Dredging the OCEAN.20: An Item Response Theory Analysis of a Shortened Personality Scale

By
Joanna L. Solomon

Thesis Submitted to the Faculty of Graduate Studies at
Saint Mary's University, Halifax, Nova Scotia
in Partial Fulfillment of the Requirements for
the Degree of Master of Science in Applied Psychology

August 2013, Halifax, Nova Scotia
(c) Joanna Solomon, 2013

Approved: Dr. Lucie Kocum
Supervisor
Assistant Professor, SMU

Approved: Dr. Vic Catano
Supervisory Committee
Professor, SMU

Approved: Dr. Damian O'Keefe
Supervisory Committee
Assistant Professor, SMU

Approved: Dr. Saad Chahine
External Examiner
Assistant Professor, MSVU

Date: August 19, 2013

Dredging the OCEAN.20: An Item Response Theory Analysis of a Shortened Personality Scale

By Joanna L. Solomon

Abstract

Valid and reliable personality assessments are important tools for personnel selection, so long as they are efficient and free from bias (Mount & Barrick, 1995). Undergraduate students (N = 503) completed the OCEAN.20, a brief 20-item self-report measure of the five factors of personality (O'Keefe, Kelloway, & Francis, 2012). Classical test theory methods had already established the scale's reliability and validity, as replicated in the present study, but item response theory analyses identified nine problematic items. Three items displayed differential item functioning, three items had a truncated range of responses, and three more items had low precision. The potential for bias or insufficient information offered by each item is cause for concern, as it could have serious consequences in determining a job applicant's fate, so it is advised that these items either be removed or revised prior to operational use. Limitations and recommendations for future research are discussed.

August 19, 2013

Table of Contents

Dredging the OCEAN.20: An Item Response Theory Analysis of a Shortened Personality Scale

Personality research in recent years has been dominated by the five-factor model (FFM), or the "Big Five," a term that refers to the five traits commonly believed to explain human behaviour and individual differences (Muck, Hell, & Gosling, 2007). These traits are Extraversion (vs. introversion), Openness to experience (vs. closedness), Agreeableness (vs. antagonism), Neuroticism (vs. emotional stability), and Conscientiousness (vs. undependability; Costa & McCrae, 1992). Multiple measures exist to assess these traits, but the so-called 'gold standard' is currently the NEO Personality Inventory (Costa & McGrae, 1992). The revised version, or the NEO-PI-R, has demonstrated impressive reliability and validity in over 20 years of research on personality and has been used to predict a wide range out of outcomes. For example, a version of the NEO was used to predict students' misconduct (Thalmayer, Saucier, & Eigenhuis, 2011) and academic performance (Kappe & van der Flier, 2010), while another study used the NEO to successfully predict positive health behaviours (Hill & Gick, 2011). Using 240 items to measure the five traits with six facets each, a major drawback to the NEO-PI-R is its notable length. The length of the test limits its applicability for many organizations and industries that may be pressed for time or may need to assess a large volume of applicants.

Shortened versions of a FFM have been proposed, with varied success (e.g., Muck, Hell, & Gosling, 2007; Rammstedt & John, 2007). A shortened version of NEO, named the NEO-FFI, was developed by Costa and McCrae (1992), but even the 60-item version is now described as "tediously long" given that organizations are facing limited

assessment times (Rammstedt & John, 2007). There is a clear desire for an abbreviated version among both practitioners and researchers, but it is often difficult to achieve the same psychometric calibres in a condensed version. Muck, Hell, and Gosling (2007) argue that it may not be so difficult to strike a compromise between test length and test quality, pointing out that a loss in reliability (i.e., Cronbach's alpha) does not always indicate a loss of predictive validity. Muck and colleagues produced a ten-item personality inventory and deemed it an efficient approximation of the extant longer measures of the same traits. To date, it does not appear to be in widespread use, perhaps due to its relative youth, the minimalism of only including ten items, or a lack of large validation studies. In the field of personnel selection, in particular, it is important to acknowledge that short versions can display satisfactory reliability and validity, despite their brevity, as time-efficient measures can overcome the obstacles preventing the use of valuable personality tests in employment settings.

A similar and successful attempt at reducing a lengthy FFM scale was made by O'Keefe, Kelloway, and Francis (2012). The Trait Self-Descriptive Inventory (TSD) was initially developed in 1994 for use with military recruits (specifically, the United States Air Force), but the complete scale comprised 163 items. Upon concluding that there was "no universally acceptable short measure of the Big Five" (O'Keefe, Kelloway, & Francis, 2012, p. 435) and in an effort to make the scale more practical for organizational use, the scale underwent principal component analysis, exploratory structural equation modeling, and confirmatory factor analyses to identify a small number of the best items to represent each factor. Three versions were compared for best fit (i.e., a 15-item version, a 20-item version, and a 25-item version). The 20-item version emerged as the

best solution, resulting in what O'Keefe and his colleagues referred to as the OCEAN.20. The OCEAN.20 was then validated and analyzed further, demonstrating acceptable reliability and predictive validity coefficients in terms of workplace deviance as an outcome (O'Keefe, Kelloway, & Francis, 2012).

There is clear evidence to support the continued use of personality testing in personnel selection, and organizations show no signs of removing the common practice (Catano, Wiesner, Hackett, & Methot, 2010). Because personality is demonstrably stable, the predictive power of a personality test is especially appealing. Schmidt and Hunter (1998) caution that using selection methods low in validity could result in millions of dollars lost through reduced performance. In response, research should focus on maximizing the benefits of existing measures and ensuring the tests are as psychometrically sound as possible.

The proposed study seeks to extend this body of knowledge on brief personality inventories, particularly for the OCEAN.20. Shortened scales are increasingly useful and certainly in demand, yet many existing measures of personality are still too unwieldy. The next step, then, is to answer the remaining questions about the scale and item properties. Although the goal now is not to reduce the OCEAN.20 any further (unless otherwise necessary), the present aim is to obtain high-functioning items and confidence in the information they provide. The OCEAN.20 has been subjected to classical test theory (CTT) analyses, and O'Keefe and colleagues (2012) have provided information about the functioning of the scale at a group level; however, more information can be gleaned from item response theory (IRT), a more modern psychometric approach. IRT was originally designed with aptitude assessment in mind, yet it has proven to be a

valuable tool for personality researchers as well (Reise & Waller, 2003), and its numerous advantages will be outlined below.

The Five Factors of Personality

Common measures of personality since the 1990s have centered on the assessment of the "Big Five" dimensions of Extraversion, Neuroticism (also referred to as "emotional stability" in much of the personnel psychology literature, although the original OCEAN.20 publication used the term Neuroticism), Agreeableness, Openness to experience (sometimes described as "intellect"), and Conscientiousness. This structural model stems from a large number of factor-analytic studies that have drawn upon a range of scales (both self- and peer- reports) in a variety of samples, which resulted in the prominence of the present-day theory of the Big Five during the 1980s. (For a review, see Thalmayer, Saucier, & Eigenhuis, 2011.) Interest in the predictive power of personality assessments has persisted over many decades; meta-analyses support the modern model of personality and work performance, in which Conscientiousness is a strong predictor of desirable workplace outcomes (Poropat, 2009). Discovering the predictive validity of personality has generated increased interest in the field and, as a result, increased demand for greater precision in associated tools.

Personality and Selection

Meta-analytic findings have demonstrated that each of the Big Five personality dimensions are able to predict aspects of job performance with relative accuracy, in addition to displaying strong concurrent validity (i.e., measures of each of the Big Five have high correlation coefficients when compared to existing measures of similar constructs); in particular, Conscientiousness correlates most strongly with performance

(with coefficients ranging from r = .22 to .31; Salgado, 1997; Egberink, Meijer, & Veldkamp, 2010; Catano et al., 2010; Oh, Wang, & Mount, 2011; Rothmann & Coetzer, 2003). In addition to predicting performance, the Big Five have displayed moderate-to-strong predictive validity coefficients with important work-related outcomes, such as counterproductive workplace behaviours (O'Keefe, Kelloway, & Francis, 2012); for example, Extraversion and Openness have been observed to correlate most highly with leadership ability (Mount & Barrick, 1995; Catano et al., 2010). Salgado (1997) confirmed that Extraversion is a valid predictor of managerial ability (r = .18). Similarly, meta-analytic results have shown that Neuroticism in particular is a valid, yet modest, predictor for law enforcement occupations (r = .10; Oh, Wang, & Mount, 2011; Salgado, 1997). Furthermore, organizations such as the military are inspired to continue using personality tests as such assessments have also displayed links to personal discipline and maintaining physical fitness, which are certainly desirable traits for new recruits (White, Young, Hunter, & Rumsey, 2008).

Incorporating personality measures into employee screening and selection has significant economic utility in that it can attenuate costs associated with turnover and low job performance (White et al., 2008). On its own, the Conscientiousness dimension of personality testing has a criterion-related validity coefficient of .31 (Mount & Barrick, 1995). Moreover, when used in conjunction with cognitive ability measures, higher criterion-related validity is achieved than either type of test alone, nearly doubling the initial value to .60 (Catano et al., 2010; Mount & Barrick, 1995). Schmidt and Hunter (1998) conducted a meta-analysis on the validity of selection measures, and they found that tests of Conscientiousness led to a validity coefficient gain of .09 (or an increase of

18%), when added or used in supplement to cognitive ability tests. Conscientiousness, and personality assessments on the whole, are nearly as valid as interviews (r = .35) and slightly more useful than reference checks (r = .23), as evidenced by meta-analytic findings (Schmidt & Hunter, 1998).

Group Differences and Bias

A frequent concern of test developers and consumers is the potential for bias at both the item and test level. Sex-based differences have been observed in American samples across most of the Big Five dimensions; for example, women tend to score higher on Neuroticism and Agreeableness (Chapman, Duberstein, Sorensen, & Lyness, 2007; Lehmann, 2006). By contrast, men tend to score higher on measures of Intellectuality or Openness to experience (effects sizes range from d = .13 to .27; Ones & Anderson, 2002). These differences are typically modest in magnitude, fit gender-role stereotypes, and are consistent across 26 cultures (Lehmann, 2006). Gender differences persist throughout the lifespan, with elderly men scoring lower on Neuroticism and Agreeableness than elderly women (Chapman, et al., 2007). Some test developers have gender-related norms for scoring personality tests (e.g., the NEO Personality Inventory); however, these norms are not typically used in selection decisions (Powell, Goffin, & Gattatly, 2011). Given the previous research into sex differences, I hypothesize:

Hypothesis 1a: For composite scores of Neuroticism, women will outscore men.

Hypothesis 1b: For composite scores of Agreeableness, women will outscore men.

Hypothesis 1c: For composite scores of Openness, men will outscore women.

Schmidt and Hunter's (1998) meta-analysis revealed there has been a lack of observed predictive bias over the years for women and minorities, but they do acknowledge that subgroup differences, such as for personality, may exist, albeit rarely or as small deviations. Although that meta-analysis occurred more than a decade ago, the inconsistent and varied results of sex differences in personality testing have continued to be tested (for examples, see Ones & Anderson, 2002). Personality measures are often recommended by experts for their lack of differential selection rates, yet Powell, Goffin, and Gallatly (2011) assert that "demographic group differences in personality traits do exist" (p. 106), and McClarty (2006) revealed numerous items with different response patterns between men and women. The discrepancy in the research about personality-related differences between men and women may be due to the varied tools; while few differences are observed when testing broad traits, the differences between groups become more apparent on narrow traits, or facets of the Big Five (Powell, Goffin, & Gallatly, 2011). Because there is the possibility of significant sex differences, and therefore, the potential for bias in the measure, it is important to identify if any items appear to favour one group over another.

Development of the OCEAN.20

The OCEAN.20 (O'Keefe, Kelloway, & Francis, 2012) was derived from the Trait Self-Descriptive Inventory (TSD), which was an in-house assessment tool administered to military recruits, comprising 163 items that covered all dimensions and facets of the Big Five. Please note that although personnel psychology literature often refers to Neuroticism as emotional stability (Catano et al., 2010), the OCEAN.20 was developed and written in terms of Neuroticism, and for that reason, it will continue to be

employed in the present paper. Using principal component analysis and exploratory structural equation modelling, O'Keefe and colleagues retained the five items with the highest factor loadings for each factor of the TSD. Next, a confirmatory factor analysis (CFA) was conducted to contrast potential models (15-, 20-, or 25- item versions of the abbreviated TSD), resulting in the OCEAN.20. To demonstrate the predictive validity, a series of studies relied on samples of Canadian military recruits, contrasting the 75-item version of the original TSD with the shortened version, named the OCEAN.20. The final result was a shortened measure of the Big Five with strong factor structure and good internal consistency, while maintaining the same predictive value of its longer version.

The scale is composed of seven adjectives and 13 sentences, and respondents are invited to endorse each item on a seven-point ordered agreement scale of 'extremely uncharacteristic of me' to 'extremely characteristic of me'. Four of the five factors consist entirely of positively-worded statement (e.g., "I like to keep my belongings neat and organized"), while the final factor consists entirely of negatively-coded items for Extraversion, essentially resulting in a measure of introversion (e.g., "I am a very shy person").

This new scale is a promising tool for both researchers and organizations, particularly for the military as new recruits were used in validation studies. In order to maximize the benefit of an instrument with such promise, further investigation may be necessary. For example, O'Keefe, Kelloway, and Francis (2012) recognized that the high observed reliability values could be coming at the expense of reduced content validity. Abbreviated scales often fall prey to concerns of abbreviated content validity, as it has been argued that 10 items, for example, may not be enough to tap all the broad and

narrow factors that longer measures can assess (Muck, Hell, & Gosling, 2007, O'Keefe, Kelloway, & Francis, 2012). Shortened selection measures are appealing for researchers and organizations alike, but they are still not widely used, perhaps due to its very recent release and therefore limited evidence of validity. One possible solution to these concerns could stem from an in-depth analysis of the items themselves, in order to ensure they are performing as desired at both the item and test level, and that the test, as a whole, is performing adequately across various trait levels.

Item Response Theory

Item response theory (IRT) is a modern perspective on testing that offers detailed test and item information not available from classical test theory methods (Embretson & Reise, 2000; Ferrando & Chico, 2007). IRT refers to a family of psychometric models and methods that evaluate items and traits empirically. Advances in psychometrics and technology have led to improved techniques for evaluating assessment instruments, extending beyond classical test theory (CTT) by providing multiple parameters to estimate ability on a latent trait. CTT is limited, primarily, by its lack of precision and inability to separate personal ability and item difficulty (Hayes, Morales, & Reise, 2000).

IRT methods estimate the latent, or unobserved, trait upon which test performance is predicated. One of the earliest models in IRT is known as the Rasch model, which allows for one parameter of item information. In the Rasch model, also referred to as one parameter logistic function or 1-PL, items are assumed to vary in difficulty, while their discrimination ability or the item's ability to separating successful from non-successful test-takers, is equal or fixed (Hayes, Morales, & Reise, 2000). This model is commonly used for dichotomous test items (responses limited to two options, such as yes/no or

correct/incorrect). Later, a second model emerged, extending the 1-PL model by adding estimations of varied item discriminations to the existing method of estimating difficulty. The 2-PL model adds another parameter into the analysis of dichotomous and polytomous tests, particularly for those where there are multiple response options, yet there is no correct choice and therefore cannot be successfully answered by guessing at random. One particular version of the 2-PL model is called the Graded Response Model, which is designed for use with ordered categorical responses, such as personality scale items (Hayes, Morales, & Reise, 2000). Another model, 3-PL, includes a third parameter that estimates the impact of chance or degree of guessing involved in responding to the question (Hayes, Morales, & Reise, 2000). Such a model is suited for ability testing, such as cognitive ability multiple choice exams, wherein test-takers can answer right simply by chance, which will affect the difficulty and discrimination of the item.

