the same block, leading to complex interactions and making the prediction of how item parameters would change extremely difficult. Nevertheless, based on this study,

possible mechanisms behind some context-induced parameter shifts are suggested and summarised in the Discussion.

Stability of trait score estimation

From a rank ordering perspective, trait score estimates for the same individuals based on different parameter sets were highly similar. Table 8 describes the correlations of scores for each of the 32 traits, the correlation of post-equating profile locations, and the profile similarities for all individuals in the sample (N=22,610). It was clear that the ordering of people at scale level as well as the similarity of whole personality profiles were preserved. The latter was important since selection decisions on comprehensive measures of personality were usually based on combinations of traits, not by comparing each individual trait.

Table 8. Comparing trait scores estimated using parameters from different instruments

Statistics	Mean	SD	Min	Max
Correlation of trait scores	.996	.002	.991	.999
Correlation of profile locations	.985	-	-	-
Profile similarity	.995	.002	.974	.999
Mean score difference by trait	−0.088	0.041	−0.183	0.005
Mean absolute score difference by trait	0.113	0.031	0.050	0.184

From an absolute difference perspective, the trait score estimates from different parameter sets were also highly similar. Table 8 describes the mean score differences and mean absolute score differences across the 32 latent traits. Reassuringly, most traits demonstrated mean differences close to zero and mean absolute differences of small magnitude. However, some traits demonstrated relatively large differences. The largest difference was seen in the “Conventional” trait, which showed mean difference of

-0.183, suggesting that respondents typically received lower scores when scored using the quad instrument parameters as opposed to triplet instrument parameters. Note that one of the five items with wording change was from the Conventional trait and removed in the scoring process. The second largest difference was seen in the "Vigorous" trait, with mean difference of -0.164. Such differences may be caused by a combination of item parameter shift across instruments, item parameter estimation error and equating error.

Discussion

The parameter invariance assumption is fundamental to the full realisation of adaptive personality assessments using the FC response format. The current study examined the effect of context on FC responding behaviour, as represented by adding one extra item per FC block. Empirically-derived item parameters, estimated independently before and after the contextual change, were compared. The threshold, loading and unique variance parameters were largely stable. Furthermore, a small proportion (less than 10%) of parameters that yielded substantial shifts, however, had little impact on the person parameter estimates. Evidence from the current study thus largely supported the parameter invariance assumption.

Nevertheless, a number of scenarios where this assumption was violated were reviewed, resulting in the identification of two recurring themes. The mechanisms behind parameter shifts are suggested below, and some recommendations for mitigating parameter shifts in adaptive FC assessments are made.

Themes in influences of context on FC item parameters

The two themes identified for parameter shifts are of particular interest to FC CAT implementations. Through understanding these themes better, appropriate test

assembly rules can be designed to mitigate their occurrences, thus reducing the likelihood of parameter shifts and enhancing the accuracy of trait estimations. With this purpose in mind, hypotheses are made for possible mechanisms behind parameter shifts due to change in context for FC items.

Theme 1: change in relative item endorsement levels

In FC blocks, some items can appear more desirable than others, either because they are more socially appealing in general, or because they are more in line with the purpose of the assessment (e.g., Donovan et al., 2003; Kam, 2013; Paulhus & Vazire, 2007). When making comparative judgements in an assessment setting, respondents are likely to be considering the desirability of items consciously or unconsciously. As a result, when item desirability within a block is not balanced, endorsement can shift towards the more desirable "right answers".

There are several factors that may intensify such desirability-induced response biases. Firstly, it is likely to occur more often in high-stakes situations, where respondents have stronger motivations to do well or appear good. Secondly, it is likely to be worsened when the desirability difference between items within the same FC block is large, thus making the perceived "right answer" more obvious to more respondents. Finally, it is likely to be more severe with smaller FC block sizes. In a block of two items, once the most desirable item is chosen to be "most" like the respondent, the other item has to be the "least" like the respondent, and the only information collected from this response is bias. But in a block of three items, the comparison between the remaining two items can still give useful information.

In terms of impact on measurement, such desirability-induced response biases introduce shifts in thresholds of the pairwise comparisons within the affected block, which can reduce the accuracy of latent trait estimation. To tackle this problem, items

should be worded neutrally or factually, so they do not sound obviously desirable or undesirable. Moreover, the relative endorsement levels of items should be estimated and controlled for in the instrument design. In a CAT setting, this translates into an additional rule in the test assembly algorithm – a numerical constraint preventing combinations of items with relative endorsement levels exceeding a certain acceptance threshold.

Theme 2: change in item discrimination levels

When considering several items simultaneously, respondents can perceive the item meaning differently to when they consider them independently. Most often, item interactions are caused by unplanned shared content between them, making their artificial similarity salient and deteriorating the original meaning of the items in relation to the attributes they indicate. Item interactions thus enhance or dilute the items' ability to measure their intended constructs, leading to shifts in item discrimination parameters.

There are several flags for identifying potential item interactions. The first clue comes from item wording – items sharing the same or synonymous keywords or phrases are likely to interact, as are items employing antonymous keywords. Furthermore, even if items do not explicitly share similar or opposite wordings, they can still have unplanned situational overlap that may lead to item interactions. The second clue comes from the constructs that the items measure – items from conceptually-similar constructs are more likely to interact than items from conceptually-distinct constructs.

In terms of measurement, item interactions can have two kinds of impacts. On one hand, when the shared context is not related to the latent constructs being measured, not only may the items have correlated residual variance caused by a common nuisance factor, but also do the items' focus shift towards that nuisance factor, thus reducing their power to measure the intended constructs. On the other hand, when the shared context is

related to the latent constructs being measured, interaction-induced item cross-loading happens. In such cases, the scoring model is no longer sufficient to model the response process. In a CAT setting, a viable solution to this problem is to prevent items that may interact from appearing in the same block. To do so, pairs of potentially interacting items need to be identified by subject matter experts and then coded in the test assembly algorithm as content "enemies" within (but not across) FC blocks.

Dealing with change in item uniqueness

Unlike the case of item thresholds and loadings, parameter shifts in item unique variances are harder to explain and to predict. This is perhaps not at all surprising because unique variances are, by definition, residual variances unexplained by the responding model. Unique variances characterise how closely the actual item responses scatter around their predicted values. While unique variances reflect certain item properties, for example how central or peripheral the item is to the measured attribute, they may also depend on environmental factors external to the items that affect the level of random variation in respondents' answers.

In terms of measurement, less random variation in answers should reduce the residual variances of items and give more accurate trait estimates. While reducing residual variances is a good thing for measurement in general, there is one complication in a CAT setting – if the unique variance of an item changes, the parameter invariance assumption is violated. And because there is no simple way to precisely quantify the extent of random influences a priori, it is challenging to construct test assembly rules that standardise unique variances across blocks.

However, in practice, change in residual variances is less of a concern compared to shifts in other item parameters in FC CAT. In order for FC CAT to be effective, a large item bank with calibrated item parameters is required. While it is not too

complicated to model a FC instrument with fixed block design as in the case of this study, it is impossible to calibrate a large item bank using FC response data because of an astronomical number of combinations in which the items can be paired together. Therefore, large item banks designed for FC CAT assessments are calibrated using SS response formats, where the residual variances are likely to be at their highest due to many response biases that affect the SS format. Consequentially, the SS-based item parameters are likely to overestimate unique variances in FC CAT. This leads to overestimation of the resulting measurement error in FC CAT. The test assembly thus operates under a worst-case scenario, making more conservative decisions regarding measurement precision, and arriving at more accurate trait estimates.

Unique variance fluctuations also have an impact on score estimation, through affecting the likelihood values of the responses. However, a small level of unique variance fluctuation is unlikely to dramatically change the score estimates. As can be seen in this study, an overall unique variance fluctuation characterized by a correlation of 0.841, together with a small number of shifts in other item parameters, still produced trait score estimates correlating to 0.991 or above. In summary, invariance of uniqueness in a FC CAT setting is given lower priority and importance compared to invariance of threshold or loading.

Limitations

One limitation of using historical data in this study is the confounding of contributions from contextual differences as well as potential sample differences in the observed parameter fluctuations. To partial out the contribution from potential sample differences, further studies need to incorporate adequate matching or randomisation designs during data collection.

The contextual difference between the two instruments used in this study is also limited in nature. Firstly, both instruments were constructed manually by experts while taking into account content requirements and best practices in measurement, so the additional item seldom introduces significant contextual shift into a FC block. Once the human factor is removed, computer-assembled FC blocks are likely to have larger impact of context, potentially leading to greater fluctuations in item parameters. Secondly, the FC block compositions were very similar across the two instruments, with three out of four items staying the same. The effect of fully shuffling the items into different blocks may lead to yet more contextual changes, and potentially larger item parameter shifts. This remains an area of research for further studies. However, the tight control over the context in this study is also its strength because it was possible to triangulate the potential causes behind the item parameter shifts, which would be much more difficult with less controlled contextual changes.

Finally, this study only focuses on measuring personality, which comprises relatively stable psychological constructs. For constructs that are more situation-dependent, contextual variations may lead to greater responding behaviour differences. Therefore, generalisations of the findings in this study to FC assessments of other constructs must be made with caution.

Conclusions

While modern IRT models provide the necessary theoretical foundation for FC CAT, a fundamental assumption in CAT is that item parameters are invariant with respect to context – items surrounding the administered item. This assumption is empirical in nature, yet there had been limited investigation into its robustness. Study 1 empirically examined the influence of context (manipulated through the addition of distractor items) on item parameter stability. The item parameter estimates with and

without manipulation were highly similar. Moreover, person trait score estimation remained very stable despite a small proportion of violations of the parameter invariance assumption. Results thus support the adoption of the parameter invariance assumption in practice.

Although infrequent, context did introduce a small number of significant item parameter shifts in this study. Therefore, while the parameter invariance assumption appears to be robust even with minor violations, it is still important to strengthen it through the incorporation of appropriate content rules. It is recommended that items within the same FC block should be constrained to have similar average endorsement levels. Also, items that may interact (i.e., change in item focus or meaning when presented together, for example due to nuisance shared context) should be prevented from appearing in the same FC block. Such content rules can be coded into the adaptive test assembly algorithm behind any FC CAT.

CHAPTER 3: FC CAT ALGORITHMS

The tailoring of questions to respondents in a CAT is governed by the logics within an automated test assembly algorithm. A CAT algorithm typically consists of four main components: 1) a trait estimator; 2) an item selector; 3) a collection of content rules; and 4) a stopping rule. The trait estimator produces estimates for the respondent's trait standings. Based on the respondent's interim trait estimates, the item selector identifies the most informative question to administer next, subject to the constraints of content rules. The content rules capture assessment design requirements and define the boundaries within which the item selector operates[6]. The stopping rule determines when the assessment terminates. Figure 4 illustrates the process flow of a CAT.

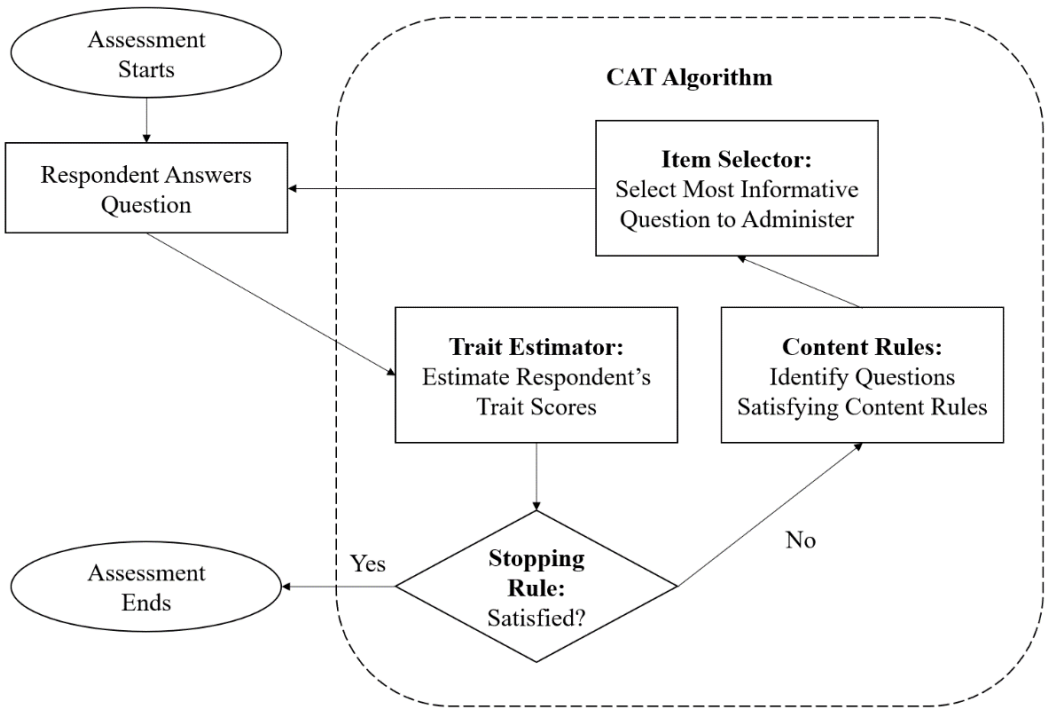


Figure 4. Process flow of a CAT

[6] The content rules are not always separable from, and may be considered an integral part of, the item selector. However, for the discussions here, a conceptual distinction is made between the mathematical criterion to be optimised for information gain (i.e., the item selector), and the content requirements that act as constraints in this optimisation process (i.e., the content rules).

Multiple methods and options exist for each of the algorithm components. As such, optimising the algorithmic design for an applied CAT is a complex problem, with design decisions heavily dependent on the purposes, needs, settings and operational constraints of the assessment program. While it would be impossible to design a one-size-fits-all FC CAT algorithm, this thesis aims to shed light on the design considerations for FC CATs, and the relative merits of different algorithmic options for FC CATs using TIRT.

This chapter is structured as follows. First, the four algorithmic components are reviewed and formulated for TIRT-based FC CATs, taking into account the unique needs and novel challenges of such assessments. Second, a simulation study (Study 2) is presented that compares the pros and cons of different trait estimators for scoring FC data. Third, a simulation study (Study 3) is presented that compares the performance of different item selectors for TIRT-based FC CAT. Finally, conclusions and practical recommendations are presented.

Algorithm Components for FC CAT

Trait Estimators

Trait estimators, also known as ability estimators or scoring methods, are mathematical algorithms for estimating a respondent's standings on the measured constructs (i.e., latent traits). Trait estimators not only determine the final scores to be reported at the end of an assessment, but also produce interim estimates to drive the item selection process forward in a CAT. An accurate and robust trait estimator is thus essential for the efficient functioning of a CAT.

A respondent's latent trait standings are estimated based on: 1) the characteristics of the administered items; 2) the respondent's responses to them; and,

optionally, 3) any existing information (i.e., prior information) about the respondent and/or the population they come from. Some trait estimators utilise the first two types of information only, including the Maximum Likelihood (ML) Estimator (Birnbaum, 1958, 1968) and the Weighted Likelihood (WL) Estimator (Warm, 1989). Other trait estimators incorporate prior information into the calculations and thus fall into the class of Bayesian estimators, including the Maximum a Posteriori (MAP) estimator (Lord, 1986; Mislevy, 1986) and the Expected a Posteriori (EAP) estimator (Bock & Mislevy, 1982). All four estimators produce point estimates for the respondent's latent trait standings. Full mathematical formulations of the trait estimators for the TIRT model are provided in Appendix B.

Theoretical comparison

The four trait estimators (ML, WL, MAP and EAP) exhibit different properties and strengths, thus making them optimal for different application scenarios. Moreover, trait estimation requirements and priorities change as a CAT progresses (van der Linden & Pashley, 2010), so some estimators are more appropriate than the others at different stages of a CAT.

The ML estimator is consistent and asymptotically efficient (Lord, 1983), and is traditionally the most widely-used trait estimator. However, the ML estimates tend to be biased outwards (i.e., the bias correlates with trait value positively), leading to overestimation of high trait values and underestimation of low trait values. Moreover, the ML estimator can be unbounded for certain response patterns (Lord, 1983), and the chance of this happening is especially high for shorter tests. Nevertheless, the bias of the ML estimator diminishes as the assessment gets longer.

The Bayesian estimators MAP and EAP, often coupled with a multivariate normal prior function around the estimated population mean, are probably the modern

favourites. With an informative prior, the Bayesian estimators tend to be biased inwards (i.e., the bias correlates with trait value negatively), pulling trait estimates towards the population mean. However, Bayesian estimates can be biased outwards if an uninformative prior is used, effectively converging towards the bias of the ML estimator when prior information diminishes (e.g., Wang, 2015). Unlike the ML estimator, both MAP and EAP are bounded when an informative prior is used, making it possible to obtain finite estimates even just after one question (Reckase, 2009). Bayesian estimates also tend to be less erratic than ML estimates, especially for shorter tests such as at the early stages of CAT (Reckase, 2009). Moreover, the prior information about the traits' covariance can enable more efficient estimation than if the traits were estimated separately (Segall, 1996). However, the utility of the Bayesian approach can be damaged by a badly chosen prior, leading to biased trait estimates and thus ultimately hindering rather than enhancing measurement (Gelman, Carlin, Stern & Rubin, 1995).

Compared to the ML, MAP and EAP estimators, there are fewer research studies that looked into the WL method. However, the available studies that benchmarked WL against ML, MAP or EAP produced very promising results. Warm (1989) compared WL against ML and MAP for the unidimensional 3PL model using a series of Monte Carlo studies, and found WL to outperform both ML and MAP over a large range of trait values in both static tests and variable-length CAT. More recently, Wang and Wang (2001) tested the WL estimator on fixed-length CAT simulations using the unidimensional generalised partial credit model (Muraki, 1992), and found it to produce more accurate results than ML, EAP and MAP. Wang (2015) compared WL against ML, MAP and EAP on fixed length tests using the multidimensional 2PL model, and found it to be the best in terms of both bias and variance in all conditions, except in the case where the prior distribution in the Bayesian estimators are identical to the generating distribution of the simulation sample (in which case the Bayesian estimators performed

better). Moreover, the WL estimator does not produce unbounded estimates as the ML estimator (Warm, 1989), nor does it require the setting of a prior (Warm, 1989), and so it is immune to the wrong prior risk of the Bayesian estimators. It is therefore very tempting to try the WL estimator out on multidimensional FC assessments using TIRT, to see whether its success elsewhere can be replicated in this new setting.

Aside from their statistical properties, the trait estimators also differ in computational procedures and complexities. The calculations for the ML, WL and MAP estimates all involve the maximisation of some score function, which is typically done by searching for zero gradient using an iterative numerical process. The score functions for ML and MAP are comparatively simple and quick to compute, whereas that for WL is significantly more complex, involving complex summations in every iteration step (i.e., updating the entire FIM and the ML bias term). The computational power and time requirement for WL is thus higher than that for ML or MAP. In contrast, the calculations for the EAP estimates are non-iterative in nature. Instead, numerical integration routines are employed to estimate the integral. In the case of unidimensional assessments, numerical integration is usually less computer intensive than iterative search, therefore the EAP estimator tends to be quicker to calculate (Bock & Mislevy, 1982). However, in the case of multidimensional assessments, the complexity of numerical integration grows exponentially as the number of dimensions increases. As a result, EAP loses its computational advantage and can become rather cumbersome for assessments with a larger number of dimensions (Segall, 1996).

A note on paradoxical results in multidimensional trait estimation

Hooker, Finkelman and Schwartzman (2009) observed that, in cognitive ability assessments using compensatory (i.e., the effect of a low score on one trait can be compensated by a high score on another trait to arrive at the same response probabilities,

see Reckase, 2009) multidimensional IRT models, “it is possible for the estimate of a subject’s ability in some dimension to decrease after they have answered a question correctly”. This phenomenon is referred to as the paradoxical results in multidimensional IRT. Finkelman, Hooker and Wang (2009) provided an explanatory example, where two candidates A and B took a test assessing both mathematical and language skills, giving identical answers to every question apart from the last one that relied heavily on language skills. Candidate A answered the last question correctly, demonstrating excellent language skills, and so the wrong answers in earlier parts of the test were likely explained by lower mathematical abilities. Candidate B answered the last question incorrectly, demonstrating lower language skills, and so the correct answers in the earlier parts of the test were supported by stronger mathematical abilities. It then followed that candidate A received a lower score on mathematical abilities than candidate B, even though intuitively a correct answer should work in the favour of candidate A on all ability dimensions. This scenario would be particularly problematic if candidate A was subsequently screened out due to not meeting a cut score on mathematical abilities while candidate B was allowed to pass, leading to the unfair situation of a wrong answer actually benefitting the candidate.

Following the initial discovery, a quick succession of studies explored this phenomenon in depth, attempting to understand the underlying mechanism and/or identify methods for avoiding such paradoxical results. Hooker, Finkelman and Schwartzman (2009) showed that this problem is unavoidable in linearly compensatory models using the ML estimator. Hooker (2010) further deduced that paradoxical results could occur when using a prior with all abilities positively correlated. Jordan and Spiess (2012) extended Hooker and colleagues’ results beyond binary linear compensatory multidimensional IRT models to ordinal models and other more general models, covering the scenarios using the ML estimator as well as Bayesian estimators. They

concluded that the paradoxical phenomenon was “highly prevalent” and called into question the general use of multidimensional IRT models because of the perceived unfairness, especially when cut scores were used to make decisions. At the same time, van der Linden (2012) showed that the paradoxical results would occur in “any multiparameter likelihood with monotone score functions”, and the paradoxical phenomenon was actually a feature of the convergence of the multiparameter ability estimator to its true values as the test lengthened. He thus argued against attempts to “fix” the perceived unfairness by modifying the ability estimates, as they would lead to “less accurate and more biased ability estimation”. Van der Linden’s (2012) view was further echoed by Reckase and Luo (2014), who showed that the paradoxical results were “not flaw in estimation”, but instead “the additional response improves the estimate of the θ-point even though the paradoxical result occurs”. These studies quickly enhanced the understanding of the paradoxical phenomenon, moving the field from considering it a detrimental artefact of multidimensional IRT to regarding it a mere feature of the multidimensional convergence process.

A couple of studies were particularly useful in providing intuitive understanding of this phenomenon. Breaking away from the model-specific algebraic investigations that dominated the study of paradoxical results to date, van Rijn and Rijmen (2012, 2015) introduced the use of graphical models instead. They realised that the paradoxical results were in fact examples of the more established “explaining-away” phenomenon in Bayesian networks (Pearl, 2009; Wellman & Henrion, 1993), also known as Berkson’s paradox in statistics (Berkson, 1946). They attributed the occurrences of paradoxical results in multidimensional IRT to the existence of a specific graphical structure called an “inverted fork” in the model, i.e., “when multiple latent variables are related to the same observed variable” (Figure 5). In multidimensional IRT terms, inverted forks occur when within-item multidimensionality exists, regardless of the exact functional

form of the multidimensional IRT model. Van Rijn and Rijmen (2015) then extended beyond multidimensional IRT models and showed that the "explaining-away" phenomenon occur in a wide class of multivariate latent variable models. With this widened understanding, they recommended treating tests of maximum performance and tests of typical performance separately in the discussion of paradoxical results, as the former had to conform to a higher level of social acceptability whereas the latter could place more emphasis on statistical optimality.

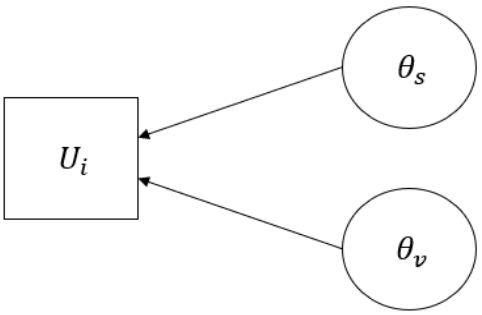


Figure 5. "Inverted fork" in MIRT

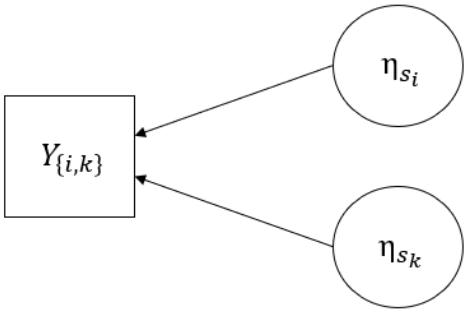


Figure 6. "Inverted fork" associated with a pairwise comparison in TIRT

The research on paradoxical results in multidimensional IRT is relevant to this thesis because the TIRT model is also a compensatory multidimensional IRT model. As "inverted forks" are inherent in the modelling of MFC responses (Figure 6), the "explaining-away" phenomenon will occur regardless of the trait estimator used or the

format of the prior function chosen (van Rijn & Rijmen, 2012). This can lead to some unintuitive results. For example, an answer to a pairwise comparison involving two traits will understandably affect the estimates of those two traits involved, but it will also affect the estimates of other traits. More specifically, updating the estimates for the two target traits will have a knock-on effect on the likelihood of the responses in earlier pairwise comparisons involving one of the two target traits and some other traits, and this rippling effect continues on until all traits are affected. The multidimensional trait estimates thus converge towards their true values in a fuzzy way that do not always conform to the explicit direction of the latest response, hence producing "paradoxical" results.

So far, research on paradoxical results in multidimensional IRT has largely focused on ability tests. While IRT models for ability and personality assessments are similar in many ways, the assessments themselves have several key differences. First, in most applications, personality assessments are tests of typical performance rather than maximum performance. For example, rather than wanting to find out how extroverted a person could possibly be, in most practical applications the aim is to instead find out how extroverted a person typically is. For this purpose, as van Rijn and Rijmen (2015) pointed out, one should focus on ensuring the most accurate estimation, rather than worrying about social acceptance of seemingly aberrant estimates. Second, personality assessments often intend to accurately recover the multidimensional, whole-person profile. The practice of using cut scores on a single personality dimension is far less justified than in the case of ability tests. Therefore, the problem of having paradoxical results occurring near a particular score for some dimension may not be a concern for personality assessments. Third, while ability test questions have definite right answers, this is not the case in personality assessments. A person can have a higher or lower standing on a personality trait, but whether one end of the trait is better than the other is

largely dependent on context. When the concept of a "better" score is removed from the picture, so is the concept of "unfair" scoring. Given their differences, the conclusions and recommendations regarding paradoxical results in ability tests require reconsideration in the setting of personality assessments.

Considering both the mathematical conditions for paradoxical results to occur, and the practical differences between ability tests and personality assessments, I argue that the concerns about paradoxical results can largely be alleviated or even removed in most personality assessments: 1) typically SS personality assessments use unidimensional items, in which case there is no within-item multidimensionality or inverted forks in the model, and so there will be no "explaining-away" phenomenon; 2) SS personality assessments using multidimensional items will give rise to "explaining-away" phenomenon, but this design is very rare in practice; 3) unidimensional FC personality assessments do not give rise to "explaining-away" phenomenon, because comparison of items from the same trait does not give rise to multidimensionality in the pairwise comparison outcomes; 4) multidimensional FC questions will give rise to "explaining-away" phenomenon, but this is not a concern given the focus on accurate estimation over pass/fail classifications. Weighing all considerations, this thesis focused on the standard trait estimators with no adjustment for potential paradoxical results in MIRT trait estimation.

Item Selectors

For a CAT to produce more accurate person scores than its non-adaptive counterparts, the item selector needs to identify which item(s) available in the pool to administer next so as to achieve the greatest information gain about the respondent, based on what is already known about them at the time. Because the item selector is

directly responsible for the adaptive assessment construction, it is arguably the most important component of a CAT algorithm.

All item selectors are based on the idea of information maximisation. However, while information maximisation is straightforward for a unidimensional test choosing one item at a time, its extension to MFC assessments presents additional complexities. First, in a multidimensional assessment, the information is dependent on the direction of consideration in the multidimensional space, and often the objective is to measure **all** the traits accurately (i.e., gaining information in multiple directions in the multidimensional space). Item selection thus becomes a multidimensional optimisation problem, and finding the best item requires the amalgamation of information from multiple directions into a single measure. Second, in a FC assessment, multiple items need to be assembled into a single FC block. The number of possible blocks to consider increases quickly as the block size n increases – there are $|R_r|$ ways to choose one item from R_r (the pool of unused items for the r^{th} question), but $\binom{|R_r|}{n} = \frac{|R_r|!}{n!(|R_r|-n)!}$ ways to construct a block of size n from the same pool of items. Finding the best FC block thus often requires extensive searches even for a small item bank. To sum up, combining the challenges introduced by multidimensionality of the intended constructs and the combinatorics of FC blocks, item selection for the MFC format is a complex and computationally intensive optimisation problem.

In order to address the multidimensionality challenge, researchers have developed a range of item selectors that reduce multidimensional information into scalar summary indices. These item selectors are mostly developed for assessments using response formats other than the MFC format, and/or developed for IRT models other than the TIRT model. Nevertheless, many of them can be extended to the measurement of personality using a MFC response format and the TIRT model. The first type of item

selectors are based on information maximisation that target a specific direction in the multidimensional trait space. Such item selectors include maximise weighted information (WI), maximise weighted core information (WCI), and maximise information in direction with minimum information (DMI; Reckase, 2009). The second type of item selectors make use of the FIM (see Mulder & van der Linden, 2009; Silvey, 1980). Such item selectors include minimise trace of the inverse FIM/ minimise total error variance (A-optimality), minimise weighted sum of entries of the inverse FIM/ minimise error variance of a linear composite (C-optimality), maximise determinant of the FIM/ minimise the volume of the confidence ellipsoid of the trait estimates (D-optimality), maximise minimum eigenvalue of the FIM/ minimise variance of the most imprecisely-estimated linear combination (E-optimality), and maximise trace of the FIM/ maximise information while ignoring contributions from correlated traits (T-Optimality). Both these types of item selectors rely on interim trait estimates and are therefore affected by their inaccuracies. As a result, they could be optimising measurement at the wrong locations, especially at the beginning of a CAT session when trait estimates are still inaccurate (e.g., Chang and Ying, 1996). The third type of item selectors bypass this problem using the Kullback–Leibler (KL) global information concept (Cover & Thomas, 2006; Kullback, 1959; Lehmann & Casella, 1998). Such item selectors include maximum item KL information (KLI-U or KLI-B; Chang & Ying, 1996; Veldkamp & van der Linden, 2002), maximum KL distance between subsequent posteriors (KLP; Mulder & van der Linden, 2010), maximum mutual information (MUI or KLB; Mulder & van der Linden, 2010; Wang & Chang, 2010, 2011; Weissman, 2007), and the continuous entropy method (CEM; Wang & Chang, 2010, 2011). Full mathematical formulations of and discussions about the item selectors for MFC assessments using TIRT are provided in Appendix D.

Theoretical comparison

Benefitting from the rapid increase in computational power and the psychometric advancement in IRT models over the last few decades, research on item selectors for CAT have progressed significantly (van der Linden & Glas, 2010). The fundamental idea of information maximisation underlies all item selectors. Initially, item selectors focused on local information, i.e., maximise information gain at the best interim score estimates. While this goal is simple in the unidimensional case, in multidimensional assessments it diverged into many ways of summarising information from multiple traits into a single scalar summary index required for item selection, thus giving rise to the differences between WI, WCI, DMI, A-, C-, D-, E-, and T-optimality (even though they all reduce to the same maximum information item selection criterion in the unidimensional case). Researchers have compared the efficiencies of FIM-based local information item selectors for various multidimensional assessments. For example, Mulder and van der Linden (2009) thoroughly examined A-, C-, D- and E-optimality theoretically and through simulations, and concluded that A- and D-optimality "lead to the most accurate estimates when all abilities are intentional, with the former slightly outperforming the latter", while C-optimality was most suited for "the measurement of a linear combination of abilities". Independently, Seo and Weiss (2015) simulated item selection in assessments using the bifactor model, and again found A- and D-optimality to outperform E-optimality.

More recently, the risk of making suboptimal decisions based on inaccurate interim trait estimates sparked a significant paradigm shift towards utilising global information measures in item selection. Mulder and van der Linden (2010) conducted a comprehensive theoretical review of the use of KL information in item selection (see Appendix D for full details and discussions). In terms of efficiency in practice, Chang

and Ying (1996) developed and simulated KLI in unidimensional CAT, showing that it tended to outperform local Fisher information maximisation. Weissman (2007) simulated MUI in unidimensional adaptive classification tests and found it to also give more accurate classifications than Fisher information maximisation. Wang and Chang (2011) simulated D-optimality, KLI, CEM and MUI in multidimensional CAT, and found MUI to be the most efficient amongst them, while D-optimality performed on par or better than CEM or KLI despite being a local information item selector.

Aside from their psychometric differences, the item selectors also differ in computational procedures and complexities. The global information item selectors (i.e., KLI, KLP, MUI, and CEM) all rely on numerical integration. As the computational complexity of numerical integration grows exponentially with increasing dimensionality, the global information methods can quickly become computationally challenging in multidimensional personality assessments that routinely involve five or more traits.

A note on selecting larger FC blocks

The item selectors (described in Appendix D) can be used to select FC blocks using three or more items, but with a couple of additional challenges: increasing computational demand, and local independence violation.

As briefly mentioned before, the first challenge of increasing computational demand arises from the growing number of ways to combine items into larger FC blocks. For example, an item bank with 100 statements gives rise to $\binom{100}{2} = 4{,}950$ unique pairs, $\binom{100}{3} = 161{,}700$ unique triplets, $\binom{100}{4} = 3{,}921{,}225$ unique quads, and so on. The exponentially increasing numbers of possibilities make searching through and choosing larger FC blocks much more computationally expensive than choosing smaller FC blocks. To overcome this challenge, content rules (see next section) can be

introduced to reduce the number of FC blocks to search through, but at the expense of making the assessment less adaptive.

The second challenge of local independence violation arises due to the shared residual variance between pairwise comparisons within the same FC block, which impacts the calculation of information measures that form the basis of all item selectors. The ideal way to handle this is to account for local dependence properly in the information calculations. However, the mathematics can get complicated fairly quickly. A non-ideal but practical way to handle this is to make a simplifying assumption of local independence and approximate the total block information by summing over contributions from all constituting pairwise comparisons (e.g., Equation 18).

In practice, there are pros and cons for using larger FC blocks in personality assessments. On one hand, larger blocks collect more pairwise comparisons per item presented, thus leading to more efficient use of the item bank for information collection. For example, assembling six items into pairs would yield three pairwise comparisons, whereas assembling six items into triplets would yield two triplets giving six pairwise comparisons in total – doubling the number of pairwise comparisons collected whilst using the same total number of items. On the other hand, because respondents need to consider more pairwise comparisons simultaneously when responding to a larger FC block, larger blocks are more cognitively demanding, making them more prone to data quality issues especially with unmotivated or unsophisticated respondents (Brown & Bartram, 2009-2011). The optimal block size for a FC personality assessment should be determined considering the practical settings of the assessment program in question, e.g., the cognition and level of motivation of the respondent population, the richness of the item bank, and any assessment time limits for response data collection.

Content Rules

Assessment assembly often needs to account for various content requirements, e.g., having a balanced mix of items measuring different personality traits. Such content requirements are realised by placing content rules on the automated item selection process[7]. More specifically, content rules prioritise content considerations over information maximisation by omitting potentially more informative FC blocks that do not conform to the content requirements. Content rules are therefore restrictive constraints that reduce the number of feasible FC blocks available for selection and thus the computational intensity of the item selection process.

Because assessment programs have different goals and requirements, content rules are often situation-dependent. Nevertheless, this section outlines some generic content rules that are applicable to many FC personality assessments. Note that the overlay of multiple content rules can lead to an overly restrictive content plan, thereby greatly reducing the freedom and effectiveness of adaptive assessment tailoring. It is therefore important to consider the collective effect of content rules on assessment assembly.

Social desirability balancing

An important appeal of the FC response format is its enhanced resistance against faking – a property that relies on the items within the same FC block to be similarly desirable (Krug, 1958). Social desirability balancing of items within the same block is thus an important content rule in many FC personality assessments. More specifically,

[7] While content rules are sometimes considered a component of the item selectors, for the sake of clarity of discussion, I make a distinction between the mathematical criterion to be optimised (i.e., the item selector) and the constraints placed around this optimisation process (i.e., the content rules).

the range of social desirability values of items within the same FC block is constrained to be within a certain threshold during the automated test assembly process. The social desirability values of items are often derived through some rating exercise, preferably structured in a way to reflect the context of the assessment program (e.g., Converse et al., 2010; Jackson et al., 2000; Krug, 1958). In lieu of such data, the items' mean utility parameters may be used as an approximation, albeit with reduced effectiveness in preventing faking (e.g., Heggestad, Morrison, Reeve, & McCloy, 2006).

Scale planning

For content validity and face validity reasons, multidimensional personality assessments often have balanced proportions of items measuring different traits (e.g., Ashton & Lee, 2009; Costa & McCrae, 1992). This can be addressed by scale planning, i.e., first determining which scales to measure in the next FC block, then choosing items for the targeted scales to construct the block.

There are many ways to implement scale planning in a CAT. A static scale plan satisfying all requirements can be pre-constructed and enforced during assessment assembly (e.g., Stark et al., 2012). Alternatively, dynamic scale planning can take into account the information collected for each scale as a CAT session progresses, in order to prioritise underperforming scale combinations in subsequent FC blocks. Underperforming scale combinations can be identified using information-based methods – criteria that are similar to those employed by item selectors, but adapted to instead model and summarise scale-level information. For example, Equations 39 and 40 show how the WCI and A-optimality item selection criteria (Equations D3 and D6) can be modified to instead choose an underperforming scale combination to focus on, with $s = v$ or $s \neq v$ depending on whether a unidimensional or multidimensional pair is

desired. Finally, it is also possible to adopt a hybrid approach where a mixture of static and dynamic scale planning techniques are used in a CAT.

$$\{s, v\} = arg\ min_{\{s,v\}} \left\{ \frac{1}{w_s} \left[CI^{\boldsymbol{\alpha}^s}_{\{i_1,k_1\}}(\widehat{\boldsymbol{\eta}}^{r-1}) + \cdots + CI^{\boldsymbol{\alpha}^s}_{\{i_{r-1},k_{r-1}\}}(\widehat{\boldsymbol{\eta}}^{r-1}) \right] \right. \\ \left. + \frac{1}{w_v} \left[CI^{\boldsymbol{\alpha}^v}_{\{i_1,k_1\}}(\widehat{\boldsymbol{\eta}}^{r-1}) + \cdots + CI^{\boldsymbol{\alpha}^v}_{\{i_{r-1},k_{r-1}\}}(\widehat{\boldsymbol{\eta}}^{r-1}) \right] \right\} \tag{39}$$

$$\{s, v\} = arg\ max_{\{s,v\}} \left\{ \left[\left(\boldsymbol{F}^{r-1}(\widehat{\boldsymbol{\eta}}^{r-1}) \right)^{-1} \right]_{s,s} + \left[\left(\boldsymbol{F}^{r-1}(\widehat{\boldsymbol{\eta}}^{r-1}) \right)^{-1} \right]_{v,v} \right\} \tag{40}$$

Stopping Rules

In a CAT, the stopping rule determines when to stop asking further questions and terminate the assessment session. The simplest stopping rule is one based on assessment length, leading to a fixed length CAT with a uniform assessment experience where all respondents see the exact same number of questions. More advanced stopping rules are based on measurement status and terminate the assessment session as soon as the collected responses have provided a level of measurement accuracy that is adequate for the intended use of the assessment scores, leading to a variable length CAT with shorter assessment sessions for some. Stopping rules based on measurement status may be placed on the maximum SEM across all traits (see Equations 20 and 21), the maximum total error variance across all traits (see A-optimality, Equation D6), the maximum volume of the confidence ellipsoid of the trait estimates (see D-optimality, Equation D9), or other similar extensions of the methods underlying the item selectors. Moreover, in order to prevent an assessment session from getting too long, the stopping rule for a variable length CAT still tends to incorporate an absolute maximum limit on the number of questions asked. The choice and formulation of the stopping rule for a CAT should be determined considering the practical requirements, constraints and priorities of the assessment program in question.

Comparing Trait Estimators for FC Assessments (Study 2)

Given the real-life impact of assessment results on human-related decisions and outcomes, it is important to estimate person scores accurately. While current research findings on trait estimator performance are highly relevant, the effectiveness of ML, MAP, EAP and especially WL has yet to be explicitly compared for TIRT-based FC personality assessments, which have several key differences compared to typical cognitive assessments that have been the focus of most trait estimator research to date.

First, FC personality assessments tend to measure a larger number of traits, often using an inseparable multidimensional FC design where no single attribute can be estimated without estimating the whole model. Such a large number of scales and the accompanying multidimensional structure may have effects on trait estimation that are rarely seen in cognitive tests with much simpler scale structures.

Second, while items in cognitive tests always have positive loadings onto the latent ability dimensions, this is not the case in personality assessments – items indicating the opposite characteristics of an intended trait (e.g., introversion rather than extraversion) will have negative item loadings. In fact, in the case of multidimensional FC assessments in particular, it has been shown that the presence of counter-indicative items can significantly improve the accuracy of trait estimation (Brown & Maydeu-Olivares, 2011). The relative performance of trait estimators can thus be very different when negatively-loading items are involved.

Finally, while local independence can be engineered when developing items for cognitive tests, the multidimensional FC format can lead to local dependencies by design. When a FC block contains three or more items, the ranking responses are decomposed into pairwise comparisons, and structured local dependencies occur between pairs involving the same items. However, trait estimation for multidimensional

FC assessments tends to ignore this local independence violation, which may affect the accuracy of the various trait estimators to different degrees.

The presence of these special features means that current results and conclusions about the effectiveness of different trait estimators as established in cognitive assessments might not generalise to multidimensional FC personality assessments. A simulation study was conducted to address this knowledge gap, incorporating a variety of assessment designs or features that have been proven consequential for trait estimation (scale relationship, item bank composition, block size, test length) or are important for content reasons (scale plan, social desirability balancing).

Method

Simulation design

A simulation study was conducted to examine the stability and accuracy of the ML, WL, MAP and EAP estimators in FC assessments. This study examined FC assessments measuring four scales for two opposing reasons. On one hand, it would be desirable to investigate assessments with many scales, in order to reflect the realistic structures of multidimensional FC personality assessments. On the other hand, in order to include the EAP estimator in this study, it was computationally challenging to include five or more scales. This study thus chose to focus on FC assessments measuring four traits, labelled s_1, s_2, s_3 and s_4. In addition to varying the trait estimators, a number of assessment design factors considered to be important for FC assessments were also simulated.

Scale relationship (3 levels)

Correlations between scales had been shown to have an impact on model convergence and identification of the latent trait metrics under the TIRT model (Brown

& Maydeu-Olivares, 2011). In order to represent the different types of psychological constructs that may be measured by a multidimensional FC assessment, three levels of scale relationship were simulated (Table 9). Then, three multivariate normal samples with mean 0 and variance 1 across all scales were simulated, with correlation matrices as described. In order to capture the performance of trait estimators at extreme true scores, large samples of 10,000 simulees were created for each scale relationship level.

Table 9. Scale relationship levels

Level	Description
Unrelated	All scale correlations are zero, simulated by a 4×4 correlation matrix with all off-diagonal entries set to 0.
Positive	All scale correlations are positive, simulated by a 4×4 correlation matrix with all off-diagonal entries set to 0.5.
Mixed	Scale correlations could be positive or negative, simulated by a 4×4 correlation matrix with entries the same as those in the positive condition, but reversing signs of the correlations associated with scales s_2 and s_4.

Item bank composition (2 levels)

The presence of negatively-loading items had been shown to significantly improve the identification of the latent trait metrics in FC assessments (Brown & Maydeu-Olivares, 2011). In order to study trait estimation with different item bank compositions, two levels of positive item proportions were simulated (Table 10). The 100% positive item bank was simulated with item mean utility randomly sampled from Uniform [−3, 3], item factor loadings randomly sampled from Uniform [0.5, 1.5], and item unique variances randomly sampled from Uniform [0.5, 2.0]. Parameters for a total of 240 items (four scales with 60 items each) were simulated using these distributions. Item parameters for the 75% positive item bank were simulated by first simulating **another** 100% positive item bank, and then reversing the item loading directions with a

25% chance (as a result, the negatively-loading items do not necessarily distribute evenly across scales). The simulated item parameters are shown in Appendix E.

Table 10. Item bank composition levels

Level	Description
100% positive	All items had positive loadings.
75% positive	75% of items had positive loadings and 25% of items had negative loadings.

Then, the two simulated item banks were respectively assembled into FC assessments. The FC blocks were constrained to be strictly multidimensional (i.e., no two items within the same block would be measuring the same scale), and each block would contain at most one negatively-loading item. These content rules reflected common practices in FC personality assessments. And apart from the other content rules outlined in this study design, the assembly of items into FC blocks was completely random (i.e., with no consideration of information optimisation).

Block size (2 levels)

Table 11. Block size levels

Level	Description
Pairs	Each block consisted of two items, leading to one pairwise comparison and no local dependencies.
Triplets	Each block consisted of three items requiring a complete ranking response, leading to three pairwise comparisons with three correlated errors among them.

A multidimensional FC block involving more than two items results in multiple pairwise comparisons with correlated uniquenesses, and a simplifying assumption of local independence is often made while estimating person scores (Brown & Maydeu-

Olivares, 2011). In order to explore trait estimation in situations with and without the violation of local independence assumption, two block size levels were simulated (Table 11).

Scale plan (2 levels)

In a FC CAT, one has the option of pre-defining the scales to be measured by each block, or leaving that decision to the item selector. In order to examine whether a balanced but fixed scale plan could have an impact on trait estimator performance, two levels of scale plan were simulated (Table 12).

Table 12. Scale plan levels

Level	Description
Fixed	Balanced scale plans were derived by creating all unique combinations of four scales of the required size (i.e., six possible combinations for pairs, four possible combinations for triplets), ordering them manually so that the different scales were evenly positioned, and then cycling through the combinations until the desired assessment length was reached. Items were then assembled into FC blocks according to the fixed scale plan.
Dynamic	No scale plan was pre-defined. In a CAT, this would allow the item selector to choose items from any scale, thus prioritising information gain over content balancing. In this study, however, there was no consideration of information optimisation during assessment assembly, so this dynamic scale plan was completely random.

Social desirability balancing criteria (2 levels)

An important content rule in FC personality assessments is the matching of item social desirability within the same block. Using the item mean utility parameters (which followed a Uniform [−3,3] distribution) as a proxy for item social desirability, two levels of social desirability balancing were examined (Table 13).

Table 13. Social desirability balancing levels

Level	Description
Lenient	Item mean utilities in the same block could differ by up to 1.
Strict	Item mean utilities in the same block could differ by up to 0.5.

Test length (4 levels)

In order to explore the amount of shrinkage of Bayesian estimators in shorter tests, the assessment length was varied by truncating the assembled instruments, so that the shorter assessments were completely nested in the longer ones (Table 14). While the number of items per scale was used as the basis for studying the effect of test length, in the case of comparing assessments with different block sizes, the number of pairwise comparisons collected should be align instead. For example, a triplet assessment with 12 items per scale gives rise to 16 triplets and 48 pairwise comparisons in total, and therefore it should be compared to a pair assessment with 24 items per scale that also gives 48 pairwise comparisons in total.

Table 14. Test length levels

Level	Description
30 items per scale	All 60 pairs / 40 triplets.
24 items per scale	The first 48 pairs / 32 triplets.
18 items per scale	The first 36 pairs / 24 triplets.
12 items per scale	The first 24 pairs / 16 triplets.

Trait estimator (4 levels with sub-levels)

Simulated responses for all conditions were scored using the ML, WL, MAP and EAP trait estimators. The Bayesian scorings were conducted using multivariate normal

priors that matched the generating distributions of the simulated samples. In addition, to include the situation where it would be desirable to use a theoretically uncontroversial prior (e.g., McDonald, 1999), the Bayesian scorings were also repeated using the identity matrix as prior (Table 15). The ML, WL, and MAP estimations were conducted in R (R Core Team, 2015), using the multiroot function in the rootSolve library (Soetaert, 2009; Soetaert & Herman, 2009) to solve their respective score functions. The EAP scoring was conducted using nine quadrature points per dimension in Mplus (Muthén & Muthén, 1998-2012) by setting ESTIMATOR=ML, LINK=PROBIT and INTEGRATION=GAUSSHERMITE(9) under the ANALYSIS command.

Table 15. Prior options for Bayesian trait estimators

Prior	Description
Matching prior	The true scale correlations for sample generation matched the prior scale correlations in Bayesian scoring, mimicking practical situations where the scale correlations had been established robustly and could be used reliably to improve scoring accuracy.
Identity prior	The identity matrix was used as the prior scale correlations in Bayesian scoring, mimicking practical situations where the scale correlations were yet to be established, or when it was not desirable to take them into account in the calculation of assessment scores.

Analysis

Crossing the different levels of scale relationship and trait estimator gave rise to 16 conditions in total – three conditions each for ML and WL (corresponding to the three scale relationship levels), and five conditions each for MAP and EAP (corresponding to the three scale relationship levels combined with the choice of matching or identity priors, see Table 16). Crossing all seven design factors thus gave rise to a total of 16 (scale relationship and trait estimator) × 2 (item bank composition) ×

2 (block size) × 2 (scale plan) × 2 (social desirability balancing criteria) × 4 (test length) = 1024 conditions.

Table 16. Crossing different levels of scale relationship and trait estimator

		Scale correlations for Bayesian scoring		
		Unrelated	Positive	Mixed
Scale correlations for sample generation	Unrelated	Identity/Matching	-	-
	Positive	Identity	Matching	-
	Mixed	Identity	-	Matching

Following the simulation and scoring of all conditions, estimated trait scores were analysed to compare the performance of different trait estimators through four statistics:

- **Scoring failure rate**: the proportion of cases where the trait estimator failed to produce a valid score for whatever reason (e.g., ML estimates can be unbounded);
- **Score outlier rate**[8]: the proportion of cases where the estimated scores were outside [–5, 5] (e.g., ML estimates can have large biases);
- **Rank ordering**: the correlations between true and estimated scores for each scale;
- **Absolute differences**: root mean square errors (RMSE) of the differences between true and estimated scores.

Note that the cases with outlier scores on any scale were then excluded from the analysis of rank ordering and absolute differences, as their inclusion may jeopardise

[8] In operational assessments, extreme outliers would likely be capped.

these statistics when comparing the performance of trait estimators for typical score ranges.

Results

Scoring failure rate

Table 17. Scoring failure (cases per 10,000) by design factors (average across conditions)

Trait estimator		ML				WL			
Test length (items per scale)		12	18	24	30	12	18	24	30
Scale correl	Unrelated	0.44	0	0	0	1.13	1.63	0.88	0.56
	Positive	0.06	0	0	0	2.69	1.44	1.19	0.94
	Mixed	0.69	0.13	0.13	0	0.88	0.25	0.19	0.06
Item bank	100%	0.29	0	0.04	0	1.13	0.83	0.38	0.33
	75%	0.50	0.08	0.04	0	2.00	1.38	1.13	0.71
Block size	Pairs	0.79	0.08	0.08	0	2.58	1.96	1.29	1.04
	Triplets	0	0	0	0	0.54	0.25	0.21	0
Scale plan	Fixed	0.33	0.08	0.04	0	1.58	1.04	0.75	0.46
	Dynamic	0.46	0	0.04	0	1.54	1.17	0.75	0.58
Social desire	Lenient	0.21	0	0.04	0	1.50	0.88	1.00	0.67
	Strict	0.58	0.08	0.04	0	1.63	1.33	0.50	0.38

The first comparison concerned the proportion of cases where the trait estimator failed to return a score. The MAP and EAP estimators successfully produced scores for all cases at all test lengths regardless of the prior chosen. The ML estimator also converged for the majority of cases, with a small scoring failure rate of up to 0.06% for some conditions with shorter test lengths. The WL estimator also failed to return scores for up to 0.20% of cases in some conditions. It was surprising that, despite reducing bias compared to the ML estimator, the scoring failure rate of the WL estimator were usually

slightly higher (Table 17). Upon closer inspection of the scoring process, it was discovered that the WL estimator was failing to return a score for a different reason than the ML estimator – WL estimator calculations involved the inversion of the FIM, and a singular FIM would cause the WL estimation to fail to return a score. Unlike the case of cognitive assessments (Warm, 1989; Wang & Wang, 2001; Wang, 2015), the multidimensional FC question design might have led to an increased likelihood of encountering a singular FIM in one of the iterations to convergence, therefore leading to a small number of cases failing to receive a score using the WL estimator. As test length increased, scoring success of both ML and WL estimators also increased.

Score outlier rate

Table 18. Score outlier rate (% of cases) by design factors (average across conditions)

Trait estimator		ML				WL			
Test length (items per scale)		12	18	24	30	12	18	24	30
Scale correl	Unrelated	14.8	6.3	3.3	1.9	1.3	0.7	0.4	0.3
	Positive	10.4	3.6	1.6	0.8	1.4	0.7	0.3	0.2
	Mixed	16.3	7.1	3.6	2.2	1.9	0.9	0.6	0.4
Item bank	100%	17.0	7.5	4.0	2.4	2.6	1.3	0.8	0.5
	75%	10.7	3.9	1.7	0.9	0.4	0.2	0.1	0.1
Block size	Pairs	21.6	9.2	4.7	2.7	2.0	1.1	0.7	0.5
	Triplets	6.1	2.2	1.0	0.6	1.0	0.5	0.2	0.2
Scale plan	Fixed	13.7	5.8	3.0	1.7	1.3	0.7	0.4	0.3
	Dynamic	14.0	5.6	2.7	1.5	1.7	0.8	0.5	0.3
Social desire	Lenient	13.7	5.5	2.7	1.5	1.5	0.8	0.5	0.3
	Strict	14.0	5.9	3.0	1.7	1.5	0.7	0.4	0.3

The second comparison concerned the proportion of cases with estimated scores exceeding the [−5, 5] range in any of the four scales. Examination of the three simulated

samples of 10,000 simulees each showed that most true scores were within [–4, 4], and only one simulee had a true score exceeding [–5, 5] in one of the four scales. Among the four trait estimators, only ML and WL returned outlier scores exceeding [–5, 5]. As expected, the WL estimator produced fewer outliers than the ML estimator (Table 18). As test length increased, score outlier rates decreased for both ML and WL estimators. It was notable that, assessments using pairs tended to produce larger proportions of outliers than assessments using triplets when the number of items per scale was the same, likely due to the pair format resulting in fewer pairwise comparisons than the triplet format with the same total number of items. However, when the total number of pairwise comparisons was aligned (i.e., triplets with 12 items per scale and pairs with 24 items per scale both lead to 48 pairwise comparisons), pair conditions on average produced less outliers than triplet conditions, likely due to reduced information gain in triplets caused by local dependencies. Moreover, assessments with 100% positive items tended to produce larger proportions of outliers, confirming previous findings that negative items help the accurate estimation of trait scores (Brown & Maydeu-Olivares, 2011). In order to prevent extreme outliers from influencing the comparison of trait estimators for typical score ranges, all cases with outlier scores in any scale were removed from subsequent analysis.

Rank ordering

As assessments are often used to create merit lists of candidates, it is important to preserve the rank ordering of individuals on the traits being measured. The third comparison concerned the correlations between true and estimated scores. Even after the removal of outliers, the ML estimator still produced the lowest score correlations amongst all trait estimators investigated (Tables 19 and 20). The WL estimator produced very similar but slightly higher correlations than the ML estimator (with

differences up to 0.03). MAP scoring with identity prior tended to produce somewhat higher (with differences ranging from –0.01 to 0.12) correlations than the WL estimator. However, when the scales were all positively correlated or when negative items were present, the differences between MAP with identity prior and WL were small (with differences ranging from –0.01 to 0.01). When a matching prior was used, the MAP estimator produced even higher correlations (up to 0.04 higher than when the identity prior was used). The EAP estimator produced virtually identical results to MAP (with differences of magnitude up to 0.001). The differences between trait estimators were most prominent in shorter tests, but gradually reduced as the test lengthened. Full results by conditions are shown in Figures 7 to 12, which confirmed the general patterns observed from Tables 19 and 20.

Table 19. Score correlations by design factors (average across conditions) – ML, WL

Trait estimator		ML				WL			
Test length (items per scale)		12	18	24	30	12	18	24	30
Scale correl	Unrelated	.73	.80	.84	.86	.75	.81	.84	.87
	Positive	.73	.80	.84	.87	.75	.81	.85	.87
	Mixed	.69	.78	.83	.85	.72	.79	.83	.86
Item bank	100%	.61	.70	.76	.80	.64	.72	.77	.80
	75%	.82	.88	.91	.93	.84	.89	.91	.93
Block size	Pairs	.66	.75	.80	.83	.69	.76	.81	.84
	Triplets	.77	.83	.87	.89	.79	.84	.87	.89
Scale plan	Fixed	.73	.79	.84	.86	.75	.80	.84	.87
	Dynamic	.71	.79	.83	.86	.73	.80	.84	.87
Social desire	Lenient	.71	.79	.83	.86	.73	.80	.84	.86
	Strict	.72	.79	.84	.87	.74	.80	.84	.87

Table 20. Score correlations by design factors (average across conditions) – MAP, EAP*

Trait estimator		MAP - Matching Prior				MAP - Identity Prior			
Test length (items per scale)		12	18	24	30	12	18	24	30
Scale correl	Unrelated	.83	.87	.89	.91	.83	.87	.89	.91
	Positive	.78	.83	.86	.89	.75	.80	.84	.87
	Mixed	.86	.89	.91	.92	.84	.88	.90	.91
Item bank	100%	.78	.82	.85	.87	.76	.81	.84	.86
	75%	.87	.90	.92	.94	.85	.89	.92	.93
Block size	Pairs	.80	.84	.87	.89	.77	.82	.86	.88
	Triplets	.85	.89	.91	.92	.84	.88	.90	.91
Scale plan	Fixed	.83	.86	.89	.90	.81	.85	.88	.90
	Dynamic	.82	.86	.89	.91	.80	.85	.88	.90
Social desire	Lenient	.82	.86	.89	.90	.80	.85	.88	.89
	Strict	.83	.86	.89	.91	.81	.85	.88	.90

* Results for EAP differed by no more than 0.001.

