ABSTRACT

ZHU, YIZHEN EGYN. An Item Response Theory Approach to Rapid Response Measurement (Under the direction of Dr. Adam Meade).

Personality assessment is widely used in organizational research today, given its association with many salient workplace outcomes. This study seeks to extend the scoring framework of the Big Five Rapid Response Measure (B5-RRM) by relating latent personality traits to observed RRM scores and response latency using item response theory (IRT). Data from 184 participants recruited from Amazon's Mechanical Turk was analyzed. The B5-RRM, a Likert-type personality scale, construct validity measures, and work-related criteria measures were administered. Results reveal that extraversion, agreeableness, conscientiousness, and neuroticism responses from the B5-RRM can be modeled using IRT. Furthermore, responses to the B5-RRM agreeableness and conscientiousness scales are positively related to response latency. The IRT-related scoring framework demonstrates acceptable construct and criterion-related validity. However, evidence regarding whether applying an IRT-based scoring framework improves the validity of the B5-RRM is inconclusive.

An Item Response Theory Approach to Rapid Response Measurement of Personality

by
Yizhen Egyn Zhu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Industrial and Organizational Psychology

Raleigh, North Carolina
2021

APPROVED BY:

Dr. Adam Meade Chair of Advisory Committee	Dr. Stephen B. Craig
Dr. Lori Foster	Dr. Mark Wilson

BIOGRAPHY

Yizhen Egyn Zhu was born in Beijing, China. She received her degree in Psychology and Mathematics at Davidson College in 2017. Egyn began her pursuit of a PhD in Industrial and Organizational Psychology at North Carolina State University following her bachelor's degree. Under the direction of Dr. Adam Meade, she conducted research focusing on psychometrics, as well as on applying big data techniques to work-related problems. In addition to academic research, Egyn also gained applied research experience through her internships with Modern Hire and the Talent Assessment team at Amazon. Egyn will be relocating to Irvine, California upon graduation to start her role as a research scientist on the Customer Behavior team at Amazon.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Over the past few decades, there has been an increased optimism in the utility of personality assessments in organizational settings, especially for employee selection (Goldberg, 1993; Hogan et al., 1996; Hogan & Ones, 1997). Such an optimism is partially a result of the growing consensus around the Big Five model of personality (Digman, 1990; Goldberg, 1990), which has helped remove one of the largest limitations in personality research - the lack of a common framework for organizing traits (Barrick & Mount, 1991; Goldberg, 1993; Hough, 1992; Hurtz & Donovan, 2000; Mount & Barrick, 1995). Early meta-analyses were able to identify somewhat stable relationships between personality traits and job performance (Barrick & Mount, 1991; Hurtz & Donovan, 2000; Tett et al., 1991). With such growing confidence in the utility of personality measurements, there is a growing need for accurate and efficient personality measurements. Self-report Likert-type scales have been one of the most popular ways to measure personality (e.g., Costa & McCrae, 1992; Goldberg et al., 2006; Gosling et al., 2003; John et al., 1991). Yet, there are challenges for typical Likert-type scale personality measures: Lengthy tests can increase boredom and irritation for participants (Donnellan et al., 2006) and faking in personality measurements, which meta-analyses have demonstrated possible (Viswesvaran & Ones, 1999), can be prevalent in selection settings (Birkeland et al., 2006).

Meade et al. (2020) introduced a new method called rapid response measurement (RRM) to address such challenges. RRM refers to a general measurement method that can be applied to multiple constructs. RRM presents stimuli in rapid succession to test takers in a computer-based format. When applied to measure the Big Five personality traits, RRM has been shown to be reliable, valid, more time-efficient, and less prone to faking compared to traditional Likert-type scales (Meade et al., 2020). The current study aims to extend the existing scoring framework of

RRM by relating latent personality traits to observed RRM scores and response latency. More specifically, the study applies item response theory (IRT; Embretson & Reise 2000; Lord, 1980) to score the Big Five Rapid Response Measure (B5-RRM; Meade et al., 2020). Additionally, no known studies have applied IRT models that account for response latency in assessing the Big Five personality traits. Therefore, this study also seeks to showcase IRT's use to model both the response and the response latency in the B5-RRM.

Personality in Organizational Research

Personality assessment is widely used in organizational research today, given its association with many salient workplace outcomes (Ployhart et al., 2005). Yet, historically, there has been a major challenge in measuring personality: The lack of a common framework to organize traits (Barrick & Mount, 1991; Goldberg, 1993; Hurtz & Donovan, 2000; Hough, 1992; Mount & Barrick, 1995). Such a problem made it difficult for early meta-analytic research to establish stable relationships between personality and other important workplace variables (Hurtz & Donovan, 2000). Because there was no stable and well-accepted taxonomy of personality traits, it was almost impossible to aggregate across studies and obtain meaningful personality correlates. As a result, early personality research was often unsuccessful in showcasing the validity and value of measuring personality in the workplace (Barrick & Mount, 1991; Hurtz & Donovan, 2000).