An individual's performance on a given item on a successful test should be predicted by latent traits (Griffith et al., 2009). This relationship, between item performance or response and the level of a latent trait, is described by item characteristic curves, which are unique functions produced by IRT. Specifically, IRT distinguishes between an item's difficulty (*b*-parameter) and the item's discriminating power (*a*-parameter), and some models include a third parameter for guessing on multiple-choice tests that are dichotomously-scored, such as achievement or educational tests (Webster & Jonason, 2013). Item difficulty, for polytomous (i.e., graded or multiple response formats, as opposed to a single correct answer, such as ratings of agreement or frequency) scales, reflects the level of the latent trait needed to be more likely to endorse the next higher value in the scale to endorse the item. For example, an individual with a low level of

Conscientiousness should be very unlikely to give a high rating to a Conscientiousness item, if the item is operating properly. The chosen rating should represent the underlying trait level for individuals, and item difficulty coefficient in IRT describes this relationship with the *b*-parameter (Griffiths et al., 2009). Item discrimination reflects the item's ability to differentiate between different individuals who may have similar levels of the latent trait, such that highly discriminating items provide more information about the respondent (Hayes, Morales, & Reise, 2000). Discrimination also reflects the strength of relation between the item and the latent trait, analogous to a factor loading in confirmatory factor analysis. In CTT, a similar goal is accomplished with item-total correlations (or point-biserial correlations) that can identify whether a high score on the item is correlated to a high score on the test as whole (Webster & Jonason, 2013).

IRT has important advantages over CTT; namely, IRT offers a substantial amount of more information about the items and traits of interest by including multiple parameters that are not tied to the sample, resulting in greater accuracy and therefore confidence in test scores (Griffiths et al., 2009). The item population parameters are invariant with respect to the ability distributions of examinees (given there is a close fit between the IRT model and the data set), enabling examinee ability estimates to be independent of the test items (Ellis & Mead, 2002). Therefore, IRT is not sample-dependent, whereas with CTT, difficulty and discrimination indices are defined and limited by the sample (Ellis & Mead, 2002). The parameter estimates that result from such a model do not depend on the sample, simply due to the fact that the standard error of measurement is variable, thus making the IRT model relatively generalizable to the entire population from which the sample was drawn, rather than being sample-specific

(Embretson & Reise, 2000). Naturally, this benefit is only applicable to an appropriate sample that approximates the population of interest, rather than a homogenous or imbalanced sample. Different heterogenous subsets of the same population have been observed to vary greatly in terms of CTT item parameter estimates, yet estimates from IRT models are robust and would be readily applicable to a second sample from the same population (Hayes, Morales, & Reise, 2000; Rouse, Finger, & Butcher, 1999). The oft-quoted notion that IRT could provide a 'sample-free' estimate of item parameters stems from some of the earliest research on the method (Wright & Panchapakesan, 1969) and is maintained by contemporary IRT experts (see Embretson & Reise, 2000).

Ferrando and Chico (2007, p. 237) summarize the difference between IRT and CTT in the assertion that IRT provides "greater accuracy in the estimation of individual trait levels." The advantage of precision comes from the additional parameters and standard error of measurement values that reflect the respondents rather than a fixed degree of error, which is used in IRT models, as compared with the functions employed by CTT. IRT is able to give weight to item difficulty, item discrimination, and guesswork of the individual, allowing for significant information over and above the raw score alone. Because IRT relies on the measurement theory of modeling latent traits, Hayes, Morales, and Reise (2000) explain clearly that raw scores do not suffice in estimations of trait scores. When looking at precision of a scale or test, IRT methods provide information, such as a test and item characteristic curves, that varies depending on the individual's trait level. CTT, on the other hand, measures precision as test reliability, which assumes the same degree of error for all individuals, regardless of their trait level (Egberink, Meijer, & Veldkamp, 2010). The other value of precision provided by CTT is

R^2 values as an indicator of the amount of error in each item, while IRT offers a value for the degree of error at each level along the continuum of the latent trait for all items (Embretson & Reise, 2000). The standard error of measurement, as provided by IRT, is derived from a function of the test-takers' abilities (Griffith et al., 2009). Cronbach's alpha, the most common measure of test reliability and internal consistency, has been used by psychometricians for decades as a quick and comprehensible measure. Scales, especially those related to decision-making and selection, are expected to demonstrate a high overall alpha in order to be considered legally defensible (Schmidt, Le, & Ilies, 2003). Nunnally (1978) recommended a minimum coefficient of .70, suggesting that high-stakes decisions should be based on scales with an alpha closer to .90. However, using alpha in this manner is far from perfect, as it varies due to the number of items included in the scale. More items, regardless of their quality, can artificially inflate coefficient alpha. Embretson and Reise (2000) describe the goal of IRT as achieving the smallest number of reliable items for a scale, and relying solely on Cronbach's alpha could lead to unnecessarily lengthy scales.

The increased precision of the measurement of reliability and error in the test can provide additional information above traditional or linear methods. For example, during the analysis of the Autobiographic Memory Test, the results demonstrated that the test had greater precision for those with low levels of the desired trait (Griffith et al., 2009). Finding variable degrees of utility of a test across differing levels of the latent trait is a unique advantage of IRT, unavailable through CFA and CTT methods, and this result is particularly relevant to personality testing, in which individuals are likely to vary significantly across the target latent traits. Related to issues of error, another important

distinction is that IRT results are not bound by the characteristics of the population; instead, IRT provides flexible estimates of sampling error such that item information or ICCs do not vary between samples (Hayes, Morales, & Reise, 2000). In contrast, CTT estimates can vary considerably between subgroups of a population and between samples.

In addition, traditional factor analysis can produce misleading results when applied to personality items (i.e., when the underlying content of items is non-linear) (Egerbink, Meijer, & Veldkamp, 2010). In CFA, factor loadings are homomorphic to item discrimination estimates, and conceptually, the intercept in CFA is similar to item difficulty, but only when the latent construct equals zero (Meade & Lautenschalger, 2004). Although similar information can be found using factor analytic methods, IRT goes further and is able to produce several *b*-parameters for each item, rather than being fixed at the latent construct's zero point. Therefore, if items differ in their *b*-parameters, or degree of difficulty, CFA has been deemed inadequate at identifying such differences between items (Meade & Lautenschalger, 2004). The main recommendation of Meade and Lautenschlager is that CFA is most useful for small sample sizes, a small number of items, or questions regarding the relationship between multiple latent factors. IRT, on the other hand, is suited for providing detailed information about specific items or single scales, especially with large sample sizes. For example, in their 2004 study on the Autobiographical Memory Test, Griffith and colleagues determined that CFA returned a well-fitting model, but IRT revealed further information, such as how the test is not "maximally informative" (p. 622) at all levels of autobiographical memory, the latent trait of interest. In other words, the difficulty and discrimination abilities of items vary across levels of the latent trait, meaning that the test provides differentially useful information

about test-takers, depending on their underlying memory skills. These findings could only be revealed through IRT.

Test scores analyzed with CTT can be interpreted as the summation of an individual's observed score on a given test and a fixed error value (that is, random degree of error, held constant across the entire sample), where the goal is to estimate one's true score by minimizing the error term as much as possible (Rouse, Finger, & Butcher, 1999). To minimize error, the traditional recommendation is to use multiple measurements of the same trait, such that individuals take parallel forms of a single trait measure and their observed score becomes stable (Hayes, Morales, & Reise, 2000). However, including an infinite number of items in order to approach one's true score through this method is impractical and unrealistic (Rouse, Finger, & Butcher, 1999). In practice, occupational psychologists attain multiple measurements through the use of structured interviews, reference checks, and other validated selection tools alongside a personality assessment, but even using multiple methods, error can never be reduced completely to zero (Catano et al, 2010). It is possible to calculate the standard error of measurement (SEM) for an entire test, to provide a snapshot of the instrument's precision. Rouse (1999) identified a problem with relying on SEM, though: the single value may not accurately represent scales that are highly reliable only for individuals with a high trait level. Reliability and precision coefficients for individuals with low trait levels are not reflected by traditional approaches of CTT in this instance. IRT is well-suited to address these concerns and limitations, as the information gleaned from IRT is not sample-dependent and scale precision can be assessed at any and every level of the

trait of interest, rather than depending on a total, summed score (Rouse, Finger, & Butcher, 1999).

The relative sample-independence of IRT allows for reliable comparison between subsets and within a given population, although it is still limited by sampling error and response bias, as with any measurement methodology. In summary, "because of these weaknesses of the CTT approach, and because IRT addresses these weaknesses, the field of personality assessment would benefit from moving away from an exclusive reliance on the CTT approach and moving toward utilizing the valuable psychometric tools provided by IRT" (Rouse, Finger, & Butcher, 1999, p. 285). The present study will build upon the recommendation of Rouse and colleagues and apply IRT techniques to a personality assessment to extend the growing body of literature on the advantages of IRT. Similar to Cronbach's alpha, a prerequisite of IRT is unidimensionality, and the established dimensions of the FFM of personality will lend themselves nicely to this type of analysis (Hays, Morales, & Reise, 2000).

CTT is also limited in its ability to identify subgroup differences, and for these reasons, IRT is valuable and preferable for high-stakes applications, such as personnel selection. CTT techniques of detecting group differences tend to be inconsistent, such that "item means are confounded by valid group differences, and item-scale correlations are affected by group variability" (Hayes, Morales, & Reise, 2000, p. 32).

IRT methods afford effective and appropriate analyses of group differences or bias. Unidimensional IRT models operate on the assumption that likelihood of endorsing an item is similar for all persons in the sample (Egberink, Meijer, & Veldkamp, 2010). However, this assumption may not hold true, as various groups may interpret or react to

an item, construct, or test in systematically different ways. Differential item functioning (DIF) is an important part of an IRT analysis, as it reveals whether the item is behaving similarly across groups. Evidence of DIF is not a sufficient indicator of bias, but it certainly can highlight potential problem areas (Wu, King, Witkiewitz, Racz, & McMahon, 2012). Typically, groups are identified by demographics, such as gender or race, as bias related to these groupings could have legal implications in the realm of personnel selection. When groups have different probabilities of endorsing an item, it may suggest some bias to the item, which can manifest in personnel selection as mistakenly selecting for one group of applicants over another group, due to their test scores that are not reflecting true differences; in other words, one group may have a higher overall mean test score because of its propensity for higher scores on certain items—*at the same level of the trait*. IRT is expertly suited to find answers to these concerns, as Egberink and colleagues (2010) provided further support for identifying psychological or measurement-related differences, rather than differences based on demographics. It is crucial to identify where the differences originate in order to attenuate the observed bias and perceived fairness.

Sheppard, Han, Colarelli, Dai, and King (2006) examined gender differences on the Hogan Personality Inventory (HPI), a personality inventory used in selection that contains 206 items, with a few constructs that may overlap with the Big Five (i.e., likeability, sociability, and prudence). They identified a total of 53 items that displayed DIF, representing approximately 30% of each subscale (Sheppard et al., 2006). Because the pattern of bias was not systematic, the moderate amount of DIF observed was deemed not to affect the test's quality, although some recommendations were made to revise

items with more gender-neutral wording. Items should behave similarly for men and women, even if there are true sex differences. Essentially, discovering differential response patterns between the sexes would suggest that each sex perceives or expresses that construct differently. If it was truly the same construct to all members of all groups, the items, under an IRT framework, would behave in the same way. Sheppard and colleagues' (2006) evidence for DIF on items tapping personality constructs, along with other examples of mean differences between men and women on various dimensions of personality assessments (e.g., Woods & Hampson, 2010; Chapman, Duberstein, Sorensen, & Lyness, 2007), is sufficient to believe the OCEAN.20 may be susceptible to similar response patterns.

Should bias exist at the item level in the OCEAN.20, there could be serious implications in terms of decision-making and selection. Although the NEO is considered the gold standard for personality testing, it relies on separate standards and norms for men and women. Digging deeper than mean differences between the sexes, McClarty (2006) discovered DIF on a large proportion of items from a version of the NEO; specifically, items such as "Sympathetic to the homeless" favoured women, in line with gender stereotypes that portray women as caretakers. Inspired by this finding, I hypothesize:

Hypothesis 2a: Differential item functioning will be observed for the tender-minded Agreeableness item "SYMPATHETIC" such that women will be favoured.

Women also consistently rated anxiety-focused items from the Neuroticism factor more favourably than men, as these items "fit with the stereotype of women being emotional and vulnerable" (McClarty, 2006, p. 68). Smith and Reise (1998) also found that items about stress were easier for women to endorse than men. In addition, men were observed to endorse items on the intellect facet of Openness to experience more often than women, showing a preference for theoretical and abstract items, such as the reverse-coded item "Am not interested in theoretical discussions" (McClarty, 2006). These findings pave the way for the following hypotheses:

Hypothesis 2b: Differential item functioning will be observed for the anxiety-focused Neuroticism item "STRESS" such that women will be favoured.

Hypothesis 2c: Differential item functioning will be observed for the abstract Openness to experience item "THEORETICAL" such that men will be favoured.

Ideally, strong items on a scale should reflect the same traits for both sexes. The potential for bias is especially worrisome when an assessment tool is being in a high-stakes situation, such as selection or career placement. If the test results consistently favour one group over another, one group is now disadvantaged and selected for with lower frequency.

The Present Study

Because IRT is believed to better reflect actual response patterns than previous testing theories (Hayes, Morales, & Reise, 2000), the present study seeks to apply

modern testing theory methods to a modern personality assessment, the newly developed OCEAN.20. CTT does not show the utility of certain items or analyze whether the item is fully capturing the trait. White and colleagues (2008) recommend that future research should address the issues of adapting personality tests for operational use, and the OCEAN.20, in particular, is in position to be exactly that sort of valuable tool. Although it was originally derived from a military assessment tool, evidence of the test's success would benefit any organization that is interested in superior selection tools. Identifying potential bias is another vital step in creating superior tools, and in response to a call from Powell, Goffin, and Gettatly (2011) to maximize prediction and minimize adverse impact, the present study seeks to extend the existing body of literature surrounding personality testing in selection settings.

Shortened scales are particularly appealing for organizational use due to their relative time- and cost-efficiency. Thalmayer, Saucier, and Eigenhuis (2011) pitted scales of various lengths against each other and, using CTT approaches to reliability (such as Cronbach's alpha), found that the shortest Big Five inventory used was a "substantially better predictor of GPA than some longer ones" (p. 1006). Not only are shortened scales preferable, but it is possible they can be just as reliable without compromising predictive power. It is hard to imagine, then, why organizations would ever opt for longer tests; as a result, there is a pressing need for a thoroughly reliable and valid brief personality scale, such as the OCEAN.20.

The present study seeks to provide further support for the newly developed OCEAN.20 personality instrument. The main research question is that of investigating the item-level properties of the scale. Given the method of item selection (principal

component analysis; O'Keefe, Kelloway, & Francis, 2012), most items are expected to have acceptable, if not quite high, discrimination values, such that they are successful differentiators between those who are high and those who are low on the trait of interest. More specifically, based on previous evidence from existing research into personality scales, I put forth the aforementioned hypotheses, anticipating that women will be favoured by Neuroticism (1a) and Agreeableness (1b) items, while men are favoured by Openness items (1c). In addition, I hypothesize differential item functioning exists for sympathy-focused (2a) and anxiety-focused (2b) items, favouring women, while an abstract Openness item will be seen to favour men (2c).