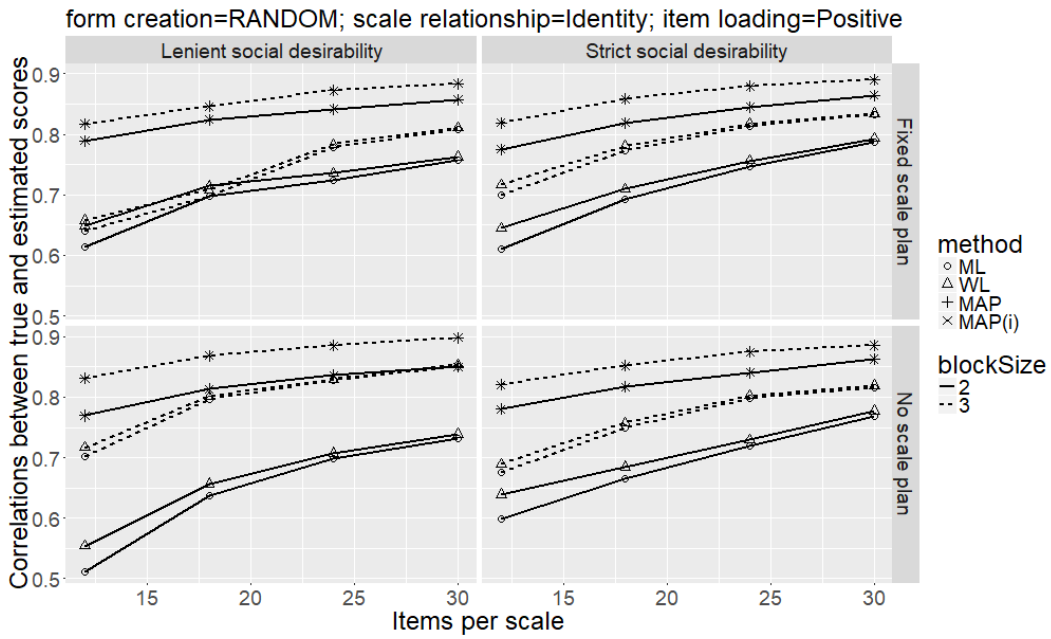


Figure 7. Score correlations – unrelated scales and 100% positive items

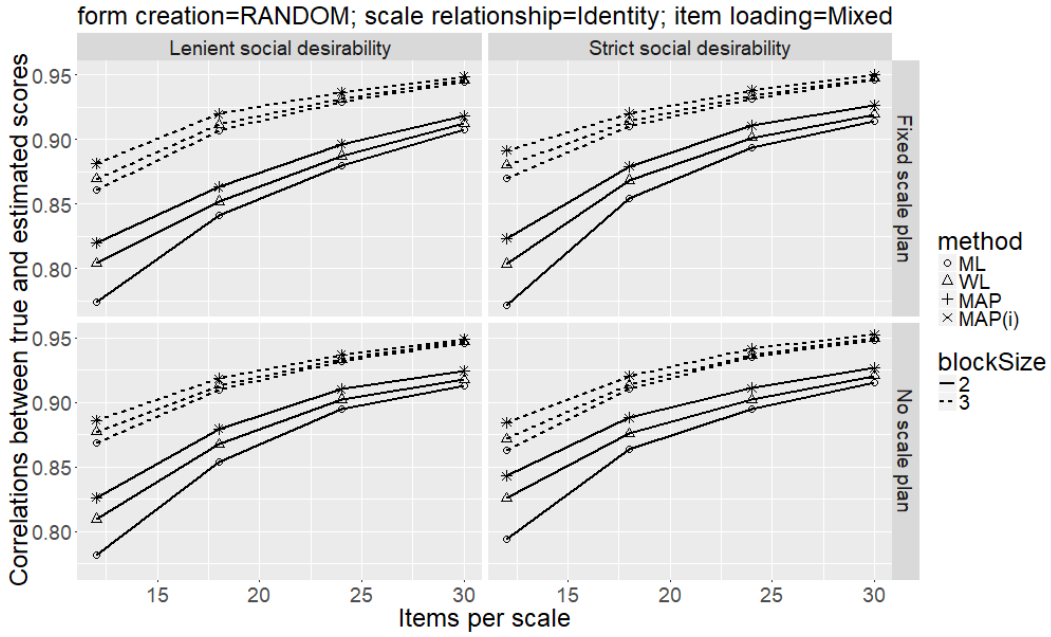


Figure 8. Score correlations – unrelated scales and 75% positive items

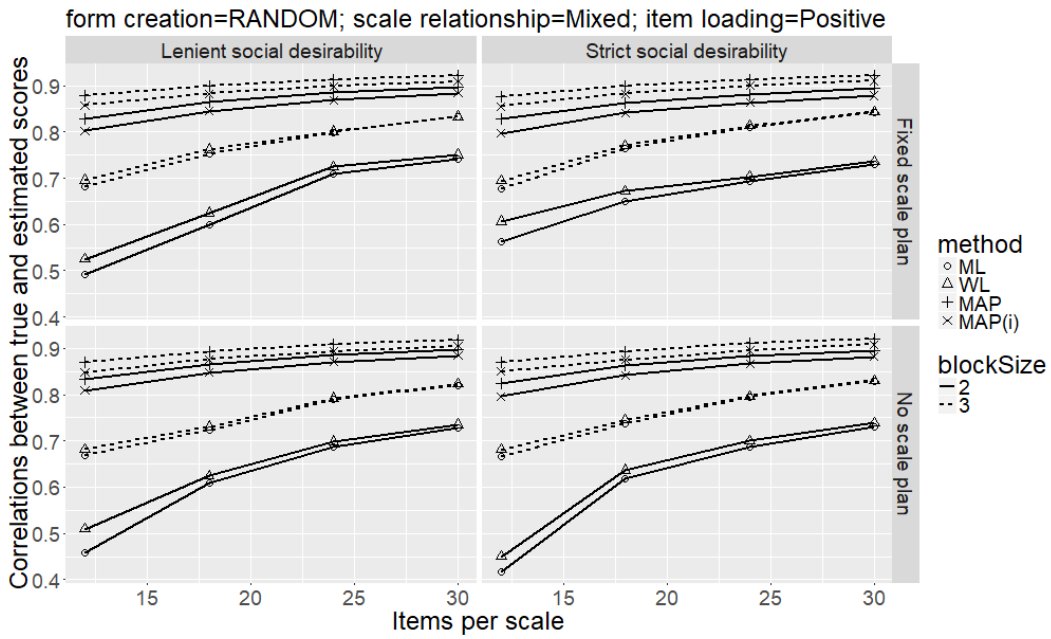


Figure 9. Score correlations – mixed scale correlations and 100% positive items

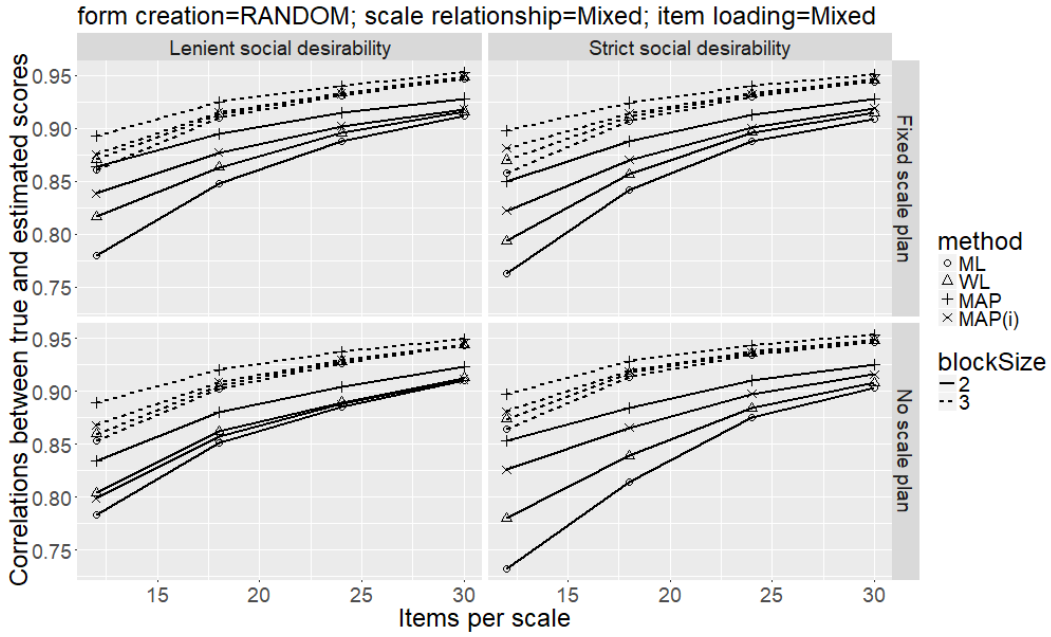


Figure 10. Score correlations – mixed scale correlations and 75% positive items

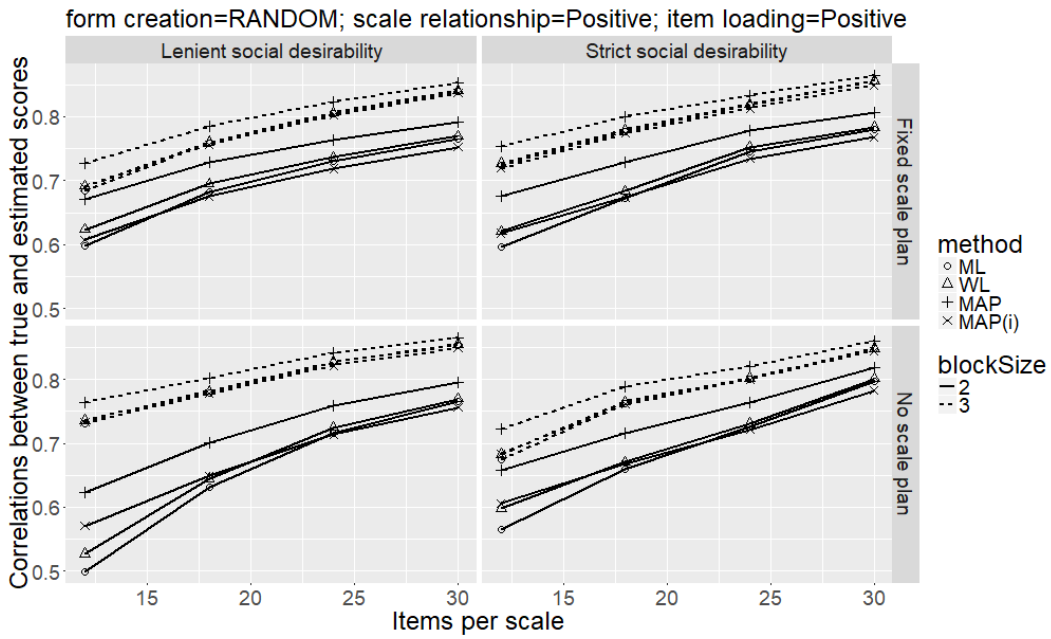


Figure 11. Score correlations – positive scale correlations and 100% positive items

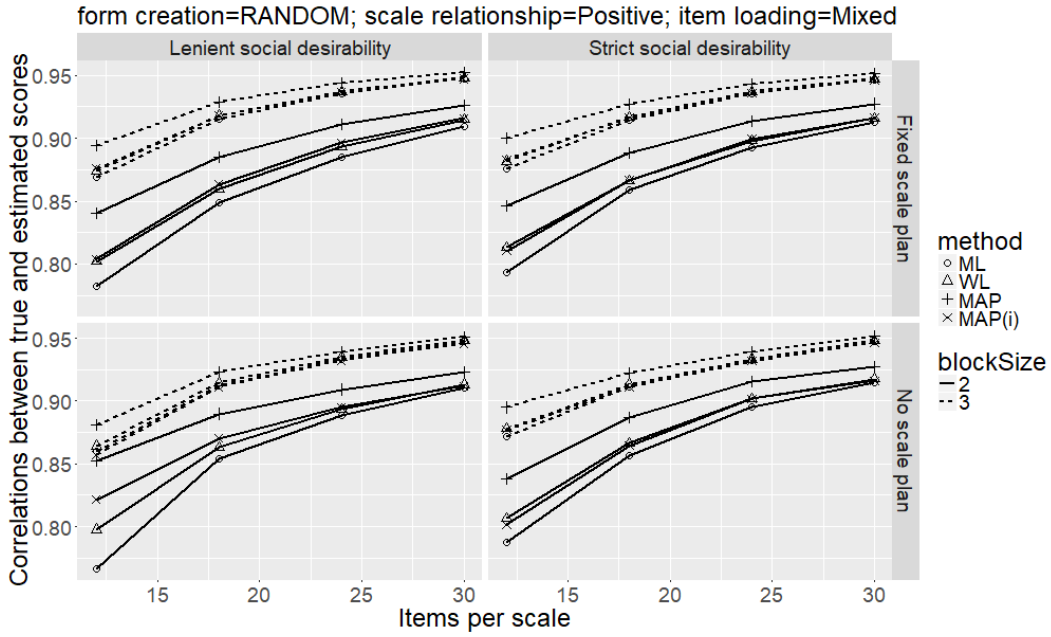


Figure 12. Score correlations – positive scale correlations and 75% positive items

Absolute differences

Apart from preserving the rank ordering of candidates, of equal practical importance is the minimisation of estimation error, which is key to accurate score norming and interpretation. The last comparison looked at absolute score estimation accuracy through the RMSEs between true and estimated scores. The performance ranking of trait estimators was the same as that based on true-estimated score correlations, with the ML estimator producing the largest RMSEs (Tables 21 and 22).

Table 21. RMSEs by design factors (average across conditions) – ML, WL

Trait estimator		ML				WL			
Test length (items per scale)		12	18	24	30	12	18	24	30
Scale correl	Unrelated	1.01	0.82	0.70	0.62	0.85	0.71	0.63	0.57
	Positive	1.03	0.83	0.69	0.60	0.87	0.72	0.62	0.55
	Mixed	1.07	0.86	0.72	0.64	0.92	0.75	0.65	0.58
Item bank	100%	1.29	1.06	0.90	0.80	1.14	0.95	0.82	0.74
	75%	0.78	0.61	0.51	0.44	0.63	0.51	0.44	0.39
Block size	Pairs	1.17	0.95	0.80	0.71	0.97	0.81	0.70	0.63
	Triplets	0.90	0.72	0.61	0.53	0.79	0.65	0.56	0.50
Scale plan	Fixed	1.01	0.83	0.70	0.62	0.86	0.72	0.63	0.56
	Dynamic	1.06	0.84	0.71	0.62	0.90	0.74	0.64	0.57
Social desire	Lenient	1.05	0.84	0.71	0.63	0.89	0.74	0.64	0.57
	Strict	1.03	0.82	0.70	0.61	0.87	0.72	0.63	0.56

In line with its theoretical rationale, the WL estimator produced notably lower RMSEs than the ML estimator (with differences ranging from 0.03 to 0.20). MAP with identity prior produced much lower RMSEs than the WL estimator (with differences ranging from 0.03 to 0.50). As in the results for true-estimated score correlations, when negative items were present, the differences between MAP with identity prior and WL were comparatively smaller (with differences ranging from 0.03 to 0.10). When a

matching prior was used, the MAP estimator produced marginally lower RMSEs than when the identity prior was used (with differences up to 0.04). The EAP estimator produced virtually identical results to MAP (with differences of magnitude up to 0.003). Similar to the results for rank ordering, the differences in RMSEs between trait estimators were most prominent in shorter tests, but gradually reduced as the test lengthened. Full results by conditions are shown in Figures 13 to 18, which confirmed the general patterns observed from Tables 21 and 22.

Table 22. RMSEs by design factors (average across conditions) – MAP, EAP*

Trait estimator		MAP - Matching Prior				MAP - Identity Prior			
Test length (items per scale)		12	18	24	30	12	18	24	30
Scale correl	Unrelated	0.55	0.49	0.45	0.42	0.55	0.49	0.45	0.42
	Positive	0.60	0.54	0.49	0.45	0.64	0.57	0.52	0.48
	Mixed	0.50	0.45	0.41	0.38	0.54	0.48	0.44	0.40
Item bank	100%	0.61	0.56	0.52	0.48	0.64	0.58	0.54	0.51
	75%	0.50	0.43	0.38	0.34	0.52	0.45	0.39	0.36
Block size	Pairs	0.59	0.53	0.48	0.45	0.62	0.56	0.51	0.47
	Triplets	0.51	0.45	0.41	0.38	0.54	0.47	0.43	0.39
Scale plan	Fixed	0.55	0.49	0.45	0.41	0.58	0.51	0.47	0.43
	Dynamic	0.56	0.49	0.45	0.41	0.58	0.51	0.47	0.43
Social desire	Lenient	0.56	0.49	0.45	0.42	0.58	0.52	0.47	0.44
	Strict	0.55	0.49	0.45	0.41	0.58	0.51	0.47	0.43

* Results for EAP differed by no more than 0.003.

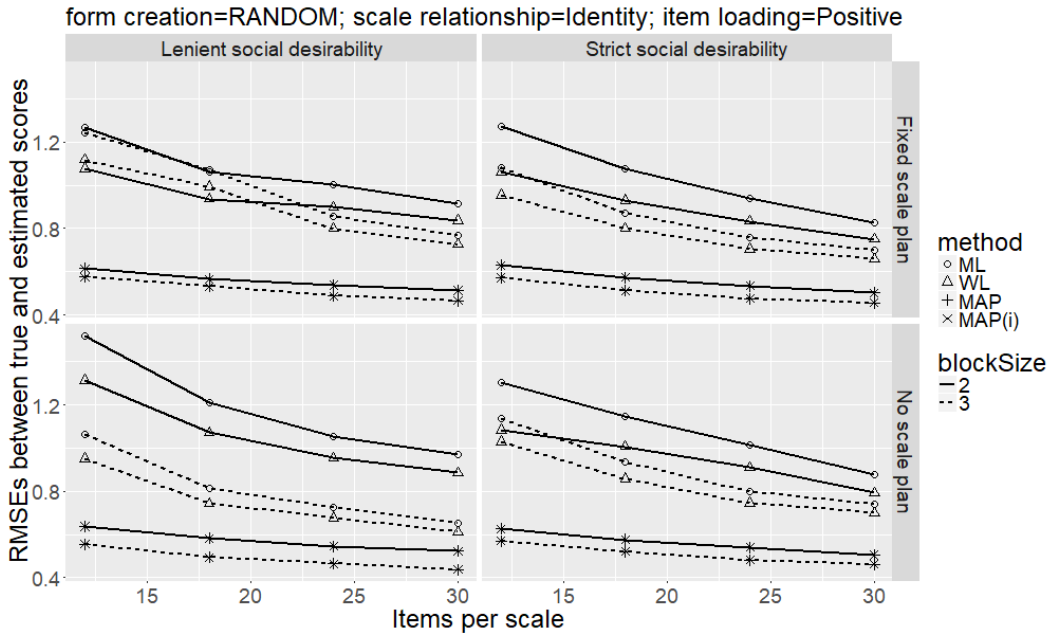


Figure 13. RMSEs – unrelated scales and 100% positive items

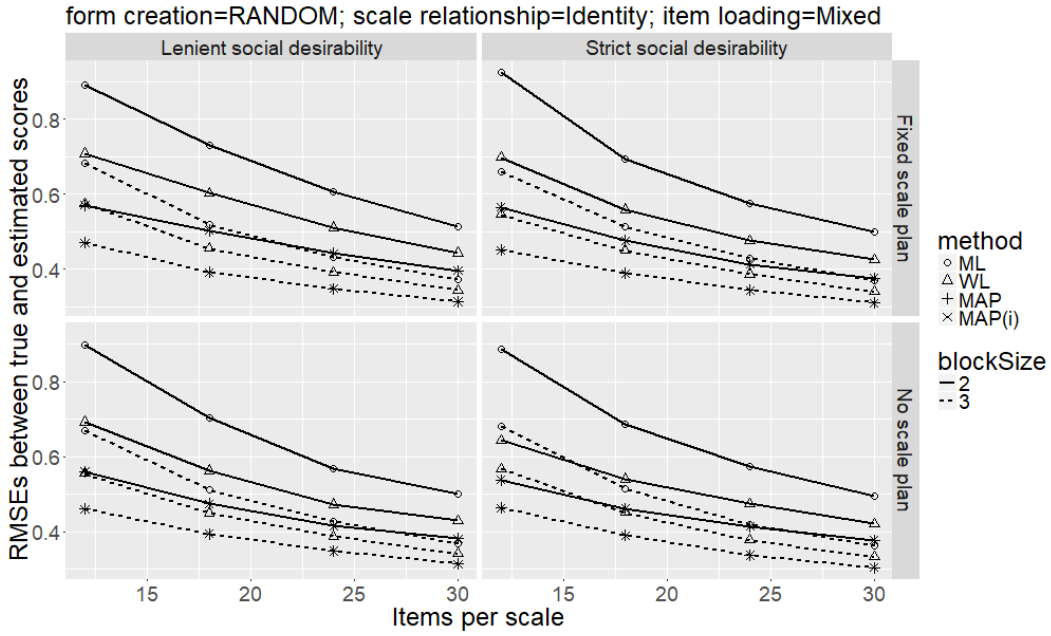


Figure 14. RMSEs – unrelated scales and 75% positive items

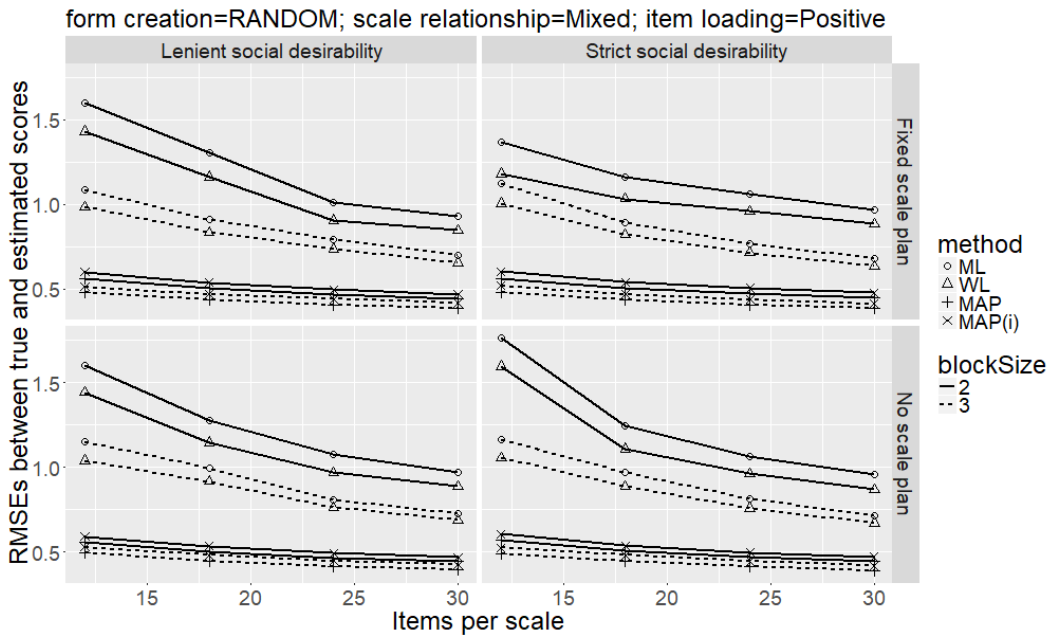


Figure 15. RMSEs – mixed scale correlations and 100% positive items

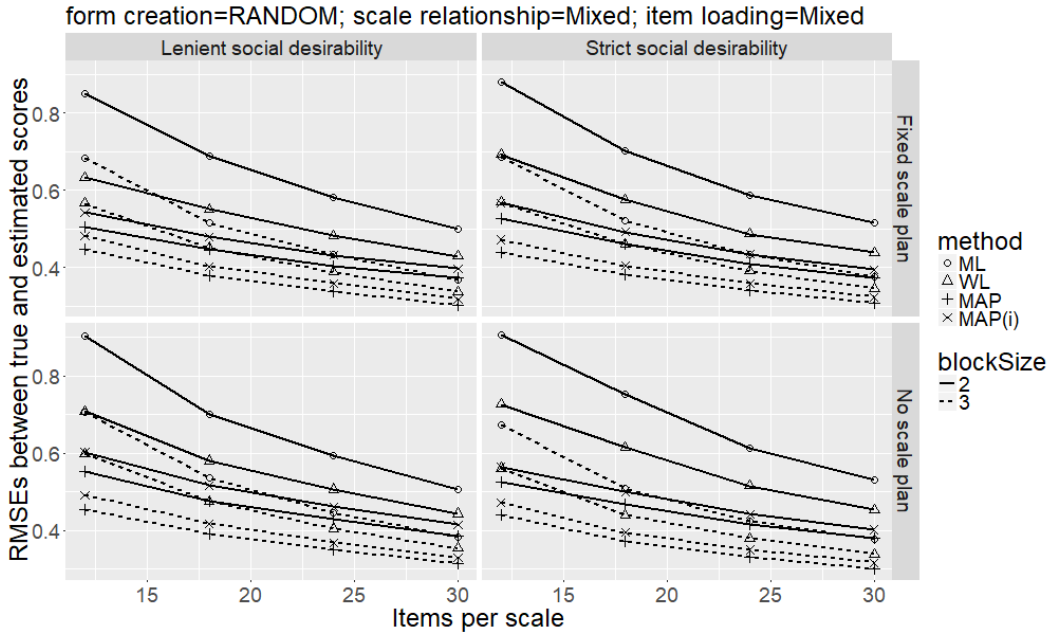


Figure 16. RMSEs – mixed scale correlations and 75% positive items

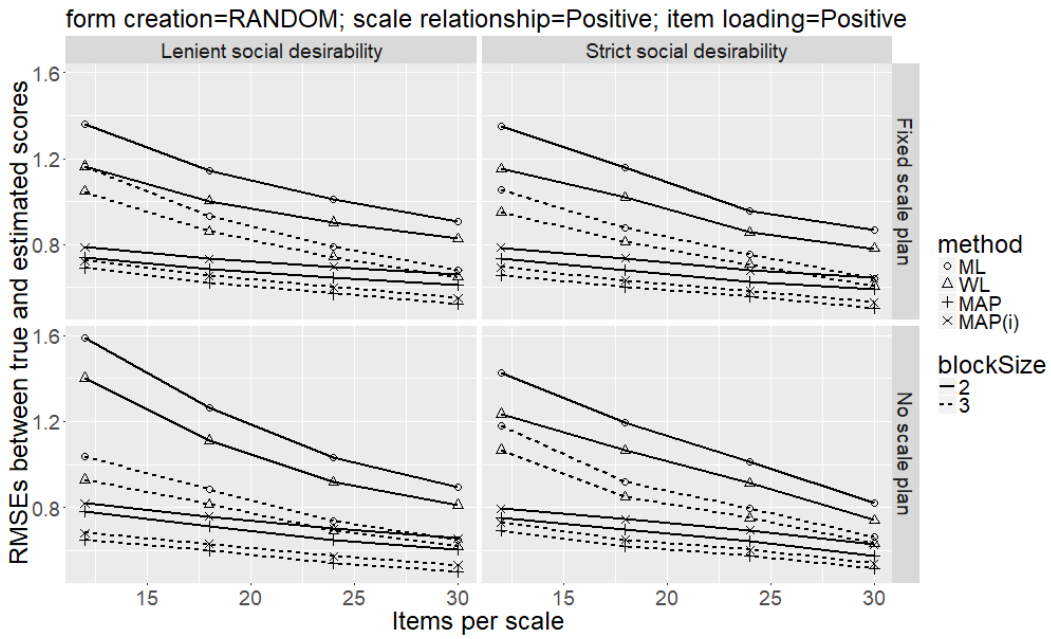


Figure 17. RMSEs – positive scale correlations and 100% positive items

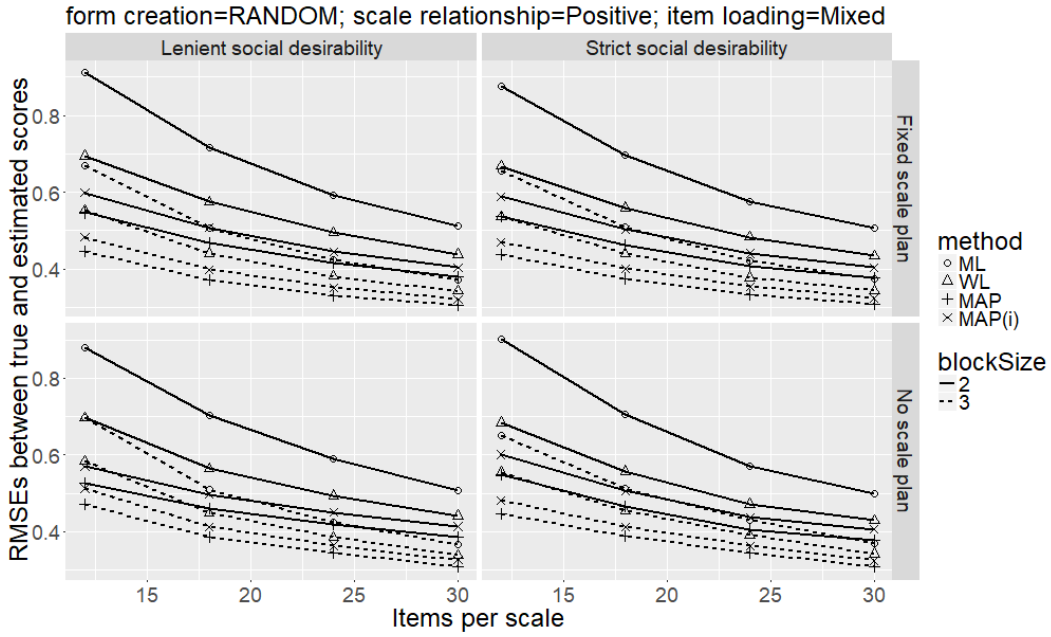


Figure 18. RMSEs – positive scale correlations and 75% positive items

Discussion

This simulation study examined the performance of the ML, WL, MAP and EAP estimators for scoring TIRT-based FC assessments. Across all conditions and after excluding outliers from the ML and WL scoring results, the Bayesian estimators (i.e., MAP and EAP) performed on par or significantly better than the non-Bayesian estimators (i.e., ML and WL). The Bayesian estimators produced no scoring failures, no outliers, and generally resulted in higher correlations between true and estimated scores as well as lower RMSEs. The performance differences between Bayesian and non-Bayesian estimators were particularly profound in shorter tests and in assessments using only positive items. The relative performance pattern and ranking of trait estimators were consistent across design conditions. The MAP and EAP estimators produced virtually identical results, and the choice of multivariate prior (scale correlation matrix) had only minor impact on the estimated scores, with results using a realistic scale correlation matrix slightly outperforming those using an identity matrix.

Contrary to prior findings that showed superior performance of the WL estimator across a number of IRT models (Warm, 1989; Wang & Wang, 2001; Wang, 2015), the WL estimator demonstrated notable weakness in the case of multidimensional FC assessments – it failed to produce scores for a proportion of respondents, which would be unacceptable in practice. Among the four trait estimators, only the WL method required inverting the FIM during its calculations, and so the WL estimator was unable to compute a person score when a singular FIM was encountered. As can be seen in Equations 24 and 25, the FIM for one pairwise comparison is always singular when more than two traits are being measured by the assessment, and it can take the summation of quite some pairwise comparisons before the total FIM finally becomes non-singular, as demonstrated by this simulation study. It may be possible to

adjust a singular FIM slightly during calculations, e.g., by using the nearPD function in R package lmf (Kvalnes, 2013) to make the FIM positive definite before attempting to invert it. Though such adjustments may compromise the scoring results in unpredictable ways and will add to the computational complexity of the WL estimator. Nevertheless, when the WL estimator successfully produced scores, as expected it outperformed the ML estimator in terms of estimation bias, giving notably lower proportion of outliers and resulted in lower RMSEs across all conditions. However, in terms of preserving the rank ordering of individuals, once outliers were removed, the WL and ML estimators performed very similarly.

The ML estimator was outperformed by MAP and EAP in every performance metric. Also, although to a lesser extent than the WL estimator, the ML estimator failed to converge and produce scores in a very small proportion of cases, therefore still rendering it unacceptable to use in practice. It also produced a large proportion of outliers at shorter test lengths, making it very undesirable for short assessments, or for early stages of a FC CAT.

Aside from the comparison of different trait estimators, results of this simulation study also confirmed earlier findings that using a mixture of positive and negative items tended to result in more accurate score estimation than using only positive items (Brown & Maydeu-Olivares, 2011), as demonstrated by higher correlations between true and estimated scores and lower RMSEs in the conditions with 75% positive items compared to the matching conditions with 100% positive items. Moreover, true scale correlations had only a small effect on trait estimation performance, with slightly worse results in the scenarios combining positive-only true scale correlations with a positively-only item bank, again echoing previous findings (Brown & Maydeu-Olivares, 2011).

Finally, when the number of pairwise comparisons was matched, pair assessments had higher score correlations and lower RMSEs than triplet assessments. This is because the local dependencies between the three pairwise comparisons in a triplet lead to less information being collected than in the case of three independent pairs. Therefore, smaller block sizes will perform better due to having less correlated errors between different pairwise comparison responses. However, collecting the same total number of pairwise comparisons from smaller blocks will take more items, thus requiring more resources during test development and longer responding times during assessment. But on the other hand, comparing three or more statements in FC blocks is more cognitively demanding than comparing one pair of statements at a time. The choice of block size for a FC assessment is ultimately a multi-faceted balancing art of maximising information gain per unit time while taking into account item properties, candidate backgrounds, and other settings and requirements of an assessment program.

Limitations

Firstly, the instruments in this study were almost randomly assembled – apart from the content rules (i.e., scale plan, social desirability balancing criteria, strictly multidimensional blocks, no more than one negative item per block), the placement of items into blocks were completely random. There was no consideration of optimal assessment design according to item characteristics. Operational assessments would almost certainly be designed better, with FC blocks carefully assembled and balanced in order to optimise information gain. As a result of this limitation, the assessments in the current study were much less efficient in score recovery than operational assessments of similar lengths. In other words, well-designed FC assessments would achieve higher true-estimated correlations as well as lower RMSEs than those seen in this study.

Nevertheless, the general trend of relative performance of different trait estimators as observed in this study would likely still hold in a more realistic assessment design.

Secondly, the item banks used in this study were simulated and the item parameters followed prescribed distributions. In reality, item bank compositions could vary significantly from one application to the next. As seen through the effects of negative items in this study, item bank composition could have a notable impact on the performance of trait estimators. Therefore, if item parameter distributions differed significantly from those used in this study, additional investigation might be needed so as to verify the choice of trait estimator with respect to the item bank in question.

Thirdly, this study only investigated strictly multidimensional FC block designs, i.e., where all pairwise comparisons consisted of items from different dimensions. However, FC assessments might instead adopt a mixed design involving both unidimensional and multidimensional comparisons. The incorporation of unidimensional comparisons was important for score estimation in FC assessments using the Multi-Unidimensional Pairwise-Preference model (MUPP; Stark, Chernyshenko & Drasgow, 2005). While unidimensional comparisons are not essential for the TIRT model, their existence could also have an effect on score estimation accuracy (Brown, 2016). Future studies might choose to quantify this effect on the different trait estimators.

Finally, this study assumed that each item measured one and only one dimension, i.e., there was no within-item multidimensionality. This assumption resulted in each item utility having only one non-zero loading, and therefore many trait estimation calculations were significantly simplified compared to the more general case involving within-item multidimensionality. However, the use of multidimensional items within FC blocks is rarely practical or desirable – good items measuring multiple dimensions

are difficult to develop and hard to calibrate accurately. Moreover, the response process involving the comparison of multiple multidimensional items within a single FC block can be significantly more cognitively complex and even confusing, and thus possibly giving rise to multiple response strategies that deviate from the simple comparison of item utilities as described by the TIRT model. Therefore, focusing on unidimensional items would be sufficient for most practical applications.

Comparing Item Selectors for FC CAT (Study 3)

The choice of item selector can have a significant impact on the efficiency of a CAT. As discussed in Study 2, TIRT-based FC personality assessments have several key features that reduce the generalisability of existing CAT research findings to them: high dimensionality with inseparable multidimensional design, item loadings in positive and negative directions, and by-design local dependencies in FC blocks involving more than two items. These special features can affect item selection as well as trait estimation. Another simulation study was thus conducted to examine the performance of the various item selectors for TIRT-based FC personality assessments.

Method

Simulation design

A simulation study was conducted to examine the efficiency of item selectors in FC CAT. Similar to Study 2, this study focused on assessments measuring four scales (labelled s_1 to s_4 respectively). Unlike Study 2, this study only explored FC assessments using pairs, which was the least computationally intensive and allowed the investigation of more conditions. Seeing the results from Study 2, the interim and final person scores were estimated using the MAP estimator with matching prior. A number of assessment

design factors considered to be important for FC CAT were also simulated. All simulations were conducted in R using code written specifically for this thesis.

Item selector (6 levels)

Six item selectors were simulated: RANDOM, WCI, A-, C-, D-, and T-optimality. The RANDOM item selector followed the content rules imposed on all CAT sessions, but otherwise chose items completely at random with no consideration of information optimisation, i.e., it did not adapt the assessment to the individual as a typical CAT would. The RANDOM item selector was introduced to provide a worst-case-scenario baseline. Indeed, CAT algorithmic research tended to adopt the RANDOM item selector as the baseline for comparison when illustrating the power of more advanced item selectors (Stark, 2011). However, in actual assessment practices, presenting items randomly without information considerations is rarely a realistic operational alternative to CAT. Therefore, as a more realistic baseline for comparison, the WCI item selector (equal weights across all scales) was included, representing the simplest (both methodologically and computationally) CAT setup. Then, A-, C- (targeting sum score across all scales), D- and T-optimality formed the focus of the investigation. An attempt was made to simulate the global information item selectors. However, their computational intensity turned out to be inhibitive when handling large numbers of pair combinations in a FC design, and therefore they were excluded from this simulation study. It would be desirable to re-visit these more advanced item selectors in the future, once computational power ceases to be a challenge.

Scale relationship (3 levels)

Three levels of scale relationship were simulated as per Study 2 (Table 9). Three multivariate normal samples with mean 0 and covariance matrices as specified were

simulated. In order to capture outliers but also enable the exploration of a large number of design conditions, a smaller sample size of 2,000 was chosen.

Item bank composition (2 levels)

Two levels of item bank composition were simulated as per Study 2 (Table 10). The assembly of items into FC assessments were determined by the item selectors while following content rules similar to Study 2 – blocks were strictly multidimensional and contain at most one negative item each.

Scale plan (2 levels)

Two levels of scale plan were simulated as per Study 2 (Table 12). With the introduction of adaptive item selectors, the dynamic scale plan represented the prioritisation of information gain during assessment assembly, whereas the fixed scale plan represented the prioritisation of content balancing considerations during assessment assembly.

Social desirability balancing criteria (2 levels)

Two levels of social desirability balancing were examined as per Study 2 (Table 13).

Test length (4 levels)

Four levels of test length were simulated as per Study 2 (Table 14 but only pairs). CAT sessions were simulated to reach the target test length of 60 pairs – the point at which half of the simulated items were administered, so that the adaptive item selection wasn't constrained towards the end due to small item bank sizes. Then, the CAT sessions were truncated to give shorter test lengths completely nested in the longer ones.

Analysis

Crossing the six design factors gave rise to a total of 6 (item selector) × 3 (scale relationship) × 2 (item bank composition) × 2 (scale plan) × 2 (social desirability balancing criteria) × 4 (test length) = 576 conditions (although only 576 ÷ 4 = 144 samples of 2,000 CAT sessions each needed simulating due to the nested test length design). Across all simulees in all conditions, the target assessment length of 60 pairs was reached successfully. In other words, the simulated item banks were deep enough and the content rules were not overly restrictive, so that the item selectors never failed to find a FC pair satisfying all content rules for the entire assessment length. Following the simulation, summary statistics were computed for each condition to quantify and compare the performance of different item selectors:

- **Rank ordering**: the correlations between true and estimated scores for each scale;
- **Absolute differences**: RMSE of the differences between true and estimated scores.

In order to summarise results across conditions, and to explore the interactions between design factors, cross-classified multilevel regressions were employed. The regression models were built on the scale-level summary statistics across conditions, i.e., on 576 conditions × 4 scales each = 2304 records. The performance statistics of true-estimated score correlations and RMSE were modelled as outcome variables, and the design factors were modelled as predictor variables. Moreover, in order to account for the dependencies between records (i.e., each of the 144 CAT session samples was generated under a unique combination of design factors, giving rise to 16 records corresponding to four nested test lengths with four scales each; the four scales also shared common settings across different CAT session samples), a cross-classified

multilevel structure (see Fielding & Goldstein, 2006) was incorporated, with CAT session samples and scales as grouping variables. Test length could also have been treated as a random effect in this design in order to control for the nesting structure within CAT session samples, but it was more useful for the purpose of this study to explore it as a design factor, and thus it was entered as a fixed effect.

A step-wise approach was adopted to arrive at the final list of significant fixed effects for each of the two outcome variables: first the base variance components model was built, followed by a model with all main effect terms for the design factors, followed by step-wise introduction of interaction terms and only retaining the ones with regression coefficients significantly different from zero. The final model was then interpreted to generate insight into how the item selectors performed under different design conditions. Details of the model setup for each outcome variable are discussed further in the Results section. The analysis was conducted in R: the cross-classified multilevel models were built using the lmer function in package lme4 (Bates, Maechler, Bolker, & Walker, 2015), with t-tests of fixed effect regression coefficient significance enabled by package lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017), and semi-partial correlations computed by package r2glmm (Jaeger, 2017) using Nakagawa and Schielzeth's (2013) approach.

Results

Descriptive statistics

The correlations and RMSEs between true and estimated scores were computed for each scale in each of the 576 conditions. Their distributions across conditions for each of the design factors are summarised in Table 23.

Table 23. True-estimated score correlations and RMSEs by design factors

Design factor	Level	Number of conditions	Correlation		RMSE	
			Mean	SD	Mean	SD
Item selector	RANDOM	384	.850	0.066	0.513	0.098
	WCI	384	.891	0.042	0.447	0.077
	A-optimality	384	.915	0.032	0.398	0.070
	C-optimality	384	.894	0.031	0.445	0.061
	D-optimality	384	.907	0.040	0.412	0.082
	T-optimality	384	.871	0.076	0.472	0.118
Scale relationship	Unrelated	768	.887	0.046	0.450	0.087
	Positive	768	.871	0.074	0.471	0.117
	Mixed	768	.905	0.031	0.423	0.065
Item bank	100%	1152	.859	0.059	0.501	0.088
	75%	1152	.917	0.031	0.395	0.067
Scale plan	Fixed	1152	.889	0.054	0.445	0.093
	Dynamic	1152	.886	0.057	0.450	0.096
Social desirability	Lenient	1152	.890	0.056	0.442	0.097
	Strict	1152	.885	0.054	0.454	0.092
Test length (items per scale)	12	576	.852	0.062	0.514	0.087
	18	576	.883	0.052	0.459	0.085
	24	576	.902	0.045	0.423	0.082
	30	576	.914	0.040	0.396	0.080
Scale	s_1	576	.891	0.050	0.445	0.088
	s_2	576	.885	0.053	0.449	0.090
	s_3	576	.897	0.050	0.430	0.087
	s_4	576	.878	0.065	0.468	0.107

Amongst the six item selectors investigated, A- and D-optimality achieved the best results on average, with A-optimality slightly outperforming D-optimality. These two item selectors were closely followed by C-optimality and WCI. T-optimality did not perform well, and RANDOM was the least effective item selector as expected. In

terms of scale relationship and item bank composition, results were in line with previous research (Brown & Maydeu-Olivares, 2011), with mixed scale correlations and mixed item loading directions achieving better results. The effects of scale plan and social desirability balancing criteria on the results were rather small. As expected, longer tests achieve better results. There were some small differences between the results across the four conceptually arbitrary scales, likely caused by random variations in simulated item content across the different scales.

Cross-classified multilevel regressions

Cross-classified multilevel regressions were used to model the variances of correlations and RMSEs between true and estimated scores from the different design factors and their interactions. In order to normalise the distribution of true-estimated score correlations for regression modelling, the Fisher-Z transformation was applied (Fisher, 1915). The transformed correlations displayed much less skewness and showed greater proximity to the normal distribution (Figures 19 and 20). The distribution of RMSEs already showed good proximity to the normal distribution (Figure 21), so no transformation was applied.

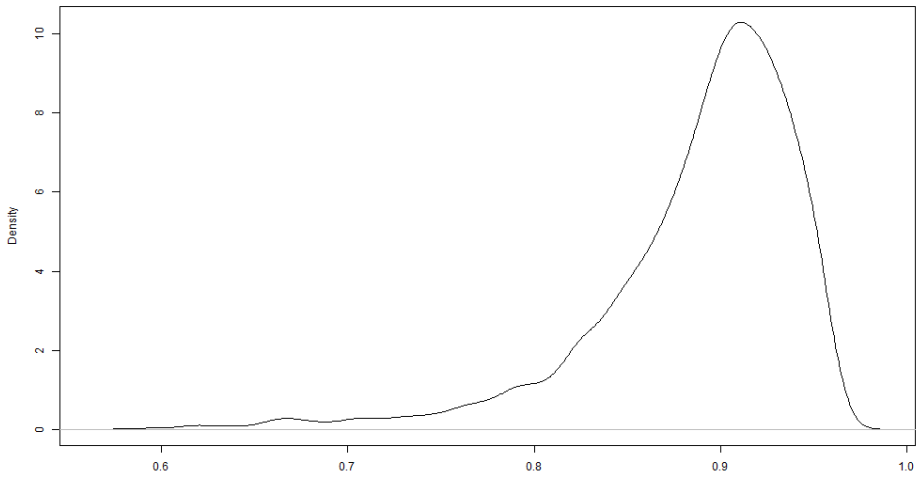


Figure 19. True-estimated correlations before Fisher-Z transformation

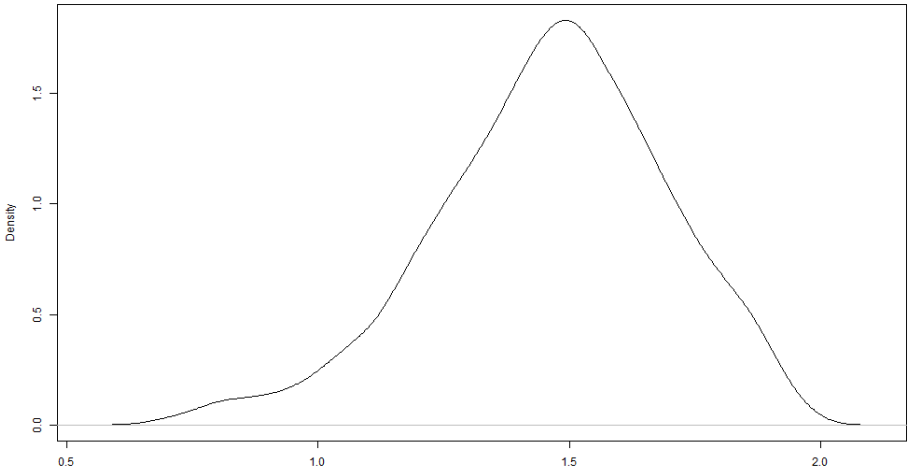


Figure 20. True-estimated correlations after Fisher-Z transformation

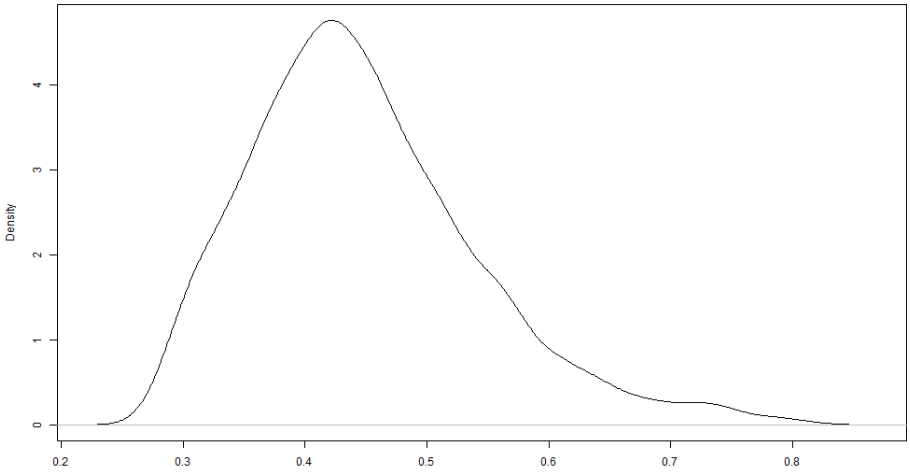


Figure 21. RMSEs with no transformation

The design factors' main effects and interactions were then explored in cross-classified multilevel models. The design factors were dummy coded into binary indicators for each level, with reference categories chosen as detailed in Table 24. The only exception was test length, which was treated as a continuous numerical variable. Test length displayed a largely linear relationship with the transformed true-estimated

score correlations (Figure 22) and RMSEs (Figure 23), and the square term was also included in the model to account for the small curvilinearity. As for the categorical factors, due to their full systematic crossing, the correlations between binary indicators for any two levels within the same design factor were a constant fully determined by the number of levels in that factor, with fewer levels leading to stronger negative correlations between binary indicators (Table 24). In order to test for potential multicollinearity, variance inflation factors (VIF; see Hair, Black, Babin, & Anderson, 2014) were computed for the binary indicators. Again, due to the systematic crossing of factors, the binary indicators for all levels within the same design factor had the same VIF values (Table 24). All VIF figures were low, indicating low likelihood for multicollinearity.

Table 24. Dummy-coding of design factors

Design factor	Reference category	Number of levels	Binary indicator correlations within design factor	VIF
Item selector	RANDOM	6	–.2	1.67
Scale relationship	Unrelated	3	–.5	1.33
Item bank	100% positive	2	N/A	1.00
Scale plan	Fixed	2	N/A	1.00
Social desirability	Lenient	2	N/A	1.00
Test length	N/A	4	N/A	1.00

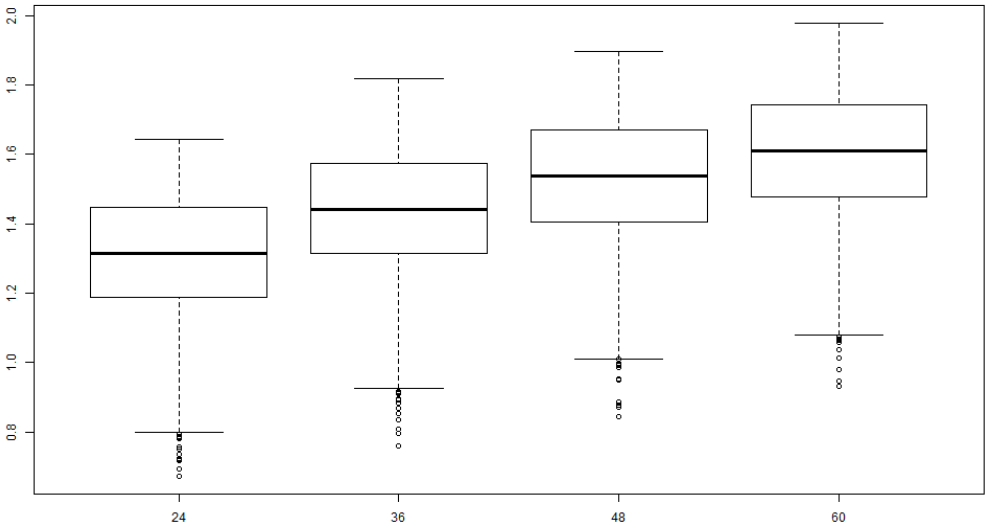


Figure 22. Transformed score correlations by test length (number of pairs)

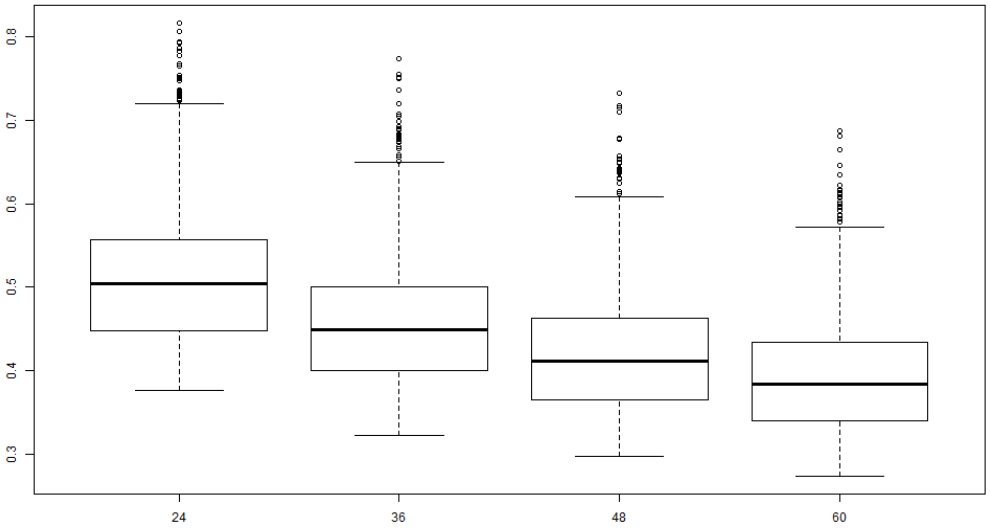


Figure 23. RMSEs by test length (number of pairs)

Following the preparation of outcome and predictor variables, cross-classified multilevel models were tested in a stepwise fashion for each of the two outcomes. As

RMSEs correlated with Fisher-Z transformed true-estimated correlations to −.994, it was not surprising that their model results were very similar (Tables 25 and 26).

The baseline variance components models showed that scales accounted for merely 2.0% and 2.6% of the variance in true-estimated score correlations and RMSEs respectively, representing the amount of random variations caused by different simulated item banks drawn from the same parameter distributions. Fixed effects were then added into the models. Not surprisingly, the models with only main effect terms found significant ($p < .05$) regression coefficients for predicting both outcomes from most of the design factors: test length, item selector, scale relationship, and item bank composition. Social desirability balancing criteria was merely marginally significant ($p = .07$) for predicting true-estimated score correlations and not significant ($p = .11$) for predicting RMSEs, while scale plan had no notable impact on either outcome. The large and varied simulated item banks likely provided sufficient content depth to counter the constraints from fixed scale plan and social desirability balancing. The effects of such content rules would likely become more apparent if the item banks were much smaller or the items had very similar parameters (i.e., effectively having a limited variety of items to choose from).

Then, interaction terms were added into the models. The regression coefficient for a binary level indicator within a categorical design factor could be interpreted as the mean difference in the outcome variable when comparing that particular level against the reference category (see Table 24). The regression coefficient for the test length main effect term could be interpreted as the slope when predicting the outcome variable using test length. And the regression coefficient for the interaction term between a design factor and test length captured the change in this slope caused by the design factor. When modelling Fisher-Z transformed true-estimated score correlations, if a design

factor was more efficient at the early stages of a test, it would increase score correlations at shorter test lengths more so than at longer test lengths, thereby reducing the slope for the test length term and resulting in a negative regression coefficient for the interaction term. Likewise, if the design factor was more effective at later stages of a test, score correlations would be boosted at longer test lengths more so than at shorter test lengths, leading to a steeper slope and a positive regression coefficient for the interaction term. When modelling RMSEs, the interaction terms could be interpreted in a similar way, but with reversed signs. Interaction terms between binary indicators were simpler to interpret. If two design factors worked well together, true-estimated correlations would be boosted (RMSEs would be reduced), and the interaction effect would be positive for predicting Fisher-Z transformed true-estimated score correlations (negative for predicting RMSEs). Likewise, if two design factors worked against each other, the coefficient for the interaction term would be negative for predicting true-estimated score correlations (positive for predicting RMSEs).

All possible two-way interactions between different design factors were explored one by one in the order shown in Tables 25 and 26. Interaction terms between all levels of two design factors were entered simultaneously at first, and the whole set was retained in the model if at least one of the levels had a regression coefficient that was at least marginally significant ($p < .10$), while the whole set was dropped if all interaction terms were insignificant ($p \geq .10$). Then, after exploring through all possible interactions, insignificant interactions for specific levels of design factors were removed until all remaining regression coefficients were at least marginally significant. The final models for the two outcome variables are presented in Tables 25 and 26.