Nevertheless, as the consensus around the Big Five as a common trait-organizing framework grew (Barrick & Mount, 1991; Digman, 1990; Goldberg, 1990), personality gained more popularity in organizational research and practices. The Big Five framework (Digman, 1990; Goldberg, 1990) organizes personality traits into five major traits: Openness to experience, conscientiousness , extraversion , agreeableness , and neuroticism (also referred to as emotional stability). Since then, the Big Five personality traits have become a ubiquitous model in modern personality assessments (Barrick et al., 2001; Oswald & Hough, 2011). In the following sections, the literature on the personality-performance relationship is reviewed, followed by a discussion of some of the remaining challenges in measuring personality.

Personality and Job Performance

Meta-analyses have found personality to be a crucial predictor of many workplace outcomes such as job satisfaction (Judge et al., 2002), organizational citizenship (Chiaburu et al., 2011), and turnover (Barrick & Mount, 1991; Zimmerman, 2008). Among these criteria, job performance is the most widely studied variable. Research concerning personality and job performance ranges from criterion development to clarifying the context-specific nature of the personality-performance relationship (Ployhart et al., 2005; Johnson & Schneider, 2013). Many meta-analyses have attempted to explicate the personality-performance relationships (e.g., Barrick & Mount, 1991; Hurtz & Donovan, 2000; Tett et al., 1991). Such studies examined both the overall correlations and occupation-specific relationships.

In terms of the overall personality-performance relationship, the meta-analytic studies averaged correlations across occupational groups. Early meta-analytic studies revealed discrepancies in findings regarding the personality-performance relationship. One meta-analysis found that conscientiousness was the strongest predictor of job performance (Barrick & Mount, 1991). Conversely, meta-analytic findings by Tett et al. (1991) found agreeableness and extraversion to be the strongest and weakest predictor of job performance, respectively. Still, other meta-analyses found moderate yet slightly different relationships between personality traits and job performance (Mount & Barrick, 1995; Salgado, 1997). These early meta-analyses were later criticized for including studies that did not use Big Five measurements of personality, which explained the mixed results (Hurtz & Donovan, 2000). Hurtz and Donovan (2000) argued that results obtained from using different personality trait taxonomies should not be aggregated. Therefore, including only studies that used explicit Big Five factors, Hertz and Donovan (2000) conducted a meta-analysis and found conscientiousness to be the strongest personality predictor of job performance, followed by emotional stability, agreeableness, and extraversion.

In addition to investigating the personality-performance relationships cross-occupationally, the above-mentioned meta-analyses also investigated this relationship across different occupational groups. Results revealed that personality traits correlated to different occupational groups differentially. For example, extraversion was strongly associated with performance in sales, customer service, and managerial jobs, but not in skilled occupations (e.g., accountants) and professional occupations (e.g., engineers; Barrick & Mount, 1991; Hurtz & Donovan, 2000). Furthermore, job complexity also moderates the relationship between conscientiousness and job performance, such that having high-complexity jobs attenuates the relationship when compared to having low-complexity jobs (Wilmot & Ones, 2019).

Trait activation theory (TAT; Tett & Burnett, 2003; Tett & Guterman, 2000) helps track the complexity of how personality traits impact job performance and workplace behaviors in general. The TAT posits that the extent to which personality traits influence behaviors is moderated by situation trait relevance (i.e., the opportunity for trait expression; Tett & Guterman, 2000). A TAT-based model of personality-performance states that the extent to which personality influences work behavior and job performance is determined by trait-relevant cues (Tett & Burnett, 2003). Accordingly, the nature of the occupation (e.g., tasks, responsibilities) serves as a cue to determine whether a personality trait is relevant to a job. In other words, different behaviors are required to perform well in different occupations, and such behaviors are associated with different personality traits. For example, a salesperson needs to talk to the customers, which is associated with trait extraversion, but an accountant does not frequently experience the same situation. As a result, extraversion is more predictive of job performance among salespeople but not accountants. When meta-analyses aggregate criterion-related correlations for extraversion across these two jobs, the resulting criterion-related validity is underestimated for sales positions and overestimated for accountants.

Overall, meta-analyses support the utility of personality in employee selection due to its recognizable relationship with job performance (Barrick & Mount, 1991; Hurtz & Donovan, 2000; Tett et al., 1991). While conscientiousness appears to be a notable predictor of performance across occupations, the other Big Five factors can also serve as important predictors for specific occupational groups.