Method

Participants

It is recommended that a minimum of 500 participants are needed to conduct reliable and valid IRT analyses in order to assure stability of the parameter estimates and reduce the standard error of those estimates (Embretson & Reise, 2000). The present study comprises a sample of undergraduate students from three medium-sized Eastern Canadian universities (N = 503). The mean age of participants was 21.11 years old (SD = 3.25) and the sample primarily identified as Caucasian (74.4%). The sample comprised 68.2% women (*n* = 343) and 30.6% men (*n* = 154), which is approximately representative of the Eastern Canadian undergraduate population. The overrepresentation of women in the sample does reflect the undergraduate population to some degree, and this imbalance does not risk influencing the analyses, since the smaller of the two comparison groups exceed the recommended minimum of 100 individuals (Scott et al., 2009).

Procedure

Undergraduate students were recruited via the school's internal research participation program (SONA), along with flyers in academic buildings and emails to alert students to the study. The study was presented as an online survey, via Qualtrics®. Students provided free and informed consent to participate in an REB-approved study (see Appendix A). Participants were asked to complete the OCEAN.20 scale and provide demographic details (see Appendix B). Upon completion, participants were presented with a feedback form, and as compensation for their time and effort, participants either received one bonus point toward their undergraduate psychology course or entered their name into a prize draw for one of three $100 dollar gift cards.

Measures

OCEAN.20. Participants completed the 20 items of the OCEAN.20 (O'Keefe, Kelloway, & Francis, 2012). The OCEAN.20 assess each of the Big Five personality factors with four items (see Table 1 for full item text), and participants endorsed items on an ordered agreement scale from "1: Extremely uncharacteristic of me" to "7: Extremely characteristic of me." Reliability, as measured by Cronbach's alpha internal consistency coefficient, was strong in this sample on all five factors: Openness ($\alpha = .78$), Conscientiousness ($\alpha = .88$), Extraversion ($\alpha = .87$), Agreeableness ($\alpha = .81$), and Neuroticism ($\alpha = .77$).

Demographics. Participants provided their sex, age, ethnicity, and current grade point average (GPA).

Results

Plan for Analysis

Data were analyzed with *SPSS 20*, *EQS 6.1*, and *IRTPRO 2.1*. First, data were cleaned and screened, and all assumptions must be checked. Because the sample comprised students earning bonus points, careful attention was paid to inattentive response patterns (see Appendix C for descriptive statistics enabling the comparison of various response times); however, due to the lack of differences resulting from varied response times, no cases were eliminated on this criteria alone. Composite scores for group differences were first compared with t-tests, under classical test theory. In terms of IRT, the graded response model (GRM) was used to analyze the OCEAN.20, due to its polytomous response options (Samuel, 2010). The GRM relies on two parameters, discrimination and difficulty, and it is the most common model for analyzing personality items (Hayes, Morales, & Reise, 2000; Samuel, 2010). For the sake of clarity, items were divided into groups by their factors and analyzed with GRM in separate unidimensional tests.

Discrimination parameter estimates essentially reflect the slope of the item characteristic curve, or the probability of endorsing the item at each level of the latent trait. Steeper curves are better able to discriminate between those high and low on the trait, while flatter curves, and therefore lower *a*-parameter values, are less sensitive to different ability levels (Baker, 2001). Guiding the interpretation of this scale, typical discrimination values range from 0.5 to 2.0, and *a*-values higher than 0.75 are considered successful discriminators between high and low performers on the latent trait (Hayes,

Morales, & Reise, 2000). Baker (2001) classifies discrimination values exceeding 1.70 as 'extremely high'.

For the difficulty parameter in a GRM for 7-point scale, there are six thresholds (for example, 'extremely uncharacteristic of me' vs. 'somewhat uncharacteristic of me' represents the first threshold, β_1). These parameters represent an item's location on the continuum of the latent trait; items that are very likely to be endorsed with only a small amount of the latent trait are considered "easy", and in turn, a "hard' item is only really functional for individuals higher on the latent trait (Baker, 2001).

In order to meet the assumptions of IRT, dimensionality of the scale was assessed with both exploratory and confirmatory factor analyses, and differential item functioning (DIF) tests were conducted. Detecting DIF is a three-step, iterative process: first, the factor's items are entered together as a broad sweep for potential group differences and candidates are identified as those with *p*-values greater than 0.05. Next, candidate items are tested while the other items of the factor are anchored, which serves as a test of non-uniform DIF. Non-uniform DIF is defined as statistically significant different *a*-parameters, or discrimination values, between the focal and reference groups. Should a difference be identified, this would indicate that the item is not related to the latent trait in the same way for men and women (Smith, 2002). The final step tests for uniform DIF, or invariance of difficulty or *b*-parameters between groups. In this step, discrimination parameters are constrained, essentially transforming the analysis into a 1-PL model, isolating the difficulty parameters so they can be evaluated properly. A difference in this step would identify unequal locations of the item for men and women, meaning the item would be easier to endorse by one group over the other (Smith, 2002).

Assumption Checking

Prior to all analyses, the data were screened and cleaned using the recommendations of Tabachnick and Fidell (2007). Outliers were identified for all four items of the Agreeableness factor, and these values were Winsorized prior to analysis. All items appeared to be normally distributed, as there was no significant skewness (absolute values did not exceed 1.20) or kurtosis (absolute values did not exceed 2.13) for any item in the OCEAN.20 scale. Shortened item names to be used throughout the paper and descriptive statistics are available in Table 1. Given that the sample drew upon more than one university, please see Appendix D for the descriptive statistics enabling the comparison of responses between the sub-samples.

Inter-item correlations for the 20 items of the scale are presented below in Table 2. Composite scores for each of the five factors on the OCEAN.20 were created, enabling a classical test theory comparison of mean group differences (see Table 3). Of note, men and women in the sample differed significantly on three out of the five factors, in line with previous findings. For Openness to experience, men (M = 4.37, SD = 1.37) scored significant higher than women (M = 3.53, SD = 1.34); t (494) = 6.37, p < .001, d =.62. In contrast, women (M = 5.75, SD = 0.82) significantly outscored men (M = 5.36, SD = 0.89) in terms of Agreeableness; t (495) = -4.80, p < .001, d = .46. Women (M = 4.64, SD = 1.28) also scored significantly higher than men (M = 3.49, SD = 1.26) on Neuroticism items, consistent with the literature and prior expectations; t (494) = -9.25, p < .001, d = .90.

Table 1.
Descriptive Statistics

Item Number	Item Name	Full Text of Item	Men		Women	
			M	*SD*	*M*	*SD*
1	SILENT	Silent	4.18	1.62	4.33	1.72
2	NEAT	Neat	4.80	1.61	4.85	1.45
3	SYMPATHETIC	Sympathetic	5.20	1.21	5.85	0.98
4	ORGANIZED	Organized	4.72	1.59	5.25	1.35
5	WITHDRAWN	Withdrawn	4.31	1.57	4.39	1.57
6	KIND	Kind	5.58	1.01	5.92	0.87
7	QUIET	Quiet	3.83	1.61	3.92	1.75
8	UNIVERSE	I have thought a lot about the origins of the universe	4.84	1.65	4.17	1.85
9	BELONGINGS	I like to keep all my belongings neat and organized	4.80	1.62	4.94	1.52
10	HEADACHES	I often have headaches when things are not going well	3.38	1.77	4.27	1.87
11	GENEROUS	I am always generous when it comes to helping others	5.33	1.16	5.55	1.14
12	STOMACH	Sometimes I get so upset, I feel sick to my stomach	3.37	1.69	4.70	1.72
13	SCIENCE	I am highly interested in all fields of science	4.27	1.78	3.53	1.78
14	PLACE	I like to have a place for everything and everything in its place	4.50	1.58	4.67	1.54
15	EVOLUTION	I am fascinated with the theory of evolution	4.63	1.71	3.94	1.90
16	STRESS	When I am under great stress I often feel like I am about to break down	3.58	1.72	4.85	1.67
17	TREAT	I always treat other people with kindness	5.36	0.98	5.77	0.94
18	FEELINGS	My feelings are easily hurt	3.62	1.55	4.72	1.63
19	SHY	I am a very shy person	4.27	1.76	4.17	1.78
20	THEORETICAL	I would enjoy being a theoretical scientist	3.73	1.86	2.50	1.56

Table 2.
Inter-item correlations

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	SILENT	1																			
2	NEAT	.13**	1																		
3	SYMPATHETIC	.01	.33**	1																	
4	ORGANIZED	.00	.69**	.37**	1																
5	WITHDRAWN	.54**	-.04	.04	.05	1															
6	KIND	.06	.22**	.59**	.26**	.12**	1														
7	QUIET	.76**	-.13**	-.07	-.07	.50**	-.01	1													
8	UNIVERSE	-.02	-.02	.08	-.05	-.10*	.03	.02	1												
9	BELONGINGS	-.09	.75**	.29**	.68**	-.08	.21**	-.10*	.06	1											
10	HEADACHES	-.05	.10*	.14**	.10*	-.14**	.10*	-.03	.04	.08	1										
11	GENEROUS	.09	.16**	.44**	.22**	.19**	.51**	.05	.09	.20**	.18**	1									
12	STOMACH	-.01	.04	.17**	.07	-.08	.13**	-.01	.04	.07	.52**	.11*	1								
13	SCIENCE	.05	.03	.02	-.01	-.03	.03	.05	.34**	.06	-.02	.07	.03	1							
14	PLACE	-.08	.53**	.19**	.51**	-.11*	.12**	-.08	.04	.68**	.16**	.20**	.12**	.15**	1						
15	EVOLUTION	.10*	-.03	.01	-.06	-.06	.00	.09*	.51**	.03	.03	.03	.07	.47**	.13**	1					
16	STRESS	-.10*	-.02	.19**	.04	-.17**	.13**	-.07	.03	.08	.48**	.11*	.55**	-.02	.17**	.04	1				
17	TREAT	.05	.22**	.49**	.28**	-.13**	.57**	.03	-.04	.25**	.12**	.56**	.12**	.00	.22**	-.06	.18**	1			
18	FEELINGS	-.20**	.04	.26**	.06	-.23**	.15**	-.22**	.05	.10*	.23**	.10*	.39**	-.03	.16**	.06	.53**	.16**	1		
19	SHY	.67**	-.08	.00	-.02	.52**	.05	.71**	-.03	-.10*	-.12**	.13**	-.08	.01	-.11*	.01	-.18**	.02	-.31**	1	
20	THEORETICAL	-.02	.02	-.10*	-.04	-.08	-.12**	.01	.36**	.03	-.07	-.09*	-.07	.67**	.13**	.47**	-.10*	-.16**	-.06	-.11*	1

Note: * denotes $p < .05$; ** denotes $p < .01$.

Table 3.
Mean differences between sexes for OCEAN.20 Big Five factors

	Men *M (SD)*	Women *M (SD)*	*t*
Openness	4.37 (1.37)	3.53 (1.34)	6.37***
Conscientiousness	4.70 (1.38)	4.92 (1.26)	-1.70
Extraversion	4.17 (1.35)	4.20 (1.46)	-0.21
Agreeableness	5.36 (0.89)	5.75 (0.82)	-4.80***
Neuroticism	3.49 (1.26)	4.64 (1.28)	-9.25***

Note: *** denotes $p < .001$.

Local dependence, or the assumption that remaining variance is not shared between pairs of items, was assessed with the LD $\chi 2$ test in *IRTPRO*. Pairs of items with values greater than 10 are considered to violate the assumption and therefore be redundant (Chen & Thissen, 1997), while values closer to zero are desirable. The assumption of local dependence was not violated in this sample, as values for all pair-wise comparisons did not exceed 8.50.

Confirmatory & Exploratory Factor Analyses

It is important to test the dimensionality of the scale prior to IRT analyses as part of the assumption-checking stage so that one is certain they are analyzing a singular latent construct (Samuel, 2010). The 20 items of the OCEAN.20 were subjected to confirmatory factor analysis (CFA) to ascertain the structure of the scale. Following the procedure of O'Keefe, Kelloway, and Francis (2012), item loadings were free to vary and factor variance was fixed to 1.0. The factors were allowed to correlate, and maximum likelihood estimates were used.

The maximum likelihood fit indices echo the results of the original publication; $\chi^2 (160) = 543.07$, $p < .001$, CFI = .906 and RMSEA = .072. These values represent

marginally good fit, despite not meeting the highest standards of fit indices (Hu & Bentler, 1999). Ideally, the chi-square value should not be significant; however, previous research has made note that large sample sizes typically result in significant chi-square statistics even in cases of acceptable model fit (Rouse, Finger, & Butcher, 1999; Hu & Bentler, 1999).

All items loaded significantly to the intended factors, confirming the basic structure of the Five Factor Model (see Table 4). The R^2 values, representing the variance in each item explained by its latent factor, were all above the recommended minimum of 0.10, which corresponds to the typical cut-off of .33 for factor loadings (Tabachnick & Fidell, 2007). In fact, the R^2 values ranged between .22 (UNIVERSE) and .82 (BELONGINGS), demonstrating adequate fit between items and their hypothesized factors. All factor loadings were significant and in the expected direction, further supporting the fit of the model.

Table 4.
Confirmatory Factor Analysis: Item Loadings and R-Squared Values

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	R^2
UNIVERSE	.47					.22
SCIENCE	.79					.62
EVOLUTION	.60					.36
THEORETICAL	.81					.66
NEAT		.84				.71
ORGANIZED		.79				.62
BELONGINGS		.91				.82
PLACE		.70				.49
SILENT (R)			.85			.73
WITHDRAWN (R)			.62			.39
QUIET (R)			.89			.78
SHY (R)			.81			.66
SYMPATHETIC				.70		.49
KIND				.78		.61
GENEROUS				.69		.47
TREAT				.76		.58
HEADACHES					.59	.35
STOMACH					.70	.50
STRESS					.81	.66
FEELINGS					.60	.36

Note: χ^2 (160) = 543.07, $p < .001$, CFI = .91, RMSEA = .07. "(R)" denotes reverse-coded item.

In addition, each factor was checked individually for unidimensionality using the exploratory factor analytic method of principal axis factoring. Openness to experience displayed the best fit with a one-factor solution; one factor accounted for 47.87% of the variance, and factor loadings for the four items ranged from .54 to .77. The four Conscientiousness items represented a unified dimension, accounting for 64.99% of the variance in the factor, and all four items loaded clearly on the single factor solution (factor loadings ranged from .68 to .92). The four Extraversion items accounted for 63.06% variance in a single factor, with factor loadings ranging from .62 to .87. Agreeableness was best described by a single factor, as the four items accounted for 53.31% of the variance and factor loadings ranged from .68 to .79. Finally, Neuroticism was also best fitted to a single factor, and the four items of Neuroticism had factor loadings that ranged from .55 to .82, accounting for 46.56% of the variance. These 20 items all displayed high loadings, easily exceeding the traditional minimum threshold of .33, demonstrating a strong fit to their respective factors (Tabachnick & Fidell, 2007). Furthermore, all five factors had a sizable proportion of their variance explained by their items, reaffirming the strength of a single factor solution and, therefore, the unidimensionality of each of the Big Five factors.

Because the factor analytic techniques sufficiently confirmed the dimensionality of the OCEAN.20, it is appropriate to proceed to an item response analysis, with confidence that each of the theorized five factors is truly a singular latent dimension.