Table 25. Cross-classified regression model with Fisher-Z-transformed true-estimated score correlations as outcome variable

	Baseline variance components only		Main effects only			Main effects and interactions		
Fixed effects	B	SE	B	SE	Semi-partial correlations	B	SE	Semi-partial correlations
(Intercept)	1.462***	0.024	0.672***	0.035		0.639***	0.028	
Test length			0.201***	0.007	.057	0.204***	0.007	.105
Test length^2			−0.015***	0.001	.016	–0.015***	0.001	.034
WCI			0.158***	0.029	.134	0.247***	0.019	.070
A-optimality			0.288***	0.029	.339	0.336***	0.019	.122
C-optimality			0.158***	0.029	.134	0.333***	0.032	.100
D-optimality			0.252***	0.029	.282	0.271***	0.023	.079
T-optimality			0.107***	0.029	.066	0.156***	0.025	.027
Scale correlation mixed			0.071***	0.02	.059	0.185***	0.018	.070
Scale correlation positive			−0.042*	0.02	.021	–0.095***	0.019	.019
Negative items			0.272***	0.017	.578	0.208***	0.019	.111
Dynamic scale plan			−0.011	0.017	.002	–0.003	0.01	.000
Strict social desirability			−0.030^	0.017	.016	–0.030**	0.009	.035

Fixed effects	Baseline variance components only		Main effects only			Main effects and interactions		
	B	SE	B	SE	Semi-partial correlations	B	SE	Semi-partial correlations
Test length × Scale correlation mixed						–0.004*	0.002	.001
Test length × Scale correlation positive						0.004^	0.002	.001
Test length × Negative items						0.026***	0.002	.032
Scale correlation mixed × Negative items						–0.178***	0.023	.173
Scale correlation positive × Negative items						0.094***	0.023	.055
Test length × WCI						–0.025***	0.003	.011
Test length × A-optimality						–0.014***	0.003	.003
Test length × C-optimality						–0.015***	0.003	.004
Test length × D-optimality						–0.015***	0.003	.004
Test length × T-optimality						–0.024***	0.003	.009
C-optimality × Scale correlation mixed						–0.057^	0.03	.012
C-optimality × Scale correlation positive						0.058^	0.031	.012
T-optimality × Scale correlation positive						–0.107***	0.027	.051
C-optimality × Negative items						–0.194***	0.026	.158
D-optimality × Negative items						0.065*	0.026	.020
T-optimality × Negative items						0.142***	0.026	.091
C-optimality × Dynamic scale plan						–0.049^	0.025	.013

Random effects	Baseline variance components only Variance	Main effects only Variance	Main effects and interactions Variance
CAT session sample	0.0384	0.0099	0.0029
Scale	0.0011	0.0011	0.0011
Residual	0.0155	0.0025	0.0022

Significance codes: < .001 '***'; .001-.01 '**'; .01-.05 '*'; .05-.1 '^'.
Test length was scaled to 1 = 6 items per scale.

Table 26. Cross-classified regression model with RMSEs as outcome variable

	Baseline variance components only		Main effects only			Main effects and interactions		
Fixed effects	B	SE	B	SE	Semi-partial correlations	B	SE	Semi-partial correlations
(Intercept)	0.448***	0.010	0.774***	0.015		0.798***	0.012	
Test length			−0.088***	0.003	.059	−0.096***	0.003	.125
Test length^2			0.007***	0.000	.020	0.007***	0.000	.042
WCI			−0.066***	0.012	.128	−0.095***	0.009	.055
A-optimality			−0.115***	0.012	.309	−0.139***	0.009	.111
C-optimality			−0.069***	0.012	.136	−0.137***	0.013	.093
D-optimality			−0.101***	0.012	.254	−0.126***	0.009	.093
T-optimality			−0.041**	0.012	.053	−0.056***	0.010	.019
Scale correlation mixed			−0.027**	0.009	.046	−0.077***	0.008	.066
Scale correlation positive			0.021*	0.009	.028	0.091***	0.010	.082
Negative items			−0.106***	0.007	.530	−0.106***	0.008	.154
Dynamic scale plan			0.005	0.007	.002	0.002	0.004	.001
Strict social desirability			0.012	0.007	.014	0.012**	0.004	.029

Fixed effects	Baseline variance components only		Main effects only			Main effects and interactions		
	B	SE	B	SE	Semi-partial correlations	B	SE	Semi-partial correlations
Test length × Scale correlation mixed						0.003**	0.001	.001
Test length × Scale correlation positive						−0.003**	0.001	.001
Test length × Negative items						−0.004***	0.001	.004
Scale correlation mixed × Negative items						0.073***	0.010	.162
Scale correlation positive × Negative items						−0.044***	0.010	.065
Test length × WCI						0.014***	0.001	.016
Test length × A-optimality						0.012***	0.001	.013
Test length × C-optimality						0.009***	0.001	.007
Test length × D-optimality						0.012***	0.001	.012
Test length × T-optimality						0.013***	0.001	.014
C-optimality × Scale correlation mixed						0.026*	0.013	.014
WCI × Scale correlation positive						−0.057***	0.012	.066
A-optimality × Scale correlation positive						−0.057***	0.012	.065
C-optimality × Scale correlation positive						−0.068***	0.014	.072
D-optimality × Scale correlation positive						−0.048***	0.012	.047
C-optimality × Negative items						0.082***	0.011	.164
T-optimality × Negative items						−0.059***	0.011	.090
C-optimality × Dynamic scale plan						0.019^	0.010	.011

Random effects	Baseline variance components only Variance	Main effects only Variance	Main effects and interactions Variance
CAT Session Sample	0.0063	0.0018	0.0005
Scale	0.0002	0.0002	0.0002
Residual	0.0025	0.0004	0.0004

Significance codes: <.001 '***'; .001-.01 '**'; .01-.05 '*'; .05-.1 '^'.

Test length was scaled to 1 = 6 items per scale.

The main effect terms were interpreted first. The test length variable was scaled to multiples of six items per scale (equivalent to 12 pairs), so that the different levels of the test length variable differed by one unit, and the square test length term wouldn't become too large for modelling. As expected, test length had a significant effect on score correlations ($B = 0.204$, $p < .001$) as well as RMSEs ($B = -0.096$, $p < .001$). The square terms of test length were also significant for both outcomes but had opposite signs to the main terms, indicating that the beneficial effect of increasing test length gradually diminished as the test converged towards the asymptote of perfect score recovery. With regards to scale relationship, it was found that having mixed scale correlations improved true-estimated score correlations ($B = 0.185$, $p < .001$) and reduced RMSEs ($B = -0.077$, $p < .001$) compared to when the scales were uncorrelated, which was in turn better than when all scales correlated positively ($B = -0.095$, $p < .001$ for score correlations; $B = 0.091$, $p < .001$ for RMSEs). Similarly, having some proportions of negatively-loading items also benefitted score recovery compared to when all items were in the positive direction ($B = 0.208$, $p < .001$ for score correlations; $B = -0.106$, $p < .001$ for RMSEs). These findings with regards to scale correlations and item loading directions were in line with previous findings (Brown & Maydeu-Olivares, 2011). It was also found that, after interaction terms were added, applying stricter social desirability balancing significantly reduced true-estimated score correlations ($B = -0.030$, $p = .002$) and increased RMSEs ($B = 0.012$, $p = .003$), which was not surprising because content rules reduce the number and variety of available FC blocks during the adaptive test construction process. Whether a scale plan was imposed, however, did not lead to any significant change in true-estimated score correlations or RMSEs. In terms of the effect of item selectors, the best methods for the baseline condition (i.e., unrelated scales, 100% positive item bank, fixed scale plan and lenient social desirability balancing) were A-optimality ($B = 0.336$, $p < .001$ for score correlations; $B = -0.139$, p

< .001 for RMSEs) and C-optimality (B = 0.333, p < .001 for score correlations; B = −0.137, p < .001 for RMSEs), followed by D-optimality (B = 0.271, p < .001 for score correlations; B = −0.126, p < .001 for RMSEs) and WCI (B = 0.247, p < .001 for score correlations; B = −0.095, p < .001 for RMSEs), while T-optimality was the worst method (B = 0.156, p < .001 for score correlations; B = −0.056, p < .001 for RMSEs) but still did significantly better than RANDOM as expected.

The interaction terms developed a more comprehensive picture of how the design factors might complement or work against each other. Interactions with test length were examined first. It was found that, especially for the earlier stages of an adaptive test, it was more beneficial to have mixed scale correlations than unrelated or positive scale relationships, or to apply a proper item selector instead of using RANDOM item selection. On the other hand, it was interesting to discover that, compared to when all items were positively-loading, the presence of negatively-loading items benefitted later stages of an adaptive test more so than the earlier stages. In TIRT score estimation, the comparison of items loading in the same direction contributed mainly to quantifying the differences between underlying scales, whereas the comparison of items loading in opposite directions contributed mainly to quantifying the sums of underlying scales (Brown & Maydeu-Olivares, 2011). When these two types of information were combined, the estimation of the true standings of scales were greatly improved, as demonstrated by the main effect term for item bank composition in the model. Thus, the phenomenon of negatively-loading items being even more effective at later stages of a CAT might have reflected the power of this second type of information in score estimation. In other words, it was beneficial to have negatively-loading items in general, and the benefit became more important at longer test lengths, because the second type of information desired by TIRT score estimation could not be effectively increased by adding more blocks where all items were positively-loading.

Nevertheless, while many interaction terms with test length were significant, their effects were all relatively small compared to the main effect terms, making them practically negligible especially for short tests.

Interactions between scale relationship and item bank composition were examined next. Brown and Maydeu-Olivares (2011) suggested that the presence of negative scale correlations worked in a similar way as having negatively-loading items in helping score recovery. So not surprisingly, their interaction terms were significant ($p < .001$). In order to study their combined effects, it was useful to combine the unstandardized regression coefficients of the main and interaction terms together (Table 27). Results showed that, while it was beneficial to have mixed scale correlations or negatively-loading items, their benefits didn't stack. In fact, the effect of scale relationship was only apparent when the item bank was 100% positive. When the item bank was 75% positive, scale relationship had very small influence on score recovery.

Table 27. Combined unstandardized regression coefficients of scale relationship and item bank composition

Model	Score correlations		RMSEs	
Item bank / Scale relationship	100% positive	75% positive	100% positive	75% positive
Unrelated	0	0.208	0	–0.106
Mixed	0.185	0.215	–0.077	–0.109
Positive	–0.095	0.207	0.091	–0.059

The models for the two outcome measures, however, diverged in the interactions between item selectors and item bank composition. When modelling score correlations, D-optimality ($B = 0.065$, $p = .015$) and T-optimality ($B = 0.142$, $p < .001$) were found to be more effective when negatively-loading items were present, but C-optimality ($B = –0.194$, $p < .001$) was a lot less effective when the item bank contained negatively-

loading items. When modelling RMSEs, no significant interaction was found with D-optimality, while T-optimality ($B = –0.059$, $p < .001$) and C-optimality ($B = 0.082$, $p < .001$) displayed similar preferences with regards to negatively-loading items as in the model for score correlations. To sum up: 1) C-optimality was more effective for an item bank where all items were positively-loading; 2) D-optimality and T-optimality were more effective when the item bank contained negatively-loading items; 3) WCI and A-Optimality displayed no interactions with item bank composition.

The interactions between item selectors and scale relationship demonstrated the greatest divergence between models. When modelling score correlations, only two item selectors had significant interaction terms: T-optimality was found to be less effective when all scale correlated positively ($B = –0.107$, $p < .001$), while C-optimality was marginally more effective for positively correlated scales ($B = 0.058$, $p = .063$) and marginally less effective for scales with mixed correlations ($B = –0.057$, $p = .064$). When modelling RMSEs, however, all but T-optimality had significant interaction terms: C-optimality was less effective for scales with mixed correlations ($B = 0.026$, $p = .042$), while WCI ($B = –0.057$, $p < .001$), A-optimality ($B = –0.057$, $p < .001$), C-optimality ($B = –0.068$, $p < .001$) and D-optimality ($B = –0.048$, $p < .001$) all worked more effectively for positively correlated scales. This finding was very interesting, and could be distilled down to two main observations: 1) regardless of outcome, C-optimality preferred having positively correlated scales over unrelated scales, and preferred unrelated scales over scales with mixed correlations; 2) for WCI, A-, D- and T-optimality, whether having positively correlated scales was beneficial depended on the outcome measure – they could reduce the effectiveness of T-optimality in estimating the rank ordering of people, but they could also enhance the effectiveness of WCI, A-, C- or D-optimality in reducing RMSEs.

Table 28. Combined unstandardized regression coefficients of scale relationship, item bank composition, and item selector for predicting true-estimated score correlations

Scale correlation	Positive items	Item Selector					
		RAN	WCI	A-opti	C-opti	D-opti	T-opti
Unrelated	100%	0.000	0.247	**0.336**	**0.333**	0.271	0.156
Unrelated	75%	0.208	0.455	**0.544**	0.347	**0.544**	0.506
Mixed	100%	0.185	0.432	**0.521**	0.461	0.456	0.341
Mixed	75%	0.215	0.462	**0.551**	0.297	**0.551**	0.513
Positive	100%	–0.095	0.152	0.241	**0.296**	0.176	–0.046
Positive	75%	0.207	0.454	**0.543**	0.404	**0.543**	0.398

Table 29. Combined unstandardized regression coefficients of scale relationship, item bank composition, and item selector for predicting RMSEs

Scale correlation	Positive items	Item Selector					
		RAN	WCI	A-opti	C-opti	D-opti	T-opti
Unrelated	100%	0.000	–0.095	–**0.139**	–**0.137**	–0.126	–0.056
Unrelated	75%	–0.106	–0.200	–0.245	–0.160	–**0.290**	–0.161
Mixed	100%	–0.077	–0.172	–**0.216**	–0.188	–0.203	–0.133
Mixed	75%	–0.109	–0.204	–0.248	–0.138	–**0.294**	–0.165
Positive	100%	0.091	–0.061	–**0.106**	–**0.114**	–0.083	0.035
Positive	75%	–0.059	–0.211	–0.255	–0.181	–**0.291**	–0.115

In order to gain a better overall understanding of the composite effects of item selectors, scale relationship, and item bank composition, their regression coefficients (both main effects and interactions) were combined and summarised for each of the two outcomes (Tables 28 and 29). In general, A- and D-optimality appeared to be most optimal. For preserving the rank ordering of people, A-optimality was the best for almost all settings, with D-optimality performing equally well when there were negatively-loading items. For reducing RMSEs, A-optimality appeared to be more

robust when all items were positively-loading, while D-optimality made better use of the presence of negatively-loading items. It was noteworthy that C-optimality with a unit-weighted sum target performed on par or better than A-optimality when the item bank was 100% positive and the scale correlations were non-negative, although this pattern might change when a different target was of interest. Across all conditions, WCI or T-optimality were consistently outperformed by some other item selectors, and RANDOM was the least effective as expected.

Discussion

This simulation study examined the performance of item selectors for TIRT-based FC CAT. A number of notable results were uncovered. First, C-optimality resulted in the largest number of interactions for predicting true-estimated score correlations as well as RMSEs: it was more effective with positively correlated scales and less effective with mixed scale correlations; it was less effective when the item bank contained negatively-loading items; and it was less effective when there were no scale plans. It was interesting that the directions of interaction effects with C-optimality were sometimes opposite to those seen in other item selectors, making it somewhat unique among them. C-optimality was designed to minimise the error variance of a particular linear combination of scale scores, which was set to the sum of all scale scores in this simulation study. The results for C-optimality in this study might be specific to the alignment between this target linear combination and the design factors. For instance, the sum of positively correlated scales would be more stable than the sum of scales with no or mixed correlations, which likely caused the interactions between C-optimality and scale relationship. Likewise, the adoption of a scale plan might have resulted in more balanced measurement across scales, leading to a sum that was marginally more stable. The interaction between C-optimality and item bank composition had a strong effect in

the prediction of both true-estimated score correlations and RMSEs, with the presence of negatively-loading items greatly reducing the effectiveness of C-optimality. This finding was somewhat surprising, given that the comparison of items with opposite loading directions contributed mainly to measuring the sum of scales (Brown & Maydeu-Olivares, 2011), which seemed to be in line with the goal of C-optimality. However, C-optimality has been found to "prefers items with discrimination parameters that reflect the weights of importance in the composite ability" (Mulder & Van der Linden, 2009). In the case of this study where all scales were assigned the same weight, C-optimality might therefore prefer to pair up items with similar factor loadings, i.e., two positive items or two negative items. The introduction of negatively-loading items might have reduced the availability of item pairs with similar loadings than when all items were positively-loading, thus potentially reducing the effectiveness of C-optimality. In order to further the understanding of C-optimality, it would be desirable to explore how it would function when different linear combination targets were applied.

Second, while the RANDOM item selector was used as the non-adaptive baseline, the WCI item selector represented the simplest adaptive baseline for comparison. In other words, the differences in performance between RANDOM and WCI could be viewed as the incremental gain due to adaptiveness of item selection, while the differences between WCI and other item selectors quantified the incremental gain achieved by better designs of the item selectors. Results generally showed bigger differences between RANDOM and WCI than between WCI and other item selectors, representing notable benefits of CAT even with a relatively simple item selector, which could be further enhanced with an item selector most suited to the situation. Interestingly, WCI actually outperformed T-optimality most of the time (Tables 28 and 29). In fact, T-optimality was always outperformed by A- and D-optimality (Tables 28 and 29), so the use of T-optimality would not be recommended in general.

Third, when operating within the design boundaries of this simulation study, A- and D-optimality demonstrated the greatest success in terms of increasing true-estimated score correlations as well as reducing RMSEs across most conditions, which was in line with findings from previous research (Mulder & van der Linden, 2009). For the recovery of rank ordering between individuals, A-optimality was more powerful, always matching or outperforming D-optimality. For the reduction of RMSEs, A-optimality was superior when working with only positively-loading items, while D-optimality was superior after negatively-loading items were introduced. In other words, A-optimality appeared to be an all-rounder that was minimally influenced by other design factors, while D-optimality was good for specific settings. Exactly which of them would work better would likely depend on the characteristics of the specific item bank and the psychological constructs being measured.

Last but not least, the good performance of A-optimality was contrary to two earlier, preliminary research studies on TIRT-based FC CAT, which found it to be less desirable than D-optimality (Brown, 2012; Lin & Brown, 2015). These preliminary studies differed from the current study in three main aspects: 1) real, limited item banks were used, with varying item parameter distributions across scales, as opposed to large simulated item banks in this study with relatively similar item parameter distributions across scales; 2) the preliminary studies allowed an item to appear multiple times to the same respondent when combined with different items into different FC blocks, whereas the current study allowed an item to appear only once; 3) the preliminary studies were conducted using existing CAT software (i.e., the MAT package in R; Choi & King, 2014), while the current study was conducted using codes written specifically for TIRT-based FC CAT, allowing the incorporation of content rules such as scale plan and social desirability balancing. Given these discrepancies with earlier studies, and in order to further understand the interactions between item bank characteristics and item selectors,

this simulation study was replicated on a real item bank developed as part of this thesis (see Chapter 4) – the scale relationship and item bank composition would be fixed, but scale plan, social desirability balancing criteria and test length could be varied alongside item selectors.

Limitations

First, limited by computational power, this study only managed to explore five item selectors (WCI, A-, C-, D- and T-optimality) with basic settings (i.e., WCI with equal weights across all scales, C-optimality targeting sum across all scales). Regrettably, although the global information item selectors showed good potential, it was not computationally feasible to simulate them for a four-dimensional FC CAT with a pool of 240 items. Future research may choose to develop simpler approximations for the global information measures to enable their use in assessments with high dimensionality.

Second, limited by scope, this study only explored FC pairs. Larger blocks were not simulated, which might show interesting new dynamics, e.g., social desirability balancing may become more restrictive with more items needing to fit into the same block. In order to explore larger FC blocks more efficiently, future research should consider how the computational intensity of selecting larger blocks could be minimised. In the current study, 240 items led to 28,680 unique pairs, which was still manageable after applying content rules. However, the same item bank would result in a total of 2,275,280 unique triplets, thus greatly increasing the computational complexity of the item selection process.

Finally, the scale and item characteristics in this study were all simulated and can be somewhat unrealistic. However, in order to address this concern, this simulation study was replicated on a real item bank in Chapter 4.

Conclusions

The measurement accuracy and efficiency of a CAT is highly dependent on the underlying automated test assembly algorithm. Optimising the design of this algorithm is therefore crucial. To shed light on this theoretically and computationally complex problem, this chapter conducted a review of key algorithmic components and formulated them for use in TIRT-based FC CAT. Furthermore, two extensive simulation studies compared the performance of trait estimators and item selectors, providing baseline guidance for design decisions in practice.

In terms of trait estimators, the ML estimator is not recommended due to its tendency to produce outliers in shorter tests as well as its risk of non-convergence, and the WL estimator is not recommended due to its potential scoring failures caused by singular FIMs. The Bayesian estimators (MAP or EAP) are recommended as the scoring method for FC assessments using the TIRT model, especially when the item bank only contains positive items, and/or when the assessment is short (including at the beginning of a CAT session, where interim score estimates are based on a limited number of responses and are key for driving the adaptive item selection process forward). Moreover, an informative prior will add to the power of Bayesian score recovery, but if in doubt, the identity prior can be adopted without losing too much estimation accuracy. The MAP and EAP estimators performed very similarly across all conditions, and therefore the choice between them is largely dependent on available software and computational efficiency. As typical FC personality assessments have at least five traits, MAP is usually more computationally efficient.

In terms of item selectors, A-optimality appears to be a good default choice that performs well across all conditions. Moreover, D-optimality may slightly outperform A-optimality when there are negative items, while C-optimality may also have a special

role depending on the match between the target linear combination and the characteristics of the scales and items. On the other hand, WCI and T-optimality are not recommended. The range of global information item selectors may outperform those investigated, but the computational power requirement was inhibitive due to the high dimensionality of personality constructs and the combinatorics complexity of FC blocks. With quick item selection run-time being essential for minimising the wait between the submission of a response and the presentation of the next question, global information item selectors remain out of reach for delivering a seamless adaptive assessment experience.

CHAPTER 4: DEVELOPING AN ADAPTIVE FC PERSONALITY ASSESSMENT

The motivation to refine algorithm designs for FC CAT was so that the measurement efficiency of FC personality assessments could be improved, leading to quicker and fairer people-related decisions in practice. However, the design of the FC CAT algorithm is only one aspect of an operational personality assessment. As an analogy, for a vehicle to reach its destination, it requires a powerful engine (the FC CAT algorithm), sufficient amount of fuel (the item bank), adequate driver steering controls (the computerised assessment delivery platform), and a map of the terrain (the psychological constructs being measured).

In order to study FC CAT methodologies in empirical practice, the last part of this thesis focused on developing a simple but operational adaptive FC personality assessment. This chapter is structured as follows. First, a model of personality is described to provide a content map for item development. Second, the development of an item bank, including empirical trialling and analysis to establish TIRT item parameters, is detailed (Study 4). Third, in order to confirm the final CAT algorithm design to use with the new item bank, a simulation study examining item selector performance is reported (Studies 5 and 5b). Fourth, empirical trial results and participant reactions of the newly developed adaptive FC personality assessment are documented (Study 6). Finally, conclusions and practical recommendations are presented.

The HEXACO Model of Personality

For decades, researches have studied the structure of personality through lexicon research. According to Lee and Ashton (2008), “the personality lexicon captures those aspects of personality that are sufficiently useful in person description to have been encoded as adjectives by generations of speakers within a given language community”,

and “fundamental personality dimensions… should be expressed within the personality lexicon by some large set of related adjectives that convey nuances and subtle variations in the expression of those dimensions.” Early research led to the Five Factor Model of personality, also known as the Big Five model or the OCEAN model, comprising of the dimensions of Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (e.g., Digman & Takemoto-Chock, 1981; Goldberg, 1990; McCrae & John, 1992; Norman, 1963; Tupes & Christal, 1961, 1992). While the Big Five model had been well-established and widely-used by the 1990s, more recent lexicon research suggested an alternative structure. The HEXACO model of personality (Ashton et al., 2004; Lee & Ashton, 2008) comprises of six dimensions, namely Honesty-Humility (H), Emotionality (E), eXtraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). The notable difference with the Big Five model is the emergence of Honesty-Humility as a separate factor in HEXACO. Full descriptions of the HEXACO factors are given in Table 30 (Lee & Ashton, 2009a).

In terms of construct validity, the HEXACO personality structure has been replicated in lexicon studies across multiple languages, including but not limited to: English (Ashton et al., 2006; Ashton, Lee, & Goldberg, 2004; Lee & Ashton, 2008), Dutch (Ashton et al., 2006; De Raad, 1992), French (Boies, Lee, Ashton, Pascal, Nicol, 2001), German (Ashton, Lee, Marcus, & De Vries, 2007), Greek (Lee & Ashton, 2009b), Hungarian (De Raad & Szirmak, 1994), Italian (Ashton et al., 2004; Ashton et al., 2006), Korean (Ashton et al., 2004; Hahn, Lee, & Ashton, 1999), Polish (Szarota, Ashton, & Lee, 2007) and Turkish (Wasti, Lee, Ashton, & Somer, 2008). In terms of criterion-related validity, McAbee, Casillas, Way and Guo (2019) summarised a large number of studies and concluded that the HEXACO personality factors have good utility in the prediction of educational and occupational outcomes. In particular, the Honesty-Humility factor demonstrated consistent predictive validity for

counterproductive student behaviours, cheating behaviours, organisational citizenship behaviours, and counterproductive work behaviours. Given its structural stability and predictive utility, this thesis thus adopted the HEXACO model of personality as the construct model in the development of the adaptive FC personality assessment.

Table 30. HEXACO personality factors

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

Item Bank Development (Study 4)

This study developed an item bank for measuring the HEXACO personality traits in a FC CAT. While personality assessments tend to use statements as items, this study instead focused on adjectives. Adjectives capture simple concepts which can be semantically compared in a FC question format. The concise nature of adjectives makes the comparative judgement process in FC questions cognitively simpler than if statements were utilised instead. Moreover, the faster comprehension and completion speed of FC adjective questions also allows quicker question progression, which helps to capitalise on the potential of adaptive testing. Finally, in terms of cross-cultural measurement, the factor structure of adjectives appear to be universal, as demonstrated by the lexicon studies that gave rise to the HEXACO model across multiple languages (e.g., see Lee & Ashton, 2008). Therefore, this study focused on building an item bank of adjectives.

Item Development

Item development started by finding a list of frequently used adjectives in the English language that would be suitable for describing personality characteristics. Lee and Ashton (2008) conducted a usage frequency rating study on a list of 1,710 adjectives from Goldberg (1982), reducing it to a subset of 449 "most familiar English personality-descriptive adjectives". For this study, Lee and Ashton's (2008) list was refined further based on the adjectives' suitability for use in self-rating FC personality questionnaires, leading to the removal of 119 items for a variety of reasons as detailed in Table 31. The remaining pool of 330 adjectives formed the initial item bank for trialling.

Table 31. Exclusion of adjectives prior to item trialling

Reason of removal	Count	Examples
The characteristic is morally wrong and self-ratings likely won't lead to honest answers.	3	Abusive Belligerent Violent
People with this characteristic are unlikely to recognise or admit they have this characteristic, leading to inaccurate self-ratings.	16	Egocentric Overconfident Unreasonable
The adjective's meaning is too ambiguous when describing personality.	50	Antagonistic Childlike Refined
The adjective is too strongly linked to sex or religion.	8	Feminine Religious Sexy
The adjective focuses on non-personality aspects of a person (e.g., looks, cognition, experience).	36	Clever Economical Knowledgeable
The adjective may cause people stress in a FC response format.	6	Anti-social Self-destructive Unstable

Item Trialling

The 330 remaining adjectives were then mapped to the HEXACO model conceptually. An initial mapping rated each adjective against each of the six factors, giving a rating of 1 (positive indicator), −1 (negative indicator), or 0 (no relationship). An adjective could have non-zero ratings on multiple factors. In addition, Lee and Ashton (2008) reported key adjective indicators for each of the HEXACO factors. Collating both the conceptual mappings from this study and Lee and Ashton's (2008) list of key indicators, a total of 100 items with unambiguous, factorially simple conceptual mappings to their respective HEXACO factors were selected to be anchors for item trialling. Each of the six factors was covered by between 14 to 20 anchor items (14, 17, 20, 16, 19 and 14 items for factors H, E, X, A, C and O respectively).

Item trialling was programmed in Qualtrics and designed so that each participant would complete 200 adjectives in total, of which 100 were anchor items, and the other 100 were randomly selected from the remaining 230 non-anchor items. The anchor items were incorporated so that for every participant there would be enough data to estimate scores for all six factors, as complete random item selection could result in low item coverage for certain factors for some participants. Items were rated using a six-point rating scale: 1) Very unlike me; 2) Somewhat unlike me; 3) A little unlike me; 4) A little like me; 5) Somewhat like me; 6) Very like me. In addition to the 200 adjectives, each participant also completed the 60-item version of the HEXACO-PI-R (Ashton & Lee, 2009). Basic background characteristics of the participants, such as gender and English proficiency level, were also collected.

Sample

Table 32. Data cleaning criteria

Data cleaning criteria	Cases
Participants who did not consent to providing data for research purposes.	279
Participants whose English proficiency level did not reach "Professional working proficiency" or higher.	175
Repeated completions by the same participants (keeping data for the first completion only).	198
Participants who completed the study too quickly (<10 minutes, indicating lack of proper consideration) or too slowly (>2 hours, indicating presence of distraction during completion).	127
Participants with unreliable response patterns (e.g., when the majority of the rating scale was never used, when a particular response option was overused, when the responses had a very small standard deviation).	23
Participants who partook in the study for reasons other than "to practice for pre-employment assessments" or "to find out more about myself".	28

A large online sample (N=2,515) was recruited in 2018 from a public-facing website specialising in pre-employment assessment practice. Participants were invited to complete the study in order to receive a personalised feedback report. Because the survey was open to any participant, extensive data cleaning was applied in order to ensure data quality, resulting in the removal of 33% of the collected cases (Table 32). Such a percentage was typical of data collected from the pre-employment assessment practice website used in this study.

Table 33. Cleaned sample demographics (N=1,685)

Sample demographics		%
Gender	Male	52.0
	Female	46.9
	Other	0.1
	Missing	1.0
Age	Up to 20	3.0
	21 to 30	34.5
	31 to 40	25.9
	41 to 50	21.6
	51 to 60	12.7
	Over 60	1.4
	Missing	0.8
English language proficiency	Native or bilingual proficiency	45.9
	Full professional proficiency	27.9
	Professional working proficiency	26.2

The final cleaned sample consisted of 1,685 cases. The sample was balanced in terms of gender, and all working ages were represented (Table 33). Nearly half (45.9%) of the sample indicated that they had “native or bilingual proficiency” in the English language, and only the participants with at least “professional working proficiency” was

retained in the sample to ensure that the interpretation of English adjectives was accurate. Most participants (54.0%) spent between 20 to 40 minutes completing the study, and the vast majority of participants (85.8%) indicated that their main reason for partaking in the research study was “to practice for pre-employment assessments”. With the random item selection in the trial design, each adjective was completed by between 675 and 1,685 participants in the sample.

Analysis and Results

Analysis was structured into four parts. First, responses to the 60-item HEXACO-PI-R instrument were analysed and compared against published results. Second, the 330 adjectives were assigned to HEXACO factors considering conceptual and empirical evidence. Third, IRT calibration was conducted to estimate parameters for the 330 adjectives and 60 HEXACO-PI-R statements. Finally, IRT scoring was conducted to estimate HEXACO factor scores for the respondents.

Properties of the HEXACO-PI-R

Responses to the HEXACO-PI-R were analysed and results were compared against properties of the same instrument published by Ashton and Lee (2009). For the HEXACO-PI-R response data across independent samples to show comparable properties, two conditions were necessary: 1) in terms of the instrument and construct, the HEXACO conceptual model needed to be stable and the HEXACO-PI-R instrument needed to be reliable; 2) in terms of the sample, participants needed to be motivated and respond conscientiously. Therefore, comparable results against published data would not only provide additional empirical support for the HEXACO conceptual model and the HEXACO-PI-R instrument for use with the pre-employment test-taker population, but also provide indication of good data quality from this study.

First, an exploratory factor analysis (EFA) was conducted to examine the factor structure of the 60 HEXACO-PI-R items. The EFA was conducted in Mplus version 8.1 (Muthén & Muthén, 1998-2012), using the ULS extraction method with OBLIMIN rotation. Six factors were extracted, which was supported by the scree plot (Figure 24). The six extracted factors corresponded one-to-one with the six conceptual factors well, with relatively simple factor structure and most items showing strongest loadings with their mapped factors (Appendix F, Table F1). The only two exceptions were item 11 ("I sometimes can't help worrying about little things") and item 35 ("I worry a lot less than most people do"), both of which were mapped to Emotionality (with pattern matrix loadings of 0.361 and −0.328 respectively), but loaded slightly stronger on Extraversion (with pattern matrix loadings of −0.405 and 0.338 respectively). The cross-loadings were understandable given the content of the items, which not only related to the tendency to be anxious (i.e., part of Emotionality), but also related to the tendency to be optimistic (i.e., part of Extraversion). Two other items, although loaded strongest on their mapped factors, also demonstrated signs of cross-loading (i.e., with pattern matrix loading magnitude differences <0.1) onto another factor. Item 9 ("People sometimes tell me that I am too critical of others") was mapped to Agreeableness (pattern matrix loading=−0.400), but also loaded onto Honesty-Humility (pattern matrix loading=−0.336). Item 32 ("I do only the minimum amount of work needed to get by") was mapped to Conscientiousness (pattern matrix loading=−0.426), but also loaded onto Honesty-Humility (pattern matrix loading=−0.335). Again, both cross-loadings were understandable given the content of the items. The good recovery of the six-factor HEXACO structure without specifying any rotation targets was encouraging, confirming the stability of the HEXACO-PI-R instrument and the HEXACO conceptual model of personality.

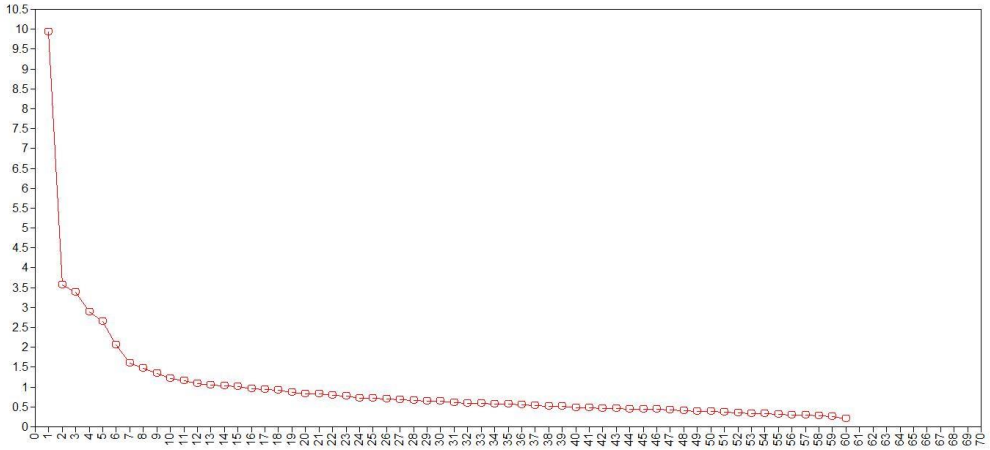


Figure 24. EFA scree plot of HEXACO-PI-R items

The initial EFA model did not specify any rotation targets when extracting the factors. However, in order to fully align the extracted factors with the HEXACO conceptual model, an exploratory structural equation model (ESEM; Asparouhov & Muthén, 2009) was constructed. The ESEM was built in Mplus version 8.1, using the ULSMV extraction method with TARGET rotation (Browne, 2001). The pattern matrix rotation target was set to minimise cross-loadings as much as possible according to the HEXACO-PI-R score key (i.e., the factor pattern loadings of items were targeted to zero unless the item was mapped to the factor by the score key). With this target rotation, all 60 items loaded strongest on their mapped factors (Appendix F, Table F2), with only four items showing secondary loadings exceeding a magnitude of 0.3 – the same four cross-loading items as identified and discussed in the EFA model with OBLIMIN rotation. The model fit for the ESEM was very good according to the Root Mean Square Error of Approximation (RMSEA = .037) and Standardized Root Mean Residual (SRMR = .034). The Comparative Fit Index (CFI = .886) was somewhat worse, but understandable given “the breadth and brevity of the scales” (Ashton & Lee, 2009).

The ESEM also estimated latent correlations between the HEXACO factors (Table 34). While observed score correlations are affected by measurement errors in the observed scores, latent correlation estimates are computed from the latent traits in the ESEM model, and therefore not affected by measurement errors. Thus, latent correlation estimates are better estimates of the true correlations between HEXACO factors than observed score correlations. Nevertheless, in order to directly compare against reported HEXACO-PI-R observed score correlations, HEXACO-PI-R scores (i.e., classical item sum scores calculated according to the score key) were computed and correlated (Table 35). The magnitudes of the observed correlations were stronger than those reported by Ashton and Lee (2009, Table 3). The generally stronger observed correlations suggested that a stronger positive manifold existed in this sample compared to Ashton and Lee's (2009) low-stakes research samples of college students and community participants. Considering that this sample was predominantly (85.8%) completing the questionnaires to practice for pre-employment assessments, this positive manifold might have resulted from typical high-stakes pre-employment assessment behaviours, such as social desirability responding and impression management.

Table 34. Latent HEXACO factor correlations from ESEM

	H*	E	X	A	C*
E	−.184				
X	.194	−.237			
A	.268	−.053	.287		
C*	.300	−.189	.342	.246	
O	.115	−.052	.230	.146	.184

* Signs of the correlations were adjusted in order to align with the definitions of the conceptual factors.

Table 35. HEXACO-PI-R observed score correlations

	H	E	X	A	C
E	−.083				
X	.195	−.284			
A	.380	−.168	.327		
C	.305	−.154	.410	.291	
O	.145	−.095	.237	.134	.181

Then, in order to examine the unidimensionality of the HEXACO-PI-R scales, a single-trait confirmatory factor analysis (CFA) model was fitted to the 10 constituting items for each factor (as indicated by the score key). The CFA models were built in Mplus version 8.1, giving model fit statistics as summarised in Table 36. The model chi-square p-values were significant for all models, indicating bad fit. However, the chi-square test of model fit is sensitive to large sample sizes. Therefore, for this study with N=1,685, the focus should be placed on RMSEA, CFI and SRMR, which are less affected by sample size. All six models had RMSEA values greater than .08, indicating bad fit (MacCallum, Browne, & Sugawara, 1996). Because the RMSEA values for the null model were all greater than .158, the CFI values were not informative for these models (Kenny, 2015). The SRMR values, however, were in satisfactory ranges, staying below .08 for all six models (Hu & Bentler, 1999). In terms of possible model modifications, no modifications suggested by Mplus had a modification index above 1, indicating that there were no simple model modifications that would significantly improve the model fit. The unsatisfactory fit of the unidimensional models may have resulted partially from the breadth of the HEXACO factors, as each of them can be further divided into four distinct subscales (see Lee & Ashton, 2009a).

Table 36. Unidimensional CFA model fits for HEXACO-PI-R items

Model	Chi-square	Degrees of freedom	P-value	RMSEA (90 percent C.I.)	CFI	SRMR
H	1068.624	35	<.001	.132 (.126, .139)	.815	.067
E	1045.719	35	<.001	.131 (.124, .138)	.780	.059
X	1338.177	35	<.001	.149 (.142, .156)	.844	.061
A	808.938	35	<.001	.115 (.108, .121)	.839	.052
C	423.499	35	<.001	.081 (.074, .088)	.931	.040
O	989.829	35	<.001	.127 (.120, .134)	.846	.059

Finally, the internal consistencies of the HEXACO-PI-R scales were examined. Coefficient omega (McDonald, 1999) was computed for each scale assuming a one-factor solution. Moreover, in order to directly compare against published internal consistency statistics of the HEXACO-PI-R, Cronbach's alpha (Cronbach, 1951) was also computed. Both reliability measures were computed in R (R Core Team, 2015) using the psych package (Revelle, 2018), and reported in Table 37. The current sample showed internal consistency statistics comparable to those collected by Ashton and Lee (2009), who reported Cronbach's alphas of .73 to .80 across samples for all scales.

Table 37. Internal consistency of HEXACO-PI-R scales

Scale	H	E	X	A	C	O
Omega	.716	.729	.821	.716	.766	.751
Alpha	.702	.728	.818	.710	.757	.751

To summarise, the responses to the 60 HEXACO-PI-R items in this sample showed good internal consistencies and demonstrated a factor structure matching the expected six-factor HEXACO conceptual model. Unidimensional CFA model fit statistics were unsatisfactory, but understandable given "the breadth and brevity of the

scales" (Ashton & Lee, 2009). Response data in this sample demonstrated signs of a stronger positive manifold than previously reported low-stakes research samples. This positive manifold likely reflected typical high-stakes pre-employment assessment behaviours. On the whole, the reported properties of the HEXACO-PI-R were largely replicated in this new sample, and when differences were observed, they were in line with expectations given the current data collection settings, motivation of participants, and sample demographics.

Mapping adjectives to HEXACO model

The next stage of the analysis focused on mapping the 330 adjectives to the six HEXACO factors, in preparation for subsequent IRT modelling. Although theoretically the TIRT model is capable of handling within-item multidimensionality, in practice it could be difficult to obtain reliable answers to FC comparisons between multiple items where each item indicates multiple traits. Therefore, the aim was to create a factorially simple item pool, sacrificing multidimensional items in the process if necessary.

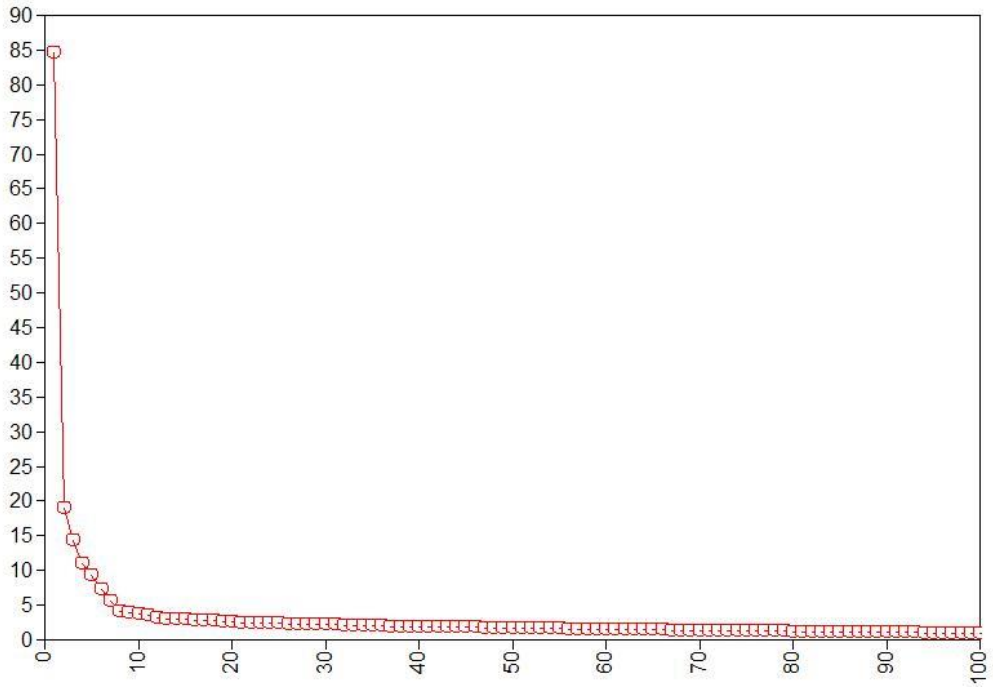


Figure 25. EFA scree plot of 330 adjectives

First, the overall factor structure of all 330 adjectives was examined. An EFA was conducted in Mplus version 8.1 (Muthén & Muthén, 1998-2012), using the ULS extraction method with OBLIMIN rotation. The scree plot (Figure 25) suggested that a six-factor solution was likely sufficient. However, examination of the oblique six-factor solution loading pattern matrix (Appendix F, Table F3) revealed many cross-loadings even for anchor items, suggesting that a positive manifold was at play, as seen earlier in the analysis of HEXACO-PI-R responses. This positive manifold, likely caused by social desirability responding and impression management in pre-employment assessment samples, appeared to have affected the HEXACO-PI-R to a smaller extent, resulting in conceptually distinct factors each indicated by homogeneous items. In the case of adjective item responses, three of the extracted factors consisted of conceptually homogeneous adjectives, which aligned roughly with the factors of eXtraversion, Conscientiousness and Openness to Experience. The other three factors were made up of conceptually heterogeneous adjectives and were more difficult to separate and summarise semantically. It was concluded that the contaminating effect of the positive manifold was too strong to meaningfully interpret this six-factor EFA solution.

It was interesting that the positive manifold in this sample affected adjectives more so than the HEXACO-PI-R statements. The reason for this contrast may be two-folds. On one hand, while the HEXACO-PI-R statements were carefully crafted to minimise bias and achieve balanced psychometric measurement (Lee & Ashton, 2004), the unedited adjectives can come with strong positive or negative linguistic connotations, making them more likely to trigger stronger social desirability responding. Furthermore, generic adjectives tend to have less nuance, complexity or context compared to longer HEXACO-PI-R statements, making them more prone to fast emotive responses (System 1, Kahneman, 2011). Therefore, it seems that adjectives, when used in a rating scale question format, may introduce unwanted artefacts into the

measurement of personality traits compared to HEXACO-PI-R statements. However, in a FC format, adjectives in the same FC block can be balanced by social desirability in order to minimise social desirability responding, and the comparative judgement format of similarly desirable characteristics will likely encourage slow conscious responses (System 2, Kahneman, 2011). The combination of adjectives and FC response format thus has the potential to retain the simplicity of adjective items, encourage more deliberate thinking, as well as removing the contamination of social desirability responding.

Following the EFA analysis, and subsequent bifactor EFA and ESEM analysis that failed to isolate the positive manifold, it was concluded that a conceptual mapping of items to the HEXACO factors would lead to the most meaningful assignment. In order to come up with this mapping, two psychometricians mapped each of the adjectives to one and only one HEXACO factor conceptually. The mappings were then compared and collated. Where the mappings agreed, the item was retained in the pool. Where the mappings disagreed, the adjective was reviewed again, and a judgement was made to either map it to one of the mapped factors, or to drop it from the pool due to its semantic multidimensionality. Some additional items were also dropped due to their negative connotations, which likely triggered a greater extent of social desirability responding compared to other items. At the end of this qualitative item review process, 299 out of the 330 adjectives were retained, covering each factor with between 24 to 82 items (51, 44, 45, 82, 53 and 24 items for factors H, E, X, A, C and O respectively). Then, this qualitative mapping was refined further quantitatively by building a CFA model for each of the HEXACO factors, using the ML estimator and treating item responses as categorical variables (i.e., effectively calibrating the items under Samejima's graded response IRT model, Samejima, 1969). Items whose standardised

loadings had magnitudes below 0.2 were removed, leaving a total of 286 adjectives (Table 38).

Table 38. Items mapped qualitatively and quantitatively to each HEXACO factor

Scale	H	E	X	A	C	O
Mapped items	46	41	41	81	53	24
Anchor items	10	14	20	19	19	9

Item calibration for TIRT

The next stage of the analysis focused on establishing measurement properties of the selected adjectives, in order to serve as an item bank for a TIRT-based FC CAT for HEXACO personality factors. Parameters pertaining to item utilities t_i needed to be estimated, namely, item mean μ_i, item factor loading λ_i, and item error variance ψ_i^2 (see Equation 3). The use of a six-point rating scale gave rise to enough variance in the adjective item responses, allowing them to be treated as continuous item utility. By aligning the arbitrary scaling of latent item utilities t_i to the response categories ranging from 1 to 6, the same scaling for item utility was enforced across different adjectives, allowing them to be meaningfully compared when eventually placed into a FC setting. With the item responses being treated as continuous variables, a simple unidimensional CFA model would provide parameters in a format that would be directly usable for TIRT modelling.

Thanks to the randomised item administration, calibration was conducted on the entire sample simultaneously (N=1,685 participants in total, each item being completed by between 675 to 1,685 participants), with no need for linking. The completion of anchor items by all participants ensured stability of the measured constructs. In addition, the 60 HEXACO-PI-R statements were included in the same calibration models, so as

to stabilise the constructs further and to obtain IRT parameters for the HEXACO-PI-R statements for subsequent analysis.

The unidimensional CFA model for each HEXACO factor was fitted independently in Mplus version 8.1 (Muthén & Muthén, 1998-2012) using a maximum likelihood estimator (ESTIMATOR = ML). Rubin (1976) showed that the use of the ML estimator ensures unbiased item parameter estimates for data that are missing completely at random (MCAR) or missing at random (MAR). In this study, the random presentation of adjectives determined by the survey randomisation algorithm ensured that responses to the non-anchor items were MCAR, so the ML estimator was adequate.

The CFA model fit statistics were then examined in order to determine whether additional model adjustments were necessary. After seeing the CFA model fit statistics for HEXACO-PI-R items earlier, it was acknowledged that the CFIs would be low given the breadth of the factors (Ashton & Lee, 2009), and a good model fit was defined to be one with RMSEA < .08 (MacCallum, et al., 1996) and SRMR < .08 (Hu & Bentler, 1999). Based on this criteria, five out of the six HEXACO scales produced satisfactory model fit without any adjustments. The model fit for Emotionality was less ideal (RMSEA = .062, SRMR = .102), and the model was reviewed and refined further by removing seven adjectives and one HEXACO-PI-R statement with 1) relatively weak and ambiguous conceptual mapping, 2) relatively large modification indices (i.e., undesirable correlations with multiple items that were not accounted for by the latent factor), or 3) relatively small magnitude of slope parameter (slope was calculated as λ_i/ψ_i, which matched the definition of the discrimination parameter in standard unidimensional IRT parameterisation). At the end of this process, IRT parameters were established for a final set of 279 adjectives. The final model characteristics are given in Table 39. The distributions of item parameters for the 279 adjectives are summarised in

Table 40. The full calibrated item bank and parameters are presented in Appendix F, Table F4.

Finally, in order to enable multidimensional MAP scoring using TIRT, the correlations between HEXACO factors were also required. For this purpose, the latent correlation estimates from the ESEM model on the 60 HEXACO-PI-R items were adopted (Table 34).

Table 39. Final calibration model characteristics

Factor	Adjectives count			RMSEA (90 Percent C.I.)	CFI	SRMR
	Total	Positive loading	Negative loading			
H	46	19	27	.041 (.040, .042)	.679	.067
E	34	23	11	.055 (.054, .057)	.658	.078
X	41	26	15	.066 (.065, .067)	.706	.076
A	81	39	42	.044 (.043, .045)	.636	.074
C	53	28	25	.048 (.047, .049)	.733	.065
O	24	20	4	.062 (.060, .064)	.631	.076
Total	279	155	124			

Table 40. Final calibrated adjectives item bank characteristics

Parameter	Statistics	H	E	X	A	C	O
μ_i	Mean	3.24	3.23	3.83	3.65	3.64	4.28
	Minimum	1.22	1.53	1.71	1.25	1.31	1.43
	Maximum	5.80	5.30	5.31	5.67	5.72	5.49
λ_i	Mean	−0.20	0.29	0.14	−0.05	−0.05	0.31
	Minimum	−0.70	−0.56	−1.07	−0.71	−0.80	−0.58
	Maximum	0.53	1.00	0.96	0.63	0.65	0.89
ψ_i^2	Mean	1.05	1.17	1.03	0.88	0.77	0.91
	Minimum	0.22	0.58	0.39	0.22	0.20	0.32
	Maximum	2.60	2.29	2.16	2.22	2.10	2.05
$\lvert\lambda_i\rvert$	Mean	0.45	0.54	0.62	0.49	0.48	0.43
	Minimum	0.15	0.24	0.29	0.23	0.25	0.19
	Maximum	0.70	1.00	1.07	0.71	0.80	0.89
λ_i/ψ_i	Mean	−0.17	0.26	0.23	0.05	−0.01	0.34
	Minimum	−0.78	−0.66	−0.95	−0.82	−1.18	−0.75
	Maximum	0.67	0.97	1.18	1.01	1.07	1.03
$\lvert\lambda_i/\psi_i\rvert$	Mean	0.49	0.52	0.68	0.61	0.61	0.49
	Minimum	0.15	0.17	0.20	0.20	0.22	0.20
	Maximum	0.78	0.97	1.18	1.01	1.18	1.03

IRT scoring

Following calibration, responses to the adjectives were scored using the newly estimated IRT parameters. The number of adjectives completed for each scale varied by participants due to the randomised item administration design, but each participant completed at least 10 adjectives in every scale. Separately, the HEXACO-PI-R items were also scored using their newly established IRT parameters. The IRT scores for HEXACO-PI-R items were based on 10 items per scale, except for the Emotionality

scale which was based on nine items only (one item was removed during IRT calibration). The distributions for the two sets of IRT scores, as well as the classical HEXACO-PI-R scores, are shown in Table 41. Both sets of IRT scores had near zero means, as would be expected when scored using IRT parameters established on the same sample. Later, when the adjectives were administered in a FC format, the mean scores would likely be lowered, due to the FC response format greatly reducing social desirability responding.

Table 41. Distributions of the three versions of HEXACO scores

Factor	HEXACO-PI-R CTT		HEXACO-PI-R IRT		Adjectives IRT	
	Mean	SD	Mean	SD	Mean	SD
H	39.9	5.38	0.0001	0.886	0.0002	0.945
E	28.2	5.89	−0.0001	0.865	−0.0005	0.946
X	38.2	5.81	0.0002	0.939	0.0000	0.976
A	36.2	5.24	0.0001	0.894	−0.0002	0.975
C	41.7	4.67	−0.0003	0.915	0.0002	0.973
O	37.7	5.68	0.0000	0.899	−0.0018	0.917

Correlations between the three versions of HEXACO scores are shown in Table 42. Not surprisingly, the HEXACO-PI-R items produced similar estimates when scored using different methods, resulting in correlations of .963 or higher (with a mean of .976) across all scales. The correspondence between scores based on HEXACO-PI-R and those based on adjectives were weaker but still showed signs of convergent and divergent validity, with scale correlations ranging from .589 to .793 (with a mean of .670) and average off-diagonal correlation of .129 when the same scoring methodology was applied.

Table 42. Correlations between the three versions of HEXACO scores

Factor	HEXACO-PI-R CTT with HEXACO-PI-R IRT	HEXACO-PI-R CTT with Adjectives IRT	HEXACO-PI-R IRT with Adjectives IRT
H	.968	.555	.589
E	.963	.541	.595
X	.995	.782	.793
A	.979	.647	.669
C	.980	.720	.746
O	.969	.568	.625
Diagonal mean	.976	.635	.670
Off-diagonal mean	.124	.127	.129

Summary

This study developed an item bank of 279 adjectives for measuring the HEXACO personality traits, covering the six scales with between 24 to 81 items each. The IRT calibration models for all six scales achieved good fit (Table 39) that surpassed the fits of the models for the 60-item HEXACO-PI-R (Table 36). Scores estimated from adjectives demonstrated moderate convergent and divergent validity against the previously validated HEXACO-PI-R. This item bank thus provided the content for driving a FC CAT for HEXACO personality traits.

Comparing Adaptive Algorithms for HEXACO Item Bank (Study 5)

This study simulated FC CAT sessions using the HEXACO adjectives item bank developed in Study 4. The purpose of this study was two-folds. First, it investigated item selector performance with a realistic item bank, to examine whether findings from Study 3 (which used simulated item banks) would still hold. Second, it examined the functioning of the new HEXACO adjectives item bank in a CAT setting, in order to

gauge its suitability and limits for practical use, and to determine the CAT algorithm settings (e.g., item selector, target test length) to adopt in a subsequent empirical study.

Method

Simulation design

A simulation study was conducted to examine the efficiency of item selectors in FC CAT using the HEXACO adjectives item bank developed in Study 4. Similar to Study 3, this study focused on FC assessments using pairs, and the interim and final person scores were estimated using the Bayesian MAP estimator with a trait correlation prior as established in Study 4 (Table 34). The assessment design factors investigated are described in this section, which largely replicated the design of Study 3. All simulations were conducted using the same R codes written specifically for this thesis.

Item selector (6 levels)

Six item selectors were simulated: RANDOM, WCI, A-, C-, D-, and T-optimality.

Scale plan (2 levels)

Two levels of scale plan were simulated as per Study 2 (Table 12).

Social desirability balancing criteria (2 levels)

Two levels of social desirability balancing were examined as per Study 2 (Table 13).

Test length (8 levels)

In order to examine the effect of test length on CAT score recovery accuracy, the assessment length was varied by truncating the simulated CAT sessions, so that the

shorter test lengths were completely nested in the longer ones. Matching the design of Study 3, test lengths of up to 30 items per scale (i.e., 30 × 6 ÷ 2 = 90 pairs) were examined. Within each CAT session for each simulee, the CAT algorithm continued to create pairwise comparisons, until the target of 90 pairs was reached, or until no remaining pairs of items would satisfy all content constraints (i.e., scale plan if there were any, social desirability balancing criteria, no two items from the same scale in each pair, and no more than one negative item in each pair). However, because the HEXACO adjectives item pool was much smaller compared to the simulated item banks (e.g., the Openness to Experience scale only had 24 items), it was anticipated that some test sessions would not reach the full length requested. It was therefore desirable to look at more levels of test length, so that comparisons could be conducted at the most meaningful test lengths. Having data at multiple test lengths also provided better information for determining the test length to adopt in the subsequent empirical CAT study. Eight levels of test length were simulated (Table 43).