Challenges in Measuring Personality

Challenges in personality measurements can still hinder researchers from accurately assessing personality. First, personality measures can become lengthy: Traditional Likert-type personality scales can contain more than 200 items and take up to 40 minutes to complete (e.g., Costa & McCrae, 1992). Such lengthy scales can be especially problematic in organizational research because participants are often required to complete multiple surveys at once, making the full battery of assessments lengthier and more boring (Donnellan et al., 2006). In low stakes testing situations, such as participating in a research study, careless responding increases as participants progress further into assessments (Bowling et al., 2020; Gibson & Bowling, 2019; Wise, 2010). The longer an assessment becomes, the more likely participants will start responding carelessly. Furthermore, low-stakes testing is associated with lower motivation levels of participants (Wise, 2010). As a result, demotivated examinees are more likely to experience increasing fatigue and decreasing efforts as they progress into surveys (Ackerman & Kanfer, 2009). Given such concerns, there appears to be a need for more efficient and innovative personality measurements to overcome the problems associated with lengthy assessments.

The potential of applicant faking is another challenge in personality assessments. Meta-analytic findings have demonstrated that, when instructed, test takers are capable of distorting their responses in socially desirable ways (Viswesvaran & Ones, 1999). While some studies suggested negligible faking behaviors in job applicants (e.g., Hough et al., 1990; Ellingson et al., 2007), a meta-analysis found that job applicants do distort their response to personality assessments to present themselves in ways that are more desirable to the employer (Birkeland et al., 2006). Furthermore, there has been a debate as to whether faking will hurt personality assessments' criterion-related validity (Birkeland et al., 2006). While several studies concluded that faking does not threaten the validity of personality measurement in real-world settings (e.g., Hogan et al., 2007; Hough et al., 1990; Ones et al., 1996), other studies suggested that the impact of faking on validity has been underestimated (e.g., Holden, 2008; Komar et al., 2008; Rosse et al., 1998). Nevertheless, there has been a consensus in the literature that faking on personality measurement can impact hiring decisions (Hough, 1998; Rosse et al., 1998), which can be especially harmful to organizations when the selection ratio is small (Mueller-Hanson et al., 2003). Thus, there have been many attempts to resist faking in personality tests: For example, imposing a time limit (e.g., Komar et al., 2010), statistically correcting for social desirability or faking (e.g., Ellingson et al., 1999; Hendy et al., 2020), and the use of forced-choice formats (e.g., Jackson et al., 2000).

Overview of Rapid Response Measurement Method

RRM is a general measurement method that attempts to overcome the above-mentioned hurdles in personality measurement (Meade et al., 2020). In RRM, each item is a single adjective that is related to the construct being measured. For example, in the B5-RRM, adjectives describe one of the Big Five personality traits (e.g., adjectives like “social”, “expressive”, and “talkative” describe extraversion). During an RRM session, respondents read adjectives on a computer screen successively and make responses using two keys to indicate if an adjective is “like me” or “not like me,” respectively (i.e., the “i” and “e” key in the desktop version; two virtual buttons in the mobile version). For each item, three variables are recorded: The adjective, the respondent’s response, and the response latency. Upon answering each item, the next adjective appears on the screen immediately. To promote rapid response, RRM assessments prompt respondents to respond to items as quickly as possible at the beginning of sessions. Respondents see a warning message if any of their responses take more than 2.5 seconds. Only responses made within 2.5 seconds are considered valid responses and are considered in trait scores calculation. Under such a setting, respondents should be able to answer items fairly quickly. In fact, Meade et al. (2020) reported that for the B5-RRM, 100 adjectives per trait (i.e., 500 items) only took seven to eight minutes to complete. Furthermore, the reliability and validity of B5-RRM were not compromised despite the quick administration, making it an efficient tool for measuring personality.

In response to obstacles related to faking in personality measurements, the B5-RRM is also more faking resistant than traditional Likert-type personality scales. Meade et al. (2020) compared honest and faking conditions within subjects. They found that even though participants in the faking condition could still increase conscientiousness, extraversion, and decrease neuroticism, there was no significant difference between the two groups in agreeableness and openness to experience. More importantly, the effect size of faking was significantly smaller in RRM compared to that in Likert-type personality measurements. The B5-RRM also had a significantly lower correlation with social desirability compared to Likert-type personality scales. Fisher et al. (2020) revealed that the "multisaturation" perspective on faking (Tett & Simonet) could be used to explain why RRM reduces faking. According to the "multisaturation" perspective, faking behavior is determined by the opportunity, motivation, and ability to fake. Fisher et al. (2020) argued that while the opportunity and motivation to fake are relatively constant within the same assessment setting, the ability to fake can be greatly reduced by RRM.

RRM is a measurement method that can be used for constructs that seek measurement efficiency and faking resistance. In addition to the Big Five personality traits, Meade et al. (2020) also validated the Workplace RRM (W-RRM), which measures more specific work-relevant personality traits, such as assertiveness and dependability. RRM has also been successfully applied to measure the dark triad personality (Valone et al., 2020) and job satisfaction (Wilgus et al., 2020). The dark triad personality traits - psychopathy, narcissism, and Machiavellianism (Paulhus & Williams, 2002) - are socially undesirable traits that test takers often wish to hide. Therefore, applying a faking resistant measurement to the dark triad personality traits can be crucial in many settings.