Item Response Analysis

Differential item functioning (DIF). DIF is an assumption of IRT, given that IRT parameter estimations will be invalid if data comes from non-equivalent groups

(Embretson & Reise, 2000). The multi-group analysis compared the responses of men and women, using men as the reference group. Following the guidelines of detecting DIF in an IRT framework as outlined by Stark, Chernyshenko, and Drasgow (2006), the first step in DIF detection requires an initial sweep of the items, looking for items that emerge as potentially non-equivalent across the two sexes. Two factors, Conscientiousness and Extraversion, did not display any items with DIF at this stage, so no further DIF analyses were conducted for these two factors. The final parameter estimates for Conscientiousness and Extraversion are presented in Tables 5 and 6, respectively, although the adequacy of each item as they pertain to the personality dimension is interpreted in more detail when discussing specific parameter estimates.

Table 5.
Discrimination and difficulty parameters: CONSCIENTIOUSNESS

Items	*a*	β_1	β_2	β_3	β_4	β_5	β_6
NEAT	3.10	-2.42	-1.47	-0.88	-0.57	0.33	1.51
ORGANIZED	2.56	-2.71	-1.83	-1.10	-0.73	0.10	1.32
BELONGINGS	5.34	-2.07	-1.35	-0.84	-0.47	0.23	1.17
PLACE	2.02	-2.59	-1.59	-0.80	-0.30	0.61	1.69

Table 6.
Discrimination and difficulty parameters: EXTRAVERSION

Item	*a*	β_1	β_2	β_3	β_4	β_5	β_6
SILENT	3.50	-2.20	-1.05	-0.28	0.02	0.60	1.40
WITHDRAWN	1.48	-3.31	-1.72	-0.64	0.09	0.80	2.11
QUIET	3.84	-1.74	-0.80	-0.03	0.27	0.81	1.58
SHY	2.65	-1.78	-1.02	-0.36	0.09	0.60	1.53

Of the four items for Neuroticism, item 18 (FEELINGS) was identified as a possible DIF item, $\chi^2(7) = 13.50$, $p = .061$, so the next step was to test it as a candidate for DIF against strong, non-DIF items as anchors. The other three items that make up the

Neuroticism factor appeared to be stable and highly discriminating (*a*-parameters range from 1.20 to 2.54, which is toward the high end of typical values; Hayes, Morales, & Reise, 2000), so these three items were used as anchors. In this step, men (*a*-parameter = 1.77) and women (*a*-parameter = 1.04) were observed to have significantly different discrimination parameters; χ^2_a (1) = 5.70, p = .017. This test is also known as non-uniform DIF, identifying whether the *a*-parameters are non-equivalent, and in this instance, the item was significantly more discriminating for women than for men. The final step assesses the presence of uniform DIF, which determines if the difficulty parameters are significantly different between groups. For item 18, no significant differences were found between the two groups in terms of difficulty; χ^2 (7) = 6.30, p = .504. With an equal amount of the trait, men are more likely, although not significantly so, to endorse the item than women, as evidenced by their higher *b*-parameter values (e.g., as seen in Table 7, β_1 is -1.78 for men and -3.45 for women). Table 7 presents the DIF analyses for Neuroticism, while Table 8 summarizes the final item parameters, with the problematic item removed.

Table 7.
Discrimination and difficulty parameters: NEUROTICISM

Items	*a*	β_1	β_2	β_3	β_4	β_5	β_6
Men							
HEADACHES	1.20	-1.51	-0.57	0.05	0.96	1.94	2.98
STOMACH	1.53	-1.36	-0.48	-0.04	0.80	1.85	3.18
STRESS	2.54	-1.48	-0.55	0.07	0.57	1.20	2.00
FEELINGS	1.77	-1.78	-0.75	-0.08	0.57	1.69	3.62
Women							
HEADACHES	1.51	-1.90	-0.80	-0.29	0.12	0.95	2.03
STOMACH	2.25	-1.93	-1.00	-0.51	-0.14	0.59	1.60
STRESS	2.36	-2.09	-1.14	-0.61	-0.23	0.51	1.43
FEELINGS	1.04	-3.45	-1.83	-1.06	-0.37	0.97	2.34

Table 8.
Discrimination and difficulty parameters: NEUROTICISM without DIF item

Items	*a*	β_1	β_2	β_3	β_4	β_5	β_6
HEADACHES	1.73	-1.74	-0.82	-0.35	0.14	0.90	1.86
STOMACH	2.60	-1.67	-0.91	-0.50	-0.06	0.62	1.51
STRESS	1.92	-2.15	-1.19	-0.60	-0.18	0.56	1.52

The factor of Openness also displayed DIF during the initial sweep. Item 8 (UNIVERSE) stood out as a candidate for DIF; χ^2 (7) = 13.80, p = .054. For the next step, item 8 was tested as a candidate, and the other three items were performing well enough to be used as anchors. UNIVERSE did not display non-uniform DIF, as tested in the second step; χ^2_a (1) = 1.40, p = .223. However, in the third step to test for uniform DIF with the 1-PL model, this item appeared to have significantly different difficulty parameters between the two groups; χ^2 (7) = 15.80, p = .027. Men had significantly higher difficulty values for the lowest level of Openness to experience (*b*-parameter values were -3.19 for men and -3.55 for women), but the opposite was true at the next threshold, and instead, women began to outscore men at β_2 (-1.84 for women and -2.37 for men). At the lowest levels of the latent trait, it appears as though men are more likely to endorse the item or give a higher score. However, this pattern reverses at the next theta threshold, in which women are more likely to award a higher score to the item, despite having an equal amount of the latent trait. Results for these DIF steps are presented in Table 9, with the final version of the factor in Table 10.

Table 9.
Discrimination and difficulty parameters: OPENNESS

Items	*a*	β_1	β_2	β_3	β_4	β_5	β_6
Men							
UNIVERSE	1.26	-3.19	-2.37	-1.21	-0.51	0.45	1.46
SCIENCE	2.50	-1.81	-1.08	-0.48	0.05	0.75	1.36
EVOLUTION	1.48	-2.63	-1.65	-1.11	-0.22	0.60	1.51
THEORETICAL	3.58	-1.16	-0.52	-0.21	0.41	0.91	1.59
Women							
UNIVERSE	1.00	-3.55	-1.84	-1.29	-0.65	0.43	1.67
SCIENCE	2.56	-1.99	-1.11	-0.66	-0.19	0.45	1.26
EVOLUTION	1.56	-2.30	-1.48	-0.93	-0.52	0.34	1.26
THEORETICAL	2.62	-1.09	-0.46	0.01	0.65	1.18	1.85

Table 10.
Discrimination and difficulty parameters: OPENNESS without DIF item

Items	*a*	β_1	β_2	β_3	β_4	β_5	β_6
SCIENCE	2.70	-1.43	-0.60	-0.11	0.37	1.02	1.75
EVOLUTION	1.38	-1.98	-1.08	-0.51	0.07	0.99	1.97
THEORETICAL	3.02	-0.61	0.01	0.41	1.00	1.48	2.12

Finally, the four items of Agreeableness were subjected to a sweep, and item 6 (KIND) was identified as a potential source of DIF; χ^2 (5) = 13.10, p = .022. When the other three items in the factor were anchored, item 6 indicated non-uniform DIF, highlighting a marginally significant difference between the discrimination values of the two groups,; χ^2_a (1) = 3.30, p = .070. The discrimination value was significantly higher for women (1.77) than men (1.68). As for uniform DIF, KIND also has marginally significantly different difficulty parameters between the two groups, as tested by the 1-PL model; χ^2 (5) = 10.08, p = .055. Women were much more likely to endorse the item than at all levels of Agreeableness (e.g., β_1 for men was -2.50, while for women it was -1.74). In summary, KIND was found to have non-equivalent discrimination and difficulty

parameters, resulting in a much more discriminating item for women than for men (*a*-parameters were 3.36 vs. 1.98, respectively). At all thresholds, women had significantly higher *b*-parameters, indicating they were more likely to assign a higher rating to this item than men, when the latent trait was held constant. The factor's DIF results can be seen in Table 11, and a final version of the factor is presented in Table 12.

Table 11.
Discrimination and difficulty parameters: AGREEABLENESS

Items	*a*	β_1	β_2	β_3	β_4	β_5	β_6
Men							
SYMPATHETIC	1.68	-2.72	-1.78	-1.11	0.14	1.81	N/A
KIND	1.98	-2.50	-1.39	-0.29	1.24	N/A	N/A
GENEROUS	2.10	-2.75	-1.92	-1.05	0.04	1.37	N/A
TREAT	2.20	-2.27	-1.21	0.11	1.61	N/A	N/A
Women							
SYMPATHETIC	1.77	-3.03	-2..04	-1.49	-0.30	1.57	N/A
KIND	3.36	-1.74	-1.38	-0.27	1.37	N/A	N/A
GENEROUS	1.85	-2.50	-1.64	-0.94	0.37	1.67	N/A
TREAT	2.34	-1.96	-1.09	-0.05	1.60	N/A	N/A

Table 12.
Discrimination and difficulty parameters: AGREEABLENESS without DIF item

Items	*a*	β_1	β_2	β_3	β_4	β_5	β_6
SYMPATHETIC	1.53	-3.42	-2.40	-1.76	-0.55	1.31	N/A
GENEROUS	2.13	-2.82	-2.02	-1.27	-0.11	1.14	N/A
TREAT	2.64	-2.33	-1.42	-0.35	1.14	N/A	N/A

Parameter estimates. Beyond looking for DIF, the *a*-parameter for each item was examined for poor discrimination; typically, item discrimination values range from 0.5 to 2.0, and *a*-values higher than 0.75 are considered successful discriminators between those high and low on the latent trait, while higher values are even better (Hayes, Morales, & Reise, 2000). For all five factors, the *a*-parameter estimates ranged from 1.38 to 5.34,

indicating excellent discriminating ability, so no items were flagged for review based on discrimination criteria. Consistent with earlier expectations, all items displayed high discrimination values, likely due to the method of item reduction (principal component analysis), as initially performed by O`Keefe, Kelloway, and Francis (2012).

The *b*-parameter is measured in terms of theta, or the latent trait, and the mean level of the latent trait is set to 0.00 and variance is fixed to 1.00, resulting in a normal distribution of theta, which typically ranges from about -3.00 to 3.00 (Hayes, Morales, & Reise, 2000). Most items had thresholds that spanned the entire range, demonstrating that the item captures both the low and high ends of that trait's spectrum. In three instances on the factor of Agreeableness (i.e., SYMPATHETIC, GENEROUS, and TREAT), however, the *b*-parameters were truncated, extending from approximately -2.50 to 0.10. (For a complete table of all final parameter estimates, see Appendix E.)

Trace lines for all items are presented in Figures 1-17. Two figures are presented for items displaying DIF, to compare the trace lines between men and women (see Figures 18-20). The trace lines, mapping the logistic function of each item's response patterns, illustrate the discrimination and difficulty parameters. The slope or steepness of each curve represents the discrimination of the item, and the left-right or lateral positioning of the curve represents whether the item was easy or difficult, graphed in relationship to the latent trait or theta (Baker, 2001). In addition to interpreting the values as presented in the tables, these figures display all possible response options, providing a clearer picture of the item's functioning in one view, which can aid in comparing and contrasting the utility of each item (Baker, 2001). One example of a strong item is BELONGINGS (see Figure 7). In this figure, each characteristic curve is distinct and has

non-overlapping peaks, indicating the clear distinctions between thresholds and a normal distribution of the *b*-parameters. The peaks of the curves are also quite high, which reflects the strong discrimination ability of this item, and this pattern of results can be seen in many of the other OCEAN.20 items as presented in the figures below. In contrast, EVOLUTION (see Figure 13) has relatively low peaks and shallow slopes, representing weaker discrimination power, and the curves overlap and are difficult to discern from one another, demonstrating that the difficulty values for this item around the mean of the latent trait are muddled and seem to a less precise fit to the model. Similar patterns of lesser fit can be seen in Figure 3, 5, 8, 9, 14.

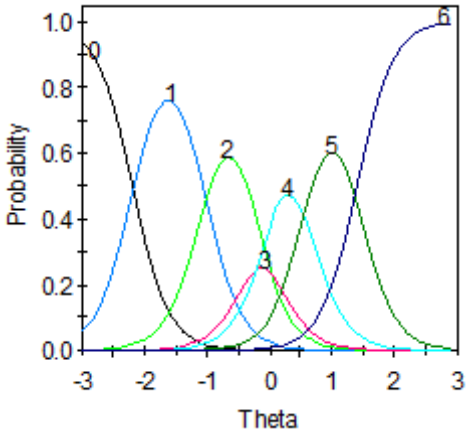


Figure 1. Trace lines for SILENT.

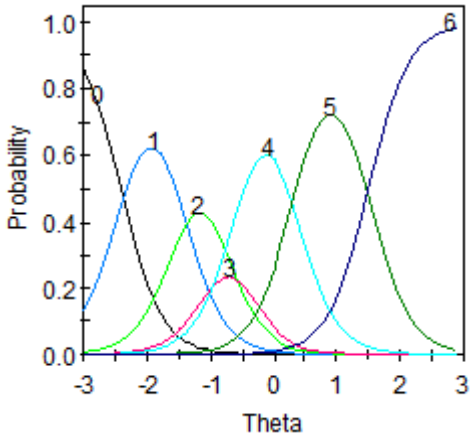


Figure 2. Trace lines for NEAT.

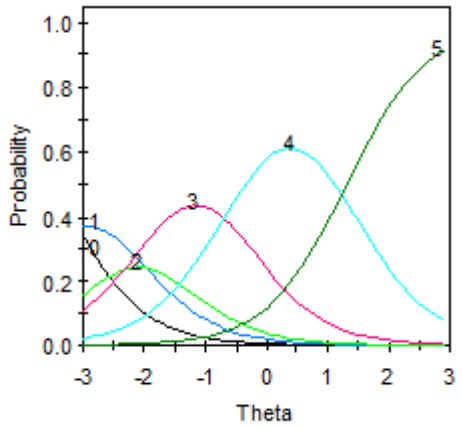


Figure 3. Trace lines for SYMPATHETIC.

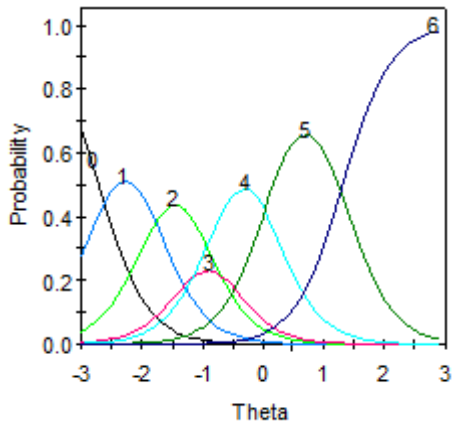


Figure 4. Trace lines for ORGANIZED.

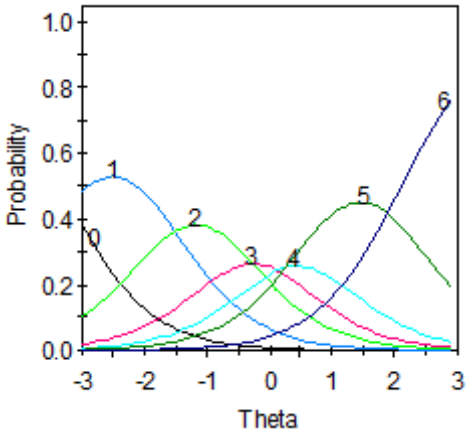


Figure 5. Trace lines for WITHDRAWN.