Table 43. Test length levels

Level	Description
9 items per scale	The first 27 pairs.
12 items per scale	The first 36 pairs.
15 items per scale	The first 45 pairs.
18 items per scale	The first 54 pairs.
21 items per scale	The first 63 pairs.
24 items per scale	The first 72 pairs.
27 items per scale	The first 81 pairs.
30 items per scale	All 90 pairs.

Analysis

Crossing the different levels of the four design factors gave rise to 6 (item selector) × 2 (scale plan) × 2 (social desirability balancing) × 8 (test length) = 192 conditions in total, with each condition being covered by a sample of 2,000 simulees generated from a multivariate normal distribution following HEXACO scale correlations established from Study 4 (Table 34). As anticipated, the maximum assessment length was not reached all the time. In other words, for some conditions or/and simulees, before the target assessment length was reached, the item bank was depleted sufficiently that no viable pairs meeting all content constraints remained. Analysis therefore examined the following summary statistics for each condition:

- **Normal test termination**: the percentage of simulees successfully reaching a given test length;
- **Rank ordering**: the correlations between true and estimated scores for each scale;
- **Absolute differences**: RMSEs of the differences between true and estimated scores.

As this study had a simpler design and a narrower focus compared to Study 3, it was concluded that cross-classified multilevel regression analysis was neither necessary nor desirable. Instead, the results were summarised and visualised graphically.

Results

Normal test termination

The percentage of simulees successfully reaching each level of test length for each condition is shown in Table 44. Not surprisingly, scale plan had the most significant effect on a limited item bank. Sessions with a fixed scale plan reached at

least 52 pairs/ 17.3 items per scale but never exceeded 68 pairs/ 22.7 items per scale, while all sessions with dynamic scales reached at least 87 pairs/ 29 items per scale. The effect of social desirability balancing was also in line with expectations, with the strict criterion leading to shorter average test lengths than the lenient criterion. Interestingly, for sessions with a fixed scale plan, the RANDOM and C-optimality item selectors led to much shorter average test lengths compared to the other item selectors. For sessions with dynamic scales, as almost all sessions reached the maximum test length, the differences between item selectors was not apparent, although there were some signs that the RANDOM and A-optimality item selectors might lead to shorter average test lengths if the sessions were allowed to continue beyond 90 pairs.

The fact that certain item selectors lead to faster item depletion was interesting and not expected prior to this simulation study being completed. How certain item selectors and the various content rules interplayed to lead to fewer remaining viable pairs was unclear without further investigation. Nevertheless, current results suggested that some item selectors may be more demanding on the size and composition of item content than others, especially when the item pool was limited after the application of stringent content rules.

Table 44. Percentage of simulees reaching each level of test length by condition

Scale plan	Social desire	Item selector	Test length			% simulees successfully reaching a test length (pairs)		
			Mean	Min	Max	63	72 / 81	90
Fixed	Lenient	RANDOM	62.5	54	68	44.9	0.0	0.0
		WCI	67.4	60	68	98.5	0.0	0.0
		A-optimality	67.3	60	68	98.9	0.0	0.0
		C-optimality	60.0	57	60	0.0	0.0	0.0
		D-optimality	66.4	60	68	90.0	0.0	0.0
		T-optimality	67.1	60	68	96.3	0.0	0.0
	Strict	RANDOM	60.8	52	68	26.9	0.0	0.0
		WCI	65.4	60	68	99.3	0.0	0.0
		A-optimality	64.6	57	68	94.0	0.0	0.0
		C-optimality	59.9	57	60	0.0	0.0	0.0
		D-optimality	64.8	60	68	95.6	0.0	0.0
		T-optimality	65.2	60	68	99.4	0.0	0.0
Dynamic	Lenient	RANDOM	90.0	89	90	100.0	100.0	100.0
		WCI/A/C/D/T	90.0	90	90	100.0	100.0	100.0
	Strict	RANDOM	90.0	87	90	100.0	100.0	99.4
		A-optimality	90.0	88	90	100.0	100.0	97.3
		WCI/C/D/T	90.0	90	90	100.0	100.0	100.0

Rank ordering and absolute differences

The correlations and RMSEs between true and estimated scores were computed for each scale in each condition and summarised graphically (Figures 26 to 33). All plots were made for test lengths up to 90 pairs, so the graphs demonstrated plateauing effects when the CAT sessions did not reach the longer lengths.

The effectiveness of different item selectors within conditions was considered first. Across all conditions, A-optimality was consistently one of the best according to

both correlations and RMSEs. There were instances where another item selector outperformed A-optimality for some scales, but the same item selector would also demonstrate notable weakness for some other scales. For example, WCI worked marginally better than A-optimality on the eXtraversion scale across multiple conditions, but it was notably worse for the Honesty-Humility scale in all conditions. In fact, the relative merit of WCI amongst the item selectors appeared to be somewhat dependent on the scale, often performing as one of the best in some scales while showing significant weakness for other scales. C-optimality exhibited similar characteristics to WCI, performing on par with A-optimality on some scales but was even worse than the RANDOM method for Emotionality when a fixed scale plan was applied. D-optimality was the next best item selector after A-optimality, often performing on par with or slightly worse than A-optimality across most conditions. The worst performing item selectors were usually RANDOM and T-optimality. The relative merits of item selectors were in line with results seen in Study 3. Based on the results of this study and Study 3, A-optimality was chosen for the subsequent empirical study.

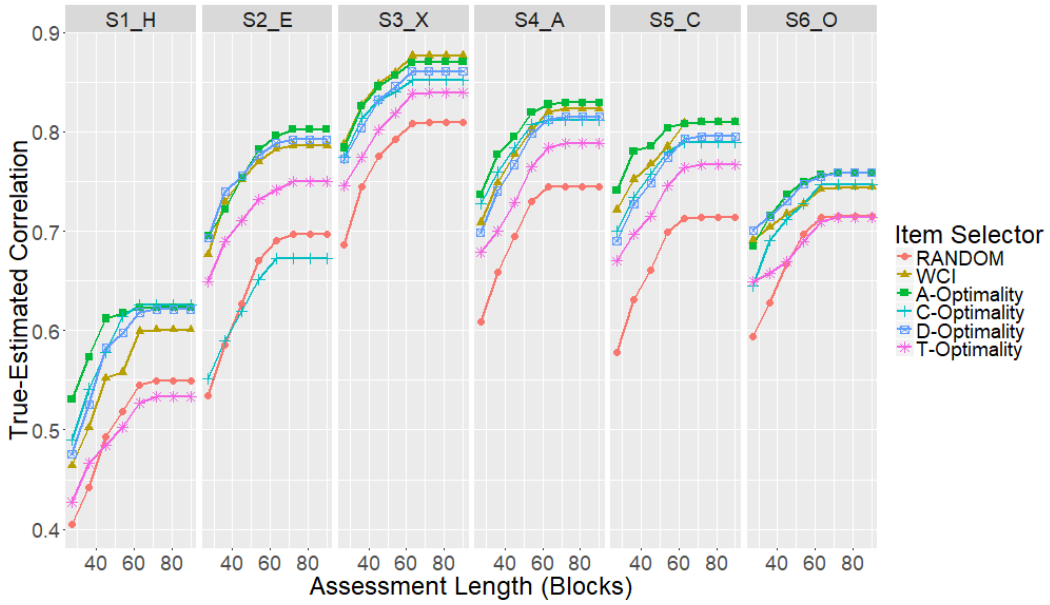


Figure 26. Score correlations – fixed scale plan and strict social desirability

Figure 27. Score correlations – fixed scale plan and lenient social desirability

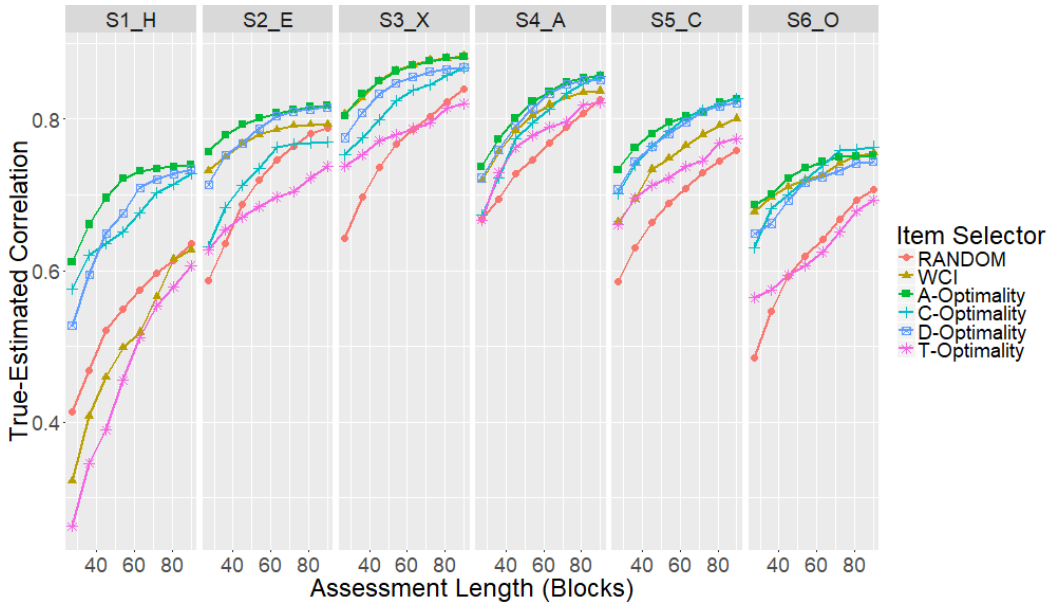


Figure 28. Score correlations – dynamic scale plan and strict social desirability

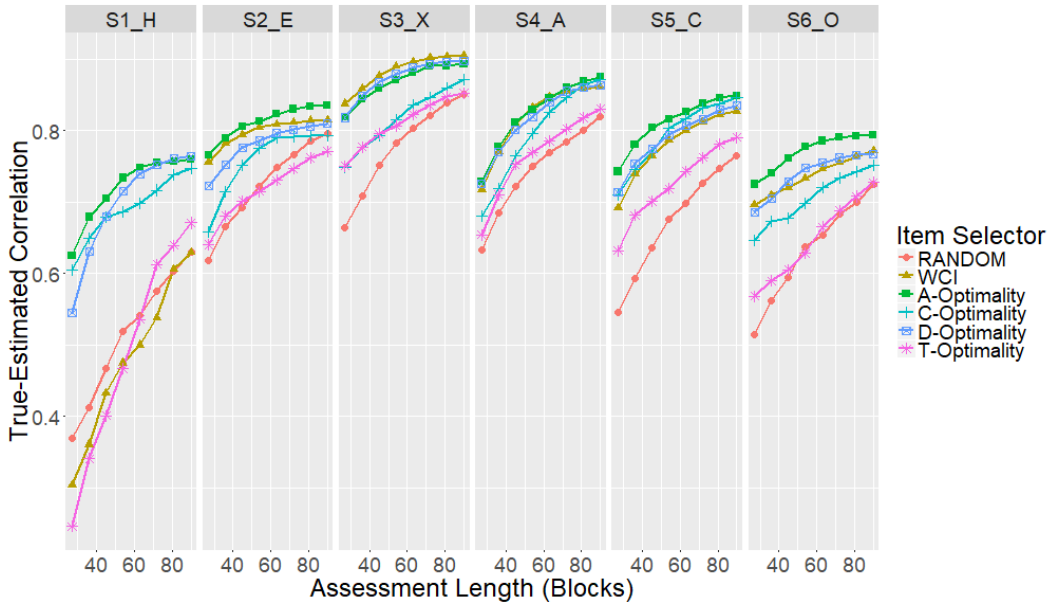


Figure 29. Score correlations – dynamic scale plan and lenient social desirability

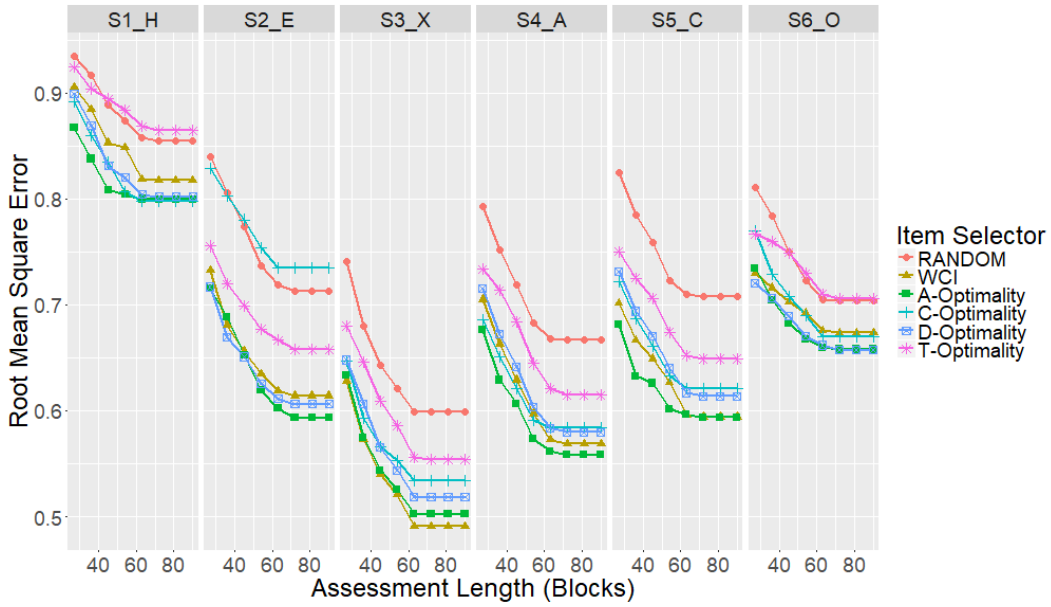


Figure 30. RMSEs – fixed scale plan and strict social desirability

Figure 31. RMSEs – fixed scale plan and lenient social desirability

Figure 32. RMSEs – dynamic scale plan and strict social desirability

Figure 33. RMSEs – dynamic scale plan and lenient social desirability

Next, the effect of scale plan was considered. Due to the CAT sessions with fixed scale plans only reaching lengths of 60-70 pairs on average, the comparison focused on results up to 60 pairs. In most situations, the presence of a fixed scale plan had a weakening effect on measurement. For example, when A-optimality was used (Figures 34 and 35), the Honesty-Humility scale was the most problematic in terms of measurement and a fixed scale plan made it much worse. This finding was contrary to results from Study 3, which found little restrictive effects of a fixed scale plan on large simulated item banks. The weakening effect found in this study likely arose from the interaction between the fixed scale plan and the very limited item bank. This effect was not surprising as a scale plan placed constraints on achievable test lengths as well as denied the item selector's freedom to prioritise measurement on underperforming scales. This limiting effect might be alleviated to some degree if the scale plan was designed with consideration for the characteristics of the available items for different scales, and/or designed in a way that didn't place as strong a limit on the scale selection for each pair. One possible way to implement this was to use a scale plan that covered only the beginning of the test in order to ensure that the minimum number of items per scale would be reached, after which the algorithm was allowed to freely choose items to enhance the measurement of underperforming scales. Nevertheless, unless there was a strong face validity argument to have a perfectly balanced scale plan, results suggested that it would be better to allow the item selector to decide which scales to test next. Given these observations, the subsequent empirical study would allow scales to be dynamically chosen.

In terms of social desirability balancing, results were as expected – a more lenient social desirability balancing criteria led to slightly better measurement in a pure theoretical sense (e.g., Figures 34 and 35 for when A-optimality was used). However, in practice, especially in high-stakes testing situations, more lenient social desirability

balancing may lead to greater impression management. Therefore, the effect of different social desirability balancing criteria would be explored further in the subsequent empirical study.

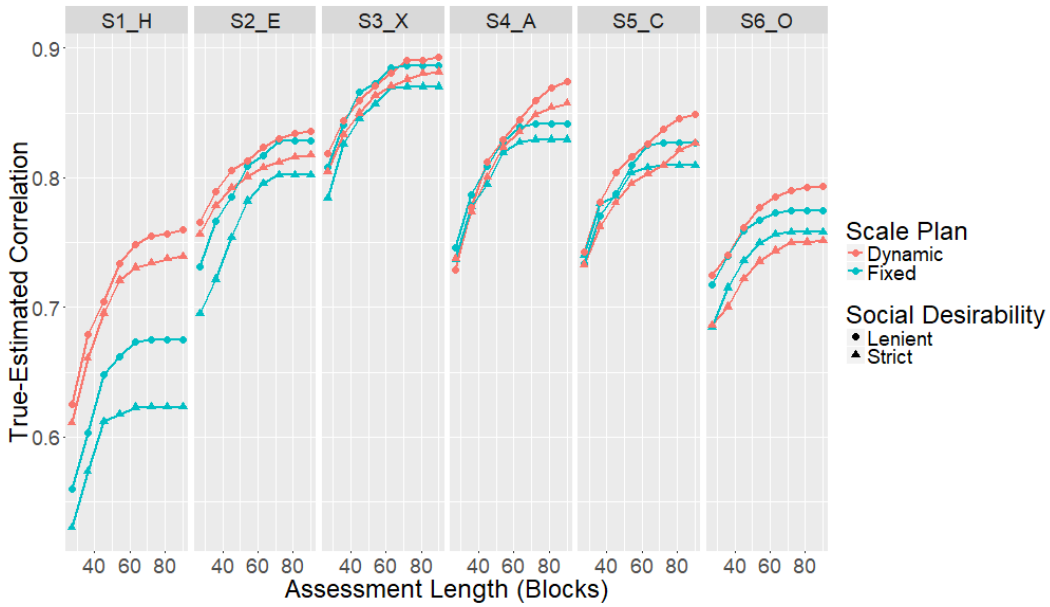


Figure 34. Correlations between true and estimated scores for A-optimality

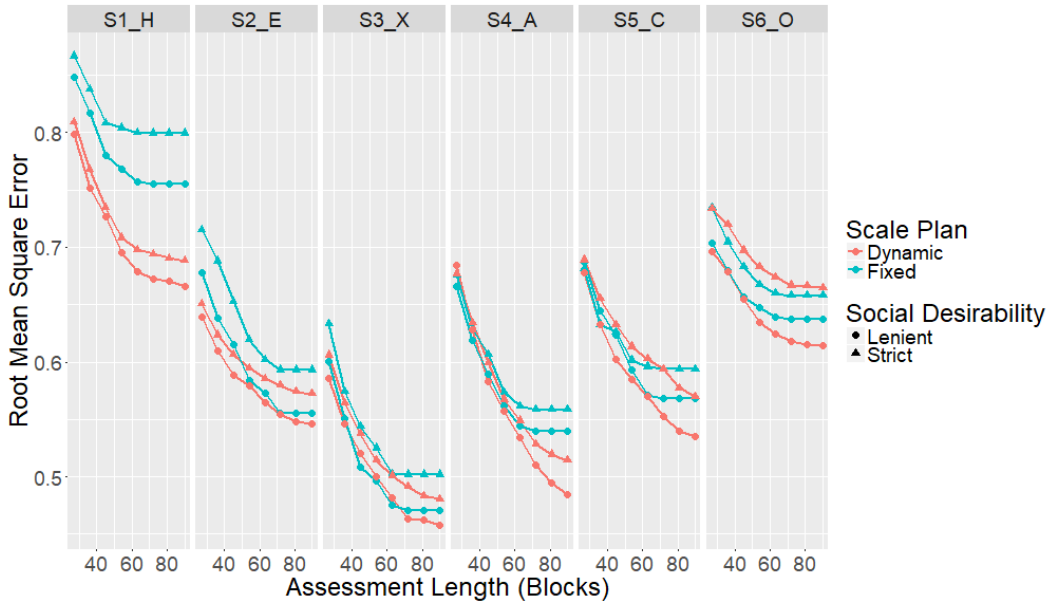


Figure 35. RMSEs between true and estimated scores for A-optimality

Next, results for different HEXACO scales were compared. The Honesty-Humility scale was clearly underperforming compared to other scales, which was interesting as it wasn't the scale with the least number of items: Honest-Humility had 46 items, eXtraversion had 41 and was the scale with the best measurement precision, Emotionality had 34, and Openness to Experience had merely 24. A closer inspection of the results against item parameter distributions (Tables 39 and 40) revealed that the approximate ranking of measurement accuracy of the six scales as determined by the simulations (i.e., X>A>C/E>O>H, Figures 34 and 35) lined up roughly with the ranking of mean $|\lambda_i/\psi_i|$ values across all items within the scales (i.e., X>A/C>E>O/H, Table 40). This observation suggested that, in an item bank for FC CAT, quantity did not make up for quality, and it was more beneficial to have a smaller pool of highly discriminating items, rather than a larger pool of items with low discrimination power. Aside from item discrimination powers, another contributing factor to the underperformance of the Honesty-Humility scale might have been its larger proportion of negative items (Table 39) in combination with the algorithmic setting of only allowing one negative item in every pair. Under this setting, a positive item could be paired with any other item, but a negative item could only be paired with a positive one. Therefore, the number of allowable pairs for a scale with high proportions of negative items would be greatly reduced. The content rule of avoiding the comparison of two negatively-loading items originated from best practices in FC assessments using statements as stimuli, where the comparison of two statements containing negations would significantly increase the cognitive load of the responding process. However, comparing two negatively-loading adjectives would likely pose no problem, as adjectives are simple concepts and do not contain negations. Considering this, and observing the difficulty in measuring the Honesty-Humility scale with the current item

bank, the subsequent empirical study would allow both adjectives in a pair to be negatively-loading.

Table 45. Simulated CAT measurement properties at 90 pairs with A-optimality, dynamic scales and no negative pairs

Social desirability	Scale	True-estimated score correlation	Reliability	RMSE
Strict	H	.74	.55	0.69
	E	.82	.67	0.57
	X	.88	.78	0.48
	A	.86	.74	0.51
	C	.83	.68	0.57
	O	.75	.56	0.66
Lenient	H	.76	.58	0.67
	E	.84	.70	0.55
	X	.89	.80	0.46
	A	.87	.76	0.48
	C	.85	.72	0.53
	O	.79	.63	0.61

Finally, the actual values of score correlations and RMSEs when using A-optimality and dynamic scales (i.e., the settings chosen for the subsequent empirical study) were considered in order to formulate test length recommendations for the subsequent empirical study. Even with a length of 90 pairs, measurement accuracy was still unsatisfactory (Table 45). The Honesty-Humility scale and the Openness to Experience scale reached true-estimated score correlations of .74 to .79, which translated to a reliability of merely .55 to .63. RMSEs were also relatively high, with four scales ending in the 0.45 to 0.60 range, and two scales ending in the 0.60 to 0.70 range. In an ideal situation, additional adjectives should be sought and calibrated to build a more discriminating item bank. However, as a short-term solution, it was

desirable to consider elongating the assessment and relaxing content constraints in order to achieve better measurement accuracy. The effects of longer test lengths and the removal of the content constraint around negative pairs were tested in a follow-up simulation study using settings matching those chosen for the subsequent empirical study.

Summary

This study simulated FC CAT sessions using the item bank of 279 adjectives for measuring the HEXACO personality traits. The relative performance of item selectors on this new item bank largely replicated findings from Study 3, with A-optimality being the most efficient, followed by D-optimality, while RANDOM and T-optimality were the worst. Although having minimal effect in Study 3, a fixed scale plan was clearly restrictive for this new item bank, leading to early test terminations and reduced measurement accuracy. Moreover, the distribution of item parameters for each scale also had a notable effect on measurement accuracy of that scale. More specifically, the scales with fewer but more discriminating items tended to outperform the scales with more items but with lower discriminations. Finally, even at the maximum test length explored, the measurement accuracy was still unsatisfactory even in the most optimal design condition. Therefore, an additional simulation study (Study 5b) was conducted to further optimise the assessment design prior to conducting the empirical FC CAT study.

Optimising the Design for HEXACO FC CAT (Study 5b)

This simulation study built upon the recommendations from Study 5, and further refined the assessment design for the subsequent empirical FC CAT study. Moreover, results from this study would provide theoretical benchmarks for comparing subsequent empirical results against.

Method

This study used the same HEXACO adjectives item pool and focused on the chosen settings for the subsequent empirical study: A-optimality, dynamic scales, and both lenient and strict social desirability as two different conditions. To further enhance measurement accuracy, this study also allowed both adjectives in a pair to be negatively-loading, which had not been explored in previous simulations. This study also allowed the test creation to continue until there was no viable pairs left, thus providing data for all achievable test lengths with the current item bank, in order to inform the choice of target test length in the subsequent empirical study.

In addition, this study explored the expected measurement differences between adaptive assessments and non-adaptive control conditions that were otherwise the same (i.e., generated using the exact same algorithmic settings but without interim score updates, effectively always targeting measurement at the average person). Both adaptive and non-adaptive measures would be included in the subsequent empirical study.

This study employed the same sample of 2,000 simulees from Study 5. The analysis strategy from Study 5 was also followed.

Results

Normal test termination

With the chosen assessment design, all CAT sessions reached at least 123 pairs (Table 46). The shortest and longest test sessions were 123 and 137 pairs respectively. Using a more lenient social desirability balancing criterion led to three extra pairs being generated on average. Compared to the test lengths reached in Study 5 using the exact same item bank, it appeared that allowing both items in a pair to be negatively-loading

greatly increased the availability of viable pairs, leading to longer achievable assessment lengths.

Table 46. Percentage of simulees reaching each level of test length by condition

Social desirability	Test length (pairs)			% simulees successfully reaching a certain test length (pairs)					
	Mean	Min	Max	123	126	129	132	135	138
Strict	129.1	123	134	100.0	93.9	69.0	6.2	0.0	0.0
Lenient	132.1	125	137	100.0	99.9	93.4	65.7	9.9	0.0

Rank ordering and absolute differences

For adaptive assessments, the correlations and RMSEs between true and estimated scores were summarised graphically (Figures 36 and 37). The effect of social desirability balancing criteria was consistent with previous findings, with a more lenient criteria leading to better measurement from a theoretical standpoint.

Compared to Study 5, allowing negatively-loading item pairs provided more viable options in adaptive item selection, leading to slightly better measurement outcomes at a test length of 90 pairs (Table 47). Although the improvement was small for most scales, it made a huge difference for the Honesty-Humility scale. This finding was in line with earlier hypothesis that the constraint around negatively-loading item pairs might have had a greater impact on the Honesty-Humility scale due to its larger proportion of negatively-loading items. After removing this constraint, the Honesty-Humility scale was no longer the worst scale. The worst scale under the new settings was Openness to Experience, which had the smallest item pool of merely 24 adjectives. Apart from the Honesty-Humility scale, the relative ranking of measurement accuracy of the other five scales were largely consistent with previous simulation findings.

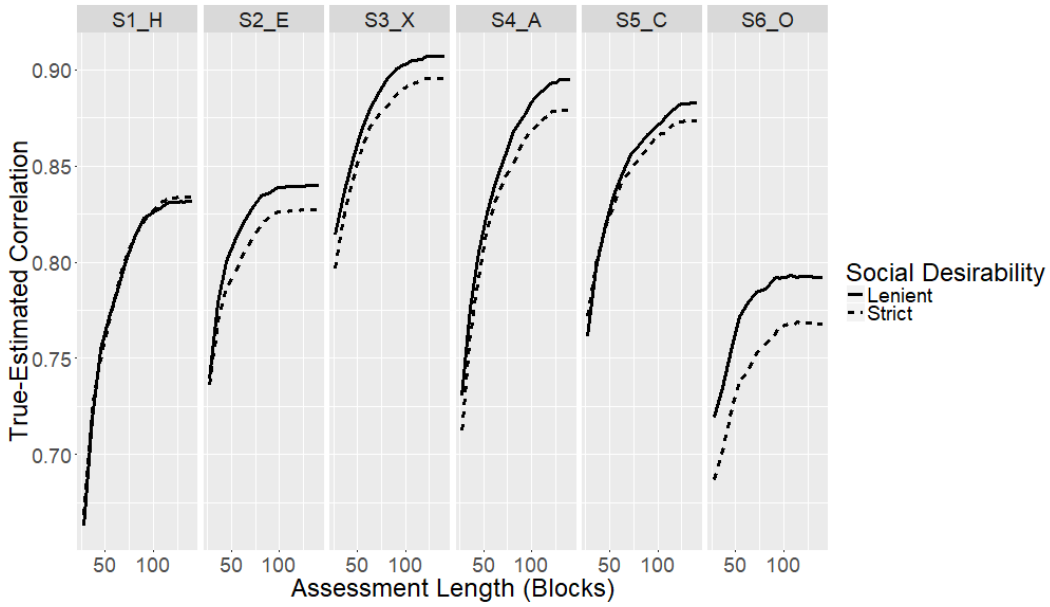


Figure 36. Correlations between true and estimated scores

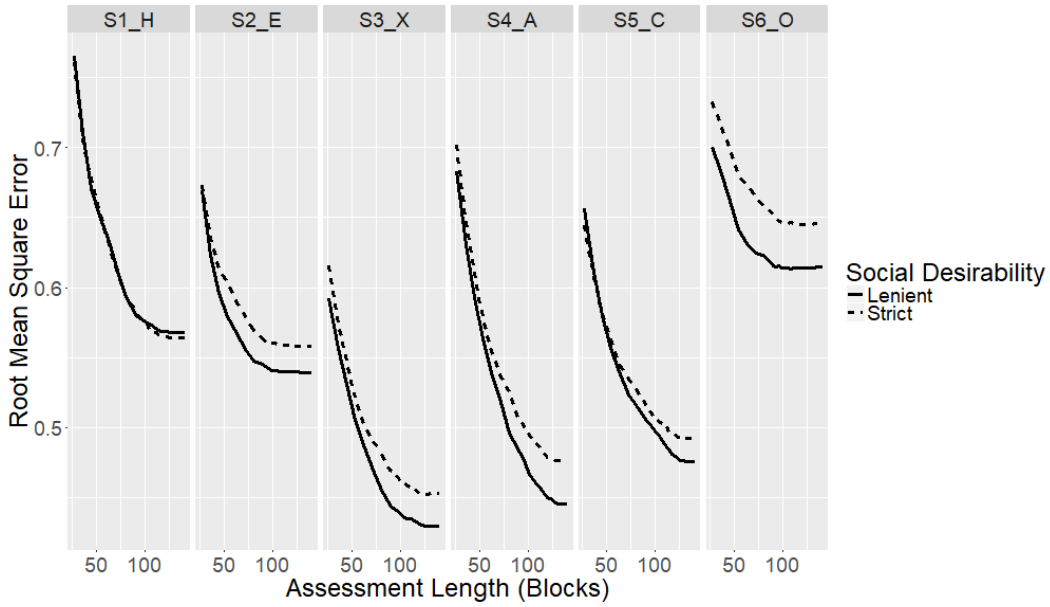


Figure 37. RMSEs between true and estimated scores

Table 47. Score correlations and RMSEs at 90 pairs with A-optimality and dynamic scales

Social desirability	Scale	No negative pairs		Allowing negative pairs	
		Correlation	RMSE	Correlation	RMSE
Strict	H	.74	0.69	.82	0.58
	E	.82	0.57	.82	0.56
	X	.88	0.48	.89	0.47
	A	.86	0.51	.86	0.51
	C	.83	0.57	.86	0.52
	O	.75	0.66	.76	0.65
	Mean	.81	0.58	.84	0.55
Lenient	H	.76	0.67	.82	0.58
	E	.84	0.55	.84	0.55
	X	.89	0.46	.90	0.44
	A	.87	0.48	.87	0.49
	C	.85	0.53	.87	0.51
	O	.79	0.61	.79	0.62
	Mean	.83	0.55	.85	0.53

Next, the correlations and RMSEs between true and estimated scores at different test lengths were considered in order to determine the target test length for the subsequent empirical study. According to Figures 36 and 37, it appeared that the additional gain in measurement accuracy was minimal from about 125 pairs onwards, so there was little reason to extending the test length beyond 125 pairs with the current item bank. Seeing that all simulated CAT sessions managed to reach 120 pairs, and that a typical respondent could comfortably complete 120 pairs of adjectives in 20 minutes (i.e., 6 pairs per minute, 10 seconds per pair), the test length for the subsequent empirical study was set to be 120 pairs. The simulated measurement properties at this test length are presented in Table 48. The Openness to Experience scale was still lacking in measurement accuracy, with reliability estimates of merely .59 and .63 for the

two social desirability conditions. The Honesty-Humility and Emotionality scales were also not optimal, with reliability estimates approaching .70. However, it was difficult to improve measurement any further with the limited item bank.

Table 48. Simulated CAT measurement properties at 120 pairs with A-optimality, dynamic scales, and allowing negative pairs

Social desirability	Scale	True-estimated score correlation	Reliability	RMSE
Strict	H	.83	.69	0.57
	E	.83	.68	0.56
	X	.90	.80	0.45
	A	.88	.77	0.48
	C	.87	.76	0.50
	O	.77	.59	0.65
Lenient	H	.83	.69	0.57
	E	.84	.70	0.54
	X	.91	.82	0.43
	A	.89	.80	0.45
	C	.88	.78	0.48
	O	.79	.63	0.61

Finally, with the target test length established, the measurement properties of non-adaptive versions of the assessments were simulated. Results are presented in Table 49. The differences in measurement efficiencies between adaptive and non-adaptive assessments were mostly small (Figures 38 and 39), with the non-adaptive conditions sometimes even doing better than the adaptive conditions. This likely resulted from having a small item bank, with the vast majority of items being used up at 120 pairs, thus greatly limiting the potential of adaptive item selection. The biggest difference was observed on the Openness to Experience scale, where the adaptive setting led to notably higher true-estimated correlations and lower RMSEs under the lenient social desirability

condition. This improvement was timely especially given that the Openness to Experience scale was the weakest measurement-wise. The practical effect of adaptive testing and different social desirability balancing criteria would be examined further in the subsequent empirical study.

Table 49. Simulated non-adaptive measurement properties at 120 pairs with A-optimality, dynamic scales, and allowing negative pairs

Social Desirability	Scale	True-estimated score correlation	Reliability	RMSE
Strict	H	.83	.68	0.58
	E	.82	.68	0.56
	X	.89	.79	0.47
	A	.88	.77	0.48
	C	.87	.76	0.50
	O	.77	.59	0.64
Lenient	H	.84	.71	0.55
	E	.82	.68	0.57
	X	.90	.80	0.45
	A	.89	.80	0.45
	C	.87	.75	0.50
	O	.77	.59	0.64

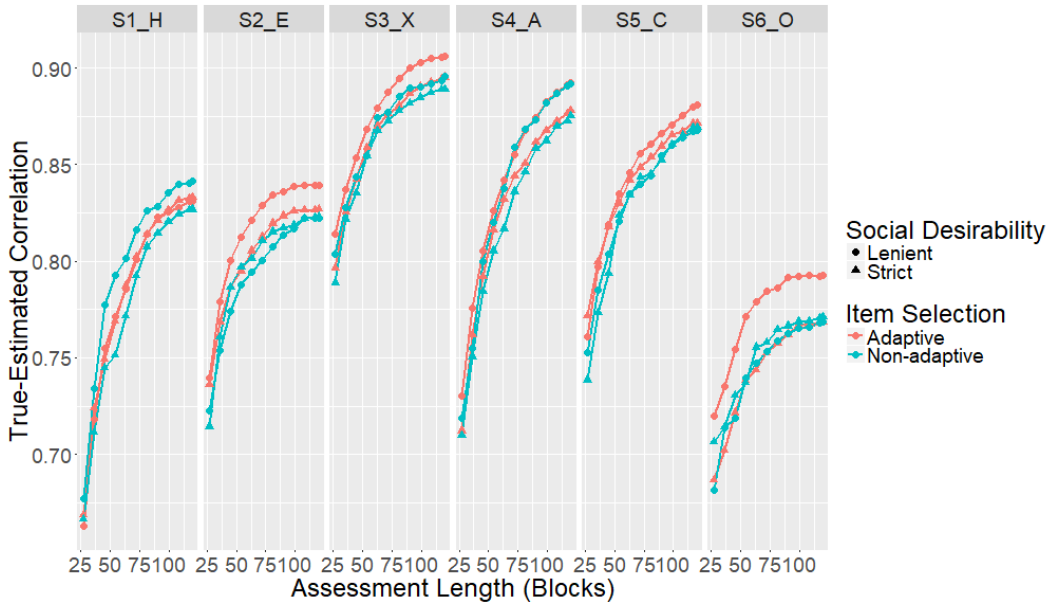


Figure 38. Correlations between true and estimated scores

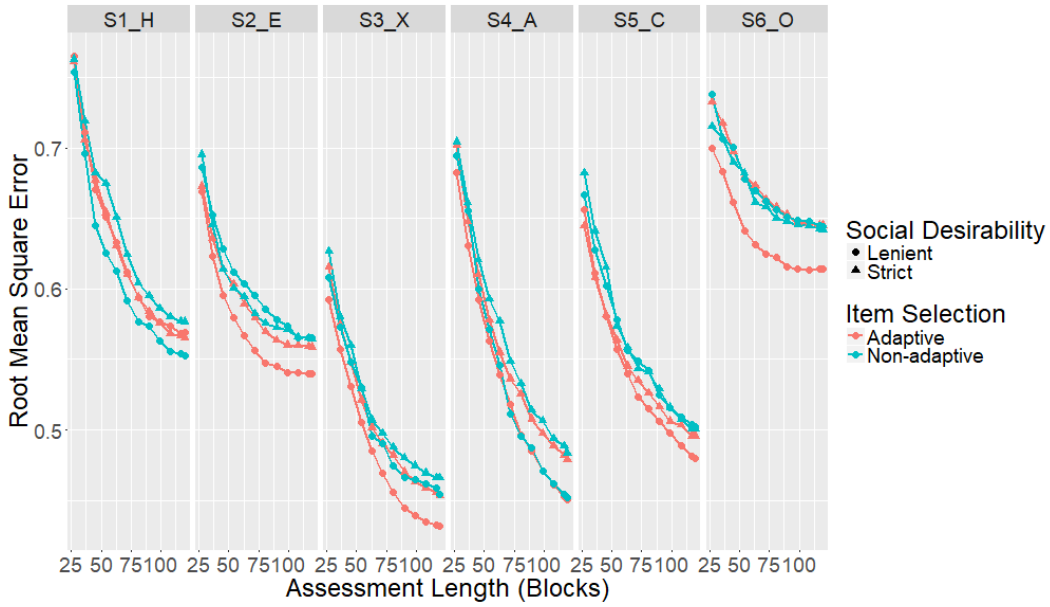


Figure 39. RMSEs between true and estimated scores

Summary

This study extended Study 5 and finalised the assessment design decisions for the subsequent empirical FC CAT study. It was discovered that allowing negatively-loading adjective pairs greatly increased the achievable test length with the current item bank, as well as improved measurement accuracy for the Honesty-Humility scale which had an item pool with a high proportion of negative items. Based on the CAT simulation results, a target test length of 120 pairs was chosen for the subsequent empirical study. The measurement differences between adaptive and non-adaptive (but otherwise optimised) assessments at the chosen test length appeared to be small, but still helpful in boosting measurement for the weaker scales especially given the limited item bank.

HEXACO FC CAT Empirical Trial (Study 6)

The final study of this thesis trialled the newly developed adaptive FC HEXACO personality assessment empirically. This study aimed to 1) explore the efficiency and utility of adaptive item selection and social desirability balancing criteria in empirical applications to identify further research questions and practical challenges; 2) examine participants' perceptions and opinions about FC and SS personality questionnaires.

Method

Sample and instruments

A large sample (N=1,440 who consented to providing data for research purposes) was recruited online in 2019 from a public-facing, pre-employment assessment practice website that was also used in the HEXACO item bank development study (Study 4). Using the same website for both studies helped to align their sampling populations, as

the characteristics of visitors to this website had been relatively stable historically. Just like Study 4, participants in this study were invited to complete the questionnaires in order to receive a personalised report.

After giving consent to partake in the research study, participants were invited to complete a FC personality instrument. The FC measures were constructed from the HEXACO adjective item bank developed in Study 4, using algorithm settings determined in Studies 5 and 5b: multidimensional pairs, A-optimality, dynamic scales, allowing both adjectives in a pair to be negatively-loading, and a target test length of 120 pairs. The investigation crossed the settings of adaptive item selection (adaptive versus non-adaptive) and social desirability balancing criteria (lenient versus strict, defined as per previous studies), giving rise to four design conditions: adaptive with lenient social desirability (AL), adaptive with strict social desirability (AS), non-adaptive with lenient social desirability (NL), and non-adaptive with strict social desirability (NS). The adaptive measures always attempted to find the best FC pair for the participants' interim trait estimates (starting from the origin), leading to initially similar but subsequently divergent questions for different participants as their trait estimates evolved. The non-adaptive measures, on the other hand, always targeted measurement at the origin (i.e., the calibration sample mean), and did not change between participants as the assessments progressed. A between-subject design was adopted – each participant was randomly routed into one of the four design conditions. Participants were not informed of the random routing and did not know which route they were assigned to.

Following the FC instrument, each participant then responded to the 60-item HEXACO-PI-R (Ashton & Lee, 2009). The administration of the HEXACO-PI-R served three purposes. First, it generated HEXACO personality scores as a personalised

report to incentivise participation. Second, it provided data to examine the construct validity of the new FC measures. Third, it offered assessment experience with the SS question format, prior to asking participants to compare the FC and SS question formats.

Following the FC and SS instruments, participants were presented with 10 feedback questions about their experience with the two questionnaires (Appendix G). It was made clear to the participants that these questions were optional and would not affect their personality reports in any way, so that only the participants who were motivated to help with the research effort would complete them. The feedback questions asked how frequently the participants noticed pairs of adjectives that were both like them or both unlike them (i.e., pairs with similar item utilities), in order to investigate whether adaptive item selection would lead to notably more difficult choices for the participants. The perception around social desirability of items was also investigated, through quantifying the perceived frequencies of FC adjective pairs with clearly unmatched social desirability, and the perceived frequencies of SS statements with clearly desirable or undesirable social connotations. Participants were then asked to compare the FC and SS instruments, indicating whether they felt that one of them: 1) was easier to complete; 2) made them think deeper about their own personality when answering; 3) gave them a better chance to describe their personality fully; 4) gave a more preferable test experience on the whole; and 5) made a fairer test for comparison between people. Finally, in order to gauge the perception of how fakable the different question formats were, participants were asked to imagine someone trying to answer the questions dishonestly in order to appear good, and rated how successful they thought that person would be in increasing their scores on the FC and SS instruments respectively.

Finally, participants were presented with six background questions (Appendix H). Gender, age and self-rated English proficiency data were collected in order to capture the characteristics of the sample. English proficiency data also helped to ensure that the final sample consisted of participants who had good understandings of the English adjectives used in the FC measures. Then, in order to understand the mindsets in which participants were completing the personality questionnaires, the questions explored whether their completion was a repeated participation, and whether their motivations to participate were associated with gaining experience for pre-employment assessments, finding out more about themselves, or something else. Repeated completions and uncommon motivations to participate could result in unnatural responding behaviours, thus introducing unpredictable contaminations to the study results.

The study website was built using javascript and integrated with R codes developed for this thesis. The website was hosted on an Amazon Web Services (AWS) server, which was chosen to provide enough computational power for running simultaneous FC CAT sessions for multiple participates without causing notable delays in adaptive item presentation. In order to monitor that this was indeed the case, the server processing time from receiving a FC response to sending the next FC question was logged. In addition, the elapsed time between the server sending a FC question till receiving the question response was also logged in order to give an estimate of the typical time participants spent considering each question. However, these response times could be inflated by bad internet connections or by participants taking breaks. Given the likelihood of such contaminations, all analysis involving response times were merely exploratory. A proper study of response times would necessitate the standardisation of study environment across participants, which was not possible with this online study. Nevertheless, the data on question generation times and participant

response times helped to reconstruct what it felt like for the respondents to complete the FC measures in this study. For each participant, the HEXACO-PI-R item response times were also logged, as well as the overall elapsed time from the first response (giving consent to participate) to the last response (submitting background questions prior to receiving personalised report).

Table 50. Data cleaning criteria

Data cleaning criteria	Cases
Participants whose English proficiency level did not reach "Professional working proficiency" or higher.	102
Completions by the same participants (keeping data for the first completion only).	87
Participants who partook in the study for reasons other than "to practice for pre-employment assessments" or "to find out more about myself".	31
Participants who completed the study too quickly (<10 minutes, indicating lack of proper consideration) or too slowly (>2 hours, indicating presence of distraction during completion).	57
Participants with unreliable response patterns (e.g. when the majority of the rating scale was never used, when a particular response option was overused, when the responses had a very small standard deviation).	13

Due to the lack of participation control in online studies, extensive cleaning was applied in order to ensure data quality (Table 50). The final cleaned sample consisted of 1,150 cases. The sample was balanced in terms of gender, and all working ages were represented (Table 51). About two fifths (39.1%) of the sample indicated that they had "native or bilingual proficiency" in the English language, a further third (32.0%) had "full professional proficiency", while the remaining (28.9%) had "professional working proficiency". Most participants (57.8%) spent between 20 to 40 minutes completing the study. Participants joined the study in order to practice for pre-employment assessments (87.4%) and/or to find out more about themselves (70.6%). With the random routing of

different FC measures, each of the four conditions was completed by between 279 to 301 participants.

Table 51. Cleaned sample demographics (N=1,150)

Sample Demographics		%
Gender	Male	51.0
	Female	44.8
	Missing	4.3
Age	Up to 20	1.8
	21 to 30	31.7
	31 to 40	32.0
	41 to 50	20.0
	51 to 60	8.7
	Missing	5.8
English language proficiency	Native or bilingual proficiency	39.1
	Full professional proficiency	32.0
	Professional working proficiency	28.9

Across all four conditions, the server was responsive in returning the next question in a timely manner despite a traffic flow of approximately 100 completions per day. Across all FC questions for all participants, most of the time (96.9% to 99.3% per condition) the next question was ready in less than one second, ensuring that the assessment experience was not hindered by excessively long wait times due to adaptive item selection. Very occasionally (0.02% to 0.09% per condition), the server had taken over 5 seconds to return the next question. This occasional delay appeared to be random and affected both adaptive and non-adaptive sessions equally. It was likely caused by server overload and would not introduce systematic bias to the comparison between adaptive and non-adaptive conditions.

Analysis strategy

Analysis explored the effect of three assessment design factors on three types of outcomes. The design factors considered were: 1) adaptive versus non-adaptive FC measures; 2) strict versus lenient social desirability balancing in FC measures; and 3) FC versus SS measures. The outcomes explored included: 1) measurement; 2) response times; and 3) participant perception. The relationships between design factors and outcomes were examined systematically. Although a small number of predictions were made, most of the analysis was exploratory.

Measurement precision and score distributions

For FC measures, SEMs were computed according to TIRT information functions (Equation 21). In general, adaptive measures were expected to achieve greater measurement precision, resulting in lower SEMs. However, based on earlier simulation results using the same item pool (Study 5b), measurement improvement due to adaptive item selection would likely only occur for some of the scales. In terms of social desirability balancing, while lenient criteria tended to lead to better measurement in a pure simulation setting (i.e., responding according to latent trait values only), analysis would explore whether that would remain the case in a practical setting with the presence of actual social desirable responding behaviours.

In terms of question format, the FC and SS measures utilised completely different item banks. As item content played a significant role in measurement, any differences in measurement precision cannot be attributed to the question format alone. Therefore, the analysis of question format on measurement focused on comparing score distributions and intercorrelations instead. For this purpose, it was useful to also consider the sample from Study 4, as it deployed adjectives in a SS format, leading to three different measurement setups across the two studies: 1) adjectives in FC format, 2)

adjectives in SS format, and 3) HEXACO-PI-R statements in SS format. As participants for both studies were recruited from the same assessment practice website using the same incentive and shared similar characteristics and motivations, it is reasonable to assume that the samples were drawn from the same population and are therefore directly comparable. The similarity of HEXACO-PI-R scores across samples would signify any sampling differences, as this instrument did not change across studies. As for the adjective-based scores, moving from SS to FC format would likely introduce some differences. More specifically, due to the removal of uniform response biases and the reduction of social desirability responding using the FC format, the resulting score means and correlations would likely be lowered, and the correlations with HEXACO-PI-R scores (which would be affected by social desirability responding as per adjectives administered in SS format) would likely be reduced.

Response time

For each participant, response times were captured for 120 pages of one FC adjective pair each, and 20 pages of three SS statements each (the response time per SS statement was then calculated as the response time for the entire page divided by three). In order to avoid the influence of outliers (i.e., excessively long response times caused by slow internet speed or participants taking a break), the analysis of response time focused on percentiles rather than means. It was anticipated that adaptive item selection would result in pairs of adjectives that had similar utilities for the participant, and strict social desirability balancing criterion would result in pairs of adjectives with more aligned average endorsement levels in the population. Both scenarios were expected to lead to more difficult choices and possibly longer response times. In terms of response format, FC adjectives were expected to require less time per question than SS adjectives because 1) reading two adjectives would likely be faster than reading a long statement;

and 2) FC pairs demanded a simple binary judgement but SS statements demanded a more complex judgement against several response categories, which would likely take longer (but would also provide more information about the respondent per response).

Participant perceptions

Response frequencies for the 10 feedback questions were summarised and compared across design conditions. As explained earlier, it was anticipated that adaptive item selection/ strict social desirability balancing would result in more difficult choices, leading to higher reported frequencies of adjective pairs that were equally like the participants/ socially desirable, as well as lower success in faking good. In terms of question format, it was anticipated that participants would find the SS format easier to complete but also easier to fake good due to its familiarity and transparency. However, the FC format was expected to provoke deeper thinking about ones' personality when responding. It remained unclear which type of measure would be perceived as giving participants a better chance to describe their personality, giving a more preferable test experience, or providing fairer comparisons between people.

Results

Measurement and adaptive item selection

Measurement precision statistics showed that the adaptive conditions tended to achieve lower SEMs compared to non-adaptive conditions with the same social desirability balancing criteria (Table 52). With lenient social desirability balancing, adaptive item selection reduced sample mean SEMs (by –0.014 to –0.002) for all six scales. With strict social desirability balancing, adaptive item selection reduced sample mean SEMs (by –0.007 to –0.004) for five out of six scales, but instead resulted in lower measurement precision for the Conscientious scale (sample mean SEM +0.002).

The full distributions of SEMs across all individuals in the sample are shown in Figure 40, which confirmed the notable but small advantage of adaptive item selection on measurement precision. Simulation Study 5b suggested that the advantage of adaptive item selection would be more prominent on the Emotionality, eXtraversion and Openness to Experience scales, which was confirmed by the empirical results.

Table 52. Sample mean SEMs by design conditions

Scale	AL (N=301)	AS (N=288)	NL (N=279)	NS (N=282)
H	0.527	0.530	0.529	0.534
E	0.502	0.517	0.516	0.524
X	0.417	0.423	0.426	0.431
A	0.400	0.408	0.402	0.411
C	0.449	0.449	0.452	0.447
O	0.605	0.614	0.615	0.619

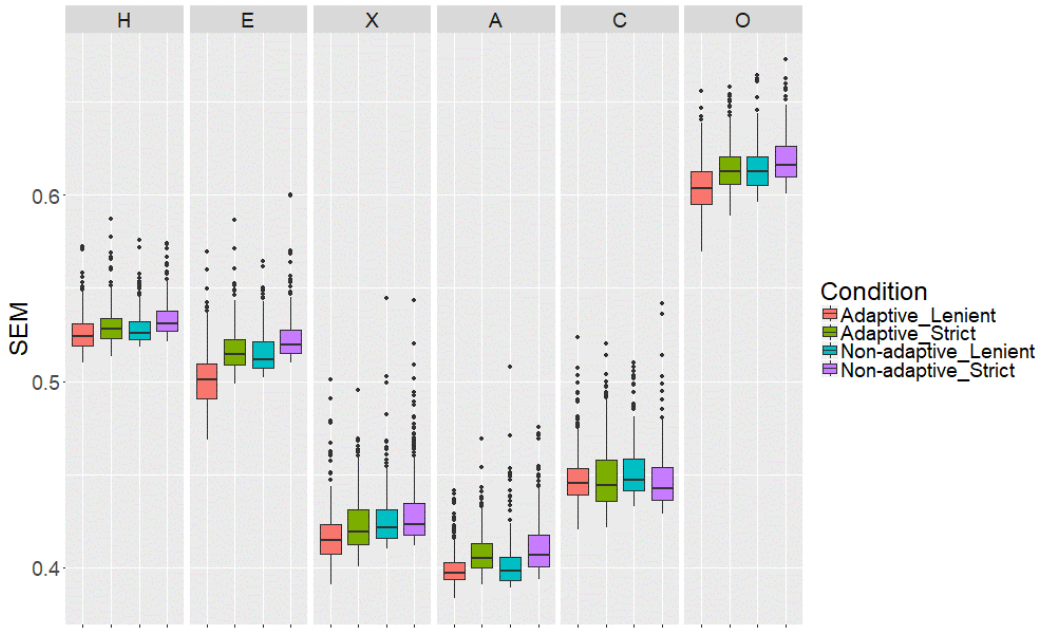


Figure 40. SEMs by design conditions

The effect of assessment length on measurement precision was also examined (Figure 41). For the initial phase of measurement (i.e., up to approximately 25 pairs), there were no visible differences between adaptive and non-adaptive conditions, likely due to having insufficient information to produce reliable interim trait estimates for driving an effective tailored assessment approach. As measurement progressed, interim trait estimates improved and adaptive item selection started to make an impact. However, for this particular study, the item pool also started drying out (i.e., each measure used 240 out of 279 items in the limited pool), thus limiting the potential of adaptive item selection towards the later phase of measurement. In the end, only Emotionality, eXtraversion and Openness to Experience scales showed visible but very small improvements when adaptive item selection was used.

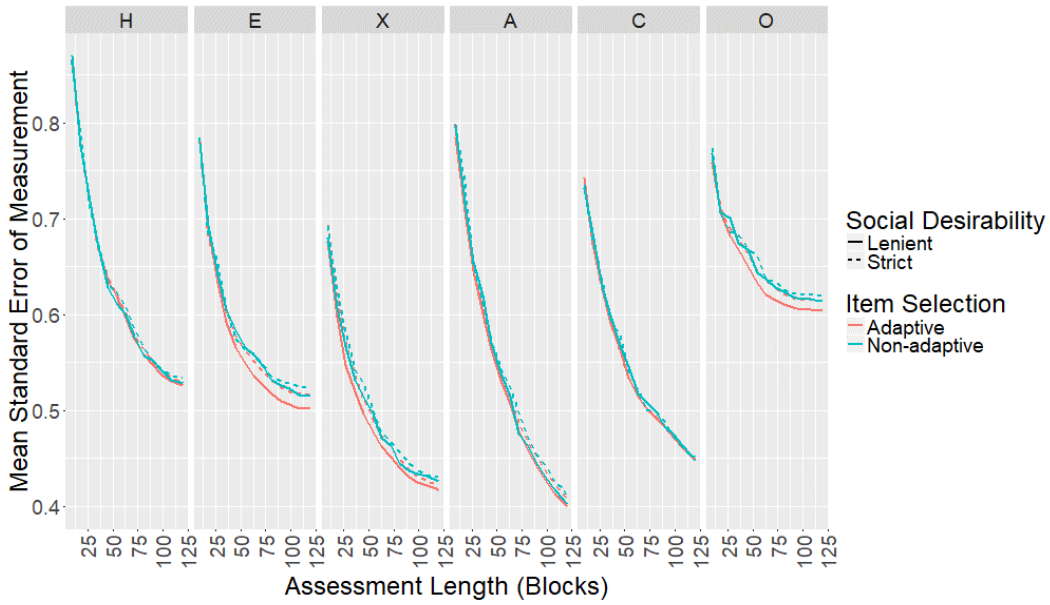


Figure 41. Sample mean SEMs by test length and design conditions

In order to understand the effect of adaptive item selection at the individual level, measurement precision was examined against estimated trait values for each scale (Figure 42). It appeared that the advantage of adaptive item selection was more

prominent for certain trait values. For example, with lenient social desirability balancing, adaptive item selection enhanced measurement for low Emotionality and high eXtraversion, but made little difference to measurement for high Emotionality or low eXtraversion. These results suggested that the relative merit of adaptive item selection in increasing measurement precision might be highly dependent on the composition of the underlying item pool, as well as the characteristics of the target candidate population.