Scoring Framework

A linearly weighted scoring framework has been used in all known applications of RRM. After reversing reverse coded items, each item score is computed by multiplying the response and the response latency weight, which gives more weight to faster responses. The final score is the average of all item scores. Response latency has been noted as an appropriate indicator of how closely an item is related to a person's self-schema (Kuiper, 1981; Markus, 1977). A self-schema is a cognitive generalization of past experiences that individuals use to organize and process self-related information, such as personality (Markus, 1977). Personality measurements like RRM involve comparing items to one's self-schemata, and the more central an item is to one's network of self-schemata, the faster the individual will be able to compare and therefore respond. Similarly, if an item is extremely unlike the individual, the individual will be able to reject the item quickly (Akrami et al., 2007; Fekken & Holden, 1992; Holden & Fekken, 1993; Holden et al., 1991; Kuiper, 1981).

Item Response Theory and RRM Scoring

The current scoring framework that RRM follows is closely related to the classical test theory (CTT). CTT assumes that any measured trait has an underlying true score, and the observed score comprises an individual's true score and error (Crocker & Algina, 1986). According to the CTT, ideally, measurement errors are random and can be canceled out when the number of items is appropriately large. However, more recently, scholars have recognized limitations in CTT and have begun to use IRT models in scale construction and item analysis (see Embretson & Reise, 2000; Hambleton et al., 1991 for full discussions on differences between IRT and CTT). Many psychometricians have utilized IRT to shorten existing scales (e.g., Ebesutani et al., 2012) or to improve the psychometric properties of the scales (e.g., Fraley et al., 2000). The current study does not seek to compare CTT and IRT scoring frameworks but to extend the current scoring framework using IRT models for the sake of future scale construction and item analysis. In the following sections, we review some basic concepts of IRT, as well as discuss how using an IRT-based scoring framework can serve as a bridge to future RRM research.

Basic Concepts of IRT. IRT is a model-based measurement that links a person's latent trait to their observed response (Embretson & Reise, 2000). An item's observed response depends on both the person's latent trait and the properties of items. For the purposes of the current study, we will focus only on dichotomous IRT models, in which items are either endorsed or rejected.

The two-parameter logistic (2PL) item response model (Birnbaum, 1968) is one of the most widely used dichotomous models (Embretson & Reise, 2000). In a 2PL model, the relationship between an item's response and a person's latent trait is represented with an item characteristic curve (ICC). ICCs are monotonically increasing curves that take an "S" shape in 2PL models. ICC represents the probability of endorsing an item as a function of the person's latent trait:

$$P(x = 1|\theta) = \frac{\exp[\mathrm{a}(\theta - b)]}{1 + \exp\,[a(\theta - b)]} \tag{1}$$

where x represents the observed response, θ represents the latent trait level, a represents the item's discrimination, and b represents the item's difficulty. In this equation, the probability of item endorsement is defined by one person parameter (θ) and two item parameters: Item difficulty (a) and item discrimination (b). Item difficulty refers to the level of the latent trait that has a probability of endorsement of 0.50. For example, if the difficulty of an item is 1, then a person with a latent trait level of 1 will have a 50 percent chance of endorsing the item. Item difficulty can be viewed as the intercept of the ICC with a horizontal line of endorsement probability equalling 0.50.

Item discrimination represents the ability of an item to differentiate between people with different levels of the latent trait. If the discrimination of an item is 0.00, then the item is incapable of differentiating latent trait levels, and the ICC will therefore be flat. A flat ICC implies no relationship between the trait and the item response. Contrastingly, when item discrimination is large, people with slightly different levels of the latent trait will have a large difference in the probability of item endorsement, resulting in a relatively steep ICC around the person parameter (θ). Item discrimination has been shown to be related to item factor loadings (Takane & de Leeuw, 1987) and item-total correlation under CTT. The more strongly the item relates to the construct being measured, the higher the item discrimination parameter.

Bridging Future RRM Research. Scoring the B5-RRM using an IRT-based framework can serve as a bridge to aid future research in both new scale construction and item analysis. IRT allows for standard errors to vary across scores (Embretson & Reise, 2000; Hambleton et al., 1991). An item that is precise for individuals with a latent trait at one level might not be accurate for people with latent traits at different levels. In other words, an item might provide more information for people at one trait level than for people at another. For example, a calculus item is highly unlikely to discriminate the mathematical abilities among a group of 4th graders because most 4th graders will likely miss the item. Such an item is more appropriate for differentiating the mathematical abilities among college students because the difficulty of the item better aligns with the test taker's ability (θ). To gain appropriate test information, IRT-based assessments must match the item to the test taker's latent trait level. For personality measurements like the B5-RRM, a desirable test should generate appropriate test information across a wide range of the personality trait levels (i.e., contain items with various difficulties). Therefore, if an IRT-based scoring framework demonstrates an appropriate fit to the model, future scale constructions and refinements using RRM can pick items with the goal of maximizing test information instead of ensuring reliability.