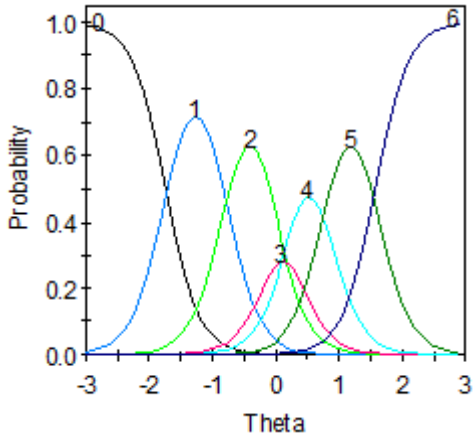


Figure 6. Trace lines for QUIET.

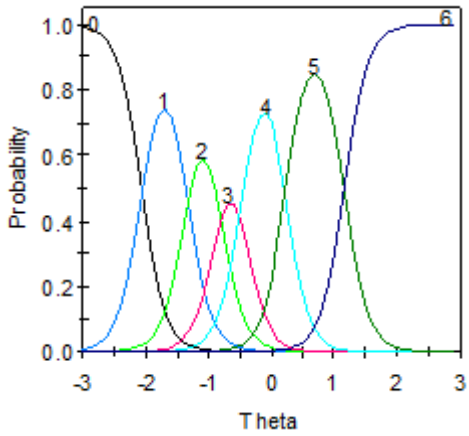


Figure 7. Trace lines for BELONGINGS.

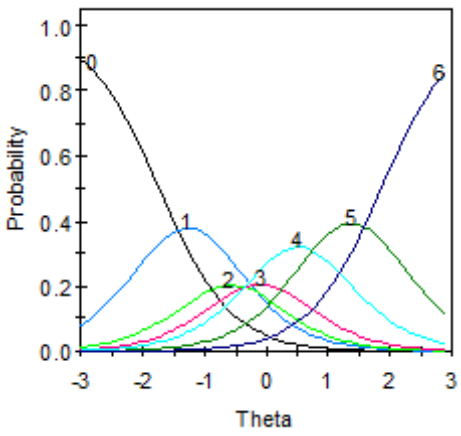


Figure 8. Trace lines for HEADACHES.

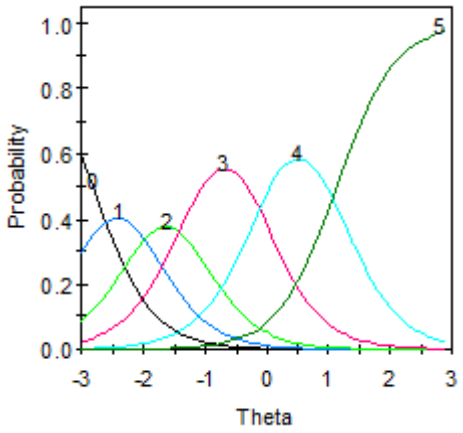


Figure 9. Trace lines for GENEROUS.

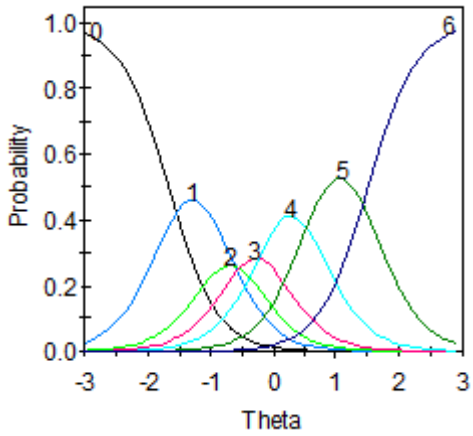


Figure 10. Trace lines for STOMACH.

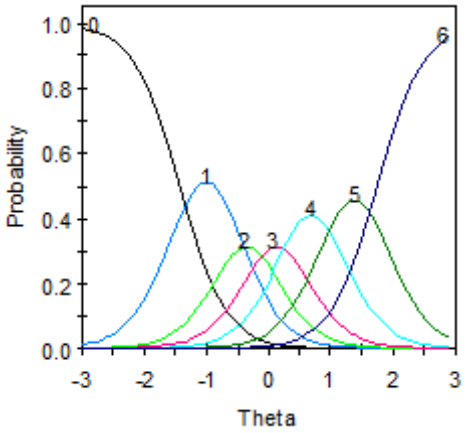


Figure 11. Traces lines for SCIENCE.

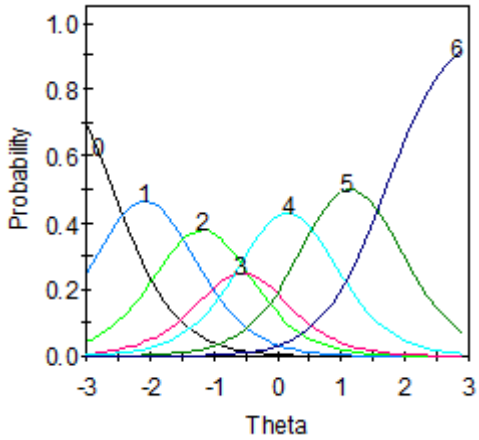


Figure 12. Trace lines for PLACE.

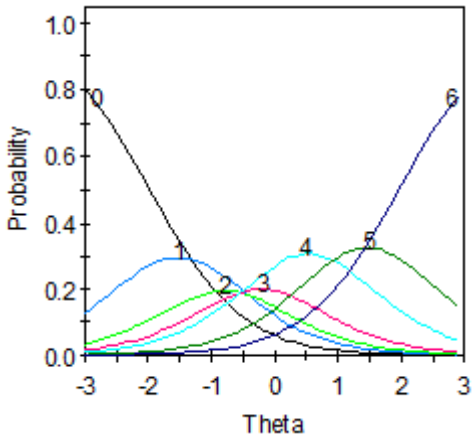


Figure 13. Trace lines for EVOLUTION.

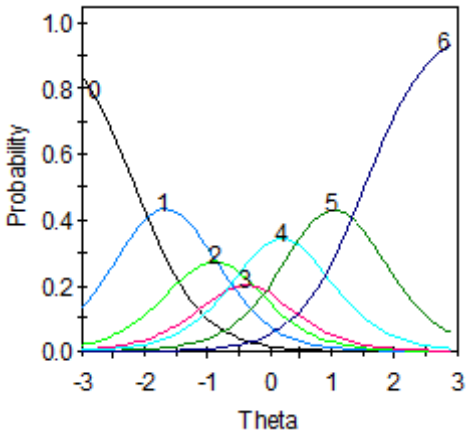


Figure 14. Trace lines for STRESS.

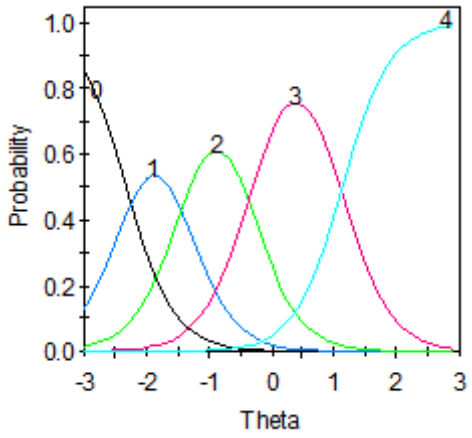


Figure 15. Trace lines for TREAT.

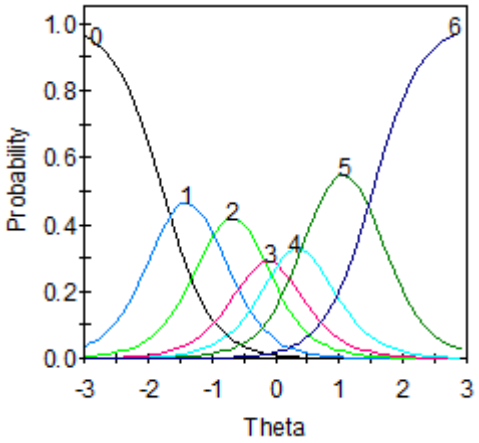


Figure 16. Trace lines for SHY.

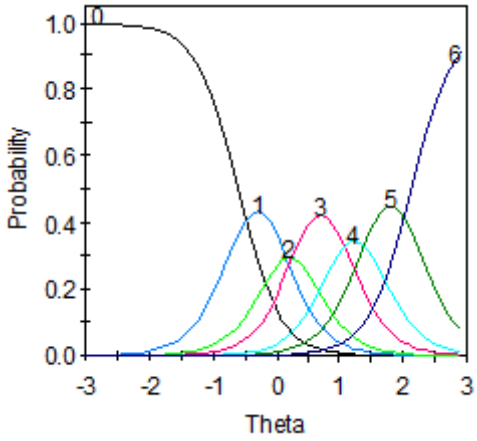
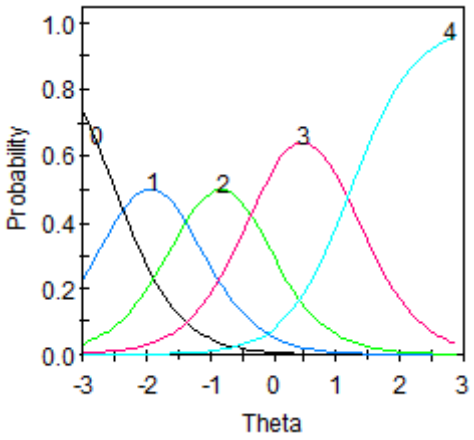


Figure 17. Trace lines for THEORETICAL.



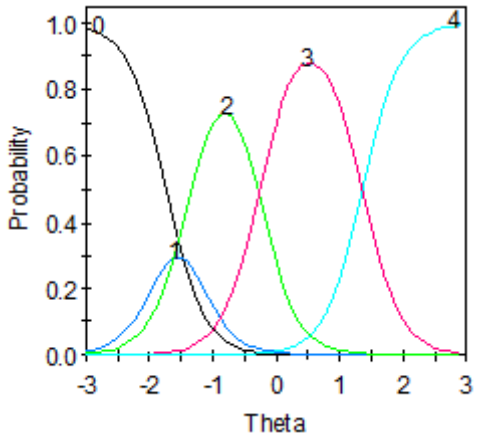


Figure 18. Trace lines for men (left) and women (right) for KIND.

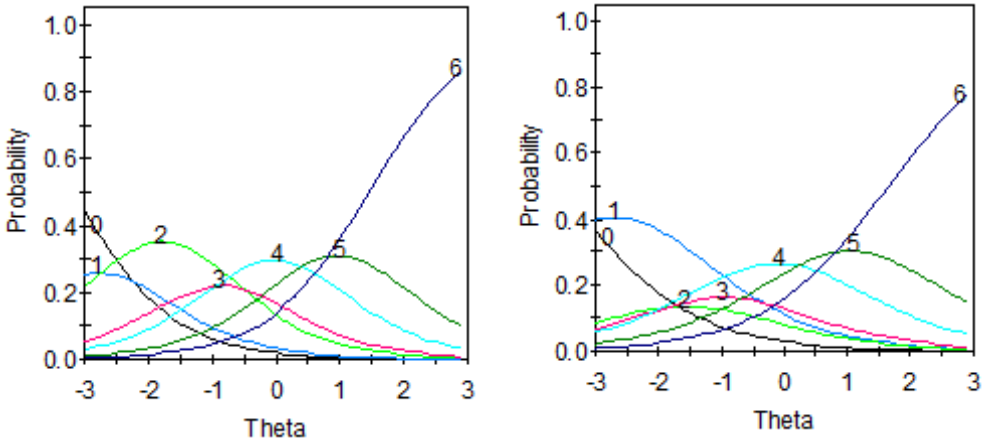


Figure 19. Trace lines for men (left) and women (right) for UNIVERSE.

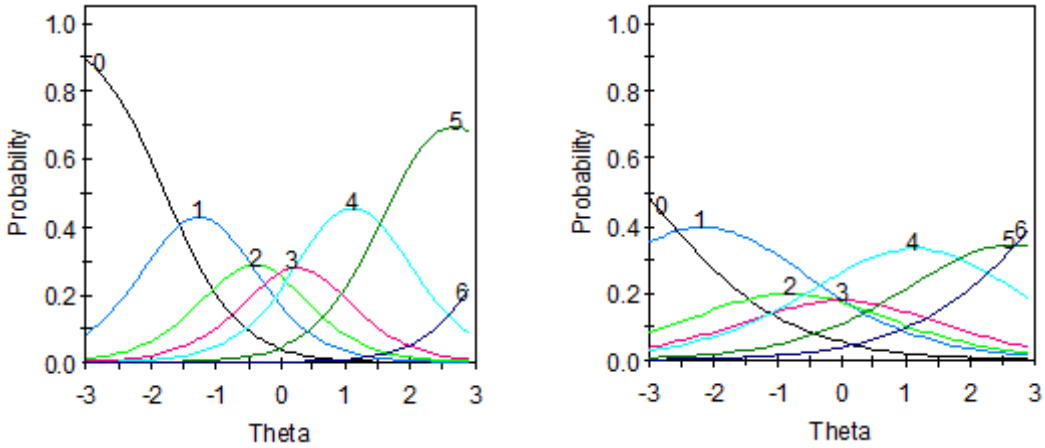


Figure 20. Trace lines for men (left) and women (right) for FEELINGS.

Three items were identified as displaying DIF, and their respective trace lines communicate the group differences quite clearly (see Figures 18, 19, and 20). For example, the slopes of the curves for women for the item FEELINGS are much steeper than those for men. For UNIVERSE, the probability of women endorsing the highest scale value is over 50% at only +1 theta (or, one standard deviation above the mean of the latent trait). It is apparent that even moderately high levels of theta resulted in maximal endorsement of the item, a clear sign this item is not meeting the gold standard.

Item information curves. In addition to parameter estimates, IRT analyses produce estimates of the amount of information afforded by each item (see Figures 21-37). The curves represent information, or the reciprocal of the amount of measurement error in each item, and these curves can highlight at what level of the latent trait the item is performing its best. Higher values along the *y*-axis indicate greater precision in identifying the individual's trait level, and items with very low and flat information curves are not contributing much useful information about the individual. Using Conscientiousness items as an example, item 9 (BELONGINGS; see Figure 27), has multiple peaks and an overall very high information line in the graph, indicating this item reliability estimates Conscientiousness across a wide range of levels of the trait. In comparison, item 14 (PLACE; see Figure 32), features no peaks, and the line barely extends beyond the *y*-axis value of 1.0, suggesting that this item does not provide as much information about the individual's Conscientiousness as the previous item did. Items with low information are said to have low precision or higher error, detracting from the interpretability and utility of the item. Based on the information curves, the following items had peaks that did not exceed 1.0, representing a low level of precision: PLACE, SYMPATHETIC, WITHDRAWN, UNIVERSE, HEADACHES, EVOLUTION, and FEELINGS. In addition, GENEROUS surpassed 1.0, but only for low levels of theta; for latent trait levels greater than 1.0, responses to this item provided very little information.

Items with DIF have two associated figures (see Figures 38-40), to display the information curves for men and for women separately. For items such as KIND (see Figure 38), it becomes immediately clear that men and women are responding differently

and therefore yielding varied information curves; men have a very low information curve, which suggests greater stability and more precision in their responses to this item.