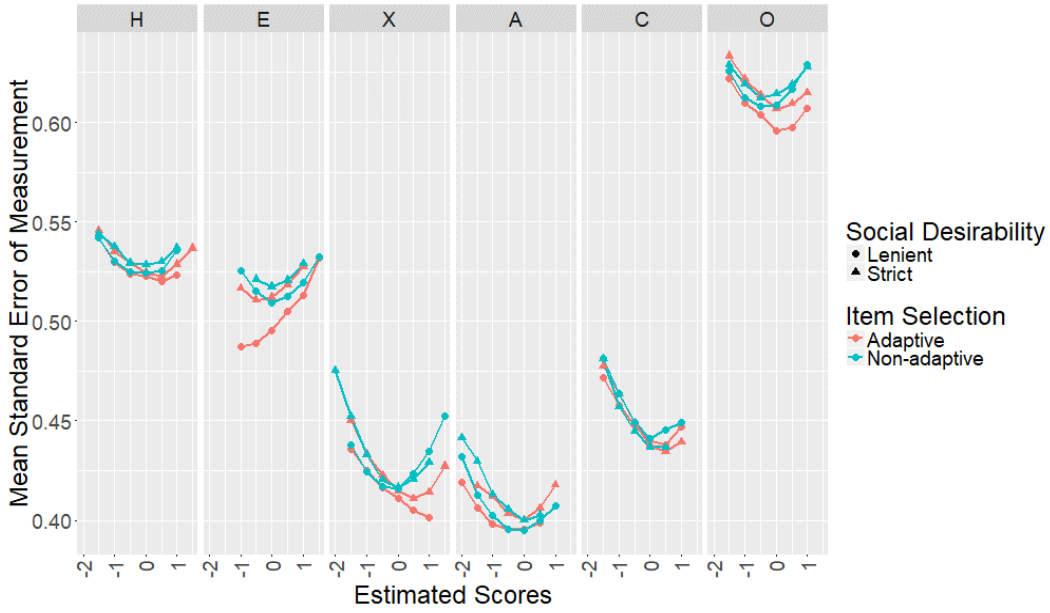


Figure 42. Sample mean SEMs by trait values and design conditions

Finally, in order to gauge the overall effect of adaptive item selection across all six scales simultaneously, measurement precision was examined against the participants' profile distance from the origin (i.e., the starting location of adaptive item selection). Figure 43 plots the profile mean SEMs (i.e., average SEM across all six scales for each participant) against the Euclidean distance between their estimated score profile and the origin. Regardless of design conditions, results showed that the score profiles further away from the origin tended to have larger SEMs compared to the score profiles nearer to the origin. This observation was not surprising, because item selection was optimised

for profiles around the origin at least initially for adaptive sessions, and at all times for non-adaptive sessions. Results also showed that adaptive item selection helped to counter this effect, by improving measurement precision for profiles further away from the origin.

Figure 43. Profile mean SEMs by distance from the origin and design conditions

Measurement and social desirability balancing criteria

Measurement precision statistics showed that lenient social desirability balancing tended to achieve lower SEMs compared to strict social desirability balancing with the same item selection method (Table 52). With adaptive item selection, lenient social desirability balancing reduced sample mean SEMs (by –0.015 to –0.001) for all six scales. With non-adaptive item selection, lenient social desirability balancing reduced sample mean SEMs (by –0.009 to –0.005) for five out of six scales, but instead resulted in lower measurement precision for the Conscientious scale (sample mean SEM +0.004). The full distributions of SEMs across all individuals in the sample are shown

in Figure 40, which confirmed the visible but small advantage of lenient social desirability balancing on improving measurement precision. Moreover, lenient social desirability balancing was sometimes required for the advantage of adaptive item selection to emerge (Figure 42), and helped such advantage to appear earlier in the assessment process (Figure 41). With lenient social desirability balancing, the difference between adaptive and non-adaptive item selection also became more prominent further away from the origin (Figure 43).

Moreover, having lenient social desirability balancing didn't lead to more desirable scores than when strict social desirability balancing criterion was applied. For non-adaptive conditions, the sample mean different effect sizes across social desirability balancing criteria were negligible (Cohen's d magnitude < 0.10 on all six factors). For adaptive conditions, the sample with lenient social desirability balancing actually received generally less desirable scores (Cohen's d = −0.250 for H, 0.180 for E, −0.178 for X, −0.163 for A, 0.079 for C, and −0.158 for O), suggesting that it wasn't affected by social desirability responding more so than the sample with strict social desirability balancing.

It appeared that the lenient social desirability balancing criterion (i.e., item mean difference of 1 or less) did not provide less resistance to score inflation than the strict social desirability balancing criterion (i.e., item mean difference of 0.5 or less). Rather, the strict social desirability balancing criterion was overly restrictive and hindered freedom of adaptive item selection in this study, reducing measurement accuracy as a result. So the more lenient criterion was more preferable for this study. However, it remains unclear whether this conclusion will still hold beyond the range of social desirability balancing values considered in this study (i.e., item mean difference of over 1) – it is plausible that larger social desirability differences in a pair would trigger

greater opportunities for social desirability responding, so it is likely that the social desirability balancing would become too relaxed after a certain range. Furthermore, the extent of social desirability responding is correlated with the stakes of the assessment (e.g., Birkeland et al., 2006), so the point at which social desirability responding becomes a problem could vary depending on the assessment setting and purpose, with high-stakes assessments demanding stricter social desirability balancing criteria.

Measurement and question format

Table 53. Score means and standard deviations by measure and study

Study	Study 4 (N=1,685)				Study 6 (N=1,150)			
Items	Adjectives		HEXACO-PI-R		Adjectives		HEXACO-PI-R	
Format	SS		SS		FC		SS	
Scale	Mean	SD	Mean	SD	Mean	SD	Mean	SD
H	0.00	0.94	0.00	0.89	–0.25	0.76	–0.23	0.91
E	0.00	0.95	0.00	0.86	0.22	0.63	0.23	0.85
X	0.00	0.98	0.00	0.94	–0.24	0.86	–0.11	0.92
A	0.00	0.98	0.00	0.89	–0.54	0.72	–0.37	0.93
C	0.00	0.97	0.00	0.92	–0.35	0.64	–0.29	0.96
O	0.00	0.92	0.00	0.90	–0.16	0.76	–0.10	0.89

The distributions of HEXACO scores from this study and Study 4 are presented in Table 53. All scores were calculated using the item parameters estimated as part of Study 4. Despite both studies having a common source of participants, on average the current study received less desirable scores (i.e., higher mean scores in Emotionality but lower mean scores in the other five traits) than Study 4, even on the same HEXACO-PI-R instrument. This indicates the presence of some differences between the samples from the two studies. Interestingly, while the SS adjective scores and SS HEXACO-PI-R scores in Study 4 had the same means, the FC adjective scores in the current study

tended to be lower than the SS HEXACO-PI-R scores for the same sample. This was likely due to the FC format preventing uniform response biases (e.g., acquiescence) and reducing social desirability responding. Moreover, the FC adjective scores demonstrated smaller variances compared to scores based on SS measures, possibly as a result of the shrinkage caused by multidimensional Bayesian scoring.

Table 54. Score correlations by measure and study

Study	Scale	H	E	X	A	C	O
Study 4	H	**.59**	–.14	.24	.42	.37	.15
(N=1,685)	E	–.47	**.60**	–.36	–.27	–.23	–.12
	X	.37	–.55	**.79**	.36	.44	.27
	A	.72	–.51	.54	**.67**	.36	.17
	C	.64	–.61	.50	.64	**.75**	.20
	O	.37	–.40	.42	.46	.44	**.63**
Study 6	H	**.33**	–.16	.24	.40	.39	.20
(N=1,150)	E	–.24	**.45**	–.33	–.21	–.19	–.19
	X	.14	–.49	**.61**	.33	.37	.34
	A	.36	.01	.21	**.34**	.31	.15
	C	.18	–.22	.27	.05	**.43**	.22
	O	.08	–.25	.33	.00	.15	**.52**

Above diagonal: HEXACO-PI-R score intercorrelations (SS format).
Below diagonal: Adjectives score intercorrelations (SS or FC format).
Diagonal: Correlations between HEXACO-PI-R and adjective measures.

The correlations between estimated scores were also considered. Table 54 presents the intercorrelations between HEXACO scores from the same measure (above diagonal for HEXACO-PI-R, below diagonal for adjectives) and the convergent correlations across different measures for the same sample (on the diagonal). The intercorrelations of HEXACO-PI-R scores were very stable across studies, with all differences having a magnitude smaller than 0.1. This was expected as the HEXACO-

PI-R instrument was identical across studies and the sampling population remained the same. The intercorrelations of adjective-based scores, however, differed significantly across studies. Adjective-based scores using the SS format (i.e., Study 4) demonstrated much stronger correlations than those using the FC format (i.e., Study 6). For example, the correlations between Agreeableness and Conscientiousness was .64 based on SS scores, but .05 based on FC scores. Similarly, the correlations between Agreeableness and Emotionality was –.51 based on SS scores, but .01 based on FC scores. It should be noted that the low intercorrelations of FC scores were not a result of the FC scores being ipsative – the average off-diagonal correlations for the FC adjective scores was small yet positive (.04), whereas ipsative scores would have resulted in negative average off-diagonal correlations. It was noted that the correlations based on FC adjective scores were more conceptually plausible than those based on SS adjectives. The inflation of intercorrelations between SS adjective scores thus suggested that a strong method factor was at play (likely an ideal employee factor given the source of the samples), making the observed scores across different traits more closely aligned than their conceptual relationships. Also, the FC response format resulted in score correlations that were more in line with (but slightly lower than) those from the HEXACO-PI-R instrument. It was interesting that, despite adopting a SS response format, the HEXACO-PI-R instrument appeared to be much less affected by the method factor than the adjectives. The reason of this difference might be the vagueness of the adjectives, which might elicit quick system 1 responses (Kahneman, 2011) and making them more prone to biases including socially desirable responding, whereas HEXACO-PI-R statements provide more context and thus likely encourage system 2 thinking (Kahneman, 2011) more so than simple adjectives. Finally, the convergent correlations between adjective and HEXACO-PI-R scores were stronger in Study 4 (.59 to .79, mean .67) than Study 6 (.33 to .61, mean .45). This result suggested that the response format had a substantial effect on

construct validity. Note that the biggest reductions in convergent validity were observed for Agreeableness (–0.33), Conscientiousness (–0.31) and Honesty-Humility (–0.26), which are consistently found to be most important in the employee selection settings and thus providing supporting evidence that the difference in construct validity is related to method (i.e. response format).

Response time

The distributions of question-level response times (seconds per FC pair) across all questions for all candidates (120 × 1150 = 138,000 data points in total) are shown in Table 55. In general, the response times were very comparable across design conditions. In line with the direction of prediction, adaptive conditions resulted in consistent (Table 56) but negligible (Table 55) increases in response time. However, contrary to prediction, strict social desirability matching showed consistent (Table 56) but negligible (Table 55) decreases in response time. One possible explanation for the latter was that, when social desirability was less balanced within a FC pair, participants spent slightly longer weighing up their own personality against social expectations; whereas when social desirability was balanced, participant only needed to consider their own personality. This observation was in line with suggestions that when candidates were presented with equally desirable or undesirable items, they would give up guessing which ones were more desirable and respond honestly. Nevertheless, the differences observed were very small and must be interpreted with caution given the limitations around data collection settings in this study. The true effect of adaptive item selection and/or social desirability balancing on response time needs to be studied in a more controlled environment than this unsupervised online study.

Table 55. FC pair response time percentiles by design conditions

Percentiles	AL (N=36,120)	AS (N=34,560)	NL (N=33,480)	NS (N=33,840)
10th	2.93	2.85	2.72	2.75
20th	3.55	3.44	3.31	3.32
30th	4.12	4.01	3.88	3.86
40th	4.73	4.61	4.51	4.40
50th	5.46	5.33	5.21	5.03
60th	6.44	6.36	6.30	5.96
70th	7.69	7.74	7.66	7.11
80th	10.12	10.45	10.21	9.18
90th	18.25	18.56	18.17	15.11

Table 56. Kruskal-Wallis rank sum test of response time by condition

Comparison	Kruskal-Wallis chi-squared	df	p-value	Epsilon squared effect size
All four conditions	302.76	3	< 0.001	0.0022
A vs N	234.84	1	< 0.001	0.0017
S vs L	64.08	1	< 0.001	0.0005
AL vs NL	76.99	1	< 0.001	0.0011
AS vs NS	158.65	1	< 0.001	0.0023
AL vs AS	13.43	1	< 0.001	0.0002
NL vs NS	51.52	1	< 0.001	0.0008

The analysis of response time with respect to question location gave rise to an interesting observation. For the FC measures, there were clear signs that response time decreased greatly as the assessment continued – the median response time was about eight seconds per pair at the start and reduced to about five seconds per pair at the end (Figure 44). The reduction in response time likely resulted from participants getting more comfortable with the FC response format, or/and participants getting less motivated towards the end of a long questionnaire. Response time for the SS measure

remained relatively stable throughout the assessment – after familiarisation with the initial couple of pages taking nearly nine seconds per item, the median response time quickly settled to around seven seconds per item till the end (Figure 45). The median response time of about seven seconds per rating scale statement was in line with the historical response times on similar content by participants from the same pre-employment assessment practice website. The stable and historically-aligned response times on HEXACO-PI-R despite it being taken after the 120-pair FC measure suggested that the likelihood of fatigue leading to reduced response times was low, and thus the decreasing response time for FC pairs was more likely attributable to increased familiarity and comfort with the FC response format.

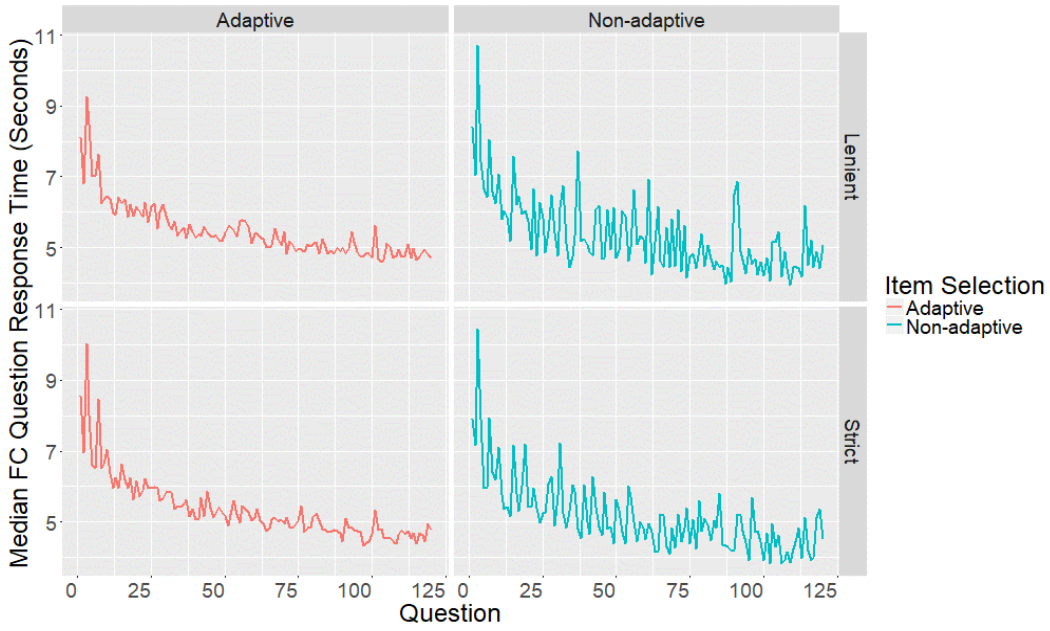


Figure 44. Median FC pair response time by question location and design conditions

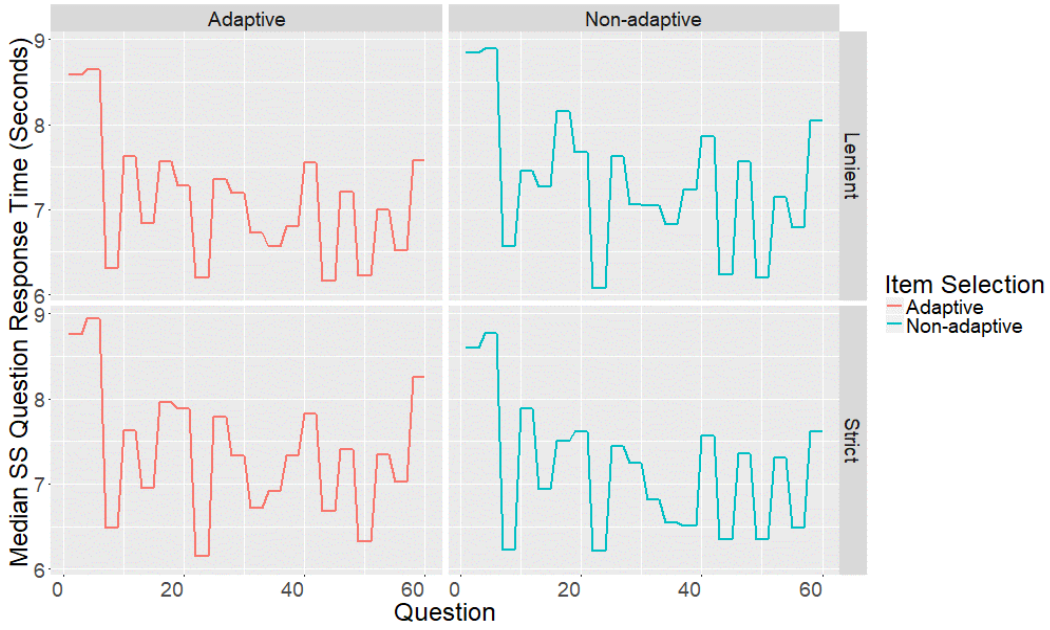


Figure 45. Median SS item response time by question location and design conditions

Participant perceptions

Despite clearly stating that the 10 feedback questions (Appendix G) were optional and inconsequential, most participants were still motivated enough to provide responses to them to help with the research effort (valid N=1,045 to 1,090 per question). Despite having "don't know" as one of the response options, all respondents indicated the approximate frequency in which they encountered adjectives with similar utility (Table 57). The same was not true for item social desirability, where a small number of respondents indicated that they "don't know" (6.4% and 6.7% for FC and SS formats respectively, Table 57). These respondents likely were not considering social desirability when answering the questionnaires. Contrary to a priori predictions, participants across different design conditions appeared to share very similar observations around item utility and social desirability, with no significant differences across conditions (Table 58).

Table 57. Participant perception around item utility and social desirability

Frequency of occurrence	FC: similar utility (N=1,045)	FC: similar social desirability (N=1,081)	SS: obvious social desirability (N=1,090)
0% of the time	1.1%	1.5%	3.3%
25% of the time	30.8%	36.7%	35.3%
50% of the time	42.8%	34.4%	30.6%
75% of the time	23.3%	18.8%	19.9%
100% of the time	2.0%	2.2%	4.1%
Don't know	0.0%	6.4%	6.7%

Table 58. Kruskal-Wallis rank sum test of participant perception of item utility and social desirability

Feedback question*	Kruskal-Wallis chi-squared	df	p-value
FC: similar utility	2.84	3	.42
FC: similar social desirability	6.70	3	.08
SS: obvious social desirability	2.46	3	.48

* For significance testing, "don't know" responses were treated as missing.

Participants also compared their experience across the 120-pair FC and 60-item SS instruments (Table 59). In terms of ease of completion, as anticipated the majority (70.0%) of respondents preferred the SS instrument, but about one in six (16.3%) preferred the FC instrument despite it containing twice as many questions, and about one in seven considered them to be the same (10.8%) or had no opinion (3.0%). Very similar percentages were observed when respondents considered which instrument gave them a better chance to describe their personality. On the other hand, as anticipated, the FC instrument was more successful than the SS instrument (63.1% versus 25.9%) in provoking respondents to think deeper about their own personality. Overall, more

respondents preferred the testing experience of the SS instrument (63.0%) than the FC instrument (13.1%), and about one in four respondents found them to be the same (17.7%) or had no opinion (6.2%). In terms of perceived fairness for comparison between people, just over half (53.5%) of the respondents considered the SS instrument to be fairer, about one in seven (13.7%) found the FC instrument fairer, about one in seven (14.0%) considered both instrument to be equally fair, and nearly one in five (18.9%) did not have an opinion. Participants across different design conditions appeared to share very similar opinions when comparing the questionnaires (Table 60).

Table 59. Participant opinions on the FC and SS questionnaires

Question	N	FC	SS	The same	Don't know
Easier to complete	1082	16.3%	70.0%	10.8%	3.0%
Think deeper about own personality	1083	63.1%	25.9%	8.0%	3.0%
Better chance to describe personality	1082	16.1%	68.2%	9.9%	5.8%
More preferable test experience	1082	13.1%	63.0%	17.7%	6.2%
Fairer for people comparison	1082	13.7%	53.5%	14.0%	18.9%

Table 60. Kruskal-Wallis rank sum test of participant opinions on the FC and SS questionnaires by condition

Question format comparison*	Kruskal-Wallis chi-squared	df	p-value
Easier to complete	1.60	3	.66
Think deeper about own personality	2.83	3	.42
Better chance to describe personality	2.00	3	.57
More preferable test experience	6.88	3	.08
Fairer for people comparison	1.20	3	.75

* For significance testing, responses were reordered so that "Rating Scale"=–1, "The same"=0, "Forced Choice"=1. "Don't know" responses were treated as missing.

Table 61. Participant opinions on question formats and faking good

Score Inflation Success	Rating Scale (N=1,090)	Forced Choice (N=1,090)
Not at all successful	20.0%	35.8%
Somewhat successful	41.9%	40.5%
Very successful	21.5%	6.1%
Extremely successful	3.9%	1.3%
Don't know	12.8%	16.3%

Table 62. Kruskal-Wallis rank sum test of participant opinions on question formats and faking good by condition

Score inflation success*	Kruskal-Wallis chi-squared	df	p-value
SS	1.82	3	.61
FC	3.14	3	.37

* For significance testing, "don't know" responses were treated as missing.

Finally, participants considered how successful a dishonest candidate might be in inflating scores for the SS and FC instruments. While most respondents considered the SS instrument to be fairer when comparing between people (Table 59), the FC instrument was considered less fakable (Table 61). About a third (35.8%) of respondents indicated that faking good on the FC instrument would be “not at all successful”, compared to one in five (20.0%) for the SS instrument. About one in five respondents (21.5%) thought faking the SS instrument could be done “very successfully”, compared to merely 6.1% who thought the same for the FC instrument. When the ratings were averaged across participants (coding “Not at all successful” to “Extremely successful” as 1 to 4, and coding “Don't know” as missing), the means were significantly different between SS and FC (SS mean = 2.105, FC mean = 1.677, t = –

16.797, df = 883, $p < 0.001$). Opinions appeared to be relatively stable across participants in different design conditions (Table 62).

Discussion

This study explored the empirical effect of adaptive item selection, social desirability balancing criteria and question format on measurement, response time, and participant perception. The analysis was largely exploratory and the results were mixed.

Adaptive item selection

It was confirmed that adaptive item selection achieved greater measurement precision than non-adaptive item selection. However, the incremental gain of adaptive item selection on measurement precision was much smaller than those reported in similar literature (e.g., Joo et al., 2019; Stark & Chernyshenko, 2007, 2011; Stark et al., 2012). One contributing factor to this was the choice of baseline reference in this study – while CAT research typically adopted random item selection with some content constraints as the baseline for comparison (e.g., Stark & Chernyshenko, 2011), this study chose a more realistic operational alternative that incorporated measurement optimisation considerations (i.e., by choosing FC pairs to maximise information gain at the population average as opposed to choosing FC pairs randomly). In other words, this study explored the practical return on investment when converting an otherwise-optimised static assessment into an adaptive one. Another contributing factor to the small adaptive advantage was the very limited item bank, with each FC assessment using up 240 out of 279 available items, thus greatly limiting the possibility and potential of adaptive item selection towards the end of the assessment sessions. Therefore, the presence of a large and varied item bank would likely be a pre-requisite for effective FC CAT. Nevertheless, the benefit of adaptive item selection on measurement precision was consistent, and became more prominent when considering

particular scales or score profiles. In particular, profiles further away from the sample mean benefitted more from adaptive item selection. There were also signs that the benefit of adaptive testing varied across different value ranges of the same trait, suggesting the presence of complex interactions between adaptive item selection, item bank composition and candidate score distributions. Such interactions made the generalisation of results across different item banks particularly difficult, and further studies with different item banks would be desirable to understand FC CAT better.

Unfortunately, adaptive item selection did not produce any notable measurement advantages at shorter test lengths. The lack of improvements at the beginning of assessment despite having plenty of items to choose from was likely due to the unreliability of interim trait estimates. Indeed, despite its bias-reducing qualities, the FC pair format elicits less information per binary response compared to a SS item with a more detailed graded response (Brown & Maydeu-Olivares, 2017). There are multiple implications of this finding in practice. At the simplest level, there might be a test length below which adaptive item selection would not be worthwhile for FC assessments. Instead, it would be more economical to delay adaptive item selection till after a certain test length has been reached (e.g., by administering a fixed optimal test first), and/or make use of other data (e.g., prior information from alternative data sources, initial SS questions) to arrive at more reliable interim trait estimates prior to converting to FC CAT for reducing SEMs for the scales that are still lacking in measurement. Alternatively, the use of larger FC blocks (e.g., triplets, quads) would result in more information gain per question than pairs (Brown & Maydeu-Olivares, 2017) while also being less demanding on the richness of the item bank (i.e., larger blocks produce more pairwise comparisons per item used), thus allowing faster convergence to reliable interim trait estimates but at the expense of greater computational complexity in item selection and higher cognitive complexity for the candidates. At a more technical level,

once computational power ceased to be a limiting factor, it will be beneficial to explore item selectors that don't rely on point estimates (e.g., KLI, KLP, MUI and CEM, all requiring intensive numerical integrations in the multidimensional trait space). The power of item selectors that consider the entire posterior distribution has been demonstrated by past research (see Chapter 3 and Appendix D) and it is reasonable to hypothesise the findings would generalise to FC CAT.

The impact of item selection methodology was largely limited to measurement precision only. Compared to static assessments, adaptive item selection had inconsequential impact on response times, and made practically no impact on participant perceptions. While candidates may hold different views about adaptive and non-adaptive assessments, the actual assessment experience appeared to be largely indistinguishable in practice.

Social desirability balancing

There is a trade-off between the strictness of social desirability balancing and the effectiveness of adaptive item selection – a more stringent social desirability balancing criterion inevitably reduces the number of acceptable FC blocks, therefore reducing the potential of adaptive item selection. In this study, the more lenient social desirability balancing criterion indeed lead to better measurement precision. However, social desirability balancing is important for ensuring resistance against faking (Krug, 1958). Therefore, the setting of the social desirability balancing criterion is a balancing act – it should be as lenient as possible, but not so lenient that there are notable "right answers" in FC blocks. The optimal threshold could be identified through an empirical study that asks participants to purposefully choose the "right answer" in FC blocks with different levels of social desirability balancing. Note that, in a realistic assessment setting, a candidate will not necessarily choose the "right answer" even if they can spot it. It is

hypothesised that whether a candidate would choose the "right answer" over their real answer depends not only on the size of the difference in social desirability of items, but also on the stakes of the assessment. Therefore, low-stakes assessments could likely afford to use more lenient criteria, while high-stakes assessments should use more stringent thresholds. For a low-to-medium stakes assessment setting as in the current study (i.e., assessment results were inconsequential for the participants, but most of them were likely answering the questions as if they were applying for a job so as to practice for their actual pre-employment assessments), the lenient criteria used was adequate, and could possibly be relaxed even further without impairing fake resistance of the FC measures. For high-stakes assessments, social desirability balancing becomes more important, and the presence of a large and varied item bank becomes necessary for effective FC CAT. In other words, for high-stakes assessments with a limited item bank, the strict social desirability balancing requirement may negate any measurement improvement potential of adaptive item selection. In such a situation, the benefits of adaptive item selection are mainly around enhancing test security, by creating different question sequences for different candidates.

Social desirability balancing criteria had inconsequential impact on response times. Also, the use of different social desirability balancing criteria led to no notable differences in participant perceptions, suggesting that the assessment experience appeared to be largely indistinguishable in practice.

Question format

In line with previous research, data showed that a strong bias affected SS responses but not FC responses (e.g., Brown et al., 2017). The SS method in both Study 4 and Study 6 greatly inflated the observed score correlations between conceptually distinct latent traits. It was discovered that quick and context-poor adjectives were

especially prone to biases in a SS format. Compared to adjectives, the HEXACO-PI-R statements were affected to a much lesser extent, but still had higher scale intercorrelations and slightly higher sample mean scores than FC measures. With a higher stakes sample than Study 4 or Study 6, the inflation effect of the SS method would likely become more prominent (e.g., Lee et al., 2019), making FC a better assessment option. Therefore, for brief item content such as adjectives, the SS response format should be avoided, and a FC format would elicit more meaningful responses. For more complex item content such as HEXACO-PI-R statements, the SS format appeared to be adequate for the current samples but would likely be disadvantaged in high-stakes samples.

In terms of assessment experience, the SS question format appeared to be the accepted status quo amongst participants currently. However, there were also signs that participants could became more comfortable with the FC format if given more exposure, as indicated by faster response times as the FC assessment progressed. Indeed, at the end of the study, more than one eighth of participants indicated a preference for the FC questionnaire despite it containing twice as many questions as the SS instrument. Nevertheless, until the FC question format becomes commonly accepted, it is important to consider measures for improving candidate experience when using FC instruments. For example, assessment instructions could provide detailed explanations and examples of how to understand and answer FC questions, and how the collected responses would be interpreted. An enquiry from a participant highlighted a common worry and confusion with the FC format – that choosing A over B would be interpreted as saying "yes" to A and saying "no" to B. Therefore, it is important to explain the relative nature of FC responses to respondents, especially when most of them are used to providing absolute responses in a SS format.

Limitations

This empirical study explored a very specific instance of TIRT-based multidimensional FC assessment – it made use of a specific HEXACO item bank; it explored the effect of only one content rule (i.e., social desirability balancing criteria); it adopted the simplest pair format which is not the most information-efficient FC design; and it adopted an item selector that relies heavily on interim point estimates of trait values. Also, the instruments were completed under only one specific assessment setting (i.e., practice for pre-employment assessments). Given the numerous design possibilities and assessment situations, it would be unwise to conclude the merits of TIRT-based FC CATs based on the findings of this one study. Nevertheless, this study provided an initial exploratory baseline for furthering research on FC CATs using the TIRT model.

Conclusions

A simple but operational adaptive FC personality assessment was developed and deployed. A well-fitting item bank of 279 adjectives was collated (Study 4) that measured the HEXACO personality model with good convergent and divergent validity, although subsequent studies showed that a larger and more varied item bank would have been more desirable for use in a CAT. A simulation study using this new item bank (Study 5) largely replicated previous findings using simulated item banks (Study 3), favouring A-optimality as the best item selector. Moreover, with a realistic item bank, the distribution of item parameters varied between scales, which resulted in variable levels of measurement accuracy across scales. In the case of this study, it appeared that having fewer but more discriminating items was more beneficial than having more items with lower discriminations. A follow-up simulation study (Study 5b) further refined the CAT algorithm design for the FC HEXACO measure, and explored the differences between adaptive and non-adaptive (but with measurement optimised for the

average person) versions of the FC HEXACO measure. Simulation showed that the advantage of adaptive item selection over non-adaptive item presentation appeared to be small on average, but CAT was helpful in boosting measurement for the weaker scales. Finally, the adaptive and non-adaptive FC HEXACO personality measures were trialled empirically (Study 6). As predicted by the simulation results (Study 5b), adaptive item selection resulted in some small gains on measurement precision on average. Moreover, certain score profiles, for example those further away from the population mean, benefitted more from adaptive item selection. However, there was no notable advantage of adaptive item selection at shorter test lengths despite having plenty of items to choose from, signalling the weakness of item selectors that rely on interim point estimates of the trait values, which could be fairly inaccurate at the beginning of the assessment. Instead, it would be more economical to deploy adaptive item selection at later parts of an assessment, and/or use larger FC blocks that give more information per question, and/or employ global information item selectors (computational power permitting) that don't rely on point estimates. Aside from its impact on measurement precision, adaptive item selection didn't appear to have any effect on response times or participant perceptions.

In line with previous research findings, empirical data (Studies 4 and 6) showed the existence of a method bias when the SS response format was used, leading to more desirable observed scores as well as inflated score correlations. Adjectives were notably more prone to this bias compared to HEXACO-PI-R statements, but remained unaffected when instead administered in a FC format with adequate social desirability balancing. For many respondents, however, the FC format wasn't preferable compared to the SS format as the familiar and accepted status quo. This was despite most respondents acknowledging that the FC format elicited deeper thinking about their personality as well as offered less opportunities for faking good. Therefore, when using

the FC response format, researchers should take care to provide participants with a clear explanation of how FC responses would be interpreted, especially when participants might be required to provide response to FC blocks consisting of only negative-sounding items.

CHAPTER 5: GENERAL DISCUSSIONS

This thesis mapped out a rough blueprint for the development of dominance FC CATs using the TIRT model. As this thesis followed a sequential structure where each chapter's conclusions informed key decisions in the next one, full results, discussions, limitations and recommendations are provided at the end of each chapter. Here, a brief summary of the key findings from each area of investigation is provided. Then, main limitations and suggestions for further research are outlined. Finally, implications for research and practice are discussed.

Thesis Summary

The development of a good FC CAT is a journey that requires considerations from many angles. This thesis investigated the key methodologies for TIRT-based FC CAT, covering research questions in essential assumption testing, CAT algorithm optimisation, and operational deployment.

From a feasibility perspective, in order to adaptively assemble items into FC blocks, the invariance of item parameters is essential. Study 1 provided empirical support for this requirement based on large operational samples, showing that person score estimation remained very stable despite minor violations to the item parameter invariance assumption. Study 1 also suggested practical remedies for minimising the effect of context when creating FC blocks, including ensuring the items had similar social desirability, and avoiding combining items that might interact semantically.

From an optimality perspective, the automated test assembly algorithm plays an important role in upholding both content and measurement requirements in a CAT. Chapter 3 systematically reviewed CAT algorithm components for TIRT-based FC CAT. Moreover, a series of intensive simulation studies were conducted to compare the

performance of trait estimators (Study 2) and item selectors (Studies 3 and 5), leading to the recommendations of the MAP trait estimator and the A-optimality item selector as best choices for TIRT-based FC CAT in general, although D- and C-optimality could potentially be more optimal for specific assessment content and setup.

From a practical perspective, Simulation (Study 5b) and empirical trialling (Study 6) demonstrated the power of CAT in improving measurement precision, and also showed that the magnitudes of such improvements were heavily dependent on both the item bank characteristics and the respondent profiles. In terms of respondent feedback, no systematic differences were found between respondents taking adaptive and non-adaptive FC assessments. However, respondents expressed a predominant preference for the assessment experience of SS questionnaires over FC assessments. Therefore, researchers and practitioners should take extra care to inform and reassure participants when deploying FC assessments.

Incidentally, while the focus of this thesis was on testing and refining the psychometric methodologies underlying TIRT-based FC CAT, in the fulfilment of this purpose an operational FC CAT for the HEXACO personality model was created (Studies 4, 5b and 6). This assessment may be used in future research studies concerning the HEXACO personality model.

Limitations and Further Research

Constrained by the scope of this thesis, the investigations have several limitations and a number of areas inviting further exploration. While study-specific limitations have been discussed in previous chapters, the overarching gaps and further research questions in psychometric methodology and empirical practice are outlined here.

Psychometric Methodology

Limited by computational power, this thesis only explored a selection of local information item selectors. Once computational power ceased to be an inhibition, or/and simplifying approximations became available for the calculations involved, the global information item selectors should be re-visited. It would also be beneficial to explore some item selector modifications. For example, measurement precision might be improved further by applying item bank stratification (Chang & Ying, 1999), preserving highly discriminating items till later in CAT sessions. Moreover, this thesis largely focused on comparing the measurement efficiency and precision of item selectors, but other aspects of item selector performance could also be important in practice. For example, item bank utilisation and item exposure control might be relevant for high-stakes FC assessments (e.g, Chen et al., 2019). The way item selectors could interact with changing item characteristics could also be informative when working with various operational item banks.

The implementation of content rules in this thesis was additive, static and absolute, i.e., the rules stacked on top of each other, remained unchanged throughout a CAT session, and stayed firmly in place even if they became too restrictive for item selection (as seen in Study 5 with a fixed scale plan). For complex assessment designs with many content rules, such an implementation could quickly become prohibitive especially with smaller item banks. A more fluid implementation considering the interplay of different content rules as well as information gain requirements would be more effective in practice. In fact, a number of content rule management heuristics have been developed, for example, the weighted deviation method (Stocking & Swanson, 1993), the shadow test approach (van der Linden, 2005), and the maximum priority index method (Cheng & Chang, 2009). Incorporating such content rule management

heuristics into FC CAT algorithms could be very beneficial for complex assessment designs or/and small item banks.

Constrained by time, this thesis only studied FC CAT using pairs. However, some operational FC assessments adopt larger FC blocks, e.g., the OPQ32 have quad and triplet versions (Bartram et al., 2006; Brown & Bartram, 2009-2011), while the Employee Selection Questionnaire-2 (Jackson, 2001) and the Gordon Personal Profile Inventory (Gordon, 1993) both use most/least quads. As larger FC blocks tend to be more efficient in gaining information, it would be beneficial to expand FC CAT methodology to larger block sizes. Joo, Lee, & Stark (2018, 2019) explored FC CAT with triplets and quads using ideal-point items modelled by the generalized graded unfolding model (Roberts, Donoghue, & Laughlin, 2000), which could inform research on the same front but instead using dominance items and associated IRT models.

Empirical Practice

While Study 1 provided reassurance on the stability of item parameters thus enabling FC CAT with item shuffling, the effect of context on item functioning in FC blocks should be investigated further. Empirical studies may examine different psychological constructs (e.g., personality vs. interest), item formats (e.g., adjectives vs. statements), or assessment settings (e.g., low vs. high stakes), seeking to verify the invariance of item parameters or to identify the conditions where this assumption would be violated. Apart from Lin et al. (2013), Study 1 (Lin & Brown, 2017), and subsequently Morillo et al. (2019), FC CAT researchers have largely taken the parameter invariance assumption for granted with no empirical justification.

In lieu of item social desirability estimates, this thesis adopted the item mean utility parameter as a proxy. These two item attributes are highly correlated – both concern the ease with which respondents endorse an item. However, they're also subtly

different – item utility is viewed from one's own perspective (i.e., "Am I X") while social desirability considers the perception in others' eyes (i.e., "Would others find X desirable"). To illustrate this divergence, consider the item "I never lie" (high social desirability), which is hard for the average person to achieve (low item mean utility). Therefore, where possible, empirical ratings of item social desirability should be collected and deployed. In order to maximise resistance against faking, rating instructions should be drafted to reflect the context in which the FC assessments would be taken (Converse et al., 2010).

Last but not least, this thesis offered only one empirical instance of FC CAT (Study 6). In order to further the understanding of FC CAT with dominance items, it would be necessary to conduct more empirical studies with varying scale constructs, item banks, assessment designs, respondent population, etc. At the time of writing, I was unable to find more reported empirical studies of FC CAT using dominance items and non-ipsative IRT scoring models.

Implications for Research and Practice

This thesis extended the literature on FC assessments using the TIRT model. In particular, it addressed some knowledge gaps regarding FC CAT using dominance items in much greater depth than previous studies on the same topic (Brown, 2012; Lin & Brown, 2015). Findings of this thesis inform research and practice around FC assessments scored using the TIRT model (e.g., the OPQ32 in Study 1; the Motivational Value Systems Questionnaire by Merk, Schlotz, & Falter, 2017), providing considerations and recommendations for the psychometric design of such assessments. Findings of this thesis also inform FC assessment development even if the TIRT model isn't adopted (e.g., see meta-analysis of FC measures by Salgado, 2014, 2015, 2017), providing empirical insight into respondent behaviours and reactions with respect to the

FC question format in general. Finally, as many personality items were developed under the dominance rather than ideal-point paradigm (e.g., the International Personality Item Pool, Goldberg et al., 2006), improving the understanding of FC CAT methodologies for dominance items opens up more opportunities for leveraging such legacy items for future FC CAT applications.

Ultimately, this thesis aims to increase the fairness and accuracy of personality assessments through CAT, which is achieved from three angles: 1) adopting the FC response format in order to reduce response biases and distortions; 2) selecting items adaptively in order to increase the accuracy of person score estimation; and 3) understanding participants' views on the FC question format in order to enhance assessment experience and engagement. As personality assessments are frequently used to drive educational, occupational, and even clinical decisions, methodologies that improve the fairness and accuracy of personality assessments even just slightly can still have large cumulative benefits when applied to a large number of assessment takers, leading to improved decision making and human cost savings.

REFERENCES

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (10th ed.). New York: Dover. Retrieved from http://people.math.sfu.ca/~cbm/aands/abramowitz_and_stegun.pdf

Ackerman, P. L., & Kanfer, R. (2006). *Test length and cognitive fatigue: Final report to the college board.* Atlanta, GA: Author.

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*(2), 163-181. doi:10.1037/a0015719

Allen-Zhu, Z., Li, Y., Singh, A., & Wang, Y. (2017). Near-optimal design of experiments via regret minimization. Paper presented at the *34th International Conference on Machine Learning*, Sydney, Australia. Retrieved from https://www.cs.cmu.edu/~aarti/pubs/ICML17_Zhu.pdf

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Chapter 1 - personality psychology and economics. In E. A. Hanushek, S. Machin & L. Woessmann (Eds.), *Handbook of the economics of education* (pp. 1-181). Amsterdam: Elsevier. doi:https://doi.org/10.1016/B978-0-444-53444-6.00001-8

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement, 19*(3), 269-290. doi:10.1177/014662169501900306

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*, 150-166.

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*, 340-345.

Ashton, M. C., Lee, K., de Vries, R. E., Perugini, M., Gnisci, A., & Sergi, I. (2006). The HEXACO model of personality structure and indigenous lexical personality dimensions in Italian, Dutch, and English. *Journal of Research in Personality, 40*(6), 851-875. doi:https://doi.org/10.1016/j.jrp.2005.06.003

Ashton, M. C., Lee, K., & Goldberg, L. R. (2004). A hierarchical analysis of 1,710 english personality-descriptive adjectives. *Journal of Personality and Social Psychology, 87*(5), 707-721. doi:10.1037/0022-3514.87.5.707

Ashton, M. C., Lee, K., Marcus, B., & de Vries, R. E. (2007). German lexical personality factors: Relations with the HEXACO model. *European Journal of Personality, 21*, 23-43. doi:10.1002/per.597

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., . . . De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*(2), 356-366. doi:10.1037/0022-3514.86.2.356

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. doi:10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2010). Multiple imputation with mplus (version 2). Retrieved from https://www.statmodel.com/download/Imputations7.pdf

Atkinson, A. C., Donev, A. N., & Tobias, R. D. (2007). *Optimum experimental designs, with SAS*. Oxford, UK: Oxford University Press. Retrieved from https://www.dawsonera.com/abstract/9780191537943

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26.

Bartram, D., Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *OPQ32 technical manual*. Thames Ditton, UK: SHL.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin, 2*(3), 47-53. doi:10.2307/3002000

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. doi:10.1111/j.1468-2389.2006.00354.x

Birnbaum, A. (1958). *On the estimation of mental ability (series report no. 15, project no. 7755-23).* Randolph Air Force Base TX: USAF School of Aviation Medicine.

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York, NY: McGraw-Hill.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444. doi:10.1177/014662168200600405

Bock, R. D., & Schilling, S. G. (2003). IRT based item factor analysis. In M. Du Toit (Ed.), *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* (pp. 584-591). Lincolnwood, IL: Scientific Software International.

Boies, K., Lee, K., Ashton, M. C., Pascal, S., & Nicol, A. A. M. (2001). The structure of the French personality lexicon. *European Journal of Personality, 15*(4), 277-295. doi:10.1002/per.411

Boyce, A. S., Conway, J. S., & Caputo, P. M. (2014). *ADEPT-15 technical documentation: Development and validation of Aon Hewitt's personality model and adaptive employee test (ADEPT-15).* New York: Aon Hewitt.

Brown, A. (2012). Multidimensional CAT in non-cognitive assessments. Paper presented at the *Paper Presented at the 8th Conference of the International Test Commission,* Amsterdam, The Netherlands.

Brown, A. (2015). Personality assessment, forced-choice. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 840-848). Oxford: Elsevier. doi:10.1016/B978-0-08-097086-8.25084-8

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika, 81*(1), 135-160. doi:10.1007/s11336-014-9434-9

Brown, A., & Bartram, D. (2009). Doing less but getting more: Improving forced-choice measures with IRT. Paper presented at the *Paper Presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology,* New Orleans, LA.

Brown, A., & Bartram, D. (2009-2011). *OPQ32r technical manual*. Surrey, UK: SHL Group.

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods, 20*(1), 121-148. doi:10.1177/1094428116668036

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational & Psychological Measurement, 71*(3), 460-502. doi:10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a thurstonian IRT model to forced-choice data using mplus. *Behavior Research Methods, 44*(4), 1135-1147. doi:10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36-52. doi:10.1037/a0030641

Brown, A., & Maydeu-Olivares, A. (2017). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal,* doi:10.1080/10705511.2017.1392247

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*(1), 111-150. doi:10.1207/S15327906MBR3601_05

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference* (2nd ed.). New York: Springer.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology,* doi:10.1037/apl0000414

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology, 56*, 453-484. doi:10.1146/annurev.psych.55.090902.141913

Chang, H., Qian, J., & Ying, Z. (2001). a-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25*(4), 333-341. doi:10.1177/01466210122032181

Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213-229. doi:10.1177/014662169602000303

Chang, H., & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222. doi:10.1177/01466219922031338

Chen, C., Wang, W., Chiu, M. M., & Ro, S. (2019). Item selection and exposure control methods for computerized adaptive testing with multidimensional ranking items. *Journal of Educational Measurement,* doi:10.1111/jedm.12252

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical & Statistical Psychology, 62*(2), 369-383.

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22*(2), 105-127. doi:10.1080/08959280902743303

Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(1), 55-77. doi:10.1207/S15328007SEM0901_4

Choi, S. W., & King, D. R. (2014). MAT: Multidimensional adaptive testing. R package version 2.2. Retrieved from https://CRAN.R-project.org/package=MAT

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267-307. doi:10.1207/s15327043hup1803_4

Christiansen, N. D., Goffin, R. D., Johnson, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47*(4), 847-860.

Clemans, W. V. (1966). *An analytical and empirical examination of some properties of ipsative measures.* (Psychometric Monograph No. 14). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN14.pdf

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement desirability ratings in forced-choice personality measure development: Implications for reducing score inflation and providing trait-level information. *Human Performance, 23*, 323-342. doi:10.1080/08959285.2010.501047

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Wilson (1991). *Journal of Occupational & Organizational Psychology, 67*(2), 89-100. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=9501231827&site=ehost-live

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO–PI–R) and NEO five-factor inventory (NEO–FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). New York: Wiley.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological), 30*(2), 248-275. Retrieved from http://www.jstor.org/stable/2984505

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/BF02310555

Davey, T., & Nering, M. L. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement, 20*(4), 405-416. doi:10.1177/014662169602000407

De Raad, B. (1992). The replicability of the Big Five personality dimensions in three word-classes of the Dutch language. *European Journal of Personality, 6*(1), 15-29.

De Raad, B., & Szirmák, Z. (1994). The search for the "Big Five" in a non-indo-European language: The Hungarian trait structure and its relationship to the EPQ and the PTS. *European Review of Applied Psychology / Revue Européenne De Psychologie Appliquée, 44*(1), 17-24.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*(1), 81-106.

Donovan, J. J., Dwight, S. A., & Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *Journal of Business & Psychology, 29*(3), 479-493. doi:10.1007/s10869-013-9318-5

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army selection and classification decisions (tech. rep. no. 1311).* Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Edwards, A. L. (1973). *Edwards personal preference schedule manual*. New York: Psychological Corporation.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*(2), 155-166.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review.* (No. RR791). Birmingham, UK: University of Birmingham.

Finkelman, M., Hooker, G., & Wang, J. (2009). *Technical report.* (No. BU-1768-M). Ithaca, NY: Department of Biological Statistics and Computational Biology, Cornell University.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507-521. doi:10.2307/2331838

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222*, 309-368. Retrieved from http://www.jstor.org/stable/91208

Friedman, H., & Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management, 9*, 114-123.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Goldberg, L. R. (1982). From ace to zombie: Some explorations in the language of personality. In C. D. Spielberger, & J. N. Butcher (Eds.), *Advances in personality assessment: Vol. 1* (pp. 203-234). Hillsdale, NJ: Erlbaum.

Goldberg, L. R. (1990). An alternative 'description of personality': The big-five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216-1229. doi:10.1037/0022-3514.59.6.1216

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84-96. doi:https://doi.org/10.1016/j.jrp.2005.08.007

Gordon, L. V. (1993). *Gordon personal Profile inventory: Manual 1993 revision.* San Antonio, TX: Pearson-TalentLens.

Gough, H. G., & Heilbrun, A. B. (1980). *The adjective check list manual*. Palo Alto, CA: Consulting Psychologists Press.

Griffin, B., & Wilson, I. G. (2012). Faking good: Self-enhancement in medical school applicants. *Medical Education, 46*(5), 485-490.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? an examination of the frequency of applicant faking behavior. *Personnel Review, 36*(3), 341-355.

Hahn, D., Lee, K., & Ashton, M. C. (1999). A factor analysis of the most frequently used Korean personality trait adjectives. *European Journal of Personality, 13*(4), 261-282. doi:10.1002/(SICI)1099-0984(199907/08)13:4<261::AID-PER340>3.0.CO;2-B

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Essex, UK: Pearson Education Limited. Retrieved

from https://is.muni.cz/el/1423/podzim2017/PSY028/um/_Hair_-_Multivariate_data_analysis_7th_revised.pdf

Hathaway, S. R., & Mckinley, J. C. (1940). A multiphasic personality schedule (Minnesota) : I. construction of the schedule. *Journal of Psychology, 10*(2), 249-254.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*(1), 9-24. doi:10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167-184. doi:10.1037/h0029780

Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "Fake-proof" measure of the Big Five. *Journal of Research in Personality, 42*(5), 1323-1333. doi:10.1016/j.jrp.2008.04.006

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2008). Computerized adaptive testing of personality traits. *Zeitschrift Für Psychologie/Journal of Psychology, 216*(1), 12-21. doi:10.1027/0044-3409.216.1.12

Hooker, G. (2010). On separable tests, correlated priors, and paradoxical results in multidimensional item response theory. *Psychometrika, 75*(4), 694-707. doi:10.1007/s11336-010-9181-5

Hooker, G., Finkelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika, 74*(3), 419-442. doi:10.1007/s11336-009-9111-6

Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden, R. M. (2006). *Development of the navy computer adaptive personality scales (NCAPS).* (No. NPRST-TR-06–2).

Millington, TN: Navy Personnel Research, Studies, and Technology Division, Bureau of Naval Personnel (NPRST/PERS-1).

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. doi:10.1080/10705519909540118

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*(6), 869-879.

Jackson, D. N. (2001). *Employee screening questionnaire - 2: Technical manual*. London, ON: SIGMA Assessment Systems. Retrieved from https://www.sigmaassessmentsystems.com/wp-content/uploads/2016/03/ESQ2-Technical-Manual.pdf

Jackson, D. N. (2002). *Employee screening questionnaire manual*. Port Huron, MI: Sigma Assessment Systems, ESQ.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*(4), 371-388.

Jaeger, B. (2017). r2glmm: Computes R squared for mixed (multilevel) models. R package version 0.1.2. Retrieved from https://CRAN.R-project.org/package=r2glmm

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*(2), 153-162.

Joo, S., Lee, P., & Stark, S. (2018). Development of information functions and indices for the GGUM-RANK multidimensional forced choice IRT model. *Journal of Educational Measurement, 55*(3), 357-372. doi:10.1111/jedm.12183

Joo, S., Lee, P., & Stark, S. (2019). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods,* doi:10.3758/s13428-019-01274-6

Jordan, P., & Spiess, M. (2012). Generalizations of paradoxical results in multidimensional item response theory. *Psychometrika, 77*(1), 127-152. doi:10.1007/s11336-011-9243-3

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: Allen Lane.

Kam, C. (2013). Probing item social desirability by correlating personality items with balanced inventory of desirable responding (BIDR): A validity examination. *Personality and Individual Differences, 54*(4), 513-518.

Kantrowitz, T. M., Grelle, D. M., & Lin, Y. (2019). Applying adaptive approaches to talent management practices. In R. N. Landers (Ed.), *The Cambridge handbook of technology and employee behavior* (pp. 131-150). Cambridge: Cambridge University Press. doi:10.1017/9781108649636.007

Kantrowitz, T. M., Tuzinski, K. A., & Raines, J. M. (2018). *Global assessment trends report*. Thames Ditton, UK: SHL.

Kenny, D. A. (2015). Measuring model fit. Retrieved from http://davidakenny.net/cm/fit.htm

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*(2), 312-320. doi:10.1037/0022-3514.55.2.312

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer.

Krug, R. E. (1958). A selection set preference index. *Journal of Applied Psychology, 42*(3), 168-170.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26. doi:10.18637/jss.v082.i13

Kvalnes, T. (2013). Lmf: Functions for estimation and inference of selection in age-structured populations. R package version 1.2.[computer software]

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96*(1), 202-210. doi:10.1037/a0020375

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 39*(2), 329-358.

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO personality inventory: Two new facet scales and an observer report form. *Psychological Assessment, 18*(2), 182-191.

Lee, K., & Ashton, M. C. (2008). The HEXACO personality factors in the indigenous personality lexicons of English and 11 other languages. *Journal of Personality, 76*(5), 1001-1054. doi:10.1111/j.1467-6494.2008.00512.x

Lee, K., & Ashton, M. C. (2009a). The HEXACO personality inventory - revised: Scale descriptions. Retrieved from http://hexaco.org/scaledescriptions

Lee, K., & Ashton, M. C. (2009b). Reanalysis of the structure of the Greek personality lexicon. *Journal of Cross-Cultural Psychology, 40*(4), 693-700. doi:10.1177/0022022109335183

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment, 25*(5), 543-556.

Lee, P., Joo, S., & Lee, S. (2019). Examining stability of personality profile solutions between likert-type and multidimensional forced choice measure. *Personality and Individual Differences, 142*, 13-20. doi:10.1016/j.paid.2019.01.022

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York, NY: Springer.

Lin, Y., & Brown, A. (2015). Response formats and trait estimation efficiency in computerized adaptive testing. Paper presented at the *5th Conference of the International Association for Computerized Adaptive Testing*, Cambridge, UK.

Lin, Y., Inceoglu, I., & Bartram, D. (2013). Towards creating forced-choice personality assessments 'on the fly': Do Thurstonian IRT assumptions hold empirically? Paper presented at the *78th Annual Meeting of the Psychometric Society*, Arnhem, The Netherlands.

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement, 77*(3), 389-414. doi:10.1177/0013164416646162

Lönnqvist, J. (2014). Increased socially desirable responding in college applicants reapplying and retesting after initial failure. *Personality and Individual Differences, 60*, S5. doi:https://doi.org/10.1016/j.paid.2013.07.157

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*(2), 233-245. doi:10.1007/BF02294018

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149. doi:10.1037/1082-989X.1.2.130

Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2013). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the NEO PI-R. *Assessment, 20*(1), 3-13.

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*(2), 247-256. doi:10.1016/S0191-8869(01)00021-6

Matthews, G. (2011). Personality and individual differences in cognitive fatigue. In P. L. Ackerman (Ed.), *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications* (pp. 209-227). Washington, DC, US: American Psychological Association. doi:10.1037/12343-000

Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods, 10*(3), 285-304. doi:10.1037/1082-989X.10.3.285; 10.1037/1082-989X.10.3.285.supp (Supplemental)

McAbee, S. T., Casillas, A., Way, J. D., & Guo, F. (2019). The HEXACO model in education and work: Current applications and future directions. *Zeitschrift Für Psychologie, 227*(3), 174-185. doi:10.1027/2151-2604/a000376

McCrae, R. R., Costa, P. T., & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment, 84*(3), 261-270. doi:10.1207/s15327752jpa8403_05

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175-215.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space.* (No. ONR83-2). Iowa City, IA: The

American College Testing Program. Retrieved from https://files.eric.ed.gov/fulltext/ED241581.pdf

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational & Organizational Psychology, 77*(4), 531-551.

Mendolia, S., & Walker, I. (2014). The effect of personality traits on subject choice and performance in high school: Evidence from an English cohort. *Economics of Education Review, 43*, 47-65. doi:10.1016/j.econedurev.2014.09.004

Merk, J., Schlotz, W., & Falter, T. (2017). The Motivational Value Systems Questionnaire (MVSQ): Psychometric analysis using a forced choice Thurstonian IRT model. *Frontiers in Psychology, 8*, 1626. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2017.01626

Miller, T. Q., Smith, T. W., Turner, C. W., Guijarro, M. L., & Hallet, A. J. (1996). A meta-analytic review of research on hostility and physical health. *Psychological Bulletin, 119*(2), 322-348.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., & Gur, R. E. (2018). Development and public release of a computerized adaptive (CAT) version of the schizotypal personality questionnaire. *Psychiatry Research, 263*, 250-256. doi:10.1016/j.psychres.2018.02.022

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Journal of Work and Organizational Psychology, 35*(2), 75-83.

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*(2), 273-296. doi:10.1007/s11336-008-9097-5

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 77-101). New York: Springer. doi:10.1007/978-0-387-85461-8

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176. doi:10.1177/014662169201600206

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's guide (seventh edition).* Los Angeles, CA: Muthén & Muthén.

Myers, I. B., McCaulley, M. H., Quenk, N., & Hammer, A. (1998). *MBTI handbook: A guide to the development and use of the myers-briggs type indicator* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133-142. doi:10.1111/j.2041-210x.2012.00261.x

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263-280.

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema, 29*(3), 390-395.