Second, IRT models allow for item-level analysis. For example, IRT models are often used to detect differential item functioning (DIF). Angoff (1993) defines DIF as “the simple observation that an item displays different statistical properties in different group settings” (p. 4). If a scale is to be used in selection, DIF analysis can be a critical step for identifying potentially biased items to ensure legal defensibility. A popular way to test for DIF is fitting IRT models to different subgroups and comparing item parameters (Camilli & Shepard, 1994; Embretson & Reise, 2000; Thissen et al., 1993). If an item’s parameters are significantly different across subgroups, then that item should be further examined to determine if it is biased.

Lastly, IRT scoring frameworks allow for the possibility of computer adaptive testing (CAT). In CAT, test items are continuously chosen from an item bank for each individual according to the individual’s estimated latent trait at the moment. CAT is more efficient than traditional testing in that it can obtain more accurate trait level estimates with the same number of items (Embretson & Reise, 2000).

In sum, an IRT-based scoring framework paves the way for future work in RRM, which includes, but is not limited to, DIF studies, the use of CAT, and more nuanced methods of scale development and refinement. As such, introducing an IRT-based measurement into RRM appears to be a logical and meaningful expansion of the RRM literature, which can shed light on future advancements in RRM.

An IRT Approach to Score RRM

The B5-RRM consists of dichotomous items. Therefore, 2PL models are appropriate choices to fit and score the B5-RRM. In fact, many studies have successfully fitted 2PL models to personality inventories (e.g., Chernyshenko et al., 2001; Ferrando & Lorenzo-Seva 2007; Ranger & Ortner, 2011). In the case of B5-RRM, indicating an adjective that is positively related to the trait is “like me,” or a negatively related adjective is “not like me,” are considered item endorsements. The opposites are considered item rejections. Item difficulty represents the latent personality trait level such that the probability of endorsing the adjective is 50 percent. And item discrimination represents the adjective’s ability to differentiate between people with different personality trait levels.

To ensure IRT model fit, Hambleton et al. (1991) recommended examining three aspects of data: Unidimensionality, local independence, and model-data fit. To assess unidimensionality, factor analyses are often conducted (Embretson & Reise, 2000). If the eigenvalues support a one-factor solution and items load well onto the single-factor model, unidimensionality may be inferred. Secondly, one way to detect local dependence is by examining Yen’s (1984) Q_3 statistic, which measures the correlation of items after controlling for people’s latent traits (θ; Embretson & Reise, 2000). Finally, a popular way to assess model-data fit is through the chi-square (χ^2) statistic, which measures the goodness-of-fit between the model and data (Embretson & Reise, 2000).

IRT Models with Response Latency

Several studies have revealed that incorporating response latency increases test information of dichotomous personality measures (e.g., Ferrando & Lorenzo-Seva 2007; Ranger & Ortner, 2011; Wang & Hanson 2005). In Ranger and Ortner’s (2011) study, incorporating response latency increased the test information of 2PL models by up to 17%. However, such IRT models have not been applied to measures that assess Big Five personality traits. The current study seeks to fill a gap in the literature by applying an IRT model that incorporates response latency to measure the Big Five personality traits. As the RRM currently incorporates response latency into the scoring of traits, an IRT model that introduces response latency is a natural extension of the current RRM.

The tenets of the distance-difficulty (DD) hypothesis (Ferrando & Lorenzo-Seva, 2007; Kuncel & Fiske, 1974) can explain how including response latency increases test information of IRT models. The DD hypothesis states that the uncertainty of endorsing an item is the highest when the item difficulty level, or item location in the case of a personality test, is close to the respondent's level of the latent trait. Response latency can be a good measure of uncertainty, such that the longer a respondent takes to respond, the more uncertain they are regarding the item. Relatedly, Ferrando and Lorenzo-Seva (2007) proposed one of the earliest IRT models that incorporates response latency to assess personality. They used 2PL parameters (i.e., item difficulty and item discrimination) to estimate the person-item distance. Then, the person-item distance was used to estimate a latency submodel, such that the larger the distance, the higher the response latency. However, in a later article, Ranger and Otner (2011) argued that this model does not differentiate between the endorsement and rejection of an item, and is, therefore, unsatisfactory. These authors contend that when a person with a high latent trait level is presented with an item located at a low trait level, the situation in which this person endorses the low-difficulty item should be considered differently from the improbable case that this person rejects the item. Ranger and Otner (2011) then proposed a model that connects the probability of responses to response latency, such that the more probable a response is, the faster the response latency will be. The current study follows Ranger and Otner's (2011) model by incorporating response latency into the 2PL model in assessing the B5-RRM.