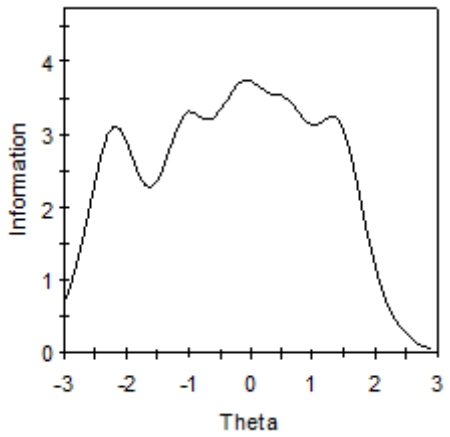


Figure 21. Information curve for SILENT

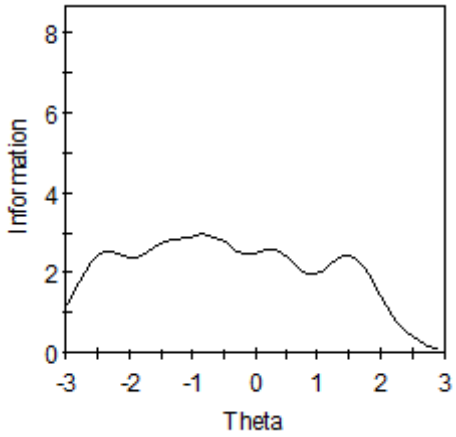


Figure 22. Information curve for NEAT

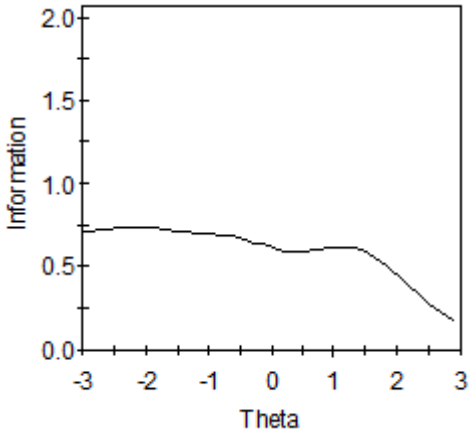


Figure 23. Information curve for SYMPATHETIC

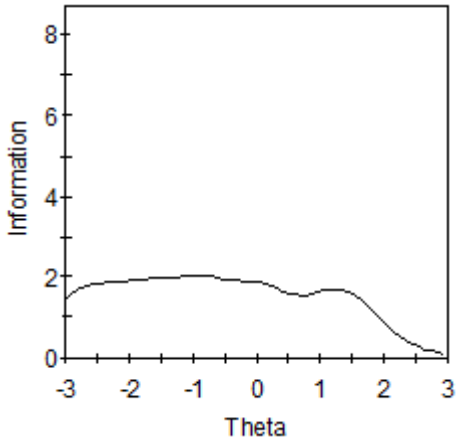


Figure 24. Information curve for ORGANIZED

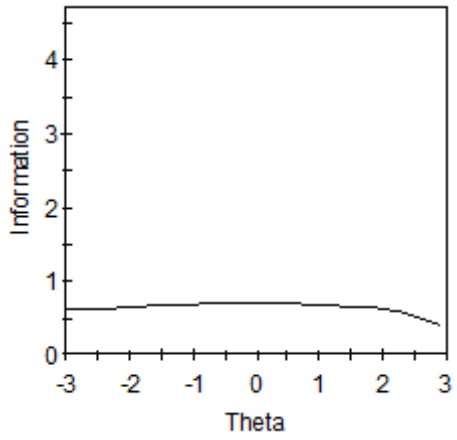


Figure 25. Information curve for WITHDRAWN

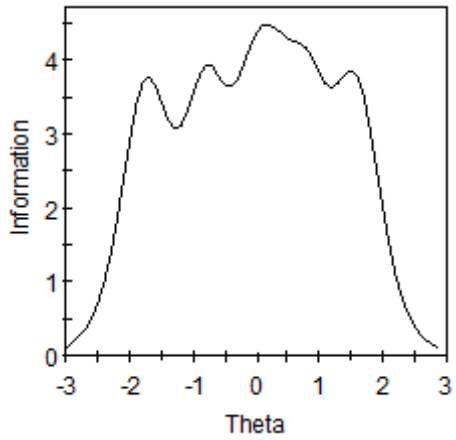


Figure 26. Information curve for QUIET

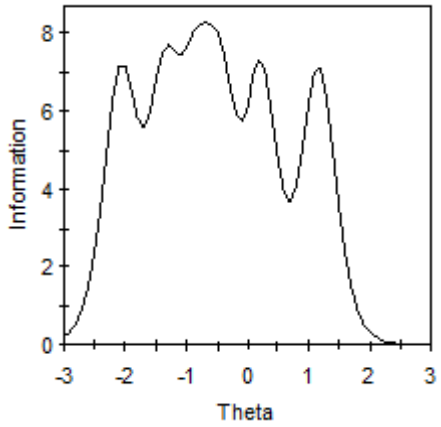


Figure 27. Information curve for BELONGINGS

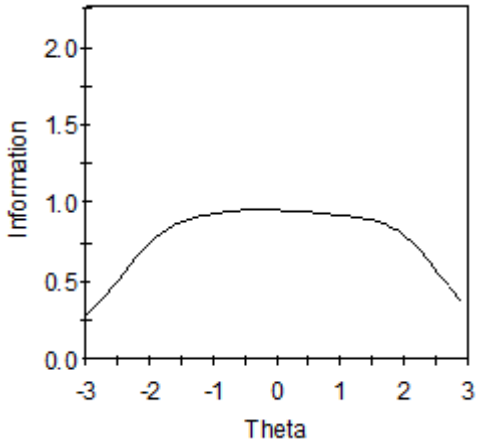


Figure 28. Information curve for HEADACHES

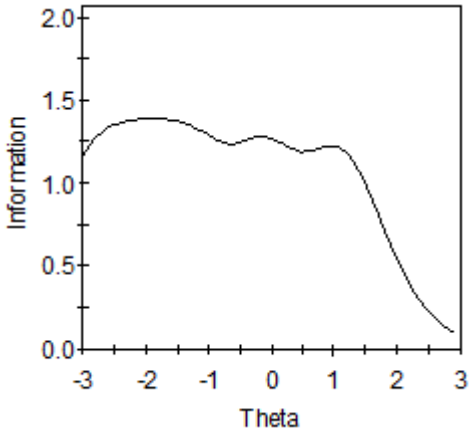


Figure 29. Information curve for GENEROUS

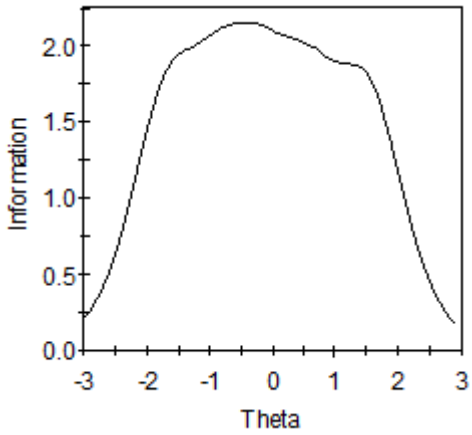


Figure 30. Information curve for STOMACH

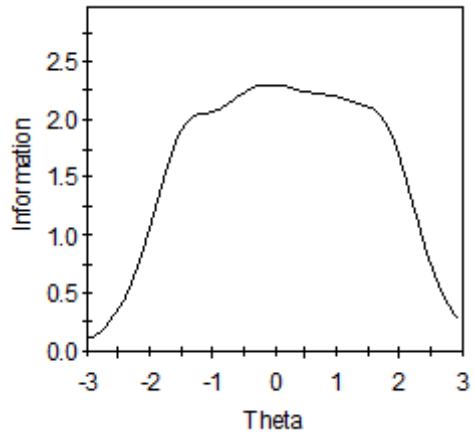


Figure 31. Information curve for SCIENCE.

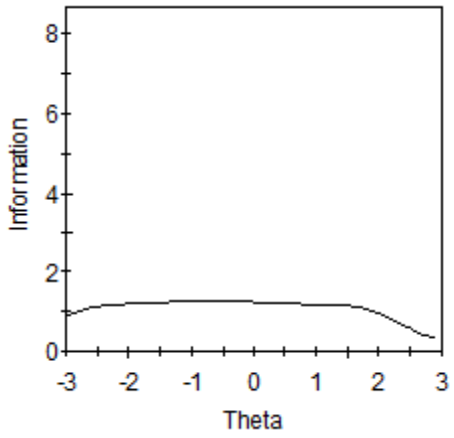


Figure 32. Information curve for PLACE.

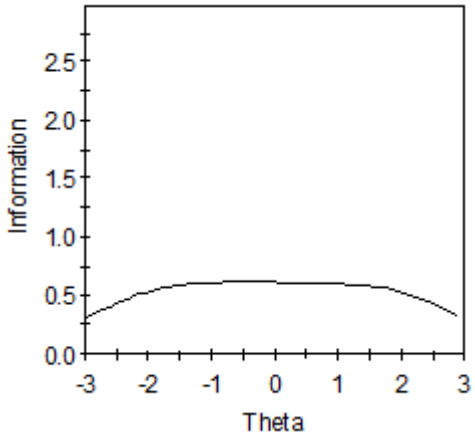


Figure 33. Information curve for EVOLUTION.

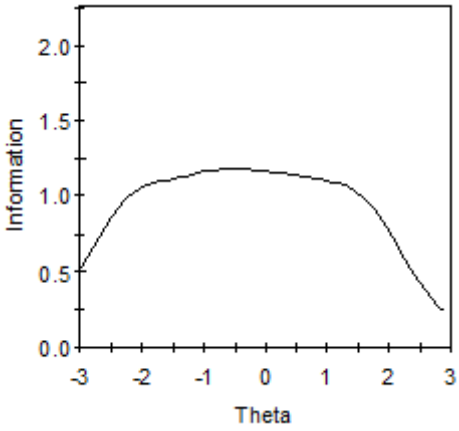


Figure 34. Information curve for STRESS.

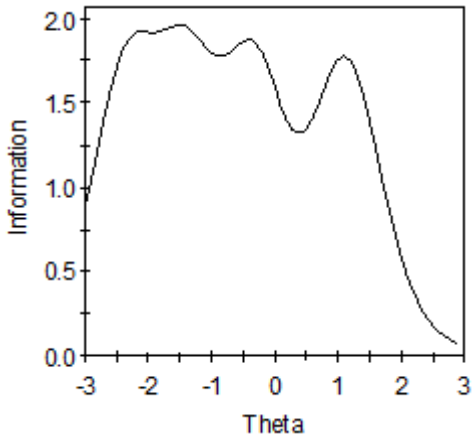


Figure 35. Information curve for TREAT.

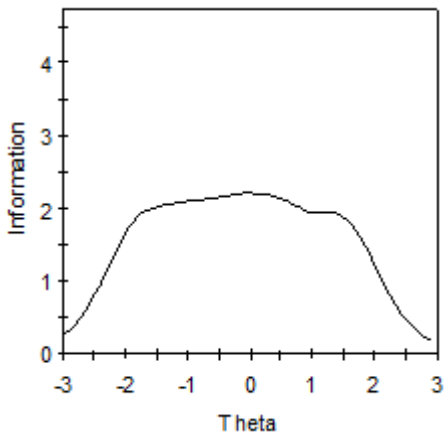


Figure 36. Information curve for SHY.

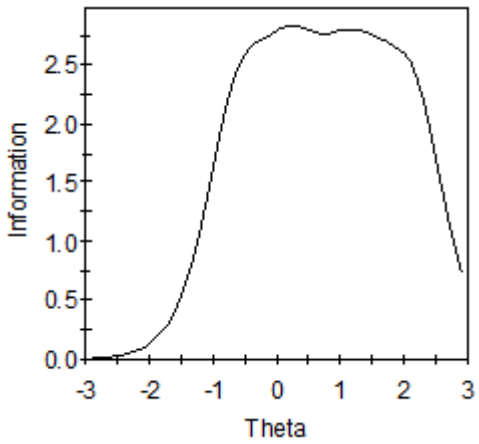


Figure 37. Information curve for THEORETICAL.

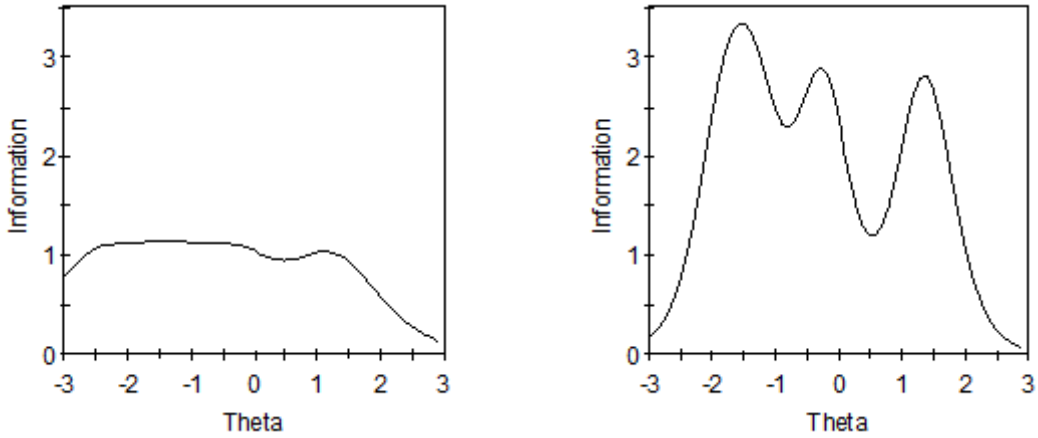


Figure 38. Information curve for men (left) and women (right) for KIND

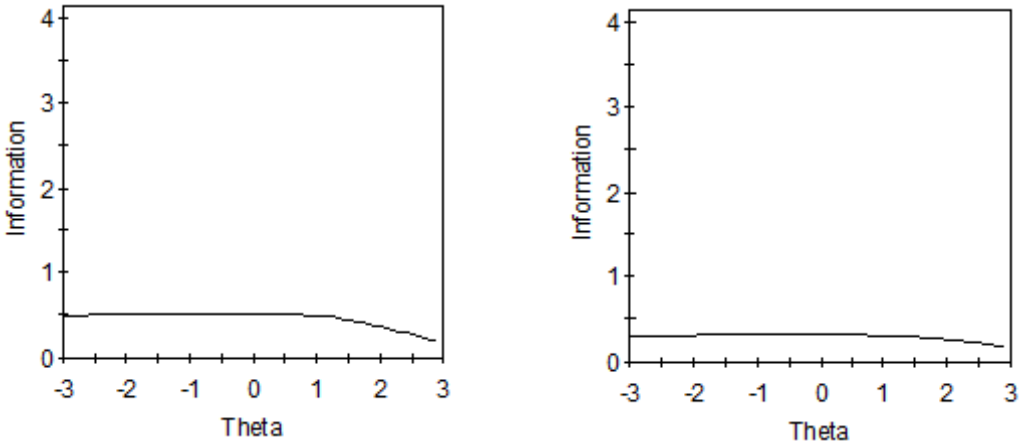


Figure 39. Information curve for men (left) and women (right) for UNIVERSE.