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

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43*(5), 971-990. doi:10.1016/j.paid.2007.03.017

O'Connell, M. S., Kung, M., & Tristan, E. (2011). Beyond impression management: Evaluating three measures of response distortion and their relationship to job performance. *International Journal of Selection and Assessment, 19*(4), 340-351. doi:10.1111/j.1468-2389.2011.00563.x

O'Neill, T. A., Lewis, R. J., Law, S. J., Larson, N., Hancock, S., Radan, J., . . . Carswell, J. J. (2016). Forced-choice pre-employment personality assessment: Construct validity and resistance to faking. *Personality and Individual Differences,* doi:10.1016/j.paid.2016.03.075

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*(4), 995-1027. doi:10.1111/j.1744-6570.2007.00099.x

Ortner, T. M. (2008). Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection & Assessment, 16*(3), 249-257.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57,* 401-421. doi:10.1146/annurev.psych.57.102904.190127

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224-239). New York: Guilford.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin, 21*(2), 100-108.

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of applicant faking on forced-choice and likert scores. *Organizational Research Methods, 22*(3), 710-739.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

Peterson, M. H., Griffith, R. L., O'Connell, M. S., & Isaacson, J. A. (2008). Examining faking in real job applicants: A within-subjects investigation of score changes across applicant and research settings. Paper presented at the *23rd Annual Conference for the Society for Industrial and Organizational Psychology*, San Francisco, CA.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338.

Pukelsheim, F. (2006). *Optimal design of experiments.* Society for Industrial and Applied Mathematics. doi:0.1137/1.9780898719109

R Core Team. (2015). R: A language and environment for statistical computing [computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, USA: Springer.

Reckase, M. D., & Luo, X. (2014). A paradox by another name is good estimation. In A. van der Ark, D. Bolt, S. M. Chow, J. Douglas & W. C. Wang (Eds.), *Proceedings of IMPS 2014*. New York, NY: Springer.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*(4), 347-364.

Revelle, W. R. (2018). *Psych: Procedures for personality and psychological research. Version 1.8.4.* Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science (Wiley-Blackwell), 2*(4), 313-345. doi:10.1111/j.1745-6916.2007.00047.x

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3-32.

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), 634-644.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. doi:10.2307/2335739

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82*(1), 30-43.

Salgado, J. F. (2002). The Big Five personality dimensions and counterproductive behaviors. *International Journal of Selection & Assessment, 10*, 117-125.

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*(3), 323-346. doi:10.1348/096317903769647201

Salgado, J. F. (2017). Moderator effects of job complexity on the validity of forced-choice personality inventories for predicting job performance. *Journal of Work and Organizational Psychology, 33*(3), 229-238.

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational & Organizational Psychology, 88*(4), 797-834. doi:10.1111/joop.12098

Salgado, J. F., & Táuriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3-30. doi:10.1080/1359432X.2012.716198

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional space. *Psychometrika, 39*, 111-121.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*(6), 966-974.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331-354. doi:10.1007/BF02294343

Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement, 75*(6), 954-978. doi:10.1177/0013164415575147

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, (27), 379-423.

SHL. (2009-2014). *Global Personality Inventory - Adaptive: Technical manual*. Surrey, UK: SHL.

SHL. (2014). *OPQ32r™ Technical manual*. Surrey, UK: SHL.

Silvey, S. D. (1980). *Optimal design: An introduction to the theory for parameter estimation*. Netherlands: Springer. doi:10.1007/978-94-009-5912-5

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT-PD project. *Journal of Personality Assessment, 93*(4), 380-389.

Soetaert, K. (2009). rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations. R-package version 1.6.[computer software]

Soetaert, K., & Herman, P. M. J. (2009). *A practical guide to ecological modelling: Using R as a simulation platform*. Netherlands: Springer.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment.* (Unpublished Doctoral Dissertation). University of Illinois at Urbana-Champaign,

Stark, S., & Chernyshenko, O. S. (2007). Adaptive testing with the multi-unidimensional pairwise preference model. Paper presented at the *2007 GMAC Conference on Computerized Adaptive Testing*, Minneapolis, MN.

Stark, S., & Chernyshenko, O. S. (2011). Computerized adaptive testing with the Zinnes and Griggs pairwise preference ideal point model. *International Journal of Testing, 11*(3), 231-247. doi:10.1080/15305058.2011.561459

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*, 184-203.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of

personality and other noncognitive assessments. *Organizational Research Methods*, *15*(3), 463-487. doi:10.1177/1094428112444611

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology*, *81*(2), 332-342. doi:10.1037/0022-3514.81.2.332

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*(3), 277-292. doi:10.1177/014662169301700308

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*(5), 429-442.

Stroud, A. H., & Sechrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs: Prentice-Hall.

Szarota, P., Ashton, M. C., & Lee, K. (2007). Taxonomy and structure of the Polish personality lexicon. *European Journal of Personality, 21*(6), 823-852. doi:10.1002/per.635

Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait.* (Unpublished Ph.D. thesis). Columbia University,

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*(4), 703-742.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 79,* 281-299.

Topping, G. D., & O'Gorman, J. G. (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences, 23*(1), 117-124. doi:10.1016/S0191-8869(97)00006-8

Trapmann, S., Hell, B., Hirn, J. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Zeitschrift Für Psychologie/Journal of Psychology, 215*(2), 132-151.

Trull, T. J., & Sher, K. J. (1994). Relationship between the five-factor model of personality and axis I disorders in a nonclinical sample. *Journal of Abnormal Psychology, 103*(2), 350-360.

Tseng, F. L., & Hsu, T. C. (2001). Multidimensional adaptive testing using the weighted likelihood estimation: A comparison of estimation methods. Paper presented at the *2001 Annual Meeting of National Council on Measurement in Education (NCME),* Seattle.

Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings. USAF technical report.* (No. 61-97). Lackland Air Force Base, TX: U.S. Air Force.

Tupes, E. C., & Christal, R. E. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality, 60*(2), 225-251.

Usami, S., Sakamoto, A., Naito, J., & Abe, Y. (2016). Developing pairwise preference-based personality test and experimental investigation of its resistance to faking effect by item response model. *International Journal of Testing, 16*(4), 288-309. doi:10.1080/15305058.2016.1145123

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42*(3), 283-302. doi:10.1111/j.1745-3984.2005.00015.x

van der Linden, W. J. (2010). Item response theory. In P. P. B. McGaw (Ed.), *International encyclopedia of education (third edition)* (pp. 81-88). Oxford: Elsevier. doi:10.1016/B978-0-08-044894-7.00250-5

van der Linden, W. J. (2012). On compensation in multidimensional response modeling. *Psychometrika, 77*(1), 21-30. doi:10.1007/s11336-011-9237-1

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York, NY: Springer.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). New York: Springer. doi:10.1007/978-0-387-85461-8

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*(3), 346-360. doi:10.1177/0022022104264126

van Rijn, P., & Rijmen, F. (2012). *A note on explaining away and paradoxical results in multidimensional item response theory.* (No. RR-12-13). Princeton, NJ: ETS. Retrieved from https://www.ets.org/Media/Research/pdf/RR-12-13.pdf

van Rijn, P., & Rijmen, F. (2015). On the explaining-away phenomenon in multivariate latent variable models. *British Journal of Mathematical & Statistical Psychology, 68*(1), 1-22. doi:10.1111/bmsp.12046

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*(2), 203-226. doi:10.2307/1165378

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*(4), 575-588. doi:10.1007/BF02295132

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197-210. doi:10.1177/00131649921969802

Waller, N. G. (1999). Searching for structure in the MMPI. In S. E. Embretson, & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 185-217). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology, 57*(6), 1051-1058.

Wang, C., & Chang, H. (2010). Item selection in MCAT - the new application of Kullback–Leibler information. Paper presented at the *75th International Meeting of the Psychometric Society*, Athens, Georgia.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika, 80*(2), 428-449. doi:10.1007/s11336-013-9399-0

Wang, C., & Chang, H. (2011). Item selection in multidimensional computerized adaptive testing: Gaining information from different angles. *Psychometrika, 76*(3), 363-384. doi:10.1007/s11336-011-9215-7

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317-331. doi:10.1177/01466210122032163

Wang, W., & Chen, P. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*(5), 295-316.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. doi:10.1007/BF02294627

Wasti, S. A., Lee, K., Ashton, M. C., & Somer, O. (2008). Six Turkish personality factors and the HEXACO model of personality structure. *Journal of Cross-Cultural Psychology, 39*(6), 665-684. doi:10.1177/0022022108323783

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67*(1), 41-58. doi:10.1177/0013164406288164

Wellman, M. P., & Henrion, M. (1993). Explaining 'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*, 287-292. doi:10.1109/34.204911

Yusoff, M. S. B. (2013). Faking good in personality and emotional intelligent tests: Self-enhancement among a cohort of medical school applicants. *Education in Medicine Journal, 5*(2), e60-e71.

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin, 63*(2), 117-124. doi:10.1037/h0021567

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika, 39*(3), 327-350. doi:10.1007/BF02291707

APPENDIX A: LIST OF MATHEMATICAL NOTATIONS

Notation	Definition
i, k, l, o	Individual items
$i_1, i_2, \ldots$	A string of questions in the order they appeared in a test
$\{i, k\}, \{i, k, l\}, \{i, k, l, o\}$	FC blocks with two, three, or four items
$\{i_1, k_1\}, \{i_2, k_2\}, \ldots$	A string of pairwise comparisons in the order they appeared in a test
n	The number of items within a FC block
s, v	Individual traits/dimensions/scales
$\{s, v\}$	A pair of traits/dimensions/scales to be measured by a FC block
r	The number of item responses from a respondent
$r-1$	The number of item responses already collected from a respondent in a CAT session
R_r	The set of unused items after administering $r-1$ questions/ when selecting the r^{th} question
$\lvert R_r \rvert$	The number of unused items after administering $r-1$ questions/ when selecting the r^{th} question
S	Number of traits/dimensions measured
s_i	The trait indicated by a unidimensional item i
$\boldsymbol{cor}$	$S \times S$ correlation matrix of latent traits
$\boldsymbol{cov}$	$S \times S$ variance-covariance matrix of latent traits
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_S)^T$	Column vector of a respondent's latent trait values – unobserved true scores
$\widehat{\boldsymbol{\theta}}^{r-1} = \left(\hat{\theta}_1^{r-1}, \ldots, \hat{\theta}_S^{r-1}\right)^T$	Column vector of a respondent's latent trait values – estimated scores after responding to $r-1$ questions
$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_S)^T$	Column vector of a respondent's latent trait values – unobserved true scores

Notation	Definition
$\boldsymbol{\eta}^* = (\eta_1^*, \dots, \eta_S^*)^T$	Column vector of a respondent's latent trait values – unobserved true scores in a different metric (Chapter 2/ Study 1)
$\boldsymbol{x}, \boldsymbol{y}$	In Chapter 2: column vectors of linear transformation coefficients between $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$ (i.e., $\boldsymbol{\eta}^* = \boldsymbol{x^T\eta} + \boldsymbol{y}$) In Appendix D: generic random variables, possibly multidimensional, can be continuous or discrete.
$\widehat{\boldsymbol{\eta}} = (\hat{\eta}_1, \dots, \hat{\eta}_S)^T$	Column vector of a respondent's latent trait values – estimated scores
$\widehat{\boldsymbol{\eta}}^{r-1} = (\hat{\eta}_1^{r-1}, \dots, \hat{\eta}_S^{r-1})^T$	Column vector of a respondent's latent trait values – estimated scores after responding to $r-1$ pairwise comparisons
$\widehat{\boldsymbol{\eta}}^{ML}, \widehat{\boldsymbol{\eta}}^{WL}, \widehat{\boldsymbol{\eta}}^{MAP}, \widehat{\boldsymbol{\eta}}^{EAP}$	Column vector of a respondent's latent trait values – estimated scores using ML/ WL/ MAP/ EAP estimators
$\eta^{composite}$	A scalar overall score calculated as a weighted sum of different trait scores
$\boldsymbol{w} = (w_1, \dots, w_S)^T$	A column vector of weights assigned to traits in item selectors (where applicable), can be a constant or a function
U_i	Binary response to a single item i
$Y_{\{i,k\}}$	Binary response to a pairwise comparison $\{i, k\}$
$\boldsymbol{Y}$	An entire response string of binary pairwise responses
$\boldsymbol{Y}^{r-1}$	Response string of the first $r-1$ binary pairwise comparisons
$p_i(\boldsymbol{\theta}) \equiv P(U_i = 1 \mid \boldsymbol{\theta})$	The probability of responding favourably to item i given latent trait vector $\boldsymbol{\theta}$ in M2PNO model
$p_{\{i,k\}}(\boldsymbol{\eta}) \equiv P\left(Y_{\{i,k\}} = 1 \mid \boldsymbol{\eta}\right)$	The probability of endorsing the first item in pairwise comparison $\{i, k\}$ given latent trait vector $\boldsymbol{\eta}$ in TIRT model

Notation	Definition
$\Phi(\cdot)$	The standard normal cumulative distribution function
$\phi(\cdot)$	The standard normal density function
$E(\cdot)$	The expectation of a random variable
$P(\cdot)$	The probability mass function of a discrete random variable
$density(\cdot)$	The probability density function of a continuous random variable
$\boldsymbol{a}_i$	Column vector of S slope parameters of item i in M2PNO model
d_i	Intercept parameter of item i in M2PNO model
t_i	An item's psychological utility values within a respondent in TIRT model (the person index is omitted in the notation)
μ_i	Mean utility of item i in TIRT model
$\boldsymbol{\lambda_i} = \left(\lambda_{i_1}, \ldots, \lambda_{i_S}\right)^T$	Column vector of S factor loadings of item i in TIRT model
$\boldsymbol{\lambda_i^*} = \left(\lambda_{i_1}^*, \ldots, \lambda_{i_S}^*\right)^T$	Column vector of S factor loadings of item i in TIRT model – in a different metric (Study 1)
$\varepsilon_i \sim N(0, \psi_i^2)$	Normally distributed error term for item i in TIRT model, with mean 0 and unique variance ψ_i^2
$\varepsilon_i^* \sim N(0, \psi_i^{*2})$	Normally distributed error term for item i in TIRT model, with mean 0 and unique variance ψ_i^2 – in a different metric (Study 1)
$\gamma_{\{i,k\}} \equiv \mu_k - \mu_i$	Threshold parameter for the pairwise comparison $\{i, k\}$ in TIRT model
$\gamma_{\{i,k\}}^*$	Threshold parameter for the pairwise comparison $\{i, k\}$ in TIRT model – in a different metric (Study 1)

Notation	Definition
$z_{\{i,k\}} \equiv \frac{-\gamma_{\{i,k\}} + (\boldsymbol{\lambda_i} - \boldsymbol{\lambda_k})^T \boldsymbol{\eta}}{\sqrt{\psi_i^2 + \psi_k^2}}$	The argument for the standard normal cumulative distribution function in TIRT model
$L(\cdot \mid \boldsymbol{\eta})$	The likelihood of the observed response(s) given latent trait vector $\boldsymbol{\eta}$
$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_S)^T$	A vector of angles with the coordinate axes, indicating a direction in the multidimensional space
$\boldsymbol{\alpha}^s$	A vector of angles with the coordinate axes, indicating the direction along trait s in the multidimensional space
$\boldsymbol{\alpha}^{min}$	A vector of angles with the coordinate axes, indicating the direction in the multidimensional trait space that has minimum information
$\nabla_{\boldsymbol{\alpha}}$	The gradient or directional derivative in the direction of $\boldsymbol{\alpha}$
$I_i^{\alpha}(\boldsymbol{\theta})$	The information from item i in direction $\boldsymbol{\alpha}$ for an individual with trait profile $\boldsymbol{\theta}$
$I_{\{i,k\}}^{\alpha}(\boldsymbol{\eta})$	The information from pairwise comparison $\{i, k\}$ in direction $\boldsymbol{\alpha}$ for an individual with trait profile $\boldsymbol{\eta}$ in TIRT model
$CI_{\{i,k\}}^{\alpha^s}(\boldsymbol{\eta})$	Core information from pairwise comparison $\{i, k\}$ for trait s for an individual with trait profile $\boldsymbol{\eta}$ in TIRT model
$I^{\alpha}(\boldsymbol{\eta})$	The total information from all responses in direction $\boldsymbol{\alpha}$ for an individual with trait profile $\boldsymbol{\eta}$
$I_{Pos}^{\alpha^s}(\boldsymbol{\eta})$	The posterior information for trait s for an individual with trait profile $\boldsymbol{\eta}$
$SEM(\hat{\eta}_s)$	The standard error of measurement associated with the estimated score for trait s
$\boldsymbol{A}$	Block-diagonal design matrix of contrasts capturing the assignment of items (columns) to blocks (with rows corresponding to pairs within blocks)

Notation	Definition
$\boldsymbol{\Lambda}$	Matrix of factor loadings of items (rows) on latent traits (columns)
$\boldsymbol{A\Lambda}$	Matrix of factor loadings of each pair (rows) on each latent trait (columns)
$(\boldsymbol{A\Lambda})_{\{i,k\}}$	Row in matrix $\boldsymbol{A\Lambda}$ with factor loadings associated with pair $\{i, k\}$
$\boldsymbol{F}_{\{i,k\}}(\boldsymbol{\eta})$	$S \times S$ Fisher Information Matrix for pair $\{i, k\}$ for an individual with trait profile $\boldsymbol{\eta}$
$\boldsymbol{F}(\boldsymbol{\eta})$	$S \times S$ Fisher Information Matrix for all responses for an individual with trait profile $\boldsymbol{\eta}$
$\boldsymbol{F}^{r-1}(\boldsymbol{\eta})$	$S \times S$ Fisher Information Matrix for the first $r - 1$ responses for an individual with trait profile $\boldsymbol{\eta}$
$S_{\boldsymbol{\alpha}}^{ML}(\boldsymbol{\eta})$	The score function for the ML estimator (i.e., the gradient of the log likelihood in direction $\boldsymbol{\alpha}$)
$S_{\boldsymbol{\alpha}}^{WL}(\boldsymbol{\eta})$	The score function for the WL estimator (i.e., the gradient of the weighted log likelihood in direction $\boldsymbol{\alpha}$)
$S_{\boldsymbol{\alpha}}^{MAP}(\boldsymbol{\eta})$	The score function for the MAP estimator (i.e., the gradient of the log posterior function in direction $\boldsymbol{\alpha}$)
$Bias_s^{ML}(\boldsymbol{\eta})$	The asymptotic bias of the ML estimator in trait s
$\boldsymbol{Bias}^{ML}(\boldsymbol{\eta})$ $\equiv \left(Bias_1^{ML}(\boldsymbol{\eta}), \ldots, Bias_S^{ML}(\boldsymbol{\eta})\right)^T$	Column vector of asymptotic bias of the ML estimator for each of the S traits
$M(\boldsymbol{\eta})$	Weight function for the WL estimator
$\frac{\partial\{\cdot\}}{\partial\boldsymbol{\eta}} \equiv \begin{bmatrix} \frac{\partial\{\cdot\}}{\partial\eta_1} \\ \cdots \\ \cdots \\ \frac{\partial\{\cdot\}}{\partial\eta_S} \end{bmatrix}$	Column vector of partial derivatives along the directions of each of the S traits
$prior(\boldsymbol{\eta})$	Density of the prior distribution of latent traits

Notation	Definition
$density(\boldsymbol{\eta}\|\boldsymbol{Y})$	Density of the posterior distribution of latent traits given responses $\boldsymbol{Y}$
$f(\cdot), g(\cdot)$	Density functions describing the distributions of some generic random variables.
$KL(f \parallel g)$	Kullback–Leibler (KL) information/ distance between two probability distributions f and g that share the same parameters
$KL^{I}_{\{i,k\}}$	KL index, quantifying the KL distance between the probabilities of response to a pairwise comparison $\{i,k\}$ at the current trait estimates and at true trait values.
$h(\boldsymbol{\eta})$	Density of the trait space to integrate over in order to account for information at different true trait values in KLI item selector
$KL^{P}_{\{i,k\}}$	KL distance between subsequent posterior distributions of the trait estimates before and after an additional response to a pairwise comparison $\{i,k\}$
$MI(\boldsymbol{x};\boldsymbol{y})$	The mutual information between two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, which is equal to the KL distance between their joint density and their product marginal densities
$MI_{\{i,k\}}$	The mutual information between the current posterior distribution of trait estimates (i.e., $density(\boldsymbol{\eta}\|\boldsymbol{Y}^{r-1})$) and the response distribution of a possible new question (i.e., $Prob\left(Y_{\{i,k\}}\|\boldsymbol{Y}^{r-1}\right)$)
$H(\boldsymbol{y})$	Shannon entropy of a random variable $\boldsymbol{y}$
$H(\boldsymbol{y}\|\boldsymbol{x})$	Conditional entropy of a random variable $\boldsymbol{y}$ given $\boldsymbol{x}$

APPENDIX B: FORMULATION OF TRAIT ESTIMATORS FOR TIRT MODEL

Classical Estimators

Classical trait estimators only incorporate information from the assessment (i.e., item characteristics and item responses) in the estimation of person scores. They make no prior assumptions about the distribution of the latent traits.

Maximum Likelihood (ML) Estimator

The traditional statistical method of maximum likelihood (Fisher, 1922) can be applied to trait estimation, giving rise to the ML estimator. The ML estimator estimates person scores by finding the vector of trait parameters that maximises the likelihood of the item responses (Birnbaum, 1958, 1968; Segall, 1996; Tam, 1992). For the TIRT model, ML estimates are achieved through maximising Equation 7, leading to Equation B1.

$$\widehat{\boldsymbol{\eta}}^{ML} = arg\ max_{\boldsymbol{\eta}} \left\{ \prod_{\{i,k\}} L\left(Y_{\{i,k\}} | \boldsymbol{\eta}\right) \right\} = arg\ max_{\boldsymbol{\eta}} \left\{ \sum_{\{i,k\}} ln\left[L\left(Y_{\{i,k\}} | \boldsymbol{\eta}\right)\right] \right\} \qquad \text{(B1)}$$

Typically, the ML estimates are calculated by setting the gradient of the log likelihood of responses to zero and solving for the values of $\boldsymbol{\eta}$, as shown in Equation B2. In this expression, the gradient of the log likelihood in direction $\boldsymbol{\alpha}$ is denoted by $S_{\boldsymbol{\alpha}}^{ML}(\boldsymbol{\eta})$, which is known as the score function for the ML estimator. When solving Equation B2, it is sufficient to consider the gradients along the directions of the trait axes, $\boldsymbol{\alpha}^s$ (Segall, 1996), leading to Equation B3.

$$S_{\boldsymbol{\alpha}}^{ML}(\boldsymbol{\eta}) \equiv \nabla_{\boldsymbol{\alpha}} \left\{ \sum_{\{i,k\}} ln\left[L\left(Y_{\{i,k\}} | \boldsymbol{\eta} \right) \right] \right\} = 0 \qquad \forall\, \boldsymbol{\alpha} \tag{B2}$$

$$S_{\boldsymbol{\alpha}^s}^{ML}(\boldsymbol{\eta}) \equiv \frac{\partial}{\partial \eta_s} \left\{ \sum_{\{i,k\}} ln\left[L\left(Y_{\{i,k\}} | \boldsymbol{\eta} \right) \right] \right\} = 0 \qquad \forall\, s \tag{B3}$$

One issue with the ML estimator is that it has notable bias, which tends to stretch the trait estimates outwards when the assessment is short (Lord, 1983). Following the work of Lord (1983) and Warm (1989) on the asymptotic bias of the ML estimator for unidimensional IRT models, and utilising the statistical properties of the ML estimator as shown by Cox and Snell (1968, equation 20), the asymptotic bias of the ML estimator for any multidimensional IRT model for dichotomous responses can be deduced (see Appendix C). Considering that the binary outcome variable modelled in the TIRT model is the pairwise comparison $\{i, k\}$, the resulting expression reads:

$$\begin{aligned} Bias_s^{ML}(\boldsymbol{\eta}) = \frac{1}{2} \sum_{v,w,x} (\boldsymbol{F}^{-1})_{vs} (\boldsymbol{F}^{-1})_{wx} \sum_{\{i,k\}} \Bigg[& -\frac{\partial^2 p_{\{i,k\}}}{\partial \eta_w \partial \eta_x} \frac{\partial p_{\{i,k\}}}{\partial \eta_v} \\ & + \frac{\partial^2 p_{\{i,k\}}}{\partial \eta_v \partial \eta_x} \frac{\partial p_{\{i,k\}}}{\partial \eta_w} - \frac{\partial^2 p_{\{i,k\}}}{\partial \eta_v \partial \eta_w} \frac{\partial p_{\{i,k\}}}{\partial \eta_x} \Bigg] \left(\frac{1}{p_{\{i,k\}}\left(1 - p_{\{i,k\}}\right)} \right). \end{aligned} \tag{B4}$$

With this bias, high scores tend to get higher and low scores tend to get lower. However, because this bias is of order $O(r^{-1})$ where r is the number of item responses (Cox & Snell, 1968), it diminishes as the assessment gets longer. Nevertheless, it prompted searches for alternative estimators with less bias.

Weighted Likelihood (WL) Estimator

Motivated by a desire to reduce bias associated with the ML estimator, Warm (1989), Tseng and Hsu (2001), and Wang (2015) developed the WL estimator, and independently showed that it is less outwardly biased than the ML estimator whilst

retaining similar variance. The formulation of the WL estimator is shown in Equation B5, which only differs from the ML estimator by an additional weight function, $M(\boldsymbol{\eta})$.

$$\widehat{\boldsymbol{\eta}}^{WL} = \arg max_{\boldsymbol{\eta}} \left\{ M(\boldsymbol{\eta}) \left[\prod_{\{i,k\}} L\left(Y_{\{i,k\}} | \boldsymbol{\eta}\right) \right] \right\} \tag{B5}$$

Similar to ML, the WL estimates are typically calculated by setting the gradient of the weighted log likelihood (also known as the score function for the WL estimator, $S_{\boldsymbol{\alpha}}^{WL}(\boldsymbol{\eta})$) to zero and solving for $\boldsymbol{\eta}$. And again, it is sufficient to consider the gradient along each trait axis (Tseng & Hsu, 2001), as shown in Equation B6.

$$S_{\boldsymbol{\alpha}^s}^{WL}(\boldsymbol{\eta}) \equiv \frac{\partial}{\partial \eta_s} \left\{ ln[M(\boldsymbol{\eta})] + \sum_{\{i,k\}} ln\left[L\left(Y_{\{i,k\}} | \boldsymbol{\eta}\right)\right] \right\} = 0 \qquad \forall\, s \tag{B6}$$

Warm (1989) observed that, when the weight function is set to a positive constant, Equation B6 is equivalent to the ML estimator; alternatively, when the weight function is set to the prior density of the latent traits, Equation B6 is equivalent to the MAP estimator (see next section). Warm then designed a weight function for the unidimensional three-parameter logistic model that removes first-order bias from the ML estimator. Warm's weight function makes no prior assumption about the latent trait distribution, and therefore the WL estimator is not Bayesian. Tseng and Hsu (2001) and Wang (2015) subsequently extended Warm's weight function to the case of multidimensional IRT models, which can be directly applied to the TIRT model (Equation B7). In Equation B7, $\boldsymbol{Bias}^{ML}(\boldsymbol{\eta}) \equiv \left(Bias_1^{ML}(\boldsymbol{\eta}), \ldots, Bias_S^{ML}(\boldsymbol{\eta})\right)^T$ denotes the column vector of ML bias values for each of the S traits. Note that in the calculations for the WL estimates, it is not necessary to deduce the functional form of $M(\boldsymbol{\eta})$. This is because the WL estimates are calculated by solving Equation B6, which

only depends on $\frac{\partial\{ln[M(\boldsymbol{\eta})]\}}{\partial \boldsymbol{\eta}}$. Moreover, using Equation B7, the relationship between score functions for the ML and WL estimators can be expressed as in Equation B8.

$$\frac{\partial\{ln[M(\boldsymbol{\eta})]\}}{\partial \boldsymbol{\eta}} \equiv \begin{bmatrix} \frac{\partial\{ln[M(\boldsymbol{\eta})]\}}{\partial \eta_1} \\ \cdots \\ \cdots \\ \frac{\partial\{ln[M(\boldsymbol{\eta})]\}}{\partial \eta_s} \\ \cdots \\ \cdots \\ \frac{\partial\{ln[M(\boldsymbol{\eta})]\}}{\partial \eta_S} \end{bmatrix} = -\boldsymbol{F}(\boldsymbol{\eta})\boldsymbol{Bias}^{ML}(\boldsymbol{\eta}) \tag{B7}$$

$$S_{\alpha^s}^{WL}(\boldsymbol{\eta}) = S_{\alpha^s}^{ML}(\boldsymbol{\eta}) - [\boldsymbol{F}(\boldsymbol{\eta})\boldsymbol{Bias}^{ML}(\boldsymbol{\eta})]_s \quad \forall\, s \tag{B8}$$

Bayesian Estimators

Bayesian trait estimators not only account for data obtained directly from the assessment, but also incorporate information gained from other sources. For example, one may hold information about the respondent population in general, and/or have prior knowledge about a particular respondent from previous assessments or interactions. Such information is captured in the prior distribution of latent traits with density $prior(\boldsymbol{\eta})$, which is typically set to be multivariate normal. Then, the posterior distribution of traits, $density(\boldsymbol{\eta}|\boldsymbol{Y})$, can be calculated as per Equation B9 (Segall, 1996). Bayesian estimators make use of the posterior distribution in the estimation of trait values.

$$density(\boldsymbol{\eta}|\boldsymbol{Y}) = \frac{L(\boldsymbol{Y}|\boldsymbol{\eta}) \times prior(\boldsymbol{\eta})}{\int L(\boldsymbol{Y}|\boldsymbol{\eta}) \times prior(\boldsymbol{\eta}) d\boldsymbol{\eta}} \tag{B9}$$

Maximum a Posteriori (MAP) Estimator

One popular Bayesian estimator is the MAP estimator, also referred to as the Bayesian Modal (BM) estimator, which estimates trait scores by finding the maximisers

of the posterior function, as described in Equation B10 (Bock & Aitkin, 1981; Lord, 1986; Mislevy, 1986; Samejima, 1969; Segall, 1996).

$$\widehat{\boldsymbol{\eta}}^{MAP} = arg\ max_{\boldsymbol{\eta}}\{density(\boldsymbol{\eta}|\boldsymbol{Y})\} = arg\ max_{\boldsymbol{\eta}}\{ln[density(\boldsymbol{\eta}|\boldsymbol{Y})]\} \qquad \text{(B10)}$$

Similar to ML and WL, the MAP estimates are calculated by setting the gradient of the log posterior function to zero and solving for $\boldsymbol{\eta}$. The score function for MAP can be deduced accordingly (Equation B11). It can be seen in this expression that, when a uniform prior is assumed (i.e., when there is no prior information), the term containing the prior function is zero and the MAP estimator reduces to the ML estimator. And again, it is sufficient to consider only the directions along the trait axes when solving Equation B11 (Segall, 1996).

$$S_{\boldsymbol{\alpha}}^{MAP}(\boldsymbol{\eta}) = S_{\boldsymbol{\alpha}}^{ML}(\boldsymbol{\eta}) + \nabla_{\boldsymbol{\alpha}}\{ln[prior(\boldsymbol{\eta})]\} = 0 \qquad \forall\ \boldsymbol{\alpha} \qquad \text{(B11)}$$

Expected a Posteriori (EAP) Estimator

Another popular Bayesian estimator is the EAP estimator, which estimates trait scores as the expected value (mean) of the posterior distribution function, as described in Equation B12 (Bock & Aitkin, 1981; Bock & Mislevy, 1982; Segall, 1996). Unlike the other trait estimators, the EAP estimates are not calculated by solving a score function. Instead, numerical integration routines, for example the Gauss-Hermite quadrature method (Abramowitz & Stegun, 1972; Stroud & Sechrest, 1966), are typically employed to approximate the integral.

$$\widehat{\boldsymbol{\eta}}^{EAP} = \int \boldsymbol{\eta} \times density(\boldsymbol{\eta}|\boldsymbol{Y})\, \boldsymbol{d\eta} \qquad \text{(B12)}$$

Full Posterior

Instead of extracting point estimates from the posterior distribution, it is sometimes possible to utilise the entire posterior function $density(\boldsymbol{\eta}|\boldsymbol{Y})$ as the

estimator. This method is only compatible with some of the more advanced item selectors (see Appendix D for details). Using the full posterior function bypasses the need to calculate point estimates until the very end of the assessment, where point estimates are typically preferred for reporting.

APPENDIX C: BIAS OF THE ML ESTIMATOR IN MULTIDIMENSIONAL IRT MODELS

Cox and Snell (1968, equation 20) deduced the general formula for the bias of the ML estimator which, in the case of multidimensional IRT models (including but not limited to the TIRT model), is as follows:

$$Bias_s^{ML}(\boldsymbol{\eta}) \equiv E[\hat{\eta}_s^{ML} - \eta_s] = \frac{1}{2}\sum_{v,w,x}(\boldsymbol{F}^{-1})_{vs}(\boldsymbol{F}^{-1})_{wx}\left(\boldsymbol{K}_{vwx} + 2\boldsymbol{J}_{w,vx}\right). \quad \text{(C1)}$$

In this expression, $Bias_s^{ML}(\boldsymbol{\eta})$ denotes the bias of the s^{th} element of the ML estimator $\widehat{\boldsymbol{\eta}}^{ML}$ for the person parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_S)^T$; $v, w, x \in \{1, \ldots, S\}$ are indices for traits; $\boldsymbol{F}$ is the total FIM for a respondent obtained from all r item responses U_i with likelihood $L(U_i|\boldsymbol{\eta})$; and $\boldsymbol{K}$ and $\boldsymbol{J}$ are three-dimensional arrays defined as follows:

$$J_{w,vx} \equiv \sum_i E\left[\frac{\partial}{\partial\eta_w} ln[L(U_i|\boldsymbol{\eta})] \times \frac{\partial^2}{\partial\eta_v\partial\eta_x} ln[L(U_i|\boldsymbol{\eta})]\right]; \quad \text{(C2)}$$

$$K_{vwx} \equiv \sum_i E\left[\frac{\partial^3}{\partial\eta_v\partial\eta_w\partial\eta_x} ln[L(U_i|\boldsymbol{\eta})]\right]. \quad \text{(C3)}$$

This bias term is of order $O(r^{-1})$ and thus tends to zero as the number of item responses increases (Cox & Snell, 1968). For any multidimensional IRT model with dichotomous responses (i.e., $U_i \in \{0,1\}$), the likelihood function has the format of $L(U_i|\boldsymbol{\eta}) = {p_i}^{U_i}{q_i}^{1-U_i}$ where $p_i \equiv P(U_i = 1|\boldsymbol{\eta})$ and $q_i \equiv 1 - p_i$, and it can be deduced that:

$$\frac{\partial}{\partial\eta_w}L(U_i|\boldsymbol{\eta}) = \frac{\partial}{\partial\eta_w}({p_i}^{U_i}{q_i}^{1-U_i}) = L(U_i|\boldsymbol{\eta})\frac{\partial p_i}{\partial\eta_w}\left(\frac{U_i - p_i}{p_i q_i}\right); \quad \text{(C4)}$$

and

$$\frac{\partial}{\partial\eta_w}\left(\frac{U_i - p_i}{p_i q_i}\right) = -\frac{\partial p_i}{\partial\eta_w}\left(\frac{U_i - p_i}{p_i q_i}\right)^2. \quad \text{(C5)}$$

It follows from Equations C4 and C5 that:

$$\frac{\partial}{\partial \eta_w} ln[L(U_i|\boldsymbol{\eta})] = \frac{1}{L(U_i|\boldsymbol{\eta})} \frac{\partial}{\partial \eta_w} L(U_i|\boldsymbol{\eta}) = \frac{\partial p_i}{\partial \eta_w} \left(\frac{U_i - p_i}{p_i q_i} \right); \tag{C6}$$

and

$$\frac{\partial^2}{\partial \eta_v \partial \eta_x} ln[L(U_i|\boldsymbol{\eta})] = \frac{\partial}{\partial \eta_v} \left(\frac{\partial p_i}{\partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right) \right)$$

$$= \frac{\partial^2 p_i}{\partial \eta_v \partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right) - \frac{\partial p_i}{\partial \eta_v} \frac{\partial p_i}{\partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right)^2; \tag{C7}$$

and

$$\frac{\partial^3}{\partial \eta_v \partial \eta_w \partial \eta_x} ln[L(U_i|\boldsymbol{\eta})] = \frac{\partial}{\partial \eta_v} \left(\frac{\partial^2 p_i}{\partial \eta_w \partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right) - \frac{\partial p_i}{\partial \eta_w} \frac{\partial p_i}{\partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right)^2 \right)$$

$$= \frac{\partial^3 p_i}{\partial \eta_v \partial \eta_w \partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right)$$

$$- \left[\frac{\partial^2 p_i}{\partial \eta_w \partial \eta_x} \frac{\partial p_i}{\partial \eta_v} + \frac{\partial^2 p_i}{\partial \eta_v \partial \eta_w} \frac{\partial p_i}{\partial \eta_x} + \frac{\partial^2 p_i}{\partial \eta_v \partial \eta_x} \frac{\partial p_i}{\partial \eta_w} \right] \left(\frac{U_i - p_i}{p_i q_i} \right)^2 \tag{C8}$$

$$+ 2 \frac{\partial p_i}{\partial \eta_v} \frac{\partial p_i}{\partial \eta_w} \frac{\partial p_i}{\partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right)^3.$$

Substituting Equations C6 and C7 into Equation C2 gives:

$$\boldsymbol{J}_{w,vx} = \sum_i E \left\{ \frac{\partial p_i}{\partial \eta_w} \left(\frac{U_i - p_i}{p_i q_i} \right) \left[\frac{\partial^2 p_i}{\partial \eta_v \partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right) - \frac{\partial p_i}{\partial \eta_v} \frac{\partial p_i}{\partial \eta_x} \left(\frac{U_i - p_i}{p_i q_i} \right)^2 \right] \right\}$$

$$= \sum_i \frac{\partial p_i}{\partial \eta_w} \frac{\partial^2 p_i}{\partial \eta_v \partial \eta_x} \left(\frac{1}{p_i q_i} \right) + \frac{\partial p_i}{\partial \eta_w} \frac{\partial p_i}{\partial \eta_v} \frac{\partial p_i}{\partial \eta_x} \left(\frac{2p_i - 1}{{p_i}^2 {q_i}^2} \right). \tag{C9}$$

Similarly, substituting Equation C8 into Equation C3 gives:

$$\boldsymbol{K}_{vwx} = \sum_i \left\{ 2 \frac{\partial p_i}{\partial \eta_v} \frac{\partial p_i}{\partial \eta_w} \frac{\partial p_i}{\partial \eta_x} \left(\frac{1 - 2p_i}{{p_i}^2 {q_i}^2} \right) \right.$$

$$\left. - \left[\frac{\partial^2 p_i}{\partial \eta_w \partial \eta_x} \frac{\partial p_i}{\partial \eta_v} + \frac{\partial^2 p_i}{\partial \eta_v \partial \eta_x} \frac{\partial p_i}{\partial \eta_w} + \frac{\partial^2 p_i}{\partial \eta_v \partial \eta_w} \frac{\partial p_i}{\partial \eta_x} \right] \left(\frac{1}{p_i q_i} \right) \right\}. \tag{C10}$$

Finally, substituting Equations C9 and C10 back into Equation C1 gives the formula for the bias of the ML estimator for multidimensional IRT models with dichotomous item responses:

$$Bias_s^{ML}(\boldsymbol{\eta}) = \frac{1}{2}\sum_{v,w,x}(\boldsymbol{F}^{-1})_{vs}(\boldsymbol{F}^{-1})_{wx}\sum_i\left[-\frac{\partial^2 p_i}{\partial\eta_w\partial\eta_x}\frac{\partial p_i}{\partial\eta_v}+\frac{\partial^2 p_i}{\partial\eta_v\partial\eta_x}\frac{\partial p_i}{\partial\eta_w}\right.$$
$$\left.-\frac{\partial^2 p_i}{\partial\eta_v\partial\eta_w}\frac{\partial p_i}{\partial\eta_x}\right]\left(\frac{1}{p_i q_i}\right). \quad \text{(C11)}$$

Note that Equation C11 is the multidimensional extension of the unidimensional ML bias formula presented by Lord (1983) and Warm (1989, equation 6). Indeed, reducing the dimensionality of $\boldsymbol{\eta}$ to one arrives at the same expression:

$$Bias^{ML}(\eta) = \frac{-\sum_i\left[\frac{\partial p_i}{\partial\eta}\frac{\partial^2 p_i}{\partial\eta^2}/p_i(1-p_i)\right]}{2F^2}. \quad \text{(C12)}$$

APPENDIX D: FORMULATION OF ITEM SELECTORS FOR TIRT FC CAT

Criteria Based on Information Maximisation

Consider first the simplest case of a unidimensional CAT (i.e., all questions in the test measure the same construct). Let $i_1, \dots i_{r-1}$ be the first $r-1$ questions in an adaptive test session, let R_r be the set of unused items up to this point, and let $\hat{\theta}^{r-1}$ be the trait estimate at this point. The classic method of selecting the r^{th} item i_r is to pick an item in R_r that maximises the total test information at $\hat{\theta}^{r-1}$ (Birnbaum, 1968; van der Linden, 2010), as shown in Equation D1.

$$i_r = arg\ max_{i \in R_r}\{I_{i_1}(\hat{\theta}^{r-1}) + \cdots + I_{i_{r-1}}(\hat{\theta}^{r-1}) + I_i(\hat{\theta}^{r-1})\}$$

$$= arg\ max_{i \in R_r}\{I_i(\hat{\theta}^{r-1})\} \quad \text{(D1)}$$

While this information maximisation method is straightforward for a unidimensional test choosing one item at a time, its extension to MFC assessments presents additional complexities (see Chapter 3). In order to address the multidimensionality challenge, researchers have developed a range of item selectors that reduce multidimensional information into scalar summary indices. The subsequent sections formulate a selection of such indices for MFC assessments using TIRT. To begin with, this section describes mathematically-simple but likely sub-optimal item selection criteria based on the idea of information maximisation. These simple item selectors, together with random item selection, can serve as worst-case benchmarks when appraising the efficiency of the more sophisticated item selectors.

Maximise Weighted Information (WI)

Equation 12 describes the information gain from a FC pair $\{i, k\}$ in the direction of $\boldsymbol{\alpha}$. Since the typical measurement goal is to optimise information gain in the direction

of all intended traits (i.e., $\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^S$), the simplest criterion is to maximise the sum of information across all traits. Moreover, weights can be assigned to each of the measured traits to indicate the relative priorities between them. Such a total weighted information criteria can be used to choose FC pairs to be present next (Equation D2). Note that the weights assigned to the measured traits, $\boldsymbol{w} = (w_1, \ldots, w_S)^T$, can be static (e.g., indicating the level of importance of each trait for the purpose of the assessment) or dynamic (e.g., prioritising the traits still lacking in measurement precision).

$$\{i_r, k_r\} = arg\ max_{\{i,k\} \in R_r} \left\{ \sum_{s=1}^{S} w_s \left[I_{\{i,k\}}^{\boldsymbol{\alpha}^s}(\hat{\boldsymbol{\eta}}^{r-1}) \right] \right\} \tag{D2}$$

Maximise Weighted Core Information (WCI)

The aforementioned WI item selector may be simplified further by considering only the core information from a FC pair (see Equation 17), giving rise to the WCI item selector (Equation D3). This way, item selection focuses only on the information gain on the traits directly involved in the FC pair, ignoring any peripheral information gain from responses to items measuring correlated traits.

$$\{i_r, k_r\} = arg\ max_{\{i,k\} \in R_r} \left\{ \sum_{s=1}^{S} w_s \left[CI_{\{i,k\}}^{\boldsymbol{\alpha}^s}(\hat{\boldsymbol{\eta}}^{r-1}) \right] \right\} \tag{D3}$$

Maximise Information in Direction with Minimum Information (DMI)

Reckase (2009) proposed to prioritise information gain in the direction of the trait space that currently has minimum information (here denoted as $\boldsymbol{\alpha}^{min}$). To apply this method, a two-step process is followed: finding the direction with minimum information (Equation D4), and then selecting items to maximise information gain in that direction (Equation D5).

$$\boldsymbol{\alpha}^{min} = arg\ min_{\boldsymbol{\alpha}}\left\{I^{\boldsymbol{\alpha}}_{\{i_1,k_1\}}(\widehat{\boldsymbol{\eta}}^{r-1}) + \cdots + I^{\boldsymbol{\alpha}}_{\{i_{r-1},k_{r-1}\}}(\widehat{\boldsymbol{\eta}}^{r-1})\right\} \quad \text{(D4)}$$

$$\{i_r, k_r\} = arg\ max_{\{i,k\}\in R_r}\left\{I^{\boldsymbol{\alpha}^{min}}_{\{i,k\}}(\widehat{\boldsymbol{\eta}}^{r-1})\right\} \quad \text{(D5)}$$

However, given all the possible directions in the multidimensional trait space, solving Equation D4 can be difficult. Reckase (2009) suggested grid-searching through directions in equally-spaced small intervals (e.g., 10-degree intervals) throughout the entire trait space in order to find an approximation for $\boldsymbol{\alpha}^{min}$. While this operation is manageable with two traits, the number of directions to search through rises quickly as the dimensionality of the trait space increases. Moreover, the focus of personality assessments tends to be precise estimation of the personality traits (i.e., gaining information along the trait axes in the multidimensional trait space), or prediction of some outcome variable using a regression model of personality traits (i.e., gaining information in a specific direction in the multidimensional trait space). If $\boldsymbol{\alpha}^{min}$ is far away from the intended directions as determined by the assessment purpose, selecting items to maximise information in the direction of $\boldsymbol{\alpha}^{min}$ can be counterproductive. A modification of this method may work better for personality assessments: select items to maximise information along the trait axis with minimum information. This simplification reduces Reckase's method to a special case of the WI item selector, i.e., setting the weights w_s to 1 for the axis with minimum information and 0 otherwise.

Criteria Based on FIM

The FIM (Equation 26) is closely related to the accuracy of trait estimations. It is therefore frequently used in the designing of item selectors. Let $\boldsymbol{F}^{r-1}(\boldsymbol{\eta})$ be the total FIM from the first $r-1$ responses.

Minimise Trace of the Inverse FIM (A-optimality)

One way to optimise measurement on all intended traits simultaneously is to minimise their total error variance, which is equivalent to minimising the trace of the inverse FIM (see Mulder & van der Linden, 2009; Silvey, 1980). Equation D6 shows how this method, often termed "A-optimality", can be applied to choosing a FC pair.

$$\{i_r, k_r\} = arg\ min_{\{i,k\}\in R_r}\left\{tr\left[\left(\boldsymbol{F}^{r-1}(\hat{\boldsymbol{\eta}}^{r-1}) + \boldsymbol{F}_{\{i,k\}}(\hat{\boldsymbol{\eta}}^{r-1})\right)^{-1}\right]\right\} \qquad \text{(D6)}$$

A-optimality is one of the most popular item selectors for multidimensional CAT. However, the calculation for the A-optimality criterion is more intensive compared to that for the WI and WCI criteria. First, while the information contributions from previous responses drop out in the WI and WCI item selectors, they are inseparable in the A-optimality criterion. As a result, after every response the FIM of all previous responses need to be re-computed based on the latest person parameter estimates. Second, to compute the A-optimality criterion, the total FIM needs to be inverted for each one of the possible FC pairs, which is a more computationally intensive operation than simple arithmetic calculations. Moreover, the number of possible FC pairs $\binom{|R_r|}{2}$ grows multiplicatively with the size of the item bank, and thus even the smallest computational delay may be exaggerated many times and become noticeable to the respondents.

With modern computational power, the calculation complexity of A-optimality is likely manageable even for FC assessments. However, the same may not be true for the more demanding item selectors described in later parts of this section. Therefore, the comparison of item selectors should consider not only their measurement efficiency, but also their computational feasibility.

Minimise Weighted Sum of Entries of the Inverse FIM (C-optimality)

Sometimes the focus of a personality assessment is to produce a scalar overall score, for example, to predict an important outcome (e.g., job performance) based on a regression equation of multiple personality traits. When the goal is to report a composite score $\eta^{composite}$ calculated as a weighted sum of different trait scores (Equation D7), the adaptive item selection process may attempt to minimise the error variance of $\eta^{composite}$, which is equivalent to minimising a weighted sum of components of the inverse FIM (see Mulder & van der Linden, 2009; Silvey, 1980). Equation D8 shows how this method, often termed "C-optimality", can be applied to choosing a FC pair.

$$\eta^{composite} = \boldsymbol{w}^T\boldsymbol{\eta} = \sum_{s=1}^{S} w_s\eta_s \tag{D7}$$

$$\{i_r, k_r\} = arg\ min_{\{i,k\}\in R_r}\left\{\boldsymbol{w}^T\left[\left(\boldsymbol{F}^{r-1}(\widehat{\boldsymbol{\eta}}^{r-1}) + \boldsymbol{F}_{\{i,k\}}(\widehat{\boldsymbol{\eta}}^{r-1})\right)^{-1}\right]\boldsymbol{w}\right\} \tag{D8}$$

When all traits are equally important, all entries of $\boldsymbol{W}$ can be set to 1, and the C-optimality criterion simplifies to minimising the sum of all entries of the inverse FIM. In this case, C-optimality is similar to A-optimality – the latter sums over the diagonal of the inverse FIM, whereas the former also includes the off-diagonal terms.

Maximise Determinant of the FIM (D-optimality)

Another way to achieve good measurement on all traits simultaneously is to minimise the volume of the confidence ellipsoid of the trait estimates, which is equivalent to maximising the determinant of the FIM (see Mulder & van der Linden, 2009; Silvey, 1980). Equation D9 shows how this method, often termed "D-optimality", can be applied to choosing a FC pair.

$$\{i_r, k_r\} = arg\ max_{\{i,k\} \in R_r} \left\{ det \left(\boldsymbol{F}^{r-1}(\widehat{\boldsymbol{\eta}}^{r-1}) + \boldsymbol{F}_{\{i,k\}}(\widehat{\boldsymbol{\eta}}^{r-1}) \right) \right\} \quad \text{(D9)}$$

D-optimality is also one of the most popular item selectors for multidimensional CAT. Similar to A-optimality and C-optimality, the information contributions from previous responses are inseparable in the D-optimality criterion, leading to more intensive calculations than the WI and WCI item selectors. But unlike A-optimality and C-optimality which had to invert $\binom{|R_r|}{2}$ matrices, D-optimality instead calculates the determinants for the same matrices. As matrix determinant calculations tend to be less intensive than matrix inversion operations, D-optimality is less computationally intensive compared to A-optimality or C-optimality.

Maximise Minimum Eigenvalue of the FIM (E-optimality)

As Atkinson, Donev, and Tobias (2007, p. 135-136) explained, E-optimality aims to minimise the variance of the most imprecisely-estimated linear combination $\boldsymbol{w}^T\boldsymbol{\eta}$ where $\boldsymbol{w}^T\boldsymbol{w} = 1$, which is equivalent to maximising the minimum eigenvalue of the FIM. Mulder and van der Linden (2009) applied this criterion to CAT, and found that it had a tendency to select bad items especially for respondents with extreme trait locations, thus contradicting the aim of adaptively choosing appropriate items for the respondent. Mulder and van der Linden (2009) therefore recommended against using E-Optimality in CAT. For this reason, E-Optimality was not explored further in this thesis.

Maximise Trace of the FIM (T-Optimality)

Another relevant method is "T-optimality", which maximises the trace of the FIM (Allen-Zhu, Li, Singh, & Wang, 2017; Pukelsheim, 2006). Equation D10 shows how T-optimality can be applied to choosing a FC pair.

$$\{i_r, k_r\} = arg\ max_{\{i,k\} \in R_r} \left\{ tr\left[\boldsymbol{F}_{\{i,k\}}(\widehat{\boldsymbol{\eta}}^{r-1}) \right] \right\} \quad \text{(D10)}$$

T-optimality is closely related to information maximisation criteria. Comparing the functional forms of $I_{\{i,k\}}^{\alpha^s}(\boldsymbol{\eta})$ (Equation 16), $CI_{\{i,k\}}^{\alpha^s}(\boldsymbol{\eta})$ (Equation 17) and $\boldsymbol{F}_{\{i,k\}}(\boldsymbol{\eta})$ (Equation 23), it can be seen that WI, WCI and T-optimality criteria are very similar, differing only in how the information from correlated traits are handled. In fact, these three item selectors are mathematically equivalent when the traits are uncorrelated and the weights are equal across traits.

In terms of computational complexity, calculating the trace of a matrix is much simpler than inverting a matrix or computing its determinant. Moreover, the information contributions from previous responses drop out in the T-optimality criterion. Therefore, the computational complexity of T-optimality is less than A-, C- or D-Optimality, but on par with WI and WCI.

Criteria Based on Kullback–Leibler (KL) Information

All item selectors described so far rely on interim trait estimates, which may be far from the true trait standings. As a result, the item selectors may be choosing items that optimise measurement at the wrong locations, especially at the beginning of a CAT session when trait estimates are still inaccurate (e.g., Chang and Ying, 1996). This phenomenon is called the attenuation paradox (Lord & Novick, 1968). In order to tackle this problem, researchers have explored methods that optimise information globally (i.e., considering information for all trait locations as opposed to focusing on interim point estimates), often utilising the Kullback–Leibler (KL) information concept (Cover & Thomas, 2006; Kullback, 1959; Lehmann & Casella, 1998). The KL information for two density functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ is defined by Equation D11 for continuous $\boldsymbol{x}$, or Equation D12 for discrete $\boldsymbol{x}$ (Cover & Thomas, 2006; Kullback, 1959; Lehmann & Casella, 1998; Mulder & van der Linden, 2010).

$$KL(f \parallel g) = E_f\left[ln\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right] = \int f(\boldsymbol{x}) ln\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} d\boldsymbol{x} \quad \text{(D11)}$$

$$KL(f \parallel g) = \sum_{\boldsymbol{x}} f(\boldsymbol{x}) ln\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \quad \text{(D12)}$$

KL information is the "distance" between two probability distributions f and g that share the same parameters $\boldsymbol{x}$ (Kullback & Leibler, 1951; Lehmann & Casella, 1998). Note that KL information is not a proper distance measure because it is not symmetrical, i.e., $KL(f \parallel g) \neq KL(g \parallel f)$ (Mulder & van der Linden, 2010). In typical applications of KL information, f is set to some prior probability distribution and g is set to the revised posterior distribution or the true probability distribution; then KL information represents the information gained when updating one's hypothesis from f to g (Burnham & Anderson, 2002). By selecting appropriate distributions to substitute into $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$, the KL information measure can be utilised in CAT.

Maximum Item KL Information (KLI-U and KLI-B)

Chang and Ying (1996) proposed the KL index (KLI) for item selection in unidimensional CAT, which was subsequently extended to multidimensional CAT by Veldkamp and van der Linden (2002). In this method, f and g are set to the probabilities of a new item response at the current trait estimates and at true trait values respectively. Then, the KL distance between these two specific density functions (denoted by KL^I) quantifies the power of the new item to differentiate between the current trait estimates and true trait values, with larger distances indicating greater discriminations and thus greater power to improve measurement accuracy (Mulder & van der Linden, 2010; Veldkamp & van der Linden, 2002). Equations D13 to D15 show how this method can be applied to FC assessments.

$$f\left(Y_{\{i,k\}}\right) = Prob\left(Y_{\{i,k\}}|\widehat{\boldsymbol{\eta}}\right) \tag{D13}$$

$$g\left(Y_{\{i,k\}}\right) = Prob\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right) \tag{D14}$$

$$\begin{aligned} KL^{I}_{\{i,k\}} &\equiv KL\left(Prob\left(Y_{\{i,k\}}|\widehat{\boldsymbol{\eta}}\right) \parallel Prob\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right)\right) \\ &= p_{\{i,k\}}(\widehat{\boldsymbol{\eta}})\left[ln\frac{p_{\{i,k\}}(\widehat{\boldsymbol{\eta}})}{p_{\{i,k\}}(\boldsymbol{\eta})}\right] + \left(1 - p_{\{i,k\}}(\widehat{\boldsymbol{\eta}})\right)\left[ln\frac{1 - p_{\{i,k\}}(\widehat{\boldsymbol{\eta}})}{1 - p_{\{i,k\}}(\boldsymbol{\eta})}\right] \end{aligned} \tag{D15}$$

$$\{i_r, k_r\} = arg\ max_{\{i,k\}\in R_r}\left\{\int_{\boldsymbol{\eta}} h(\boldsymbol{\eta}) \times KL^{I}_{\{i,k\}}\, d\boldsymbol{\eta}\right\} \tag{D16}$$

Because the true trait values $\boldsymbol{\eta}$ are unknown, Equation D15 is integrated over the trait space with some density function $h(\boldsymbol{\eta})$ to arrive at a global information criterion as shown in Equation D16. The choice of the density function $h(\boldsymbol{\eta})$ gives rise to different varieties of the KLI item selector. Chang and Ying (1996) proposed setting $h(\boldsymbol{\eta}) = 1$ over a confidence region of the trait estimates that shrinks as the adaptive test progresses (and $h(\boldsymbol{\eta}) = 0$ elsewhere), giving rise to the KLI-U (uniform) item selector. Chang and Ying (1996) and Veldkamp and van der Linden (2002) also suggested incorporating the likelihood variations of $\boldsymbol{\eta}$ at different locations of the trait space by setting $h(\boldsymbol{\eta}) = density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1})$, i.e., the posterior density of the trait estimates after considering the previous $r - 1$ responses $\boldsymbol{Y}^{r-1}$, giving rise to the KLI-B (Bayesian) item selector.