Research Questions

The current study investigates the appropriateness of applying IRT-based scoring frameworks to the B5-RRM. First, the study evaluates whether B5-RRM responses fit appropriately to the 2PL model. Thus, research question (RQ) 1 asks

RQ1. Does the B5-RRM demonstrate an appropriate fit to the 2PL model?

In addition to the 2PL model, the study is also interested in fitting the RRM data to Ranger and Otner's (2011) model, which allows for an investigation of whether response latency provides information beyond examinee responses. Therefore, RQ2 asks

RQ2a. Does the B5-RRM demonstrate an appropriate fit to the 2PL response latency model?

RQ2b. How much does incorporating response latency increase test information of the 2PL model?

The study also examines whether IRT-based measurement models for RRM demonstrate appropriate construct validity. Furthermore, the study seeks to test whether the construct validity is comparable to those of the original RRM scoring model. The following RQs guide the investigation of construct validity.

RQ3a. Does IRT-based scoring frameworks of the B5-RRM demonstrate convergent validity?

RQ3b. Is the convergent validity of the IRT-based RRM scores different from the original RRM scores?

RQ4a. Does IRT-based scoring frameworks of the B5-RRM demonstrate discriminant validity?

RQ4b. Is the discriminant validity of the IRT-based RRM scores different from the original RRM scores?

Finally, the current study investigates the criterion-related validity of the IRT-based RRM scoring framework. The following RQs guide the investigation on criterion-related validity.

RQ5a. Does IRT-based scoring frameworks of the B5-RRM demonstrate criterion-related validity?

RQ5b. Is the criterion-related validity of the IRT-based RRM scores different from the original RRM scores?

Method

Sample

The study recruited 615[1] participants from Amazon's Mechanical Turk (MTurk). To be included in the study, participants must be over 18 years old, employed full-time (i.e., work at least 35 hours per week), and work within the United States. These criteria were enforced through MTurk's worker screening system. The study did not screen for native language; therefore, non-native English speakers might be present in this sample.

Following Meade and Craig's (2012) guidelines, four criteria were used to identify and remove careless responses. First, participants who incorrectly answered the instructed response items were removed (n = 103). Second, participants who had a Maximum Longstring (i.e., answered a large number of consecutive items with the same response) larger than 99% of the

[1] The sample size required was estimated to be 500. The proposal (Appendix M) explains how this number was obtained.

sample were removed (*n* = 7). Third, the Mahalanobis distance was used to identify and remove participants whose response pattern (confidence = .999, *n* = 75). Lastly, "psychological synonyms" (Goldberg, 2000, cited in Johnson, 2005) were used to identify and remove careless respondents. "Psychometric synonyms" are the 30 pairs of distinct items with the highest correlations. If a participant's responses to these synonyms are not highly correlated, then there is a high likelihood that they are responding carelessly. As Johnson (2005) recommends, frequency curve was used to decide the cutoff point for this study. If a participant had a negative synonym correlation, they were considered as careless respondents (*n* = 203). Additionally, after removing careless respondents, 44 participants did not provide valid responses to the B5-RRM, resulting in invalid or missing RRM responses.

After removing careless participants, the final sample consisted of 184 participants (29.92% of the original sample).[2] The original and the final samples had similar demographic distributions. Of the final participants, 100 (54.35%) were male, 83 (45.11%) were female, and 3 (0.54%) were non-binary; 122 (66.30%) self-reported to be non-Hispanic Caucasian, 11 (4.98%) Hispanic, 15 (8.15%) African American, 8 (4.35%)) Asian, 21 (11.41%) Native American, 1 (0.54%) Pacific Islander, 4 (2.17%) multi-racial, and 2 (1.09%) did not respond. The average age of the final sample was 39.42 (*SD* = 12.03). The majority of participants had bachelor's degrees (55.43%) and had worked for more than 10 years (57.61%).

---

[2] The current study observed an unusually large number of careless respondents. Because recent research identified declining data qualities from MTurk (e.g., Chmielewski & Kucker, 2020; Kennedy et al., 2020), this study utilized extensive screening to ensure the quality of data.

Measures

B5-RRM

The study distributed the B5-RRM assessment following Meade et al.'s (2020) guidelines for RRM administration. Each Big Five factor presented 30 adjectives to the participants. Because RRM can repeat adjectives, only the first valid response (i.e., response made within 2.5 seconds) to each adjective was used in this study. The sample's split-half reliabilities were $r_{xx} = .83$ for extraversion, $r_{xx} = .87$ for agreeableness, $r_{xx} = .84$ for conscientiousness, $r_{xx} = .83$ for neuroticism, and $r_{xx} = .55$ for openness. Appendix A shows the list of adjectives used in the B5-RRM.