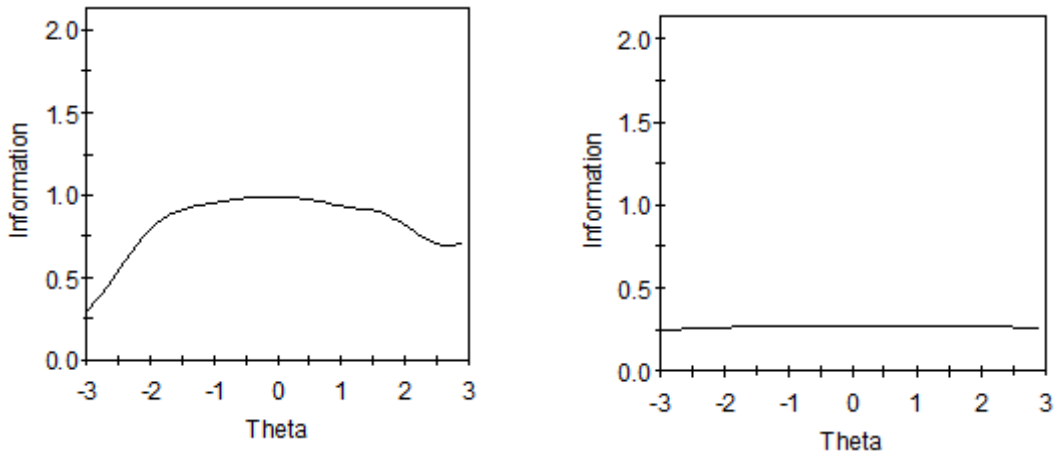


Figure 40. Information curve men (left) and women (right) for FEELINGS.

Discussion

Personality scales are important tools for decision-making in organizational selection, and shortened scales are especially valuable in terms of saving time and money. The present study sought to shed light on one such scale, the OCEAN.20, by applying a modern test theory approach to better understand the psychometric properties and potential for differential impact.

Findings from the CTT approach of comparing mean differences between men and women for each factor support Hypotheses 1a, 1b, and 1c. These results are consistent with the existing body of literature on personality and sex differences, and reinforcing the trends indicates this study's sample is dependable and provided meaningful responses, which is sometimes a concern with student samples. Reaffirming the sex differences that exist in personality items is important, as it indicates sex differences are stable across generations and cultures, since these findings have been repeated in other situations, and

therefore the items are worthy of detailed investigation for bias prior to operational use (e.g., Lehmann, 2006; Ones & Anderson, 2002).

Of greater interest, however, IRT methods helped identify three items that function differently for men and women (KIND [Agreeableness], UNIVERSE [Openness to experience], and FEELINGS [Neuroticism]). These differences indicate that women will provide higher ratings of their own kindness as compared to men, despite being equally kind in true personality, which could result in women yielding a higher score on Agreeableness and being selected over men more often (in the event this item is deemed desirable or necessary for the role and used in selection testing). The potential for bias is clear and unavoidable in the item KIND, and subject-matter experts should investigate what changes could be made to improve the item's functioning. Similar to the results of McClarty's (2006) study looking for DIF in NEO items, items tapping concern for others were more likely to be endorsed by women than similarly agreeable men. The observed DIF in KIND falls in line with the previous research into behavioural sex differences and could be explained as women fitting into "the gender stereotype of women as communal caretakers" (McClarty, 2006, p. 67).

Men and women also differed in terms of difficulty for item 8, UNIVERSE, but the pattern was inconsistent over the full range of the trait. Overall, this may not translate into the item favouring one gender over the other, but it is important to note that their probability of responding varies and differs, despite having equal levels of Openness to experience. To explain these results by relying on past research is complex, given how the probability of endorsement of the item is not steady across all levels of the latent trait. Typically, men express their Openness to experience by endorsing items that deal with

novelty and exploration, while women who are high in Openness tend to favour abstract ideas and aesthetics (Smith, 2002). The OCEAN.20 item, asking about the origins of the universe, spans both concepts of exploration and abstract ideas, which could serve to explain the complicated pattern of results that favour neither women nor men overall. It is also possible that this would be differentially perceived not by men and women, but rather by various religious or spiritual groups, as different belief systems may provide explanations that supersede the gender differences that may exist in Openness items.

Item 18, FEELINGS, had non-equivalent difficulty parameters for men and women such that women needed to have significantly greater amounts of Neuroticism to endorse the item as highly as men. In previous research on DIF in Neuroticism items, there are gender differences in terms of manifestation of distress; women tend to endorse items that reflect vulnerability, while men endorse items that reflect irritability (Smith, 2002). The present findings, in which men are more likely to endorse an item about feelings and emotions, seems contrary to past research, but it is important to remember that women significantly outscored men on a composite of Neuroticism items, so this finding does not negate previous studies. Rather, having one's feelings easily hurt is not a desirable concept, and women in the sample could be overcompensating for what they perceived to be an exceedingly feminine trait (e.g., anxious and/or emotional; McClarty, 2006).

The results the DIF analyses do not directly support Hypotheses 2a-c. However, consistent with McClarty's (2006) interpretation of NEO items displaying DIF, a tender-minded item from the Agreeableness factor did in fact favour women, partially supporting Hypothesis 2a. Women were expected to be favoured by an anxiety-related item, as stated in Hypothesis 2b, yet the findings presented here favour men instead and

do not support that hypothesis. Perhaps the hypothesis could have been better realized by a Neuroticism item that directly inquires about anxiety, rather than the OCEAN.20's item STRESS, which is a common state for undergraduate students throughout a typical semester. Hypothesis 2c was only partially supported, as men were favoured by an Openness item that was inquiring about something quite abstract and theoretical, though this was only true for the very lowest level of the trait as evidenced by the complex pattern of results. The recommendation from Embretson and Reise (2000) is that DIF observed in fewer than 10% of the items of a scale is unlikely to be problematic. In this instance, DIF has been observed for three out of 20 items, or for 15% of the scale, which raises concerns that necessitate further review.

Furthermore, the remaining three items of Agreeableness (i.e., SYMPATHETIC, GENEROUS, and TREAT) are potentially problematic, as they do not effectively capture the full range of the trait. Strong items span the full range, providing information about both the low and high ends of the trait's spectrum. In the OCEAN.20, SYMPATHETIC, GENEROUS, and TREAT are observed to be unable to measure the true range of Agreeableness, and they are flagged for removal or content review by subject-matter experts prior to future use. By limiting the range of the trait, employers who continue to use these items are failing to capture the array of kindness that actually exists in a population of job candidates, reducing the ability to glean meaningful information from this factor.

In addition to the parameters, item information curves were also evaluated, as these curves indicate the degree of measurement error across the levels of the trait. Higher values on the *y*-axis in these curves represent greater precision in measurement, and in

turn, increasingly reliable items. While CTT assumes an equal level of measurement error for all items and all levels of the latent trait, the item information curve illustrates just how much detail is extracted through IRT methods. Based on the information curves, it might be prudent to remove items affording the lowest amount of information (peaks that did not exceed 1.0), such as SYMPATHETIC, WITHDRAWN, UNIVERSE, HEADACHES, EVOLUTION, and FEELINGS. In addition, GENEROUS provided very little information for those with higher levels of Agreeableness. All of the aforementioned items would be solid candidates for further review, since minimal information is gleaned through the current phrasing of the item.

Recommendations

To summarize the recommendations resulting from the IRT analysis, nine items have been flagged for review: three items displaying DIF (KIND, UNIVERSE, and FEELINGS), three items with limited difficulty values (SYMPATHETIC, GENEROUS, and TREAT), and three more items with low information (WITHDRAWN, HEADACHES, and EVOLUTION). These nine items are recommended for subject-matter expert review prior to future use of the scale. It may be possible to revise the items to improve their psychometric properties, perhaps by removing any gender-stereotype-prone language, for example, or it may be preferable to replace the items with other reliable and valid items that fit the factor. Using the scale as-is could result in less than ideal selection decisions, as one gender could be selected for more often than the other, causing an unfounded bias. Items with low precision also raise one's risk of rejecting qualified candidates, or vice versa, since the amount of error in responses to the item reduces how reliable and dependable those items can be. Additionally, given the DIF

observed, it would be prudent to also test the full version of the TSD, as it is currently in operational use in North America. Evidence of DIF in these items points to the potential for DIF in the TSD, which would be valuable to identify to minimize any potential adverse impact.

As a comment on the scale in general, the five traits measured here are measuring just one facet of each of the five factors from the NEO-PI-R. The narrow factors could be due to the development process; by using CTT and selecting the highest loading items that clustered together well when creating the OCEAN.20, items are likely to be tapping similar and narrow aspects of each broad factor. The resulting single facet, in most cases, displays strong psychometric properties, but users of the scale should be cautious and acknowledge that only one facet of a broader trait is represented by the items of the OCEAN.20. In this case, for example, the OCEAN.20 items for Conscientiousness exclusively tap into the facet for 'order,' which may not be best suited predictor of academic or workplace success on its own, despite Conscientiousness typically being a strong predictor of performance in general (O'Keefe, Kelloway, & Francis, 2012; Egberink, Meijer, & Veldkamp, 2010). IRT methods could be used to select the best performing item from each facet of the Big Five factors, resulting in a scale that may better assess the broader factors. The demand for such a study and modification could be revealed if the narrow facets as measured by the current OCEAN.20 fail to predict broadly defined markers of successful job performance.

Contributions and Implications

An important strength of the present study is the application of modern test theory to personality items; using IRT in this way is still relatively new for the field of

personality testing and especially for personnel selection (Rouse, Butler, & Finger, 1999). IRT, at its core, assumes that responses to items reflect the individual's ability on the latent trait, and although ability might not be the right phrasing for the field of personality, it certainly seems appropriate to consider items as reflections of the individual's trait level. Samuel and colleagues (2010) proposed that IRT would be particularly useful for personality researchers by uncovering the amount of information collected by existing instruments at each level of a trait. In a unique study, Samuel and colleagues (2010) used IRT to compare results from the NEO-PI-R and a measure of personality disorders, and they demonstrated the shared latent constructs underlying both measures. In other words, abnormal personality traits appeared to exist at either very low or very high levels of the traditional Big Five traits. IRT methods highlighted the overlap between two scales, revealing new information about the structure of the latent traits that could not be as easily assessed with classical approaches. Numerous authors have recommended IRT be used outside of educational testing (e.g., Fraley, Waller, & Brennan, 2000; Wetzel, Bohnke, Carstensen, Ziegler, & Ostendorf, 2013; Wu et al., 2012), especially because IRT identifies maximally informative items and can be used in developing efficient scales. Rather than the fixed amount of error for each item, as assumed by classical test theory methods such as factor analysis, IRT lets each item have a unique error term, increasing precision and enabling the results to be independent of the sample's characteristics.

Identifying the best performing items and removing poor items (such as those that do not provide valuable information about the respondent's latent trait or those that do not operate consistently across demographic groups) results in both time- and cost-

effective scales. In high-stakes situations, such as when one's fate is being determined as it is in personnel selection, shortened scales are desirable. Even more desired, though, are high-quality scales that guide decisions, and IRT is capable of assisting in that process. In the present study, a personality scale for selection purposes was assessed with these methods, and new information was uncovered that was unavailable through classical test theory alone. More precisely, the presence of differential item functioning in three of the OCEAN.20 items is an important finding; for scores to be comparable across groups, items must be working in the same way (Embretson & Reise, 2000). DIF, one of the main advantages of IRT, is more sensitive to group differences than CTT due to the more precisely defined value of theta (Embretson & Reise, 2000).

In order to avoid artificially distorting results, users of the test are reminded that items with the potential for adverse impact should be removed, so that comparisons between genders can be conducted with confidence (Wetzel et al., 2013). Items that are found to work differently for different groups are problematic and a potential source of bias, and the items identified in the present study deserve a second look from content experts. As discussed previously, two types of DIF exist and were tested in the OCEAN.20: uniform and non-uniform DIF. The practical implications of uniform DIF are more readily managed than issues associated with non-uniform DIF. For uniform DIF, if one gender is more likely to endorse the item or find it easier, this discrepancy can be resolved simply by adding or subtracting a constant to either group's scores such that they can be compared on equal grounds. Non-uniform DIF, however, reveals a difference in the discrimination ability between the groups, which cannot be corrected in a straight-

forward manner and is therefore a bigger cause for concern to researchers and practitioners alike.

Limitations

Of course, despite these contributions, the present study is limited by its reliance on undergraduate students. IRT is robust in that results are not tied to the sample in same way as CTT results tend to be, but broadening the sample to its population of undergraduate students still has its drawbacks. Students are not necessarily representative of a general, working population, so in order to extend these results and improve generalizability, future research should consider sampling a more diverse population. Diversity in age, ethnicity, education, and employment field would be valuable in applying these findings to other settings. At the time of writing, the scale has been administered only to military recruits and undergraduate students, so it would be ideal to see how it performs for adults seeking employment, as the scale is designed for organizational selection purposes. Although IRT is considered sample-independent, this does not imply that the findings generalize without hesitation to any population; instead, provided the data fits the model adequately, IRT parameters are only considered to be invariant for various subsets drawn from the same population (Ellis & Mead, 2002).

In addition, future research should examine the impact of changing the wording of the items to reflect personality in a work context. As it stands, respondents are asked to provide information about their personality in general, but since the OCEAN.20 stands to be a useful organizational tool, perhaps some additional predictive validity can be gained from framing the items in terms of an organizational context. For example, one's workplace personality could differ from their day-to-day personality, and by rephrasing

the question to inquire specifically about one's behaviours at work, precision could be improved.

Another potential limitation is that the scale relies on self-report data. Respondents looking to be hired may be more motivated to present the best version of their selves or provide responses they believe the employer would find attractive. Additionally, self-report data is limited due to being a source of shared error, since all participant information stems from the same source (Thalmayer, Saucier, & Eigenhuis, 2011). In the present study, responses were kept anonymous and confidential, and the stakes were low since it was clearly presented as research, so it is unlikely to be a serious issue here, but practitioners who wish to use this scale should keep the issues of self-report scales in mind. In personnel selection, multiple indicators are always recommended in order to make the most informed decision about high-stakes scenarios, rather than relying on results of a single measure (Catano, Wiesner, Hackett, & Methot, 2010).

Additionally, IRT as a methodology has some limitations. A major drawback is the need for quite large sample sizes, which can be a resource issue for many researchers and organizations. In order to gain confidence in the findings, sample sizes close to 1000 participants are typically ideal (Reise & Yu, 1990), and the GRM in particular requires a minimum of $N = 500$. Although there were 503 participants in present study, simulations studies show that DIF can disappear as sample size increases (Meade & Lautenschlager, 2004; Reise & Yu, 1990). The findings presented here are therefore qualified by such simulations and, although they should be considered adequate, larger samples may reveal a slightly different pattern of results. Another limiting factor of IRT is the strict assumptions; demonstrating veridical unidimensionality, local independence, and the

absence of sub-group differences can be rigorous and difficult to achieve in non-simulated samples. Additionally, as a relatively new technique, it receives little attention and can be difficult to interpret without extensive training, as well as requiring some expertise on alternative software that are not particularly user-friendly.

Future Research

Moving forward, it is recommended that, at minimum, the six most problematic items (three with DIF and three with truncated difficulty parameters) be resolved prior to the OCEAN.20's next administration. These items are problematic to the extent that they do not provide sufficient information about the respondent and could disadvantage an entire demographic group. Future studies could examine whether other items from the original TSD have strong psychometric properties and a lack of DIF and therefore could serve as valuable replacements. Should no suitable replacements be found, the problematic items should be reviewed thoroughly by subject-matter experts to identify the source of the DIF and how these items could be revised to mitigate or eliminate psychometric issues. Even if the goal is not to replace items in the OCEAN.20, reviewing the TSD for DIF and other problematic items would be a meaningful endeavour, as the same items from the OCEAN.20 (and that have now been identified as problematic) appear in the TSD, which is currently being used as a selection tool.