The KLI item selectors have a couple of desirable features compared to those based on the FIM. In terms of measurement efficiency, KLI-U had been shown to outperform FIM-based methods[9] in unidimensional CAT, especially at the early stages when the person location estimates were inaccurate (Chang & Ying, 1996). In terms of

[9] In the unidimensional case, A-, C-, D- and T-optimality are all equivalent.

dimensional simplicity, KLI is always a scalar regardless of the number of traits measured, whereas the size of the FIM grows with the dimensionality of the test and thus an additional step of dimension-reduction is required to produce a scalar summary index suited for item selection (Mulder & van der Linden, 2010). The measurement efficiency and dimensional simplicity of KLI, however, come with significant computational complexity due to the integration in Equation D16. This integration is often approximated using Gauss-Hermite quadrature (Abramowitz & Stegun, 1972; Stroud & Sechrest, 1966). Because the computational intensity of numerical integration increases exponentially with the number of traits being measured, the computational power demands of KLI in the case of multidimensional personality assessments are likely significantly higher than that in the case of ability assessments with only one or two dimensions. Furthermore, the computational power demands of KLI are further intensified in the case of FC personality assessments where even a small item bank can lead to a large number of possible FC pairs to consider, all of which require integration when computing the KLI criteria for choosing the best pair to present next.

Maximum KL Distance Between Subsequent Posteriors (KLP)

Mulder and van der Linden (2010) suggested that the KL distance between subsequent posterior distributions of the trait estimates can be utilised in item selection. They proposed setting f and g respectively to the posterior distributions of the trait estimates before and after an additional response. Then, the KL distance between these two specific density functions (denoted by KL^P) quantifies how much the new response changes the posterior distribution of the trait estimates, with larger distances indicating greater power to refine the trait estimates. Equations D17 to D21 show how this method can be applied to FC assessments.

$$f(\boldsymbol{\eta}) = density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}) \tag{D17}$$

$$g(\boldsymbol{\eta}) = density\left(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}, Y_{\{i,k\}}\right) = \frac{Prob\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right) density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1})}{Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)} \tag{D18}$$

$$Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right) = \int_{\boldsymbol{\eta}} Prob\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right) density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1})\, d\boldsymbol{\eta} \tag{D19}$$

$$\begin{aligned} KL^{P}_{\{i,k\}} &\equiv KL\left(density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}) \parallel density\left(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}, Y_{\{i,k\}}\right)\right) \\ &= \int_{\boldsymbol{\eta}} density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}) ln \frac{density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1})}{density\left(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}, Y_{\{i,k\}}\right)} d\boldsymbol{\eta} \\ &= \int_{\boldsymbol{\eta}} density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}) ln \frac{Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)}{Prob\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right)} d\boldsymbol{\eta} \end{aligned} \tag{D20}$$

$$\{i_r, k_r\} = arg\, max_{\{i,k\} \in R_r} \left\{ \sum_{Y_{\{i,k\}}} \left[Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right) \times KL^{P}_{\{i,k\}} \right] \right\} \tag{D21}$$

Updating of the posterior distribution of the trait estimates after an additional response is an iterative process utilising Bayes' theorem, as described by Equations D18 and D19 (Mulder & van der Linden, 2010). And because the subsequent response $Y_{\{i,k\}}$ is unknown, Mulder and van der Linden (2010) suggested taking the expectation over all possible values of $Y_{\{i,k\}}$ to arrive at the KLP criterion, as shown in Equation D21.

Mulder and van der Linden (2010) demonstrated algebraically that KLP and KLI-B are closely related and only differ in how the item response probabilities are computed: KLI-B estimates it based on the current trait estimate (i.e., $Prob\left(Y_{\{i,k\}}|\widehat{\boldsymbol{\eta}}\right)$), whereas KLP estimates it conditional on the existing response string (i.e., $Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)$). They thus concluded that KLP is theoretically more robust and less prone to inaccurate interim trait estimates than KLI-B. However, as the calculations for

$Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)$ involve yet another integration, the computational complexity of KLP is much higher than that of KLI-B, especially for MFC personality assessments.

Maximum Mutual Information (MUI or KLB)

Weissman (2007) suggested making use of a special version of the KL information measure – the mutual information measure – for adaptive item selection. The mutual information between two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by Equations D22 and D23 for the continuous and discrete variables respectively (Cover & Thomas, 2006). From its algebraic expressions, it can be seen that the mutual information between two random variables is the KL distance between their joint density and their product marginal densities (Equation D24; Cover & Thomas, 2006).

$$MI(\boldsymbol{x};\boldsymbol{y}) = \int_{\boldsymbol{y}}\int_{\boldsymbol{x}} density(\boldsymbol{x},\boldsymbol{y}) ln \frac{density(\boldsymbol{x},\boldsymbol{y})}{density(\boldsymbol{x})density(\boldsymbol{y})} d\boldsymbol{x}\, d\boldsymbol{y} \qquad \text{(D22)}$$

$$MI(\boldsymbol{x};\boldsymbol{y}) = \sum_{\boldsymbol{y}}\sum_{\boldsymbol{x}} Prob(\boldsymbol{x},\boldsymbol{y}) ln \frac{Prob(\boldsymbol{x},\boldsymbol{y})}{Prob(\boldsymbol{x})Prob(\boldsymbol{y})} \qquad \text{(D23)}$$

$$MI(\boldsymbol{x};\boldsymbol{y}) = KL\big(density(\boldsymbol{x},\boldsymbol{y}) \parallel density(\boldsymbol{x})density(\boldsymbol{y})\big) \qquad \text{(D24)}$$

Mutual information is a measure of the amount of information that $\boldsymbol{x}$ and $\boldsymbol{y}$ provide about each other (Mulder & van der Linden, 2010). Mutual information is equal to zero when $\boldsymbol{x}$ and $\boldsymbol{y}$ are not related, i.e., when $density(\boldsymbol{x},\boldsymbol{y}) = density(\boldsymbol{x})density(\boldsymbol{y})$. The closer $\boldsymbol{x}$ and $\boldsymbol{y}$ are related to each other, the larger their mutual information. Weissman (2007) observed that the mutual information between the current posterior distribution of trait estimates (i.e., $density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1})$) and the response distribution of a possible new question (i.e., $Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)$) indicates the match between the question's operational range in the trait space and the posterior distribution of traits, with larger values indicating better match. Weissman (2007) thus

proposed the MUI item selector, which attempts to maximise this mutual information measure. Equations D25 and D26 show how this method can be applied to FC assessments.

$$MI_{\{i,k\}} \equiv MI\left((\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}); \left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)\right)$$

$$= \sum_{Y_{\{i,k\}}} \int_{\boldsymbol{\eta}} density\left(\boldsymbol{\eta}, Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right) \times ln \frac{density\left(\boldsymbol{\eta}, Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)}{density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}) Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)} d\boldsymbol{\eta}$$

$$= \sum_{Y_{\{i,k\}}} \int_{\boldsymbol{\eta}} Prob\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right) density(\boldsymbol{\eta}|\boldsymbol{Y}^{r-1}) \times ln \frac{Prob\left(Y_{\{i,k\}}|\boldsymbol{\eta}\right)}{Prob\left(Y_{\{i,k\}}|\boldsymbol{Y}^{r-1}\right)} d\boldsymbol{\eta} \qquad \text{(D25)}$$

$$\{i_r, k_r\} = arg\ max_{\{i,k\} \in R_r} \{MI_{\{i,k\}}\} \qquad \text{(D26)}$$

Mulder and van der Linden (2010) demonstrated algebraically that MUI and KLP are almost the same except that KLP considers the KL distance between the **current** and **new** posteriors, whereas MUI considers the KL distance between the **new** and **current** posteriors. In other words, they differ only because the KL distance measure is not symmetrical. And because of this interpretation within the framework of KL information measures, the MUI item selector is also known as the KL information with Bayesian update method, or KLB (Wang & Chang, 2010, 2011). Although MUI and KLP have similar computational complexities, Mulder and van der Linden (2010) predicted that MUI would be more robust to interim trait estimation errors than KLP.

Continuous Entropy Method (CEM)

Wang and Chang (2010, 2011) suggested making use of some other special measures within the framework of KL information, namely Shannon entropy measures, for adaptive item selection. Shannon entropy of a random variable $\boldsymbol{y}$ is given by

Equation D27, while the conditional entropy[10] of a random variable $\boldsymbol{y}$ given $\boldsymbol{x}$ is described by Equation D28 (Cover & Thomas, 2006; Shannon, 1948). It can be deduced algebraically that Shannon entropy $H(\boldsymbol{y})$ differs from the negative KL distance between the distribution of $\boldsymbol{y}$ and the uniform distribution by a mere constant (Equation D29), and that Shannon entropy measures are related to the mutual information measure (Equation D30; Cover & Thomas, 2006).

$$H(\boldsymbol{y}) = -\int density(\boldsymbol{y}) ln\big(density(\boldsymbol{y})\big)\, d\boldsymbol{y} \tag{D27}$$

$$H(\boldsymbol{y}|\boldsymbol{x}) = \int_{\boldsymbol{x}} \int_{\boldsymbol{y}} density(\boldsymbol{x},\boldsymbol{y}) ln \frac{density(\boldsymbol{x})}{density(\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{y}\, d\boldsymbol{x} \tag{D28}$$

$$H(\boldsymbol{y}) = -KL\big(density(\boldsymbol{y}) \parallel density(uniform)\big) - ln(density(uniform)) \tag{D29}$$

$$H(\boldsymbol{y}) - H(\boldsymbol{y}|\boldsymbol{x}) = MI(\boldsymbol{x};\boldsymbol{y}) \tag{D30}$$

Shannon entropy "measures the uncertainty inherent in the distribution of a random variable", with smaller values indicating more concentrated distributions and larger values indicating more uniform distributions (Wang & Chang, 2011). As the goal of CAT is to make the posterior distribution of trait estimates as precise and concentrated as possible, Wang and Chang (2010, 2011) proposed the CEM item selector, which attempts to minimise the expected entropy (following the administration of a new question) of the posterior distribution of trait estimates. Equation D31 shows how this method can be applied to FC assessments.

[10] Also referred to as "continuous entropy" or "differential entropy".

$$\{i_r, k_r\} = arg\ min_{\{i,k\} \in R_r} \left\{ E_{Y_{\{i,k\}}} \left[H\left(\boldsymbol{\eta} | \boldsymbol{Y}^{r-1}, Y_{\{i,k\}}\right) \right] \right\}$$

$$= \sum_{Y_{\{i,k\}}} Prob\left(Y_{\{i,k\}} | \boldsymbol{Y}^{r-1}\right) \times H\left(\boldsymbol{\eta} | \boldsymbol{Y}^{r-1}, Y_{\{i,k\}}\right) \quad \text{(D31)}$$

Wang and Chang (2011) observed that CEM and MUI are closely related and only differ in terms of the baseline used: CEM uses the uniform distribution as the baseline, whereas MUI uses the current posterior distribution as the baseline. With a more realistic baseline, Wang and Chang (2011) expected MUI to be more robust than CEM, and subsequently confirmed their hypothesis via simulations using two different item bank conditions.

Criteria Modifications and Extensions

So far, this appendix described the most basic formulations of the item selectors. This section develops them further through the application of three common modifications and extensions.

Incorporating Prior Information

Often, some information about the respondents is already available prior to the assessments taking place. For example, the distribution of trait scores in the respondent population is often quantified as part of the assessment development process. Moreover, information about a specific respondent may be available from past assessments and/or other data sources. Such prior information can be incorporated into most item selectors. For example, Segall (1996) showed how a multivariate normal prior distribution (characterised by a variance-covariance matrix $\boldsymbol{cov}$) of latent traits can be incorporated into D-optimality, leading to the Bayesian extension of D-optimality that minimises the volume of the Bayesian credibility ellipsoid of the trait estimates (Equation D32).

$$\{i_r, k_r\} = arg\ max_{\{i,k\}\in R_r}\left\{det\left(\boldsymbol{F}^{r-1}(\hat{\boldsymbol{\eta}}^{r-1}) + \boldsymbol{F}_{\{i,k\}}(\hat{\boldsymbol{\eta}}^{r-1}) + \boldsymbol{cov}^{-1}\right)\right\} \quad \text{(D32)}$$

For item selectors based on information maximisation or the FIM, prior information can be added to total information or the FIM alongside the information contributions from assessment responses. For item selectors based on KL information, the prior distribution of trait values can be used as the initial baseline for posterior updates. Table D1 describes how this Bayesian extension can be applied to the item selectors described in this appendix.

Table D1. Bayesian extensions of item selectors

Item selector	Modification and effect
WI, WCI	Prior information only adds a constant to the total information and has no real effect on the maximisation process. Bayesian extension is thus redundant.
DMI	Prior information can potentially change the direction with minimum information, but the prior information term drops out in the calculations for the item selection criterion.
A-, C-, D-, E-optimality	The prior information matrix is added to the total FIM before computing the FIM's inverse/eigenvalues.
T-optimality	Prior information only adds a constant matrix to the FIM and has no real effect on the maximisation of the trace. Bayesian extension is thus redundant.
KLI-U	KLI-U only concerns the KL information of future questions and ignores any prior information. The Bayesian extension is irrelevant for this item selector.
KLI-B, KLP, MUI/KLB, CEM	Prior information is used as the initial baseline for updating the posterior density of trait estimates.

Prior information can also be incorporated into the trait estimator of a CAT, leading to different interim trait estimates and thus indirectly affecting item selection decisions even if Bayesian extension isn't incorporated into the item selector. Note that the inclusion of a prior in trait estimation is an independent decision from the inclusion of prior information in item selection – it is technically possible to add a prior to both, neither, or either one but not the other.

Incorporating Likelihood or Posterior Weighting

While the adoption of a global information measure is one way to address the attenuation paradox caused by interim point estimates (Lord & Novick, 1968), another way around this problem is likelihood (if using a frequentist approach) or posterior (if using a Bayesian approach) weighting. More specifically, the item selection criterion is weighted by the likelihood or posterior of the trait distribution, and then integrated over the trait space. Veerkamp and Berger (1997) and van der Linden and Pashley (2010) respectively presented likelihood-weighted and posterior-weighted maximum information item selection criteria for unidimensional CAT. Their methodologies may be extended to multidimensional CAT in a similar way. For example, the frequentist and Bayesian versions of D-optimality may be likelihood-weighted and posterior-weighted as shown in Equations D33 and D34 respectively.

$$\{i_r, k_r\} = arg\ max_{\{i,k\}\in R_r} \left\{ \int \left[det\left(\boldsymbol{F}^{r-1}(\boldsymbol{\eta}) + \boldsymbol{F}_{\{i,k\}}(\boldsymbol{\eta}) \right) \times L(\boldsymbol{Y}^{r-1}|\boldsymbol{\eta}) \right] d\boldsymbol{\eta} \right\} \tag{D33}$$

$$\{i_r, k_r\} = arg\ max_{\{i,k\}\in R_r} \left\{ \int \left[det\left(\boldsymbol{F}^{r-1}(\boldsymbol{\eta}) + \boldsymbol{F}_{\{i,k\}}(\boldsymbol{\eta}) + \boldsymbol{cov^{-1}} \right) \times L(\boldsymbol{Y}^{r-1}|\boldsymbol{\eta}) \times prior(\boldsymbol{\eta}) \right] d\boldsymbol{\eta} \right\} \tag{D34}$$

Following the same principle, likelihood or posterior weighting can be applied to item selectors that rely on local information measures. However, likelihood or posterior weighting would be unnecessary for item selectors that rely on global information measures. Table D2 describes how likelihood or posterior weighting can be applied to the item selectors described in this appendix.

Table D2. Likelihood or posterior weighting of item selectors

Item selector	Modification and effect
WI, WCI, DMI, A-, C-, D-, E-, T-optimality	Multiply the item selection criterion by the likelihood or posterior of the trait distribution, then integrating over the trait space before taking the maximum or minimum.
KLI-U	Likelihood-weighted KLI-U is equivalent to KLI-B without prior information, and posterior-weighted KLI-U is equivalent to KLI-B with prior information incorporated.
KLI-B, KLP, MUI/KLB, CEM	Global information measures are utilised by design, which incorporate weighting by the current trait density already. So additional likelihood or posterior weighting is unnecessary.

Incorporating Item Bank Stratification

As described by Davey and Nering (2002), items with high discriminations are intense "spotlights" that focus on measuring a small region in the trait space, whereas items with low discriminations are less-intense "floodlights" that give less targeted information but over a larger region in the trait space. Information-based item selection criteria have a tendency to favour items with larger discrimination parameters (e.g., Mulder & van der Linden, 2009), thus using up the "spotlight" items too early and leaving only "floodlight" items that provide limited local information towards the end

of a CAT. This tendency also results in an overexposure of "spotlight" items across many respondents while the "floodlight" items remain under-used.

Observing this issue, Chang and Ying (1999) proposed forcing the usage of "floodlight" items early on to roughly locate the respondent, and then utilise the "spotlight" items to get precise measurement at targeted locations when the interim trait estimates become more accurate. This is achieved by stratifying the item bank by item discrimination parameters and blocking subsets of items from use according to the current measurement stage and status.

Item bank stratification does not change the functional form of the item selection criteria and can be applied to all of the item selectors described in this appendix. For FC adaptive personality assessments with each item indicating one and only one trait, each item has only one non-zero discrimination parameter, thus allowing the direct application of existing item bank stratification methods originally designed for unidimensional CAT (Chang & Ying, 1999; Chang, Qian, & Ying, 2001). Moreover, as item bank stratification reduces the number of available items to search through at each item selection step, the computational intensity of item selection will be reduced.

APPENDIX E: STUDY 2 SIMULATED ITEM BANKS

Table E1. *Simulated item bank – 100% positive*

Item	Scale	Mean utility	Loading	Unique variance
I1	S1	−0.70717	1.044764	1.709273
I2	S1	1.776714	0.521984	1.159328
I3	S1	−2.97055	0.597848	1.918757
I4	S1	−0.78604	0.837094	1.465483
I5	S1	1.763968	0.943131	1.775363
I6	S1	−0.37533	1.03983	0.746428
I7	S1	0.055738	1.347173	1.261252
I8	S1	−1.4333	0.530354	1.711655
I9	S1	2.109475	0.814683	1.374312
I10	S1	2.222574	1.381934	1.589197
I11	S1	2.408814	1.251946	1.111888
I12	S1	2.858291	0.635261	0.723369
I13	S1	0.877976	0.998649	1.681363
I14	S1	−2.52309	1.135166	1.158618
I15	S1	1.738301	1.056274	1.83603
I16	S1	−2.23214	1.256136	0.858689
I17	S1	−1.49682	0.677328	0.642739
I18	S1	1.338091	1.334109	0.980379
I19	S1	2.542173	0.509001	1.722056
I20	S1	0.254422	1.456636	0.641377
I21	S1	−1.52033	0.626646	0.621635
I22	S1	1.286766	0.957425	1.307904
I23	S1	−1.46239	1.291039	1.301192
I24	S1	0.487458	0.825032	1.498967
I25	S1	−0.47645	0.994172	1.485816
I26	S1	−0.30307	0.762203	1.740826
I27	S1	−2.29736	0.904577	1.586327
I28	S1	2.566231	1.268938	1.998251
I29	S1	−0.76049	0.535093	1.0373
I30	S1	−2.1732	1.101808	1.734048

Item	Scale	Mean utility	Loading	Unique variance
I31	S1	−2.70143	1.470858	0.605768
I32	S1	2.44367	1.130876	1.657164
I33	S1	−0.01648	0.980106	1.180476
I34	S1	−1.04665	1.487247	1.166657
I35	S1	−0.20002	1.396372	0.974691
I36	S1	2.373314	0.578374	1.49604
I37	S1	−1.9373	1.307447	0.668385
I38	S1	−1.01221	1.30304	1.333969
I39	S1	1.006367	1.127418	0.852887
I40	S1	−2.95298	1.294949	1.926699
I41	S1	1.35794	0.606184	1.292774
I42	S1	−2.6723	1.455788	1.536501
I43	S1	1.541513	0.667618	1.153456
I44	S1	1.774449	0.506914	1.583871
I45	S1	1.022068	0.63724	0.92261
I46	S1	−1.28199	0.748686	0.533322
I47	S1	−2.00839	1.256875	0.753026
I48	S1	1.395604	1.034006	1.312848
I49	S1	−1.7708	0.68934	0.679371
I50	S1	−2.75252	1.186829	1.231235
I51	S1	1.940873	1.162276	0.711712
I52	S1	−1.12319	1.215094	1.955734
I53	S1	0.887513	0.541971	1.558018
I54	S1	2.695767	1.244847	1.225814
I55	S1	−0.84728	1.476441	1.543151
I56	S1	2.724339	1.052975	0.683382
I57	S1	−1.43594	0.918003	1.451041
I58	S1	2.166961	0.764956	1.121029
I59	S1	1.785916	1.425376	1.432784
I60	S1	−0.30565	1.164398	1.375679
I61	S2	−1.00086	1.314333	0.656758
I62	S2	−0.23282	1.098962	0.539763
I63	S2	−0.70935	1.112086	0.858855

Item	Scale	Mean utility	Loading	Unique variance
I64	S2	2.251604	1.231069	1.945818
I65	S2	0.80617	0.609337	1.817204
I66	S2	2.962207	0.655697	1.676377
I67	S2	0.93262	1.307993	1.695263
I68	S2	1.077605	1.20564	1.797696
I69	S2	2.63775	1.491399	0.741045
I70	S2	−1.44849	1.071221	0.640828
I71	S2	2.497214	1.052483	1.055205
I72	S2	−1.36475	0.654232	0.669546
I73	S2	2.389784	1.114059	1.416965
I74	S2	2.106327	1.412757	1.262848
I75	S2	0.199509	0.777704	1.498026
I76	S2	0.695199	1.471867	1.864336
I77	S2	−2.11685	0.786797	1.426781
I78	S2	−1.48568	1.244354	1.55485
I79	S2	−0.05293	1.354789	1.013078
I80	S2	−2.36611	0.780531	1.641007
I81	S2	−2.81852	1.477087	1.580083
I82	S2	0.768142	1.0957	0.7764
I83	S2	−1.95379	1.268915	0.915201
I84	S2	1.201951	1.244745	0.630123
I85	S2	−1.13315	0.536401	1.1212
I86	S2	0.156165	0.723865	1.110339
I87	S2	−0.57256	1.233276	1.325592
I88	S2	−1.17561	0.965104	0.60785
I89	S2	0.597976	0.88356	1.684607
I90	S2	−1.95029	1.162302	1.367657
I91	S2	−0.82385	1.236888	1.421738
I92	S2	1.793103	0.929132	1.52211
I93	S2	−2.46621	0.80587	1.210501
I94	S2	−2.20879	1.394776	1.91838
I95	S2	0.110063	0.826422	1.347841
I96	S2	2.805448	1.212081	1.54258

Item	Scale	Mean utility	Loading	Unique variance
I97	S2	−2.04789	0.688374	0.931942
I98	S2	−1.34274	1.164239	1.176277
I99	S2	−1.83278	1.290975	1.982132
I100	S2	−1.83442	0.989969	1.181699
I101	S2	−0.84331	0.750039	1.361267
I102	S2	−1.96162	1.479281	1.780192
I103	S2	2.088592	1.160997	1.384894
I104	S2	2.292348	1.460164	1.510745
I105	S2	2.45239	1.363981	1.630625
I106	S2	−1.61573	0.800252	1.17007
I107	S2	−0.6609	1.140738	1.793021
I108	S2	−2.92597	1.181677	0.956483
I109	S2	−0.93999	1.424051	1.818617
I110	S2	−2.88119	0.808724	0.605688
I111	S2	2.332869	0.993394	1.130501
I112	S2	1.288734	0.998845	1.32662
I113	S2	−1.23537	0.55261	1.346665
I114	S2	1.332569	1.066582	1.567246
I115	S2	1.998022	1.210246	1.594587
I116	S2	−2.93306	0.783299	1.784913
I117	S2	2.203721	0.868696	1.987004
I118	S2	−2.89067	1.078161	1.603016
I119	S2	−1.46551	1.253589	1.015734
I120	S2	2.050612	0.788272	1.295783
I121	S3	1.905365	1.320729	1.5379
I122	S3	2.874162	1.298887	0.869254
I123	S3	2.102413	1.004574	1.325582
I124	S3	1.113874	1.456304	1.724762
I125	S3	0.571454	0.990924	0.855801
I126	S3	0.761949	0.792522	1.663985
I127	S3	2.840616	1.088611	0.529747
I128	S3	2.063539	1.389637	1.85313
I129	S3	−0.7873	0.975724	0.602414

Item	Scale	Mean utility	Loading	Unique variance
I130	S3	−2.86111	0.946514	0.971073
I131	S3	1.010706	1.449034	0.631331
I132	S3	−2.50178	1.176475	1.356271
I133	S3	1.780011	1.274073	1.097156
I134	S3	1.431167	1.115502	0.536032
I135	S3	−0.48227	0.652287	1.817361
I136	S3	2.906664	1.451132	0.959309
I137	S3	0.774475	0.844575	1.514362
I138	S3	−2.40328	0.904077	1.541425
I139	S3	2.3731	0.631884	1.655328
I140	S3	−2.09769	1.399702	0.959637
I141	S3	0.989295	0.757439	1.666095
I142	S3	−2.68112	1.293044	1.715216
I143	S3	−0.65609	0.741059	0.800702
I144	S3	−1.15864	1.260875	0.501656
I145	S3	1.325911	1.106558	0.82537
I146	S3	−2.46422	1.148801	0.611531
I147	S3	−1.77634	0.653253	1.396422
I148	S3	−0.48226	0.753627	1.386698
I149	S3	−1.88242	1.414389	1.133791
I150	S3	−2.63865	0.80737	0.95742
I151	S3	0.858229	0.59095	1.841757
I152	S3	−1.38814	1.412835	0.959257
I153	S3	−1.26254	0.821435	1.866857
I154	S3	2.573805	0.965535	0.537992
I155	S3	2.822335	0.599631	1.504073
I156	S3	1.030967	0.509947	0.924916
I157	S3	0.633957	1.0376	1.099161
I158	S3	2.188902	1.11063	0.646553
I159	S3	−1.38696	1.031026	1.959634
I160	S3	2.95891	0.680724	1.676718
I161	S3	2.746844	1.379173	0.768446
I162	S3	−1.02651	0.739163	1.061255

Item	Scale	Mean utility	Loading	Unique variance
I163	S3	−2.16705	1.005682	1.894634
I164	S3	−0.00974	0.539527	1.145638
I165	S3	−2.50124	1.062885	0.875363
I166	S3	2.541305	1.499678	1.393498
I167	S3	−2.05582	0.783394	0.968895
I168	S3	−2.15554	0.996658	1.674064
I169	S3	0.779925	1.02982	1.646921
I170	S3	2.021863	1.181353	1.701442
I171	S3	−1.6939	1.226192	1.580581
I172	S3	−0.79724	0.690048	1.28086
I173	S3	−0.66997	1.46952	1.732528
I174	S3	0.927324	0.920868	1.596561
I175	S3	−2.78193	1.392045	1.43742
I176	S3	1.343639	0.747108	1.046478
I177	S3	−2.73642	0.700285	0.555626
I178	S3	1.411008	0.554969	1.740477
I179	S3	−1.22854	1.323831	1.11128
I180	S3	0.066916	1.170432	1.651138
I181	S4	2.701035	1.29944	1.835487
I182	S4	1.629159	1.082557	1.768946
I183	S4	2.982132	0.524793	0.584199
I184	S4	0.34643	1.099608	1.388931
I185	S4	2.199607	0.654908	0.577527
I186	S4	−2.16101	1.127358	1.839987
I187	S4	−1.28504	1.179491	1.598822
I188	S4	−1.42704	1.289826	1.042475
I189	S4	0.377099	0.530906	0.651169
I190	S4	−1.33169	0.698314	1.215724
I191	S4	−2.95414	0.94246	0.523479
I192	S4	1.071671	0.884617	0.725637
I193	S4	−0.85087	1.013881	1.929788
I194	S4	0.083232	1.220094	1.662475
I195	S4	−1.08012	1.088174	0.843882

Item	Scale	Mean utility	Loading	Unique variance
I196	S4	−2.07003	1.054425	1.53246
I197	S4	1.362294	0.90027	0.664637
I198	S4	−1.62554	0.67286	1.406437
I199	S4	2.665035	0.621725	1.535995
I200	S4	0.360646	1.396951	1.423617
I201	S4	0.894452	1.133047	0.565219
I202	S4	−2.17174	1.025278	1.864178
I203	S4	−1.83806	1.155104	1.649784
I204	S4	−2.46192	0.562721	0.661366
I205	S4	−1.44658	0.717503	1.239353
I206	S4	−1.82096	1.135589	1.9552
I207	S4	−0.44104	0.800768	1.325154
I208	S4	2.554475	0.515687	1.282952
I209	S4	−0.75691	1.178059	1.857148
I210	S4	0.669504	1.159846	1.673772
I211	S4	0.382614	1.052489	1.948729
I212	S4	1.245584	1.361889	1.182306
I213	S4	2.529668	0.78575	1.829132
I214	S4	−2.14579	1.147671	0.629497
I215	S4	0.324265	0.626727	1.681962
I216	S4	−1.6184	0.690967	1.20852
I217	S4	−1.51799	0.555603	1.70861
I218	S4	−0.23399	0.814726	1.354336
I219	S4	−1.04584	0.578794	0.895708
I220	S4	−2.44882	0.795116	1.26398
I221	S4	−1.73837	1.043629	1.135328
I222	S4	−1.25779	1.216859	1.16813
I223	S4	1.615201	1.05802	1.148718
I224	S4	0.584745	1.222584	0.944466
I225	S4	−2.0807	1.234347	1.420496
I226	S4	2.959551	0.756777	1.48828
I227	S4	1.411305	0.61726	1.015189
I228	S4	1.978808	0.681078	1.179577

Item	Scale	Mean utility	Loading	Unique variance
I229	S4	0.873955	0.673808	0.839654
I230	S4	2.349058	0.739181	1.431902
I231	S4	−1.52837	0.808565	0.656232
I232	S4	2.94154	0.633979	1.429061
I233	S4	−1.3363	1.20652	0.644636
I234	S4	−1.90734	0.886839	0.803271
I235	S4	−0.90322	0.668134	1.863185
I236	S4	1.062545	0.707674	1.871761
I237	S4	0.064844	0.876164	0.776131
I238	S4	−2.72946	0.657057	1.101567
I239	S4	−0.69719	0.55783	0.689128
I240	S4	−1.30336	1.088996	0.949722

Table E2. *Simulated item bank – 75% positive*

Item	Scale	Mean utility	Loading	Unique variance
I1	S1	−1.52727	1.039403	1.158735
I2	S1	1.331374	−0.91953	0.733098
I3	S1	−0.58906	0.515618	1.731037
I4	S1	−0.72761	−1.24808	1.643265
I5	S1	−1.64405	1.482661	1.62885
I6	S1	−2.62161	−1.41998	1.591264
I7	S1	−2.8082	1.256383	1.139884
I8	S1	−1.80533	1.280572	0.964826
I9	S1	2.758434	1.238582	1.616543
I10	S1	−1.24985	1.457736	1.903991
I11	S1	1.646294	1.249631	0.944753
I12	S1	−1.52598	1.305556	0.574035
I13	S1	2.6042	0.817222	0.576856
I14	S1	−0.58646	1.074111	1.707827
I15	S1	−1.01932	1.482211	1.869893
I16	S1	1.710495	−1.3259	0.853066
I17	S1	0.964381	−0.6784	0.613116
I18	S1	2.825593	1.359718	0.684994
I19	S1	1.011978	0.860682	0.64319
I20	S1	−0.2748	1.363666	1.416607
I21	S1	0.215266	0.84509	1.709916
I22	S1	2.275884	1.039317	1.954364
I23	S1	−0.49419	−0.65985	1.162402
I24	S1	1.004494	−1.27036	1.156624
I25	S1	1.941083	0.60721	1.820822
I26	S1	2.810882	0.802103	1.792857
I27	S1	2.628163	1.135359	1.079096
I28	S1	−2.59655	1.404104	1.912003
I29	S1	−1.0306	−0.9757	1.576904
I30	S1	−1.73559	0.93288	1.326837
I31	S1	−0.00019	1.038492	1.315302
I32	S1	0.108303	0.951533	0.670451

Item	Scale	Mean utility	Loading	Unique variance
I33	S1	0.350351	1.363981	1.758647
I34	S1	−1.39254	−0.91415	0.622721
I35	S1	−2.51538	1.094876	0.856303
I36	S1	−2.05713	0.507796	0.72451
I37	S1	1.721832	1.455338	0.79499
I38	S1	1.750965	1.024268	1.051979
I39	S1	−0.41764	1.182759	0.730891
I40	S1	−2.11907	1.175938	1.92906
I41	S1	2.698202	0.958158	0.554602
I42	S1	−1.51648	−0.651	1.680723
I43	S1	−0.06162	−1.26978	1.934597
I44	S1	0.287151	0.716776	1.759265
I45	S1	−2.89733	0.834485	1.896444
I46	S1	2.119246	0.743339	1.472789
I47	S1	0.146531	0.747357	1.789578
I48	S1	0.502056	1.375629	1.097286
I49	S1	−2.24892	−1.42912	1.358657
I50	S1	0.640592	−1.38087	1.269794
I51	S1	−2.55463	0.912866	0.955385
I52	S1	1.369237	0.665562	1.804924
I53	S1	−0.29076	1.21312	1.851356
I54	S1	−1.36037	0.838684	1.025571
I55	S1	−0.45134	−1.45697	0.726835
I56	S1	0.182962	0.898078	1.415622
I57	S1	−1.56885	−1.00966	1.322713
I58	S1	1.544609	0.659854	0.703412
I59	S1	2.638707	−1.12778	0.558842
I60	S1	1.75356	0.671386	1.82035
I61	S2	2.270468	−1.13654	0.766203
I62	S2	−2.68809	1.472265	1.007288
I63	S2	0.21971	0.829099	1.817985
I64	S2	0.71706	−0.51349	0.880107
I65	S2	0.686074	0.53948	0.80413

Item	Scale	Mean utility	Loading	Unique variance
I66	S2	−0.42309	−1.24902	0.585186
I67	S2	2.549205	1.402838	1.997247
I68	S2	2.348675	1.199985	1.184162
I69	S2	1.021643	1.109731	1.68759
I70	S2	−1.54393	0.833863	1.046161
I71	S2	1.918327	0.515441	1.120971
I72	S2	−0.90593	0.754402	1.394209
I73	S2	−0.24919	0.599489	1.969739
I74	S2	0.331937	0.886202	1.156745
I75	S2	2.84811	0.536462	1.895636
I76	S2	−0.36394	0.89772	1.303994
I77	S2	2.088846	1.356582	1.680671
I78	S2	1.383212	0.904306	1.580151
I79	S2	1.955943	0.76776	1.973709
I80	S2	−2.38697	0.818926	0.898461
I81	S2	−0.85129	1.013973	0.550442
I82	S2	−2.53092	0.985726	1.433393
I83	S2	−2.12505	0.873522	0.518149
I84	S2	1.22488	−0.57855	1.974627
I85	S2	2.486688	0.9343	1.709765
I86	S2	−0.46111	−1.32246	0.536707
I87	S2	−1.15122	0.888431	1.528376
I88	S2	−0.4695	0.99081	1.876348
I89	S2	−1.22126	−1.47161	1.815302
I90	S2	−2.68641	1.48779	1.065039
I91	S2	−2.82429	1.263535	0.540071
I92	S2	−2.35335	1.308078	0.996951
I93	S2	1.512981	0.610702	1.710313
I94	S2	1.011369	0.940638	0.739884
I95	S2	−1.54662	1.454184	1.712722
I96	S2	−1.78402	−0.66835	1.695852
I97	S2	−2.48505	1.139585	0.863075
I98	S2	−2.18959	1.351204	1.150824

Item	Scale	Mean utility	Loading	Unique variance
I99	S2	−0.61388	−0.53336	1.833551
I100	S2	0.619604	0.785451	0.958615
I101	S2	−0.59885	1.484778	0.817624
I102	S2	1.871789	1.232816	1.457406
I103	S2	−2.663	0.545272	0.965919
I104	S2	1.571305	−1.20088	0.680592
I105	S2	−0.84133	0.963063	1.325005
I106	S2	−2.19197	−1.08178	0.598026
I107	S2	−2.79643	0.604584	1.637894
I108	S2	2.470946	1.105365	0.819337
I109	S2	−1.3028	1.139573	0.877839
I110	S2	1.76606	1.17867	1.653632
I111	S2	2.997332	0.628633	1.548733
I112	S2	0.823125	0.911534	0.821786
I113	S2	−0.99374	1.074148	1.949712
I114	S2	−2.97283	1.380048	1.326228
I115	S2	−2.50635	0.67551	0.858234
I116	S2	−0.28211	−1.19068	0.975327
I117	S2	−2.41377	1.113249	0.762466
I118	S2	2.141457	0.649766	0.937073
I119	S2	0.276858	0.820212	0.568652
I120	S2	0.779602	−0.6836	1.344755
I121	S3	2.535792	1.453253	0.535136
I122	S3	2.490829	0.944041	1.80183
I123	S3	1.980719	1.354037	1.663047
I124	S3	−1.26817	0.693196	1.506293
I125	S3	0.266677	0.587056	0.599414
I126	S3	−1.47982	0.868405	1.405504
I127	S3	2.242694	0.698696	0.708201
I128	S3	−2.48644	−1.21475	1.881475
I129	S3	2.748967	−0.96499	1.137041
I130	S3	−2.76339	1.397327	1.449094
I131	S3	−1.65051	−1.48462	1.377449

Item	Scale	Mean utility	Loading	Unique variance
I132	S3	0.845993	1.313858	1.225321
I133	S3	0.699062	0.793094	0.752191
I134	S3	−1.81036	1.302708	1.890411
I135	S3	−2.28634	1.166052	0.800122
I136	S3	−0.10869	−0.82389	1.408314
I137	S3	−2.60994	0.639796	1.794373
I138	S3	−0.25695	−1.28879	1.080076
I139	S3	−2.00339	1.087328	1.161993
I140	S3	−2.31721	−0.91783	1.436085
I141	S3	−2.23688	1.074456	0.946314
I142	S3	−0.69464	0.503893	1.930048
I143	S3	−2.71091	1.44118	0.735169
I144	S3	−1.84128	0.695828	1.170296
I145	S3	−1.02725	0.96807	1.118704
I146	S3	1.172695	−0.73896	1.824352
I147	S3	2.543087	0.921182	1.169723
I148	S3	−2.14213	1.029017	0.738513
I149	S3	−2.88165	0.917719	0.572146
I150	S3	−0.30827	1.476473	0.669292
I151	S3	−2.82943	−0.77353	1.329324
I152	S3	1.892158	1.14584	1.024494
I153	S3	−1.98173	1.450897	0.701069
I154	S3	0.810805	−0.69769	1.66749
I155	S3	−1.31419	−0.84914	1.050131
I156	S3	−0.10915	0.611127	1.347795
I157	S3	−2.32751	0.973604	1.505251
I158	S3	1.11898	0.907873	1.132963
I159	S3	1.836153	−1.25121	1.496931
I160	S3	2.11375	0.579597	0.914551
I161	S3	2.38556	0.610642	1.624932
I162	S3	−0.08318	1.267537	1.51979
I163	S3	−0.22453	1.189735	1.662027
I164	S3	−2.27341	1.439811	1.292087

Item	Scale	Mean utility	Loading	Unique variance
I165	S3	−1.90162	0.902644	1.090572
I166	S3	0.944469	1.25216	0.991138
I167	S3	−1.7722	0.532685	0.515995
I168	S3	1.785324	1.219448	1.387513
I169	S3	1.13149	0.667863	1.508287
I170	S3	−2.33367	0.767753	0.888653
I171	S3	−2.14927	−1.28336	0.623142
I172	S3	−1.55096	1.021074	1.174626
I173	S3	−2.25179	1.370759	1.510455
I174	S3	1.833031	0.531363	0.859513
I175	S3	−1.99124	−0.57301	1.883557
I176	S3	−0.65385	0.787546	1.426762
I177	S3	−0.39236	−0.92806	1.889425
I178	S3	−1.60506	1.472942	1.265203
I179	S3	−1.08686	0.656271	1.92996
I180	S3	−1.99206	−1.13726	1.817828
I181	S4	1.275777	0.791714	0.806782
I182	S4	0.217087	1.216128	0.926474
I183	S4	0.686916	1.041199	1.797586
I184	S4	−2.03226	1.150862	1.752832
I185	S4	−0.76593	0.899802	1.1507
I186	S4	−2.61463	−0.51316	1.279348
I187	S4	−2.29918	0.96255	0.576761
I188	S4	1.084747	−1.19687	0.964577
I189	S4	1.627327	−0.5366	0.802577
I190	S4	−2.63024	0.597939	1.363758
I191	S4	2.858081	−0.99071	1.050736
I192	S4	−0.14245	1.244697	1.79236
I193	S4	2.022958	1.175586	0.627854
I194	S4	2.874978	−1.10837	0.745153
I195	S4	−0.88976	−0.73809	1.661603
I196	S4	−0.31446	0.849581	1.533003
I197	S4	−2.99484	−1.1282	1.570524

Item	Scale	Mean utility	Loading	Unique variance
I198	S4	0.548677	1.404308	0.922622
I199	S4	1.376283	−0.93021	0.577918
I200	S4	−2.10283	−1.24263	1.033832
I201	S4	0.24247	−0.77914	0.823449
I202	S4	−1.45395	1.479635	1.255664
I203	S4	−2.35301	0.524987	0.806084
I204	S4	2.191876	0.711672	1.767907
I205	S4	−2.90609	−0.97831	1.325674
I206	S4	1.321593	1.099337	0.579617
I207	S4	−1.02297	−0.65417	1.708322
I208	S4	0.830063	1.154141	1.415372
I209	S4	0.457997	0.539002	1.604893
I210	S4	−2.37279	1.276379	1.199899
I211	S4	−2.6664	0.73332	1.219893
I212	S4	2.264197	0.765962	0.742091
I213	S4	0.459508	1.406902	1.006315
I214	S4	−0.67339	0.750876	0.922533
I215	S4	−2.86929	−0.78608	1.932197
I216	S4	−2.76023	1.34352	1.708117
I217	S4	2.652934	1.120324	1.995337
I218	S4	−0.18875	0.936407	1.987321
I219	S4	1.826791	1.095951	1.179323
I220	S4	−0.06983	0.62004	1.958971
I221	S4	−2.79142	1.273217	1.974166
I222	S4	−1.5147	−0.85242	1.024309
I223	S4	0.084557	−0.81269	1.111171
I224	S4	1.14671	−0.54603	1.534478
I225	S4	0.025774	−0.74487	1.211677
I226	S4	2.108342	1.387284	1.511095
I227	S4	−0.93176	1.119688	0.946089
I228	S4	−0.95914	1.236624	1.44023
I229	S4	1.636763	1.253559	1.480816
I230	S4	−1.66918	0.826947	1.078861

Item	Scale	Mean utility	Loading	Unique variance
I231	S4	−1.48774	0.98209	0.77515
I232	S4	1.281074	−1.48737	1.635362
I233	S4	−0.84713	1.148947	1.435073
I234	S4	1.849931	−1.10379	0.600523
I235	S4	1.201282	−1.05583	0.870676
I236	S4	1.22601	1.088623	1.366947
I237	S4	0.23315	−0.70052	1.565296
I238	S4	−2.80445	−1.24715	1.967237
I239	S4	−2.2235	0.714197	1.491673
I240	S4	−2.78304	1.027425	0.586466

APPENDIX F: STUDY 4 ANALYSIS RESULTS

Table F1. *EFA pattern matrix loadings of HEXACO-PI-R items*

Item	Mapped Scale	H*	E	X	A	C*	O
I1	O	−0.110	−0.031	0.028	0.036	0.024	−0.644
I2	C	0.005	0.020	0.073	0.089	0.601	−0.124
I3	A	0.124	−0.092	0.120	0.462	−0.005	0.035
I4	X	0.189	−0.099	0.421	0.102	0.110	−0.031
I5	E	−0.009	0.392	−0.173	0.048	0.067	−0.141
I6	H	0.355	0.064	−0.127	0.051	0.063	0.046
I7	O	0.005	−0.147	0.090	−0.081	0.016	0.524
I8	C	−0.063	−0.011	0.225	0.013	0.527	0.078
I9	A	−0.336	−0.038	−0.053	−0.400	0.067	0.038
I10	X	−0.060	0.025	−0.573	0.114	−0.073	−0.109
I11	E	−0.161	0.361	−0.405	−0.130	0.175	0.028
I12	H	−0.502	−0.097	−0.183	−0.081	−0.147	−0.017
I13	O	−0.027	0.096	−0.070	0.113	−0.060	0.718
I14	C	−0.058	−0.002	0.077	−0.019	−0.586	−0.132
I15	A	−0.243	0.029	−0.117	−0.400	−0.075	0.038
I16	X	−0.030	0.192	0.520	0.099	−0.092	0.055
I17	E	−0.050	0.578	0.121	−0.039	−0.126	−0.024
I18	H	0.342	0.065	−0.072	0.162	−0.032	0.105
I19	O	0.026	0.023	−0.062	0.101	−0.114	−0.283
I20	C	−0.136	0.164	0.020	−0.093	−0.494	−0.068
I21	A	−0.191	0.189	−0.008	−0.441	−0.166	−0.026
I22	X	0.118	−0.016	0.566	0.184	0.126	−0.007
I23	E	0.001	0.553	−0.004	0.079	−0.052	0.088
I24	H	−0.476	−0.044	0.011	−0.107	−0.011	−0.040
I25	O	0.041	−0.011	−0.044	−0.042	−0.008	0.615
I26	C	−0.212	0.097	−0.166	0.017	−0.480	0.068
I27	A	0.107	0.007	0.168	0.542	−0.036	0.064
I28	X	−0.210	0.046	−0.547	0.031	−0.039	−0.024
I29	E	−0.058	0.354	−0.200	0.073	0.133	−0.164

Item	Mapped Scale	H*	E	X	A	C*	O
I30	H	−0.501	0.087	0.065	−0.039	−0.079	−0.080
I31	O	−0.151	0.086	0.022	0.094	−0.081	−0.487
I32	C	−0.335	−0.019	−0.194	0.130	−0.426	−0.051
I33	A	0.031	0.083	0.000	0.363	−0.103	0.073
I34	X	−0.171	0.063	0.694	0.068	0.040	0.009
I35	E	−0.005	−0.328	0.338	0.206	−0.201	−0.029
I36	H	0.409	−0.004	0.141	0.002	0.123	0.080
I37	O	−0.208	0.035	0.179	0.129	0.168	0.470
I38	C	−0.137	0.102	−0.088	0.176	0.478	0.041
I39	A	−0.050	0.072	−0.027	0.436	0.152	0.002
I40	X	−0.197	0.228	0.496	0.307	0.033	0.058
I41	E	−0.153	−0.499	−0.020	0.182	0.218	0.065
I42	H	−0.447	0.020	0.073	−0.088	−0.025	−0.079
I43	O	−0.055	−0.085	0.070	−0.091	−0.144	0.320
I44	C	−0.211	0.174	−0.129	−0.031	−0.543	−0.031
I45	A	−0.022	−0.211	0.009	0.343	0.063	0.071
I46	X	−0.119	0.037	−0.583	0.016	−0.079	−0.034
I47	E	0.021	0.528	0.037	0.059	0.063	0.059
I48	H	−0.575	−0.003	0.159	−0.148	0.025	−0.027
I49	O	0.106	−0.096	−0.126	−0.085	−0.100	−0.448
I50	C	−0.231	0.101	−0.034	−0.083	0.552	0.050
I51	A	−0.115	0.041	−0.206	0.496	0.001	−0.028
I52	X	−0.202	0.213	−0.555	0.000	−0.104	0.047
I53	E	−0.042	−0.458	0.158	0.096	−0.056	0.178
I54	H	0.409	−0.004	−0.063	0.115	0.005	0.053
I55	O	−0.038	0.084	0.000	0.054	0.027	−0.635
I56	C	−0.104	0.070	−0.055	0.012	−0.500	0.087
I57	A	−0.198	0.095	0.009	−0.375	−0.256	−0.012
I58	X	−0.212	−0.049	0.632	−0.092	0.096	0.043
I59	E	−0.235	−0.596	−0.041	−0.024	0.024	−0.065
I60	H	−0.558	−0.146	−0.143	−0.006	−0.223	−0.064

* Signs of the loadings were reversed to align with the conceptual definitions.

Table F2. *ESEM pattern matrix loadings of HEXACO-PI-R items*

Item	Mapped Scale	H*	E	X	A	C*	O
I1	O	−0.136	−0.035	0.045	0.026	0.033	−0.643
I2	C	−0.027	0.049	0.080	0.090	0.609	−0.120
I3	A	0.087	−0.110	0.117	0.474	−0.026	0.023
I4	X	0.156	−0.125	0.409	0.100	0.119	−0.039
I5	E	0.001	0.415	−0.117	0.045	0.054	−0.134
I6	H	0.355	0.074	−0.118	0.065	0.064	0.034
I7	O	0.020	−0.151	0.053	−0.076	0.016	0.524
I8	C	−0.086	0.006	0.219	0.010	0.535	0.085
I9	A	−0.306	−0.022	−0.069	−0.414	0.084	0.055
I10	X	−0.045	0.056	−0.559	0.128	−0.088	−0.110
I11	E	−0.128	0.409	−0.361	−0.131	0.166	0.044
I12	H	−0.483	−0.091	−0.195	−0.090	−0.155	−0.002
I13	O	−0.001	0.101	−0.078	0.124	−0.085	0.722
I14	C	−0.045	−0.041	0.081	−0.027	−0.592	−0.135
I15	A	−0.205	0.041	−0.123	−0.413	−0.060	0.051
I16	X	−0.045	0.157	0.541	0.084	−0.099	0.059
I17	E	−0.029	0.577	0.191	−0.054	−0.140	−0.011
I18	H	0.337	0.064	−0.062	0.177	−0.040	0.091
I19	O	0.016	0.018	−0.047	0.100	−0.117	−0.287
I20	C	−0.110	0.140	0.040	−0.105	−0.502	−0.063
I21	A	−0.150	0.193	0.007	−0.460	−0.151	−0.011
I22	X	0.078	−0.049	0.564	0.178	0.128	−0.012
I23	E	0.020	0.562	0.064	0.073	−0.075	0.099
I24	H	−0.466	−0.041	0.004	−0.122	−0.012	−0.023
I25	O	0.067	−0.005	−0.066	−0.033	−0.016	0.616
I26	C	−0.183	0.084	−0.154	0.013	−0.500	0.074
I27	A	0.068	−0.015	0.178	0.553	−0.066	0.054
I28	X	−0.186	0.081	−0.536	0.039	−0.054	−0.017
I29	E	−0.052	0.381	−0.148	0.071	0.121	−0.156
I30	H	−0.493	0.085	0.076	−0.057	−0.088	−0.061
I31	O	−0.168	0.079	0.050	0.084	−0.083	−0.484
I32	C	−0.321	−0.031	−0.190	0.127	−0.449	−0.045

Item	Mapped Scale	H*	E	X	A	C*	O
I33	A	0.016	0.073	0.017	0.371	−0.130	0.069
I34	X	−0.199	0.023	0.697	0.046	0.041	0.018
I35	E	−0.037	−0.372	0.300	0.205	−0.202	−0.039
I36	H	0.398	−0.008	0.137	0.011	0.134	0.067
I37	O	−0.209	0.037	0.169	0.129	0.151	0.481
I38	C	−0.153	0.137	−0.073	0.180	0.467	0.050
I39	A	−0.078	0.078	−0.008	0.444	0.125	0.001
I40	X	−0.224	0.201	0.525	0.293	0.011	0.068
I41	E	−0.183	−0.497	−0.080	0.192	0.220	0.060
I42	H	−0.440	0.019	0.075	−0.104	−0.028	−0.061
I43	O	−0.038	−0.096	0.047	−0.091	−0.145	0.322
I44	C	−0.179	0.157	−0.106	−0.039	−0.560	−0.024
I45	A	−0.051	−0.218	−0.012	0.354	0.048	0.063
I46	X	−0.094	0.071	−0.572	0.027	−0.092	−0.031
I47	E	0.034	0.540	0.101	0.052	0.046	0.069
I48	H	−0.567	−0.005	0.154	−0.170	0.025	−0.004
I49	O	0.100	−0.098	−0.124	−0.086	−0.085	−0.456
I50	C	−0.234	0.142	−0.026	−0.088	0.557	0.067
I51	A	−0.136	0.049	−0.187	0.508	−0.034	−0.030
I52	X	−0.166	0.250	−0.525	0.006	−0.124	0.057
I53	E	−0.061	−0.482	0.097	0.102	−0.052	0.169
I54	H	0.401	−0.005	−0.061	0.130	0.005	0.037
I55	O	−0.062	0.083	0.031	0.044	0.034	−0.635
I56	C	−0.080	0.047	−0.048	0.009	−0.515	0.089
I57	A	−0.161	0.089	0.013	−0.391	−0.244	0.000
I58	X	−0.231	−0.081	0.617	−0.114	0.108	0.054
I59	E	−0.252	−0.604	−0.113	−0.021	0.038	−0.069
I60	H	−0.546	−0.149	−0.159	−0.016	−0.235	−0.049

* Signs of the loadings were reversed to align with the conceptual definitions.