Likert-Type Personality

A Likert-type personality measurement was included to examine the convergent and discriminant validity of the B5-RRM. The current study used the 50-item version of the International Personality Item Pool - Five-Factor Model (IPIP-FFM; Goldberg, 1992, 1999; Goldberg et al., 2006). Even though longer IPIP scales exist, this shorter version is less likely to cause fatigue and careless responding in participants (Bowling et al., 2020; Gibson & Bowling, 2019; Wise, 2010). Furthermore, studies have shown that the 50-item IPIP is not any less valid than longer versions of the IPIP (Cucina et al., 2005; Lim & Ployhart, 2006). The reliabilities were $\alpha = .84$ for extraversion, $\alpha = .81$ for agreeableness, $\alpha = .84$ for conscientiousness, $\alpha = .88$ for neuroticism, and $\alpha = .74$ for openness (also referred to as intellect openness). Appendix B shows the IPIP-FFM 50-item version. The scale used a 5-point Likert scale ranging from *very inaccurate* (1) to *very accurate* (5).

Construct Validity Measures

In addition to the IPIP, this study also collected constructs that have been shown to relate to various Big Five personality traits to further examine construct validity. Two self-report surveys measuring general self-efficacy (GSE; Appendix C) and cultural intelligence (CQ; Appendix D), respectively, as well as one delay discounting task that measures trait delay of gratification, were administered.

General Self-Efficacy. GSE, or the self-judgment of how an individual performs across different situations (Chen et al., 2001), has been shown by a meta-analysis to relate to the Big Five personality traits (Judge et al., 2002). Judge et al. (2002) found that the GSE construct has communality with neuroticism. The New General Self-Efficacy Scale (NGSE; Chen et al., 2001) was used to measure GSE. The NGSE contained 8 items (α = .90) and was rated on a 5-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5). GSE was expected to be related to neuroticism, conscientiousness, and extraversion (Judge et al., 2002).

Cultural Intelligence. CQ, defined as an individual's ability to effectively handle culturally diverse situations (Earley & Ang, 2003), has also been related to personality (Ang et al. 2006; Sahin et al. 2012). Personality traits such as extraversion and openness can influence how an individual perceives international experiences and help develop CQ (Sahin et al., 2012). CQ was measured with the Cultural Intelligence Scale (CQS; Ang et al., 2007). This scale consisted of four factors: metacognitive CQ (4 items; α = .80), cognitive CQ (6 items; α = .94), motivational CQ (5 items; α = .81), and behavioral CQ (5 items; α = .87). The CQS was measured on a 5-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5). According to a study, different factors of CQ were expected to relate to different personality traits (Ang et al., 2006). Metacognitive CQ was expected to correlate with conscientiousness and openness, cognitive CQ with extraversion and openness, motivational CQ with extraversion and openness, and behavioral CQ with extraversion, agreeableness, neuroticism, and openness.

Delay of Gratification. The delay discounting task served as a behavioral measure to assess delay of gratification, which is related to conscientiousness (Roberts et al., 2012). The current study utilized a short delay discounting task, the 5-trial adjusting delay task (Koffarnus & Bickel, 2014). The task presented to the participants a series of 5 questions between some amount of a delayed commodity (i.e., $1000) and half that amount (i.e., $500) available immediately (e.g., “Which would you rather have? $500 now or $1000 in 3 weeks?”). The amount of delay was updated in real-time according to answers to the previous questions. Final trait scores were obtained using the parameter table Koffarnus and Bickel (2014) provided in their study. The lower the trait score, the longer delay a person is willing to take for a given gratification. The trait score was expected to correlate with conscientiousness.

Criterion-Related Validity Measures

Self-report scales that measure job performance, organizational citizenship behavior (OCB), counterproductive work behavior (CWB), organizational commitment (OC), and turnover intentions, respectively, were used to assess the criterion-related validity of the B5-RRM. Appendices E-I show items that appeared in the following scales, respectively.

Job Performance. Job performance was measured with the Individual Work Performance Questionnaire (IWPQ; Koopmans et al., 2013, 2014). The IWPQ consisted of three factors: task performance (5 items; α = .79), contextual performance (8 items; α = .85), and CWB (5 items; α = .93). Items were rated on a 5-point Likert scale ranging from *never* (1) to *always* (5). Meta-analytic results revealed that task performance and CWB should correlate with conscientiousness, agreeableness, and neuroticism (Hurtz & Donovan, 2000; Grijalva &

Newman, 2015). Contextual performance was expected to relate to extraversion, agreeableness, and conscientiousness (LePine & Van Dyne, 2001).

Organizational Citizenship Behavior. OCB was measured with the 10-item version of the OCB-Checklist (OCB-C; Spector et al., 2010). In this sample, parallel analysis (Horn, 1965) suggested for a two-factor solution and an exploratory factor analysis (EFA) with oblique rotation revealed the two factors to be interpersonal-related (6 items; α = .88) and organization-related (4 items; α = .79), respectively (see Appendix F for details). Items were rated on a 5-point Likert scale ranging from *never* (1) to *every day* (5). According to a meta-analysis, OCB was expected to correlate with agreeableness and conscientiousness (Chiaburu et al., 2012; Organ et al., 2006).