Future researchers should continue applying IRT to novel assessment situations to maximize scale performance in all fields. Industrial/organizational psychology often features high-stakes testing scenarios, and it would be extremely valuable to extend these methods and results when analyzing other scales. The potential for differential item functioning in personality items is high, given all the previous research on gender

differences and the findings of the present study, and even subtle item bias can have serious implications. Practitioners can use the results of the present study to feel confident in the items of this shortened personality scale and make informed decisions when hiring applicants.

References

Baker, F. B. (2001). *The basics of Item Response Theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Catano, V. M., Wiesner, W. H., Hackett, R. D., & Methot, L. L. (2010). *Recruitment and selection in Canada* (4th ed). Toronto, ON: Nelson Canada.

Chapman, B. P., Duberstein, P. R., Sörensen, S., & Lyness, J. M. (2007). Gender differences in Five Factor Model personality traits in an elderly cohort. *Personality and Individual Differences, 43*(6), 1594-1603. doi:10.1016/j.paid.2007.04.028

Chen, W-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. doi:10.2307/1165285

Christal, R. (1994, November). *The Air Force Self Descriptive Inventory* (Final R&D Status Report). United States Air Force.

Costa, P. T, Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13,* 653-665

Egberink, I. J. L., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality, 44,* 232-244.

Ellis, B. B., & Mead, A. D. (2002). Item analysis: Theory and practice using classical and modern test theory. In S. G. Rogelberg (Ed.), *Handbook of Research Methods in Industrial and Organizational Psychology* (pp. 325-343). Oxford, UK: Blackwell.

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

Ferrando, P. J., & Chico, E. (2007). The external validity of scores based on the two-parameter logistic model: Some comparisons between IRT and CTT. *Psicologia, 28*, 237-257.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78*(2), 350-365. doi:10.1037//0022-3514.78.2.350

Griffith, J. W., Sumner, J. A., Debeer, E., Raes, F., Hermans, D., Mineka, S., ... Craske, M. G. (2009). An item response theory/confirmatory factor analysis of the Autobiographical Memory Test. *Memory, 17*(6), 609-623. doi:10.1080/09658210902939348

Hayes, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38,* 1128-1142.

Hill, E. M., & Gick, M. L. (2011). The big five and cervical screening barriers: Evidence for the influence of conscientiousness, extraversion and openness. *Personality and Individual Differences, 50*(5), 662-667. doi:10.1016/j.paid.2010.12.013

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

Kappe, R., & van der Flier, H. (2010). Using multiple and specific criteria to assess the predictive validity of the Big Five personality factors on academic performance. *Journal of Research in Personality*, *44*(1), 142-145. doi:10.1016/j.jrp.2009.11.002

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.), pp. 230-251. New York, NY: Guilford.

Lehmann, R., Denissen, J. J. A., Allemand, M., & Penke, L. (2012). Age and gender differences in motivational manifestations of the Big Five from age 16 to 60. *Developmental Psychology*. doi:10.1037/a0028277

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analysis methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361-388. doi:10.1177/1094428104268027

McClarty, K. (2006). A feasibility study of a computerized adaptive test of the International Personality Item Pool NEO (Doctoral dissertation). University of Texas at Austin, TX.

Mount, M. K., & Barrick, M. R. (1995). The Big Five personality dimensions: Implications for research and practice in Human Resources Management. In G. R. Ferris (Ed.), *Research in Personnel and Human Resources Management, 13* (pp. 153-200). Greenwich, CT: JAI Press.

Muck, P. M., Hell, B., & Gosling, S. D. (2007). Construct validation of a short Five-Factor Model instrument: A self-peer study on the German adaptation of the Ten-Item Personality Inventory (TIPI-G). *European Journal of Psychological Assessment, 23*(3), 166-175. doi:10.1027/1015-5759.23.3.166

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

O'Keefe, D., Kelloway, E. K., & Francis, R. (2012). Introducing the OCEAN.20: A 20-item Five-Factor personality measure based on the Trait Self-Descriptive Inventory. *Military Psychology, 24*, 433-460.

Oh, I., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96*(4), 762-773. doi:10.1037/a0021832

Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology, 75*(3), 255-276. doi:10.1348/096317902320369703

Poropat, A. E. (2009). A meta-analysis of the Five-Factor Model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338. doi:10.1037/a0014996

Powell, D. M., Goffin, R. D., & Gellatly, I. R. (2011). Gender differences in personality scores: Implications for differential hiring rates. *Personality and Individual Differences, 50*(1), 106-110. doi:10.1016/j.paid.2010.09.010

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*, 203-212.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*(2), 164-184. doi:10.1037/1082-989X.8.2.164

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27* (2), 133-144.

Rothmann, S., & Coetzer, E. P. (2003). The big five personality dimensions and job performance. *Journal of Industrial Psychology, 29*, 68-74.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 *PSY*-5 scales. *Journal of Personality Assessment, 72*(2), 282-307.

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. doi:10.1037/0021-9010.82.1.30

Samuel, D. B., Simms, L. J., Clark, L. A., Livesley, W. J., & Widiger, T. A. (2010). An item response theory integration of normal and abnormal personality scales. *Personality Disorders: Theory, Research, and Treatment, 1*(1), 5-21. doi:10.1037/a0018136

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

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods, 8*(2), 206-224. doi:10.1037/1082-989X.8.2.206

Scott, N. W., Fayers, P. M., Aaronson, N. K, Bottomly, A., de Graeff, A., Groenvold, M., ... Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales, *Journal of Clinical Epidemiology, 62* (3), 288-295, doi:10.1016/j.jclinepi.2008.06.003

Sheppard, R., Han, K., Colarelli, S. M., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment, 13*(4), 442-453. doi:10.1177/1073191106289031

Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin, 28,* 754-763. doi:10.1177/0146167202289005

Smith, L. L., & Reise, S. P. (1998). Gender differences on Negative Affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology, 75* (5), 1350-1362.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306. doi:10.1037/0021-9010.91.6.1292

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Educational, Inc.

Thalmayer, A., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six personality questionnaires. *Psychological Assessment, 23*(4), 995-1009. doi:10.1037/a0024165

Webster, G. D., & Jonason, P. K. (2013). Putting the "IRT" in "Dirty": Item response theory analyses of the Dark Triad Dirty dozen – An efficient measure of narcissism, psychopathy, and Machiavellianism. *Personality and Individual Differences, 54,* 302-306.

Wetzel, E., Bohnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences, 34*(2), 69-81. doi:10.1027/1614-0001/a000102

White, L. A., Young, M. C., Hunter, A. E., & Rumsey, M. G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology, 1*, 291-295.

Woods, S. A., & Hampson, S. E. (2010). Predicting adult occupational environments from gender and childhood personality traits. *Journal of Applied Psychology, 95*(6), 1045-1057. doi:10.1037/a0020600

Wu, J., King, K. M., Witkiewitz, K., Racz, S. J., & McMahon, R. J. (2012). Item analysis and differential item functioning of a brief conduct problem screen. *Psychological Assessment 34*(2), 444-454. doi:10.1037/a0025831

Appendix A
Research Ethics Board Certificate of Approval

SAINT MARY'S UNIVERSITY
One University. One World. Yours.

RESEARCH ETHICS BOARD

Department Office
T 902.420.5728
F 902.496.8772
E ethics@smu.ca

Certificate of Ethical Acceptability for Research Involving Humans

This is to certify that the Research Ethics Board has examined the research proposal:

SMU REB File Number:	12-399
Title of Research Project:	Analysis of a Shortened Personality Scale
Faculty, Department:	Science, Psychology
Faculty Supervisor:	Dr. Lucie Kocum
Student Investigator:	Joanna Solomon

and concludes that in all respects the proposed project meets appropriate standards of ethical acceptability and is in accordance with the Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans (TCPS 2) and Saint Mary's University relevant policies.

Approval Period: February 6, 2013 – February 6, 2014*

Post-approval Reporting Requirements

ADVERSE EVENT
Adverse Event Report: http://www.smu.ca/academic/reb/forms.html
Adverse events must be immediately reported (no later than 1 business day).
SMU REB Adverse Event Policy: http://www.smu.ca/academic/reb/policies.html

MODIFICATION
FORM 2: http://www.smu.ca/academic/reb/forms.html
Research ethics approval must be requested and obtained prior to implementing any changes or additions to the initial submission, consent form/script or supporting documents.

YEARLY RENEWAL*
FORM 3: http://www.smu.ca/academic/reb/forms.html
Research ethics approval is granted for **one year only**. If the research continues, researchers can request an extension one month before ethics approval expires.
FORM 4: http://www.smu.ca/academic/reb/forms.html
Research ethics approval for course projects is granted for **one year only**. If the course project is continuing, instructors can request an extension one month before ethics approval expires.

CLOSURE
FORM 5: http://www.smu.ca/academic/reb/forms.html
The completion of the research must be reported and the master file for the research project will be closed.

*Please note that if your research approval expires, no activity on the project is permitted until research ethics approval is renewed. Failure to hold a valid SMU REB Certificate of Ethical Acceptability or Continuation may result in the delay, suspension or loss of funding as required by the federal granting Councils.

On behalf of the Saint Mary's University Research Ethics Board, I wish you success in your research.

Dr. Jim Cameron, Ph.D.
Chair, Research Ethics Board, Saint Mary's University

Appendix B
Survey Items and Demographic Questionnaire

DEMOGRAPHICS (4 items)

1. What is your sex? *(Male, Female)*
2. What is your age? *(open-ended text box for numerical responses only)*
3. Which ethnic background do you identify with? *(Caucasian, African, Middle Eastern, Asian and Pacific Islander, First Nations, South/Southeast Asian, More than one, Other)*
4. What is your current GPA? *(selection of all options from 0.0 to 4.3)*

The OCEAN.20 Personality Scale (20 items; O'Keefe, Kelloway, & Francis, 2012)

"Using the following rating scale (1 = extremely uncharacteristic; 7 = extremely characteristic), decide how well each adjective or statement describes you. Please reply to all adjectives and statements. Give your first impression of how characteristic each phrase is of you. Don't spend too long on deciding what your answer should be. Answer all questions, even if you are not entirely sure of your answer. Answer honestly. Please respond as honestly and accurately as you can."

1. Silent
2. Neat
3. Sympathetic
4. Organized
5. Withdrawn
6. Kind
7. Quiet
8. I have thought a lot about the origins of the universe
9. I like to keep all my belongings neat and organized
10. I often have headaches when things are not going well
11. I am always generous when it comes to helping others
12. Sometimes I get so upset, I feel sick to my stomach
13. I am highly interested in all fields of science
14. I like to have a place for everything and everything in its place
15. I am fascinated with the theory of evolution
16. When I am under great stress I often feel like I am about to break down
17. I always treat other people with kindness
18. My feelings are easily hurt
19. I am a very shy person
20. I would enjoy being a theoretical scientist

Appendix C
Comparing Response Times Reliability

Table 13.
Cronbach's Alpha Coefficients by Response Time

	Response Time < 3 Min.	Response Time > 3 min.
Openness	.82	.77
Conscientiousness	.88	.87
Extraversion	.89	.85
Agreeableness	.83	.79
Neuroticism	.79	.76

Note: For participants completing the survey in less than 3 minutes, n =216. For participants completing the survey in more than 3 minutes, n = 316. Four participants took more than 1.5 hours to complete the survey.

Appendix D
Statistics Split by School

Table 14.
Descriptive Statistics by University

Item Number	Item or Factor Name	SMU		Non-SMU	
		M	*SD*	*M*	*SD*
1	SILENT	4.33	1.65	4.06	1.88
2	NEAT	4.90	1.46	4.49	1.63
3	SYMPATHETIC	5.65	1.08	5.56	1.16
4	ORGANIZED	5.10	1.40	4.98	1.67
5	WITHDRAWN	4.43	1.51	4.00	1.78
6	KIND	5.82	0.95	5.77	0.82
7	QUIET	3.96	1.68	3.56	1.85
8	UNIVERSE	4.36	1.76	4.55	1.98
9	BELONGINGS	4.99	1.47	4.38	1.83
10	HEADACHES	4.03	1.88	3.82	1.87
11	GENEROUS	5.53	1.14	5.23	1.17
12	STOMACH	4.24	1.80	4.52	1.88
13	SCIENCE	3.75	1.86	3.90	1.86
14	PLACE	4.66	1.52	4.40	1.66
15	EVOLUTION	4.13	1.87	4.30	1.84
16	STRESS	4.46	1.76	4.45	1.95
17	TREAT	5.69	0.94	5.38	1.09
18	FEELINGS	4.35	1.68	4.44	1.65
19	SHY	4.21	1.75	4.17	1.90
20	THEORETICAL	2.87	1.71	2.93	1.90
	Openness	3.78	1.37	3.92	1.50
	Conscientiousness	4.91	1.26	4.56	1.46
	Extraversion	4.24	1.37	3.95	1.65
	Agreeableness	5.66	0.86	5.48	0.89
	Neuroticism	4.28	1.38	4.32	1.36
	Age (in years)	21.04	3.18	21.51	3.59

Note: SMU = Participants from Saint Mary's University, n = 421 (30.9% male, 68.4% female; 72.4% Caucasian); Non-SMU = Participants from Dalhousie University and Mount Saint Vincent University, n = 82 (29.3% male, 67.1% female; 84.1% Cauasian).

Appendix E
Final Parameter Estimates

Table 15.
Final discrimination and difficulty parameters for the OCEAN.20

Number	Item Name	*a*	β_1	β_2	β_3	β_4	β_5	β_6
1	SILENT	3.50	-2.20	-1.05	-0.28	0.02	0.60	1.40
2	NEAT	3.10	-2.42	-1.47	-0.88	-0.57	0.33	1.51
3	SYMPATHETIC	1.53	-3.42	-2.40	-1.76	-0.55	1.31	N/A
4	ORGANIZED	2.56	-2.71	-1.83	-1.10	-0.73	0.10	1.32
5	WITHDRAWN	1.48	-3.31	-1.72	-0.64	0.09	0.80	2.11
6	KIND							
	Men	1.98	-2.50	-1.39	-0.29	1.24	N/A	N/A
	Women	3.36	-1.74	-1.38	-0.27	1.37	N/A	N/A
7	QUIET	3.84	-1.74	-0.80	-0.03	0.27	0.81	1.58
8	UNIVERSE							
	Men	1.26	-3.19	-2.37	-1.21	-0.51	0.45	1.46
	Women	1.00	-3.55	-1.84	-1.29	-0.65	0.43	1.67
9	BELONGINGS	5.34	-2.07	-1.35	-0.84	-0.47	0.23	1.17
10	HEADACHES	1.73	-1.74	-0.82	-0.35	0.14	0.90	1.86
11	GENEROUS	2.13	-2.82	-2.02	-1.27	-0.11	1.14	N/A
12	STOMACH	2.60	-1.67	-0.91	-0.50	-0.06	0.62	1.51
13	SCIENCE	2.70	-1.43	-0.60	-0.11	0.37	1.02	1.75
14	PLACE	2.02	-2.59	-1.59	-0.80	-0.30	0.61	1.69
15	EVOLUTION	1.38	-1.98	-1.08	-0.51	0.07	0.99	1.97
16	STRESS	1.92	-2.15	-1.19	-0.60	-0.18	0.56	1.52
17	TREAT	2.64	-2.33	-1.42	-0.35	1.14	N/A	N/A
18	FEELINGS							
	Men	1.77	-1.78	-0.75	-0.08	0.57	1.69	3.62
	Women	1.04	-3.45	-1.83	-1.06	-0.37	0.97	2.34
19	SHY	2.65	-1.78	-1.02	-0.36	0.09	0.60	1.53
20	THEORETICAL	3.02	-0.61	0.01	0.41	1.00	1.48	2.12