Table F3. *EFA pattern matrix loadings of 330 adjectives*

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A1	Abrasive	0.442	0.332	−0.116	−0.154	−0.022	0.004
A2	Abrupt	0.229	0.338	0.012	−0.159	0.138	0.060
A3	Absent-minded	0.010	0.277	0.141	0.102	0.531	−0.057
A4	Accommodating	0.073	−0.164	0.008	0.618	−0.065	0.039
A5	Adaptable	−0.069	−0.177	−0.090	0.322	−0.149	0.344
A6	Adventurous	0.221	−0.109	−0.312	0.189	0.079	0.409
A7	Affectionate	−0.029	0.063	−0.194	0.662	0.039	0.008
A8	Aggressive	0.292	0.364	−0.096	−0.252	−0.009	0.131
A9	Agreeable	0.025	−0.056	0.068	0.435	−0.053	0.152
A10	Aloof	0.397	0.117	0.384	−0.164	0.073	−0.040
A11	Altruistic	−0.052	0.010	0.011	0.151	0.174	0.366
A12	Ambitious	0.221	−0.025	−0.293	0.033	−0.371	0.289
A13	Analytical	0.017	−0.067	0.156	−0.021	−0.219	0.539
A14	Animated	−0.008	0.149	−0.376	0.252	0.119	0.227
A15	Anxious	0.164	0.383	0.308	0.297	0.098	−0.178
A16	Approachable	−0.145	−0.037	−0.275	0.438	−0.105	0.101
A17	Argumentative	0.183	0.385	−0.046	−0.113	0.133	0.213
A18	Arrogant	0.219	0.392	−0.012	−0.375	0.158	0.148
A19	Articulate	−0.196	0.045	−0.111	0.095	−0.170	0.369
A20	Artistic	0.115	−0.060	0.017	0.359	0.115	0.326
A21	Assertive	0.055	0.129	−0.230	−0.051	−0.253	0.399
A22	Authoritative	−0.031	0.320	−0.111	−0.181	−0.275	0.175
A23	Bashful	0.107	0.217	0.344	0.276	0.051	−0.130
A24	Big-hearted	−0.079	−0.020	−0.228	0.680	0.015	−0.001
A25	Bigoted	0.741	0.040	0.060	−0.028	−0.073	−0.034
A26	Bitter	0.334	0.298	0.188	−0.261	0.087	−0.085
A27	Blunt	0.085	0.388	−0.020	−0.221	−0.023	0.265
A28	Bold	0.256	−0.019	−0.247	0.007	−0.115	0.495
A29	Bossy	−0.013	0.569	−0.153	−0.148	−0.133	0.030
A30	Brave	0.185	−0.102	−0.323	0.112	−0.151	0.445

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A31	Bubbly	0.043	0.045	−0.573	0.366	−0.026	−0.106
A32	Bull-headed	0.070	0.453	−0.077	−0.181	0.114	0.118
A33	Calculating	0.521	−0.112	0.061	−0.043	−0.174	0.280
A34	Callous	0.671	0.034	0.021	−0.139	0.041	−0.028
A35	Calm	0.049	−0.450	0.111	0.222	−0.071	0.289
A36	Candid	−0.087	0.064	0.007	0.030	−0.168	0.355
A37	Carefree	0.135	0.002	−0.207	0.161	0.303	0.068
A38	Careful	0.236	−0.143	0.154	0.360	−0.520	−0.007
A39	Careless	0.167	0.162	0.044	−0.101	0.524	−0.032
A40	Casual	0.168	−0.018	0.078	0.370	0.173	0.074
A41	Cautious	0.302	−0.068	0.269	0.339	−0.380	0.012
A42	Charitable	−0.010	−0.128	−0.041	0.545	−0.111	0.190
A43	Chatty	−0.017	0.294	−0.529	0.274	0.096	−0.160
A44	Cheerful	−0.015	−0.114	−0.473	0.387	−0.051	0.021
A45	Civil	−0.125	−0.025	0.111	0.357	−0.231	0.270
A46	Clingy	0.494	0.205	0.117	0.184	0.132	−0.181
A47	Closed-minded	0.403	0.203	0.127	−0.220	−0.105	−0.329
A48	Cold	0.456	0.086	0.302	−0.324	0.057	0.193
A49	Cold-hearted	0.484	0.015	0.126	−0.415	0.002	0.124
A50	Compassionate	−0.169	−0.023	−0.065	0.687	−0.043	0.043
A51	Complaining	0.125	0.452	0.107	−0.061	0.219	−0.222
A52	Complex	−0.011	0.333	0.300	0.024	0.161	0.261
A53	Compliant	−0.101	−0.020	0.027	0.261	−0.186	−0.096
A54	Compulsive	0.216	0.352	−0.071	0.100	0.152	−0.028
A55	Conceited	0.668	−0.034	−0.108	−0.041	0.018	−0.070
A56	Condescending	0.454	0.106	−0.004	−0.099	0.057	0.015
A57	Confident	0.174	−0.185	−0.468	0.000	−0.384	0.347
A58	Conscientious	−0.396	0.032	0.116	0.106	−0.224	0.341
A59	Conservative	0.167	0.113	0.280	0.122	−0.399	−0.158
A60	Considerate	−0.234	0.020	−0.023	0.599	−0.094	0.124
A61	Conventional	0.271	−0.006	0.048	0.181	−0.538	−0.340

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A62	Cooperative	−0.068	−0.112	−0.092	0.435	−0.205	0.090
A63	Courageous	0.237	−0.095	−0.238	0.157	−0.199	0.450
A64	Courteous	−0.327	0.028	0.070	0.432	−0.079	0.202
A65	Cowardly	0.274	0.136	0.235	−0.043	0.262	−0.263
A66	Crabby	0.321	0.435	0.145	−0.125	0.019	−0.080
A67	Crafty	0.278	0.028	−0.049	0.082	0.076	0.224
A68	Creative	0.108	−0.120	−0.089	0.324	0.190	0.511
A69	Cunning	0.197	0.142	0.003	−0.171	0.154	0.223
A70	Curious	−0.182	0.247	0.010	0.186	0.084	0.478
A71	Daring	0.226	0.020	−0.264	0.030	0.057	0.493
A72	Deceitful	0.515	0.088	−0.027	−0.129	0.164	−0.029
A73	Deceptive	0.515	0.103	0.016	−0.159	0.153	−0.075
A74	Decisive	−0.161	0.026	−0.168	−0.100	−0.306	0.381
A75	Deep	0.199	0.000	0.138	0.230	0.123	0.436
A76	Defensive	0.410	0.283	0.143	0.066	−0.008	−0.145
A77	Defiant	0.270	0.393	−0.046	−0.110	0.049	0.120
A78	Demanding	−0.033	0.545	−0.097	−0.295	−0.118	0.259
A79	Dependable	−0.461	0.220	0.043	0.126	−0.225	0.120
A80	Detached	0.246	0.105	0.371	−0.133	0.181	0.120
A81	Determined	−0.020	0.097	−0.135	0.204	−0.477	0.363
A82	Devious	0.621	0.017	−0.041	−0.080	0.125	0.081
A83	Diligent	−0.260	0.039	0.055	0.149	−0.462	0.227
A84	Diplomatic	−0.245	−0.101	0.092	0.158	0.028	0.269
A85	Direct	0.105	0.250	−0.135	0.046	−0.340	0.418
A86	Discreet	−0.437	0.170	0.209	0.106	−0.040	0.237
A87	Dishonest	0.413	0.065	0.025	−0.189	0.376	−0.072
A88	Disorganized	0.017	0.126	0.107	0.039	0.746	0.005
A89	Disrespectful	0.229	0.210	−0.050	−0.360	0.292	−0.078
A90	Distant	0.244	0.169	0.490	−0.106	0.166	0.152
A91	Dominant	0.132	0.465	−0.168	−0.233	−0.238	0.259
A92	Domineering	0.305	0.427	−0.171	−0.290	−0.121	0.076

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A93	Down-to-earth	−0.070	−0.051	0.080	0.405	−0.266	0.067
A94	Dull	0.169	0.220	0.452	−0.157	0.158	−0.126
A95	Dynamic	0.124	0.014	−0.341	0.172	−0.185	0.431
A96	Easygoing	0.035	−0.226	−0.153	0.507	0.168	0.074
A97	Efficient	−0.081	−0.047	−0.058	0.100	−0.450	0.347
A98	Egotistical	0.401	0.213	−0.045	−0.288	0.131	0.011
A99	Emotional	0.033	0.372	0.003	0.437	0.085	−0.193
A100	Empathetic	−0.298	0.063	0.017	0.463	0.086	0.161
A101	Energetic	0.141	−0.077	−0.463	0.186	−0.233	0.270
A102	Enthusiastic	−0.091	0.022	−0.448	0.299	−0.103	0.206
A103	Ethical	−0.281	0.060	0.031	0.293	−0.112	0.390
A104	Expressive	0.000	0.177	−0.460	0.284	−0.074	0.262
A105	Extroverted	0.080	0.056	−0.659	0.044	0.021	0.133
A106	Faithful	−0.123	0.034	−0.023	0.425	−0.349	0.071
A107	Fearful	0.132	0.241	0.293	0.128	0.002	−0.337
A108	Fearless	0.206	−0.083	−0.256	−0.045	−0.047	0.492
A109	Flexible	0.004	−0.200	−0.064	0.433	−0.064	0.236
A110	Flighty	0.485	0.094	−0.020	0.052	0.298	−0.066
A111	Flippant	0.342	0.203	0.047	−0.047	0.246	−0.037
A112	Forceful	−0.128	0.419	−0.139	−0.296	−0.036	0.219
A113	Forgetful	0.215	0.105	0.095	0.193	0.411	−0.074
A114	Forgiving	0.024	−0.232	−0.062	0.524	−0.039	0.083
A115	Frank	0.127	0.187	−0.077	0.010	−0.243	0.432
A116	Friendly	−0.036	−0.128	−0.350	0.538	−0.079	−0.003
A117	Frivolous	0.298	0.259	−0.068	0.083	0.266	−0.179
A118	Fussy	−0.009	0.476	0.081	−0.004	0.027	−0.153
A119	Generous	−0.050	−0.029	−0.109	0.632	−0.120	0.069
A120	Gentle	0.242	−0.237	0.119	0.629	−0.061	0.052
A121	Giving	−0.034	0.002	−0.110	0.612	−0.117	0.031
A122	Gloomy	0.365	0.220	0.357	−0.003	0.179	−0.022
A123	Good-hearted	−0.084	0.041	−0.110	0.646	−0.123	0.099

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A124	Good-natured	−0.078	−0.139	−0.133	0.588	−0.053	0.103
A125	Gracious	0.082	−0.209	−0.062	0.555	−0.110	0.177
A126	Greedy	0.243	0.275	−0.019	−0.240	0.198	0.028
A127	Grumpy	0.095	0.489	0.267	−0.083	0.178	−0.056
A128	Gullible	0.339	0.211	0.142	0.242	0.151	−0.248
A129	Happy-go-lucky	0.229	−0.062	−0.274	0.372	0.091	−0.046
A130	Hard-headed	0.066	0.409	−0.006	−0.201	0.040	0.130
A131	Hard-working	−0.143	0.067	−0.094	0.144	−0.536	0.038
A132	Harsh	0.167	0.416	0.062	−0.380	−0.012	0.157
A133	Heartless	0.334	0.085	0.171	−0.485	0.071	0.146
A134	Helpful	−0.098	0.043	−0.041	0.582	−0.278	0.121
A135	High-strung	0.441	0.347	0.035	0.090	−0.002	−0.110
A136	Honest	−0.189	0.012	0.065	0.339	−0.363	0.143
A137	Hospitable	−0.148	−0.027	−0.160	0.534	−0.014	0.157
A138	Hostile	0.505	0.340	−0.009	−0.217	−0.066	−0.041
A139	Hot-tempered	0.110	0.586	0.015	−0.126	0.070	−0.027
A140	Humble	0.201	−0.180	0.220	0.500	−0.202	0.122
A141	Idealistic	0.265	0.031	0.001	0.309	0.093	0.262
A142	Illogical	0.262	0.099	−0.015	0.058	0.270	−0.295
A143	Imaginative	0.012	−0.016	−0.107	0.243	0.139	0.493
A144	Immature	0.233	0.158	−0.012	−0.025	0.495	−0.177
A145	Impartial	−0.160	0.006	0.130	−0.032	0.002	0.338
A146	Impatient	−0.176	0.549	0.059	−0.145	0.191	−0.034
A147	Impersonal	0.487	−0.207	0.287	−0.165	0.015	0.079
A148	Impolite	0.290	0.118	0.105	−0.269	0.198	−0.152
A149	Impressionable	0.179	0.109	−0.112	0.182	0.029	−0.130
A150	Impulsive	0.106	0.372	−0.177	0.107	0.330	0.058
A151	Inconsiderate	0.379	0.107	0.010	−0.397	0.221	0.071
A152	Inconsistent	0.146	0.172	0.047	0.030	0.480	−0.084
A153	Indecisive	0.110	0.235	0.240	0.203	0.403	−0.197
A154	Independent	−0.099	0.016	−0.046	0.138	−0.132	0.380

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A155	Individualistic	−0.062	0.145	0.117	0.036	0.100	0.174
A156	Industrious	−0.289	0.036	−0.005	0.014	−0.195	0.419
A157	Inefficient	0.144	0.087	0.133	−0.059	0.529	−0.158
A158	Informal	−0.275	0.167	0.037	0.083	0.406	0.106
A159	Ingenious	0.012	−0.080	−0.011	−0.001	−0.016	0.467
A160	Inhibited	0.144	0.220	0.417	0.121	−0.048	−0.120
A161	Innovative	0.001	−0.099	−0.098	0.142	0.031	0.621
A162	Inquisitive	−0.227	0.165	−0.008	0.084	0.084	0.346
A163	Insecure	−0.013	0.421	0.371	0.247	0.257	−0.223
A164	Insensitive	0.364	0.027	0.122	−0.399	0.125	0.088
A165	Insightful	−0.187	−0.042	0.028	0.085	0.030	0.610
A166	Insincere	0.358	0.005	0.071	−0.286	0.208	−0.142
A167	Intense	0.204	0.258	−0.042	−0.078	0.028	0.328
A168	Introspective	0.129	0.016	0.375	0.126	0.100	0.383
A169	Introverted	0.049	0.065	0.750	0.079	0.104	0.089
A170	Intuitive	−0.246	0.091	0.066	0.124	0.016	0.467
A171	Irrational	0.342	0.194	0.007	0.005	0.318	−0.179
A172	Irresponsible	0.190	0.054	0.032	−0.048	0.587	−0.039
A173	Irritable	0.072	0.576	0.193	−0.083	0.150	−0.039
A174	Jolly	0.093	−0.088	−0.507	0.382	−0.053	0.021
A175	Kind	−0.056	−0.064	−0.021	0.705	−0.137	0.033
A176	Kind-hearted	−0.108	0.015	−0.008	0.698	−0.113	0.089
A177	Law-abiding	−0.266	0.054	0.038	0.247	−0.432	−0.021
A178	Lazy	0.168	0.073	0.217	0.025	0.567	−0.023
A179	Lenient	0.095	−0.101	0.102	0.393	0.118	0.004
A180	Lethargic	0.282	0.171	0.248	0.202	0.413	−0.107
A181	Light-hearted	−0.144	−0.014	−0.144	0.282	0.109	0.120
A182	Lively	0.187	−0.046	−0.582	0.368	−0.131	0.118
A183	Logical	0.054	−0.192	0.207	0.036	−0.292	0.533
A184	Loud	0.185	0.482	−0.469	0.086	0.048	−0.086
A185	Loving	0.051	0.021	−0.182	0.749	−0.069	−0.030

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A186	Loyal	−0.134	0.076	−0.017	0.444	−0.321	0.040
A187	Manipulative	0.372	0.181	−0.039	−0.145	0.115	0.088
A188	Materialistic	0.340	0.206	−0.058	−0.051	−0.013	−0.069
A189	Meek	0.461	−0.082	0.237	0.248	0.013	−0.030
A190	Melodramatic	0.254	0.418	−0.121	0.205	0.166	−0.167
A191	Messy	0.005	0.153	0.098	0.057	0.663	0.070
A192	Methodical	−0.086	0.009	0.103	−0.071	−0.471	0.292
A193	Meticulous	−0.131	0.080	0.107	0.024	−0.438	0.232
A194	Mild	0.153	−0.141	0.246	0.321	0.037	0.052
A195	Mischievous	0.072	0.251	−0.038	−0.049	0.372	0.127
A196	Modest	0.050	−0.147	0.279	0.317	−0.118	0.148
A197	Moody	0.140	0.469	0.229	0.085	0.132	−0.079
A198	Moral	−0.210	0.052	0.024	0.176	−0.333	0.245
A199	Narrow-minded	0.262	0.250	0.133	−0.107	0.014	−0.271
A200	Negative	0.037	0.321	0.346	−0.042	0.234	−0.180
A201	Nervous	0.052	0.310	0.368	0.277	0.135	−0.268
A202	Noisy	0.043	0.435	−0.421	−0.038	0.248	−0.149
A203	Nonchalant	0.202	−0.153	0.010	0.033	0.273	0.145
A204	Nosey	−0.021	0.476	−0.114	−0.067	0.149	−0.040
A205	Objective	−0.088	−0.010	0.102	−0.046	−0.198	0.565
A206	Old-fashioned	0.185	0.165	0.238	0.140	−0.249	−0.196
A207	Open-minded	−0.119	−0.174	−0.111	0.348	0.114	0.466
A208	Opinionated	0.165	0.437	−0.070	0.113	0.023	0.156
A209	Opportunistic	0.261	0.016	−0.163	0.005	−0.079	0.073
A210	Optimistic	0.097	−0.196	−0.368	0.253	−0.147	0.227
A211	Organized	0.043	−0.021	−0.108	0.094	−0.693	0.017
A212	Original	0.260	−0.124	−0.079	0.165	−0.060	0.512
A213	Outgoing	0.093	0.025	−0.646	0.235	−0.059	0.089
A214	Outspoken	0.100	0.214	−0.374	0.077	0.046	0.425
A215	Overbearing	0.531	0.221	−0.090	−0.155	0.024	0.031
A216	Oversensitive	0.056	0.494	0.197	0.270	0.137	−0.188

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A217	Passive	0.272	0.047	0.312	0.304	0.090	−0.191
A218	Patient	0.118	−0.430	0.162	0.410	−0.100	0.161
A219	Peaceful	0.147	−0.380	0.155	0.488	−0.012	0.179
A220	Perceptive	−0.208	0.115	0.048	0.139	−0.026	0.412
A221	Perfectionistic	0.060	0.321	0.058	0.230	−0.418	0.086
A222	Persistent	−0.056	0.146	−0.015	0.067	−0.317	0.449
A223	Pessimistic	0.135	0.241	0.323	−0.015	0.135	−0.127
A224	Philosophical	−0.005	−0.059	0.161	0.249	0.147	0.387
A225	Picky	0.131	0.451	0.144	−0.033	−0.073	0.216
A226	Playful	0.009	0.086	−0.253	0.443	0.277	0.160
A227	Pleasant	−0.077	−0.153	−0.200	0.523	−0.131	0.100
A228	Polite	−0.138	−0.136	0.050	0.520	−0.237	0.093
A229	Pompous	0.627	0.019	−0.111	−0.137	0.024	0.011
A230	Practical	−0.041	−0.017	0.014	0.206	−0.276	0.335
A231	Pretentious	0.549	0.156	−0.069	−0.158	−0.003	−0.034
A232	Prompt	−0.203	0.135	−0.048	0.056	−0.421	0.131
A233	Proper	0.224	−0.095	−0.017	0.215	−0.409	0.197
A234	Proud	0.018	0.217	−0.221	0.044	−0.231	−0.031
A235	Quick-tempered	0.042	0.597	0.005	−0.016	0.014	−0.044
A236	Quiet	0.109	−0.052	0.726	0.104	−0.044	0.051
A237	Rambunctious	0.142	0.284	−0.173	−0.027	0.170	−0.005
A238	Rash	0.278	0.367	−0.068	−0.066	0.284	−0.003
A239	Rational	−0.211	−0.014	0.066	0.076	−0.114	0.481
A240	Realistic	−0.009	0.018	0.027	0.122	−0.376	0.295
A241	Reasonable	−0.114	−0.194	0.087	0.430	−0.131	0.302
A242	Rebellious	0.021	0.372	−0.086	−0.122	0.413	0.247
A243	Reckless	0.261	0.188	−0.065	−0.124	0.445	0.065
A244	Relaxed	0.061	−0.358	−0.079	0.230	0.157	0.234
A245	Reliable	−0.337	0.056	−0.071	0.230	−0.414	0.157
A246	Resentful	0.337	0.283	0.215	0.032	0.053	−0.070
A247	Reserved	0.077	0.014	0.639	0.123	−0.081	0.031

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A248	Resilient	−0.406	0.074	0.003	−0.036	−0.061	0.477
A249	Resourceful	−0.172	0.021	−0.057	0.079	−0.199	0.524
A250	Respectful	−0.185	−0.022	0.004	0.446	−0.343	0.184
A251	Responsible	−0.167	0.041	−0.043	0.174	−0.605	0.184
A252	Restless	0.106	0.436	0.032	−0.057	0.103	0.071
A253	Rough	0.355	0.283	−0.001	−0.262	0.078	0.104
A254	Rugged	0.101	0.135	0.000	−0.221	0.122	0.287
A255	Ruthless	0.201	0.294	−0.116	−0.372	−0.009	0.215
A256	Scatterbrained	0.249	0.238	0.072	0.175	0.499	−0.108
A257	Scheming	0.563	0.064	−0.050	−0.132	0.105	0.061
A258	Secretive	0.382	0.099	0.218	0.060	0.045	0.115
A259	Self-assured	0.105	−0.180	−0.279	−0.016	−0.263	0.397
A260	Self-centered	0.361	0.222	0.051	−0.247	0.166	0.082
A261	Self-confident	0.109	−0.320	−0.362	−0.073	−0.128	0.446
A262	Self-conscious	0.370	0.037	0.307	0.274	0.012	0.096
A263	Self-disciplined	0.016	−0.073	−0.039	0.115	−0.615	0.188
A264	Self-indulgent	0.522	0.137	−0.062	0.034	0.144	0.052
A265	Selfless	−0.172	0.014	0.067	0.390	−0.095	0.155
A266	Self-reliant	−0.189	0.175	0.065	0.117	−0.246	0.374
A267	Self-righteous	0.621	0.066	0.027	0.063	−0.120	0.078
A268	Sensitive	−0.059	0.313	0.204	0.351	0.086	−0.045
A269	Sentimental	−0.100	0.318	0.053	0.493	0.034	−0.216
A270	Serious	0.119	0.120	0.295	0.010	−0.363	0.227
A271	Short-tempered	0.073	0.606	0.056	−0.099	0.075	−0.052
A272	Shy	0.044	0.116	0.678	0.206	0.035	−0.088
A273	Simple	0.404	−0.107	0.097	0.373	−0.157	0.066
A274	Sincere	−0.176	−0.013	0.010	0.397	−0.138	0.142
A275	Skeptical	0.073	0.361	0.278	−0.056	0.075	0.171
A276	Sloppy	0.338	0.102	0.112	−0.009	0.502	−0.090
A277	Sly	0.393	0.071	0.014	−0.130	0.186	0.167
A278	Sneaky	0.413	0.183	0.018	−0.145	0.204	0.047

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A279	Snobbish	0.345	0.275	0.012	−0.164	0.012	−0.064
A280	Sociable	0.021	−0.010	−0.640	0.299	−0.088	−0.023
A281	Social	0.025	−0.015	−0.680	0.262	−0.061	−0.006
A282	Spineless	0.490	−0.062	0.220	0.084	0.188	−0.194
A283	Spontaneous	0.244	0.110	−0.419	0.280	0.099	0.228
A284	Stern	0.273	0.172	0.003	−0.176	−0.267	0.109
A285	Straightforward	−0.155	0.248	−0.069	0.033	−0.369	0.276
A286	Stubborn	−0.104	0.594	0.086	−0.108	0.135	0.090
A287	Studious	0.064	−0.027	0.136	0.109	−0.305	0.378
A288	Stuffy	0.487	0.125	0.239	−0.087	0.058	−0.039
A289	Sympathetic	−0.105	0.000	0.012	0.747	0.028	0.071
A290	Talkative	−0.017	0.275	−0.633	0.245	0.100	−0.071
A291	Temperamental	0.238	0.470	0.052	−0.022	0.095	−0.060
A292	Tense	0.085	0.512	0.293	0.002	0.047	−0.041
A293	Thorough	−0.261	0.118	0.071	0.092	−0.480	0.114
A294	Tidy	0.124	−0.027	−0.087	0.138	−0.617	−0.106
A295	Timid	0.245	0.112	0.489	0.242	0.104	−0.177
A296	Tolerant	−0.105	−0.243	0.047	0.413	0.130	0.236
A297	Touchy	0.268	0.344	0.003	0.158	0.143	−0.185
A298	Tough	0.022	0.203	−0.087	−0.116	−0.089	0.427
A299	Traditional	0.232	0.074	0.101	0.189	−0.461	−0.237
A300	Trustworthy	−0.216	0.007	0.022	0.334	−0.348	0.151
A301	Truthful	−0.233	−0.091	−0.021	0.292	−0.295	0.186
A302	Unapproachable	0.197	0.188	0.255	−0.342	0.128	0.043
A303	Unassuming	−0.266	0.020	0.098	0.018	0.055	0.169
A304	Uncompromising	0.126	0.204	0.069	−0.225	0.060	0.131
A305	Unconventional	−0.195	0.071	0.015	−0.093	0.512	0.515
A306	Uncooperative	0.295	0.159	0.134	−0.369	0.240	−0.038
A307	Underhanded	0.608	0.050	−0.005	−0.087	0.141	−0.100
A308	Understanding	−0.034	−0.134	−0.021	0.672	−0.035	0.180
A309	Undisciplined	−0.002	0.179	0.085	−0.083	0.586	0.032

Item	Adjective	H(-) /E(-)	E(+) /A(+)	X(-)	A(+) /E(+) /H(+)	C(-)	O(+)
A310	Unemotional	0.135	−0.179	0.120	−0.460	0.012	0.251
A311	Unfeeling	0.324	0.040	0.165	−0.395	0.061	0.151
A312	Unforgiving	0.114	0.418	0.166	−0.342	−0.024	−0.091
A313	Unfriendly	0.109	0.142	0.334	−0.488	0.197	0.017
A314	Unimaginative	0.216	0.158	0.137	−0.133	−0.132	−0.419
A315	Uninhibited	−0.052	0.127	−0.249	−0.111	0.268	0.129
A316	Unkind	0.271	0.154	0.072	−0.533	0.107	−0.045
A317	Unreliable	0.230	−0.102	0.088	−0.066	0.643	−0.032
A318	Unruly	0.346	0.140	−0.197	−0.211	0.306	0.020
A319	Unsympathetic	0.293	0.132	0.135	−0.459	0.076	0.069
A320	Untidy	−0.106	0.115	0.118	0.020	0.693	0.111
A321	Vain	0.419	0.201	−0.088	−0.042	0.128	−0.023
A322	Verbal	−0.122	0.332	−0.394	0.170	0.032	0.299
A323	Vibrant	0.115	−0.049	−0.542	0.259	−0.075	0.192
A324	Vindictive	0.535	0.156	0.072	−0.161	0.173	0.080
A325	Vocal	0.015	0.213	−0.452	0.101	−0.091	0.268
A326	Warm	−0.195	−0.040	−0.224	0.628	−0.015	−0.028
A327	Warm-hearted	−0.130	−0.009	−0.129	0.741	−0.042	−0.020
A328	Well-mannered	−0.123	−0.041	0.018	0.489	−0.257	0.136
A329	Whiny	0.347	0.305	0.113	0.067	0.200	−0.176
A330	Withdrawn	0.248	0.168	0.552	0.023	0.234	0.052

Table F4. *Final calibrated item parameters for 279 adjectives*

No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A25	Bigoted	H	1.510	−0.487	0.659
A33	Calculating	H	3.709	−0.458	2.595
A36	Candid	H	4.548	0.306	1.448
A55	Conceited	H	1.931	−0.555	1.236
A56	Condescending	H	1.990	−0.551	1.469
A69	Cunning	H	2.326	−0.544	1.951
A72	Deceitful	H	1.300	−0.392	0.406
A73	Deceptive	H	1.513	−0.550	0.586
A82	Devious	H	1.622	−0.580	0.837
A85	Direct	H	4.778	0.150	0.954
A86	Discreet	H	4.658	0.319	1.638
A87	Dishonest	H	1.217	−0.351	0.233
A93	Down-to-earth	H	5.357	0.344	0.612
A98	Egotistical	H	1.786	−0.599	0.841
A103	Ethical	H	5.542	0.347	0.424
A106	Faithful	H	5.515	0.391	0.508
A126	Greedy	H	1.715	−0.583	0.818
A136	Honest	H	5.725	0.313	0.245
A140	Humble	H	4.953	0.284	0.952
A145	Impartial	H	4.290	0.226	2.269
A166	Insincere	H	1.384	−0.415	0.404
A177	Law-abiding	H	5.428	0.380	0.723
A187	Manipulative	H	1.880	−0.639	1.011
A188	Materialistic	H	2.628	−0.634	1.673
A196	Modest	H	4.728	0.263	1.123
A198	Moral	H	5.419	0.376	0.596
A205	Objective	H	5.009	0.186	0.988
A229	Pompous	H	1.696	−0.533	0.907
A231	Pretentious	H	1.799	−0.539	0.876
A255	Ruthless	H	1.942	−0.481	1.272
A257	Scheming	H	1.868	−0.696	1.134
A258	Secretive	H	2.740	−0.519	2.056

No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A260	Self-centred	H	1.994	−0.697	0.982
A264	Self-indulgent	H	2.706	−0.665	1.785
A265	Selfless	H	4.464	0.529	1.634
A267	Self-righteous	H	2.855	−0.627	2.433
A274	Sincere	H	5.628	0.304	0.351
A277	Sly	H	1.850	−0.595	1.235
A278	Sneaky	H	1.584	−0.604	0.602
A279	Snobbish	H	1.734	−0.498	0.888
A285	Straightforward	H	5.104	0.295	0.736
A300	Trustworthy	H	5.802	0.241	0.216
A301	Truthful	H	5.647	0.341	0.257
A303	Unassuming	H	3.822	0.274	2.238
A307	Underhanded	H	1.553	−0.521	0.538
A321	Vain	H	1.837	−0.551	1.018
A15	Anxious	E	2.896	0.996	1.060
A30	Brave	E	4.833	−0.463	0.668
A34	Callous	E	1.610	0.260	1.013
A35	Calm	E	5.049	−0.418	0.683
A46	Clingy	E	1.876	0.623	1.052
A51	Complaining	E	1.881	0.648	0.710
A63	Courageous	E	4.946	−0.401	0.707
A65	Cowardly	E	1.529	0.408	0.584
A71	Daring	E	4.255	−0.325	1.298
A80	Detached	E	2.220	0.396	1.473
A99	Emotional	E	3.452	0.613	1.428
A107	Fearful	E	2.402	0.669	1.074
A108	Fearless	E	4.220	−0.486	1.270
A130	Hard-headed	E	2.743	0.307	2.290
A135	High-strung	E	2.270	0.579	1.488
A153	Indecisive	E	2.120	0.736	0.976
A154	Independent	E	5.304	−0.270	0.702
A160	Inhibited	E	2.536	0.514	1.477
A163	Insecure	E	2.216	0.882	0.865

No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A190	Melodramatic	E	1.943	0.593	1.187
A197	Moody	E	2.193	0.719	1.033
A201	Nervous	E	2.834	0.891	1.019
A216	Oversensitive	E	2.347	0.855	1.060
A244	Relaxed	E	4.497	−0.411	1.175
A248	Resilient	E	5.077	−0.366	1.130
A259	Self-assured	E	4.872	−0.562	0.719
A262	Self-conscious	E	4.208	0.436	1.912
A266	Self-reliant	E	5.187	−0.276	0.834
A268	Sensitive	E	3.695	0.601	1.816
A269	Sentimental	E	3.884	0.415	1.896
A292	Tense	E	2.456	0.797	1.023
A297	Touchy	E	2.408	0.620	1.566
A298	Tough	E	4.092	−0.241	1.918
A329	Whiny	E	1.612	0.540	0.679
A10	Aloof	X	1.943	−0.632	0.938
A14	Animated	X	4.140	0.528	1.414
A21	Assertive	X	4.684	0.414	1.086
A23	Bashful	X	2.569	−0.478	1.674
A28	Bold	X	4.352	0.434	1.251
A31	Bubbly	X	4.222	0.838	1.225
A43	Chatty	X	4.068	0.570	1.487
A44	Cheerful	X	5.202	0.558	0.464
A57	Confident	X	5.107	0.614	0.409
A90	Distant	X	2.272	−0.734	1.126
A94	Dull	X	1.724	−0.601	0.598
A95	Dynamic	X	4.909	0.545	0.589
A101	Energetic	X	5.133	0.608	0.430
A102	Enthusiastic	X	5.307	0.520	0.421
A104	Expressive	X	4.749	0.593	0.692
A105	Extroverted	X	4.009	0.960	1.289
A122	Gloomy	X	1.705	−0.524	0.822
A169	Introverted	X	2.897	−1.068	1.256

No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A174	Jolly	X	4.820	0.697	0.653
A180	Lethargic	X	1.772	−0.474	0.948
A181	Light-hearted	X	4.328	0.311	1.533
A182	Lively	X	4.968	0.684	0.386
A184	Loud	X	2.744	0.336	1.853
A210	Optimistic	X	5.165	0.575	0.613
A213	Outgoing	X	4.780	0.806	0.592
A214	Outspoken	X	4.168	0.577	1.451
A217	Passive	X	2.710	−0.488	1.860
A223	Pessimistic	X	2.233	−0.627	1.440
A236	Quiet	X	3.233	−0.810	1.230
A247	Reserved	X	3.586	−0.733	1.544
A252	Restless	X	2.972	−0.289	2.160
A261	Self-confident	X	5.023	0.566	0.630
A272	Shy	X	2.824	−0.889	1.259
A280	Sociable	X	5.086	0.721	0.444
A281	Social	X	5.005	0.757	0.412
A290	Talkative	X	4.261	0.691	1.194
A295	Timid	X	2.352	−0.711	1.303
A322	Verbal	X	4.526	0.502	1.222
A323	Vibrant	X	4.806	0.707	0.568
A325	Vocal	X	4.513	0.591	0.922
A330	Withdrawn	X	2.065	−0.794	0.831
A1	Abrasive	A	1.942	−0.493	1.073
A2	Abrupt	A	2.302	−0.617	1.186
A4	Accommodating	A	5.150	0.503	0.550
A5	Adaptable	A	5.485	0.447	0.403
A8	Aggressive	A	2.044	−0.542	1.498
A9	Agreeable	A	4.847	0.368	0.826
A16	Approachable	A	5.493	0.480	0.447
A17	Argumentative	A	3.017	−0.620	1.969
A24	Big-hearted	A	5.141	0.567	0.663
A26	Bitter	A	1.497	−0.515	0.470

No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A27	Blunt	A	3.060	−0.544	2.038
A29	Bossy	A	2.871	−0.447	1.815
A32	Bull-headed	A	2.377	−0.630	1.639
A45	Civil	A	5.364	0.379	0.538
A50	Compassionate	A	5.309	0.548	0.415
A53	Compliant	A	4.565	0.331	1.881
A60	Considerate	A	5.390	0.483	0.366
A62	Cooperative	A	5.614	0.346	0.267
A64	Courteous	A	5.376	0.436	0.526
A66	Crabby	A	1.621	−0.517	0.577
A76	Defensive	A	2.912	−0.547	1.546
A77	Defiant	A	2.477	−0.580	1.691
A78	Demanding	A	3.225	−0.524	2.109
A84	Diplomatic	A	4.883	0.345	1.124
A89	Disrespectful	A	1.245	−0.338	0.269
A96	Easygoing	A	5.028	0.472	0.904
A100	Empathetic	A	5.074	0.402	1.055
A109	Flexible	A	5.384	0.439	0.343
A112	Forceful	A	3.093	−0.441	2.219
A114	Forgiving	A	5.093	0.570	0.535
A116	Friendly	A	5.528	0.449	0.281
A118	Fussy	A	2.388	−0.404	1.752
A119	Generous	A	5.265	0.514	0.386
A120	Gentle	A	4.872	0.512	0.776
A121	Giving	A	5.246	0.456	0.457
A123	Good-hearted	A	5.514	0.435	0.299
A124	Good-natured	A	5.520	0.440	0.270
A125	Gracious	A	5.046	0.573	0.509
A127	Grumpy	A	1.900	−0.661	0.750
A132	Harsh	A	1.985	−0.631	0.973
A137	Hospitable	A	5.307	0.491	0.514
A138	Hostile	A	1.482	−0.431	0.630
A139	Hot-tempered	A	1.859	−0.664	0.915

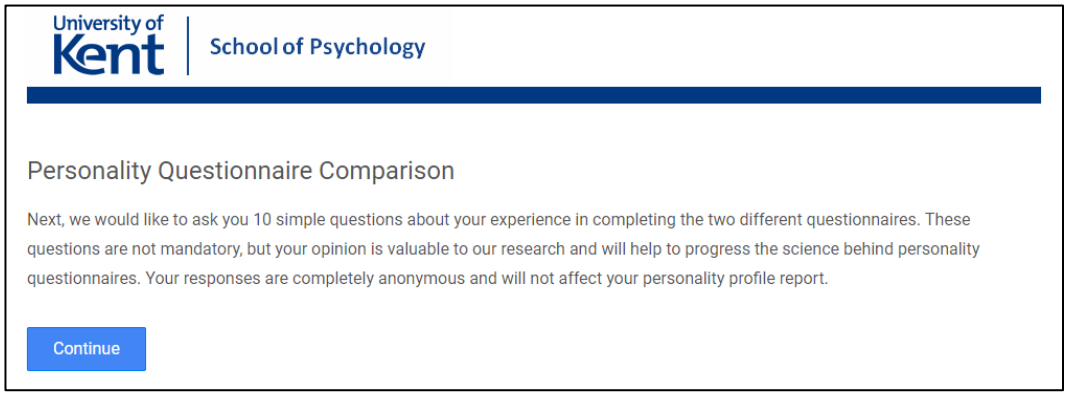
No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A146	Impatient	A	2.524	−0.666	1.481
A148	Impolite	A	1.290	−0.302	0.374
A151	Inconsiderate	A	1.494	−0.485	0.466
A173	Irritable	A	2.011	−0.707	0.894
A175	Kind	A	5.456	0.482	0.275
A176	Kind-hearted	A	5.400	0.479	0.396
A179	Lenient	A	4.156	0.231	1.312
A185	Loving	A	5.312	0.489	0.388
A208	Opinionated	A	3.674	−0.306	2.136
A218	Patient	A	4.983	0.632	0.805
A219	Peaceful	A	5.019	0.504	0.647
A225	Picky	A	3.142	−0.458	1.911
A227	Pleasant	A	5.323	0.538	0.282
A228	Polite	A	5.623	0.419	0.238
A235	Quick-tempered	A	2.344	−0.650	1.461
A246	Resentful	A	1.972	−0.492	1.182
A250	Respectful	A	5.665	0.381	0.220
A253	Rough	A	1.871	−0.575	0.863
A271	Short-tempered	A	1.888	−0.670	0.892
A275	Sceptical	A	3.419	−0.498	1.627
A284	Stern	A	3.126	−0.331	1.992
A286	Stubborn	A	2.919	−0.705	1.690
A289	Sympathetic	A	5.207	0.524	0.463
A291	Temperamental	A	2.136	−0.599	1.355
A296	Tolerant	A	5.215	0.426	0.545
A302	Unapproachable	A	1.532	−0.521	0.666
A304	Uncompromising	A	2.543	−0.449	1.852
A306	Uncooperative	A	1.313	−0.387	0.270
A308	Understanding	A	5.447	0.462	0.288
A312	Unforgiving	A	1.709	−0.600	0.651
A313	Unfriendly	A	1.381	−0.468	0.325
A316	Unkind	A	1.319	−0.425	0.295
A318	Unruly	A	1.642	−0.445	0.829

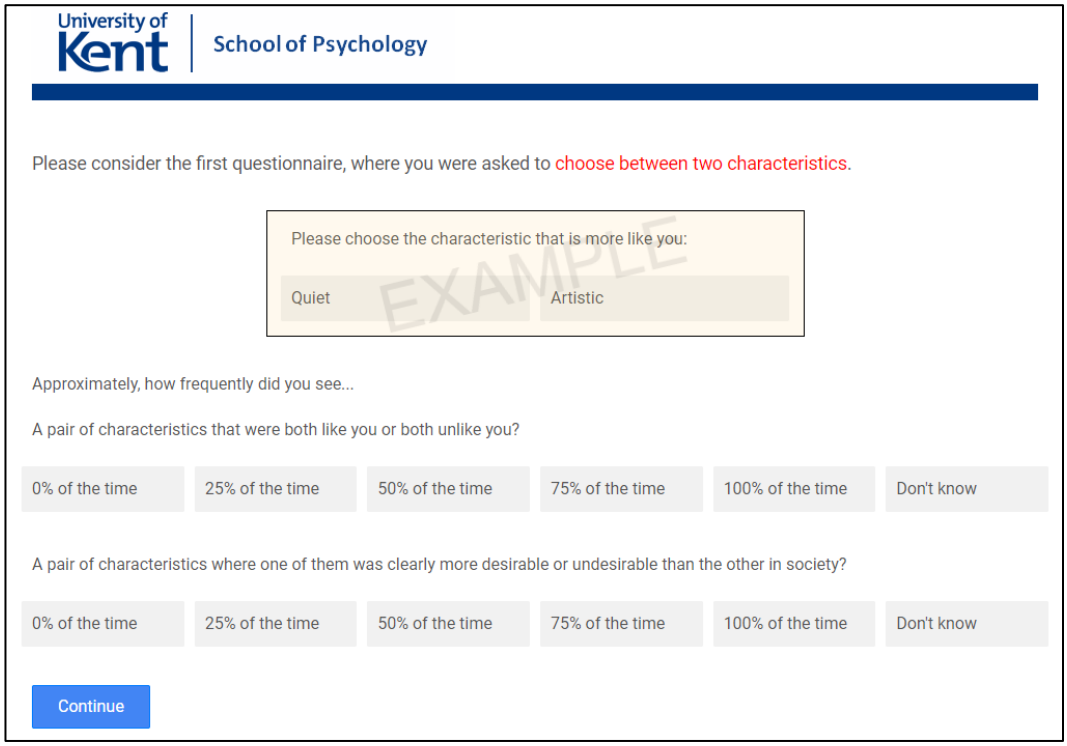
No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A319	Unsympathetic	A	1.534	−0.519	0.473
A324	Vindictive	A	1.525	−0.490	0.740
A326	Warm	A	5.160	0.593	0.493
A327	Warm-hearted	A	5.319	0.580	0.383
A328	Well-mannered	A	5.592	0.354	0.343
A3	Absent-minded	C	1.972	−0.720	0.833
A12	Ambitious	C	5.258	0.448	0.634
A38	Careful	C	5.146	0.394	0.550
A39	Careless	C	1.642	−0.597	0.513
A41	Cautious	C	4.516	0.268	1.194
A54	Compulsive	C	2.726	−0.404	1.771
A58	Conscientious	C	5.283	0.411	0.742
A79	Dependable	C	5.204	0.286	1.763
A81	Determined	C	5.458	0.459	0.316
A83	Diligent	C	5.404	0.495	0.446
A88	Disorganised	C	1.722	−0.799	0.459
A97	Efficient	C	5.417	0.504	0.330
A110	Flighty	C	1.836	−0.503	1.051
A111	Flippant	C	1.813	−0.489	1.017
A113	Forgetful	C	2.500	−0.684	1.754
A117	Frivolous	C	1.996	−0.482	1.182
A131	Hard-working	C	5.699	0.350	0.275
A142	Illogical	C	1.512	−0.302	0.612
A144	Immature	C	1.572	−0.615	0.565
A150	Impulsive	C	2.803	−0.498	1.684
A152	Inconsistent	C	1.736	−0.569	0.711
A156	Industrious	C	5.053	0.420	0.981
A157	Inefficient	C	1.494	−0.526	0.372
A158	Informal	C	3.665	−0.329	2.098
A171	Irrational	C	1.567	−0.450	0.604
A172	Irresponsible	C	1.334	−0.436	0.434
A178	Lazy	C	1.669	−0.708	0.698
A183	Logical	C	5.368	0.386	0.492

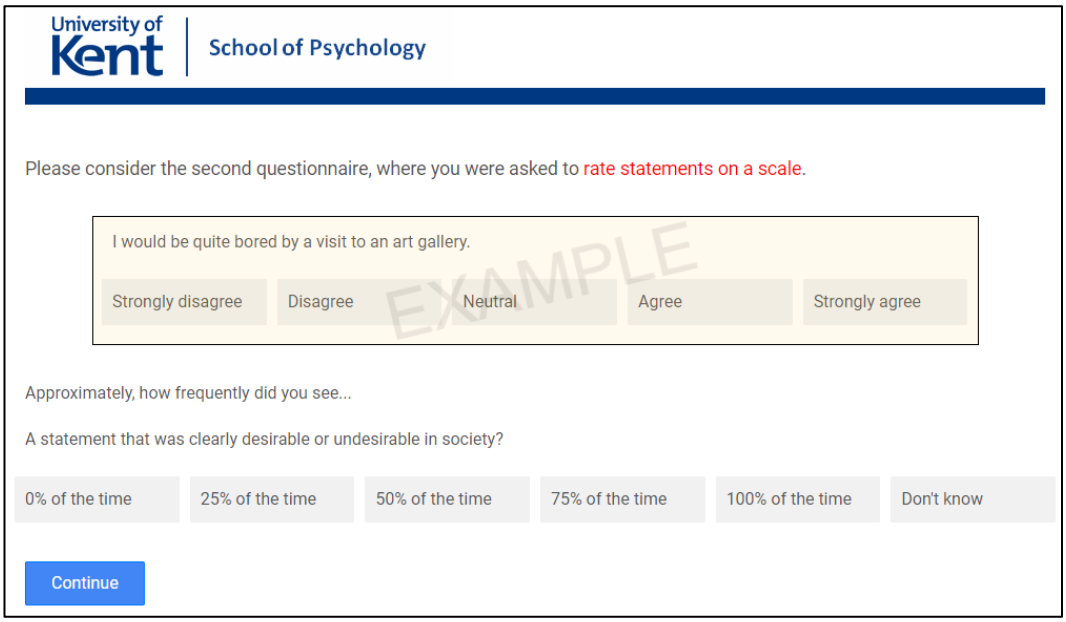
No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A191	Messy	C	1.881	−0.784	0.791
A192	Methodical	C	5.017	0.456	0.749
A193	Meticulous	C	4.918	0.534	1.064
A211	Organised	C	5.293	0.646	0.365
A221	Perfectionistic	C	4.663	0.369	1.259
A222	Persistent	C	5.159	0.360	0.746
A230	Practical	C	5.351	0.355	0.430
A232	Prompt	C	5.023	0.449	0.903
A233	Proper	C	4.925	0.409	0.885
A238	Rash	C	1.898	−0.542	0.905
A239	Rational	C	5.261	0.315	0.766
A240	Realistic	C	5.375	0.331	0.444
A241	Reasonable	C	5.431	0.299	0.332
A243	Reckless	C	1.738	−0.559	0.791
A245	Reliable	C	5.716	0.299	0.296
A251	Responsible	C	5.689	0.408	0.202
A256	Scatterbrained	C	1.913	−0.787	0.880
A263	Self-disciplined	C	5.334	0.595	0.350
A270	Serious	C	4.433	0.246	1.220
A276	Sloppy	C	1.541	−0.580	0.401
A293	Thorough	C	5.371	0.453	0.479
A294	Tidy	C	4.951	0.640	0.818
A309	Undisciplined	C	1.539	−0.558	0.561
A317	Unreliable	C	1.306	−0.373	0.362
A320	Untidy	C	1.887	−0.713	0.878
A13	Analytical	O	5.274	0.361	0.666
A19	Articulate	O	5.115	0.384	0.693
A20	Artistic	O	4.026	0.889	1.275
A47	Closed-minded	O	1.486	−0.347	0.568
A68	Creative	O	4.930	0.709	0.472
A70	Curious	O	5.293	0.303	0.665
A75	Deep	O	4.353	0.399	1.573
A143	Imaginative	O	5.038	0.600	0.517

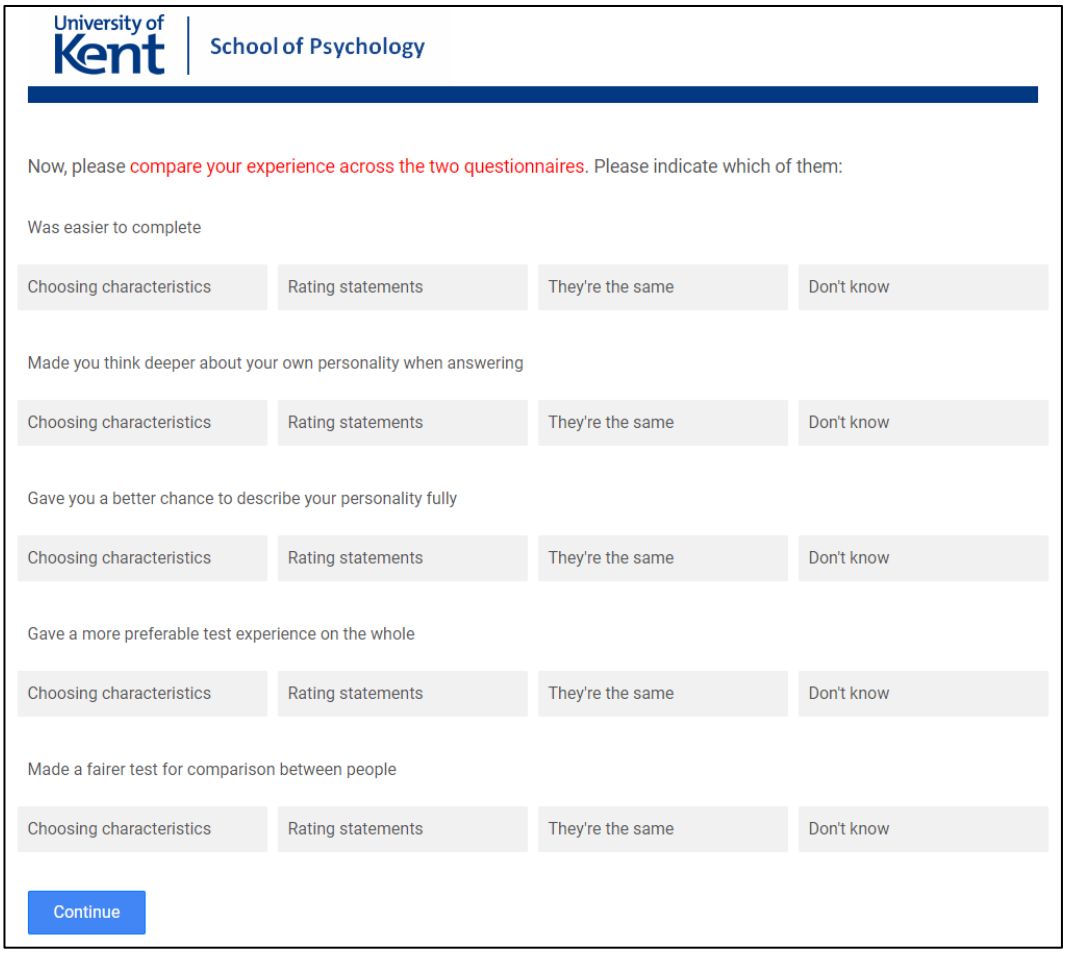
No.	Adjective	Scale	μ_i	λ_i	ψ_i^2
A159	Ingenious	O	4.414	0.468	1.389
A161	Innovative	O	4.994	0.640	0.507
A162	Inquisitive	O	5.173	0.273	1.189
A165	Insightful	O	5.041	0.466	0.901
A168	Introspective	O	4.190	0.311	2.053
A170	Intuitive	O	4.874	0.359	1.158
A199	Narrow-minded	O	1.433	−0.258	0.573
A207	Open-minded	O	5.493	0.299	0.316
A212	Original	O	4.928	0.445	0.639
A220	Perceptive	O	5.093	0.320	0.816
A224	Philosophical	O	4.325	0.520	1.299
A249	Resourceful	O	5.363	0.324	0.404
A287	Studious	O	4.940	0.492	0.906
A288	Stuffy	O	1.677	−0.191	0.875
A305	Unconventional	O	3.428	0.347	1.807
A314	Unimaginative	O	1.763	−0.584	0.612

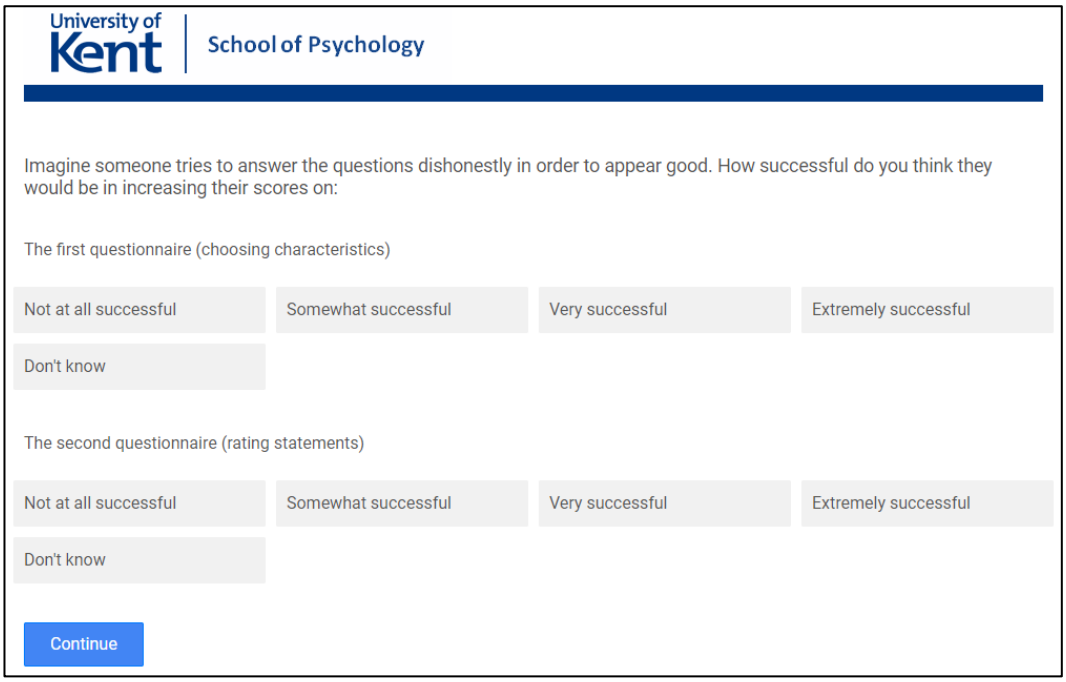
APPENDIX G: STUDY 6 PARTICIPANT FEEDBACK QUESTIONS



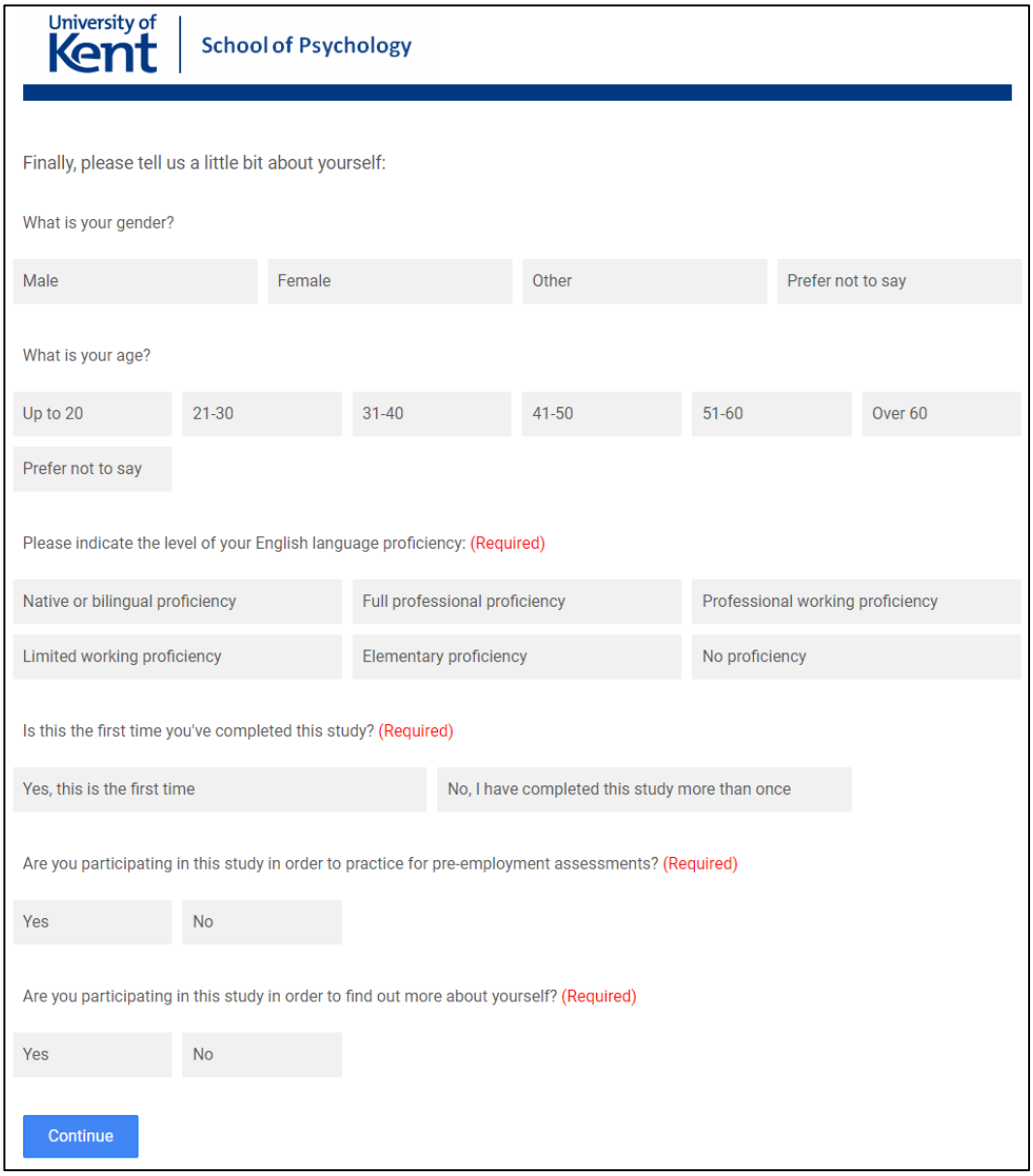



University of Kent
School of Psychology
Please consider the second questionnaire, where you were asked to rate statements on a scale.
I would be quite bored by a visit to an art gallery.
EXAMPLE
Strongly disagree
Disagree
Neutral
Agree
Strongly agree
Approximately, how frequently did you see...
A statement that was clearly desirable or undesirable in society?
0% of the time
25% of the time
50% of the time
75% of the time
100% of the time
Don't know
Continue


University of Kent
School of Psychology
Now, please compare your experience across the two questionnaires. Please indicate which of them:
Was easier to complete
Choosing characteristics
Rating statements
They're the same
Don't know
Made you think deeper about your own personality when answering
Choosing characteristics
Rating statements
They're the same
Don't know
Gave you a better chance to describe your personality fully
Choosing characteristics
Rating statements
They're the same
Don't know
Gave a more preferable test experience on the whole
Choosing characteristics
Rating statements
They're the same
Don't know
Made a fairer test for comparison between people
Choosing characteristics
Rating statements
They're the same
Don't know
Continue


University of Kent
School of Psychology
Imagine someone tries to answer the questions dishonestly in order to appear good. How successful do you think they would be in increasing their scores on:
The first questionnaire (choosing characteristics)
Not at all successful
Somewhat successful
Very successful
Extremely successful
Don't know
The second questionnaire (rating statements)
Not at all successful
Somewhat successful
Very successful
Extremely successful
Don't know
Continue

APPENDIX H: STUDY 6 PARTICIPANT BACKGROUND QUESTIONS



University of Kent | School of Psychology

Finally, please tell us a little bit about yourself:

What is your gender?

Male | Female | Other | Prefer not to say

What is your age?

Up to 20 | 21-30 | 31-40 | 41-50 | 51-60 | Over 60

Prefer not to say

Please indicate the level of your English language proficiency: (Required)

Native or bilingual proficiency | Full professional proficiency | Professional working proficiency

Limited working proficiency | Elementary proficiency | No proficiency

Is this the first time you've completed this study? (Required)

Yes, this is the first time | No, I have completed this study more than once

Are you participating in this study in order to practice for pre-employment assessments? (Required)

Yes | No

Are you participating in this study in order to find out more about yourself? (Required)

Yes | No

Continue