Counterproductive Work Behavior. CWB was measured with the 10-item version of the CWB-Checklist (CWB-C; Spector et al., 2010; α = .96). Items were rated on a 5-point Likert scale ranging from *never* (1) to *every day* (5). Meta-analytic results suggested that CWB should be associated with agreeableness, conscientiousness, and neuroticism (Grijalva & Newman, 2015).

Organizational Commitment. The current study measured OC with the Organizational Commitment Questionnaire (OCQ; Allen & Meyer, 1990). The OCQ consisted of three factors: affective commitment (6 items; α = .82), continuance commitment (6 items; α = .83), and , normative commitment (6 items; α = .81). Items were measured on a 5-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5). Affective commitment was expected to correlate with extraversion, continuance commitment with extraversion, conscientiousness, neuroticism, and openness, and normative commitment with extraversion and agreeableness (Erdheim et al., 2006).

Turnover Intentions. The Turnover Intentions Scale (TIS; Bothma & Roodt, 2013) was used to measure turnover intentions. The TIS consisted of 6 items (α = .73) and was rated on 5-point Likert scales (see Appendix I for anchors). Meta-analytic results suggested turnover intentions to relate to agreeableness, conscientiousness, and neuroticism (Zimmerman, 2008).

Table 1 shows all the above-mentioned expected correlates of the Big Five personality traits.

Procedure

Upon recruitment, participants were first prompted to sign a consent form. Participants first completed the B5-RRM and the IPIP-FFM in random order, followed by all construct validity and criteria measures. All measures were completed under a low stakes setting that prompted for honest responses by asking participants to select the responses that best described themselves. Lastly, demographic variables were recorded (i.e., gender, race, age, education, and work experience). Moreover, following Meade and Craig's (2012) recommendations in screening for careless responses, two instructed response questions (e.g., "Respond with 'disagree' for this item") were included. The first screening question was embedded within the IPIP-FFM, and the second within the OCQ. After completing all measures and passing both instructed response screening questions, participants were compensated $0.50. All measures were recorded using Qualtrics, and RRM assessments were administered through an assessment tool called PerSight, which was embedded in Qualtrics.

Results

2PL Model Fit

RQ1 asked to evaluate the fit of RRM data to the 2PL model. Each Big Five personality trait was examined separately following Hambleton et al.'s (1991) guidelines. First, EFA was used to examine unidimensionality. Using the RRM responses (i.e., 0's and 1's without response latency), EFA supported the unidimensionality assumption in extraversion, agreeableness, conscientiousness, and neuroticism. Table 2 shows the eigenvalues and variance explained by the first two factors of each personality trait. For extraversion, agreeableness, conscientiousness, and neuroticism, only the eigenvalue of Factor 1 was larger than 1 (Kaiser, 1960), and the difference between eigenvalues of Factor 1 and Factor 2 was large enough to result in notable changes in slopes (Cattel, 1966). For openness, even though the first eigenvalue was the only value larger than 1, the difference between the first two eigenvalues was relatively small. The Scree plot of openness did not show a notable drop in slope between Factor 1 and 2 (Figure 1), thus not supporting the unidimensionality assumption (Cattel, 1966). Therefore, the study only used extraversion, agreeableness, conscientiousness, and neuroticism to fit IRT models.

To further ensure unidimensionality, single-factor EFAs were conducted for extraversion, agreeableness, conscientiousness, and neuroticism. Factor loadings were examined: Adjectives with factor loadings smaller than 0.30 were removed from the scales in further analyses (see Appendix J factor loadings of all adjectives). For extraversion and conscientiousness, all adjectives had factor loadings larger than 0.30, thus further supporting the unidimensionality assumption. On the other hand, three adjectives from the agreeableness scale had factor loadings smaller than 0.30: Tolerant (λ = 0.29), uncritical (λ = -0.30), and undemanding (λ = 0.08). However, these three adjectives did not uniformly load onto a second factor, therefore supporting the unidimensionality of agreeableness. Lastly, all six of the negatively coded neuroticism adjectives had poor factor loadings. A two-factor EFA with oblique rotation revealed that four of the negative adjectives loaded onto a second factor (see Appendix J for EFA results). Even though unidimensionality was not supported for the full neuroticism scale, all positively coded adjectives appeared to load well onto a single factor. Therefore, for the sake of this study, the 10 positive adjectives were used to calculate the neuroticism RRM scores.[3]

Stone's (2000) chi-square (χ^2) statistic was also used to examine the fit between the 2PL model and the data. Stone's (2000) χ^2 uses parametric bootstrapping to overcome possible imprecise latent trait estimates in short assessments and to reduce Type I error (Müller, 2020). Items with significant χ^2's ($p < 0.05$) were removed from the scales in further analyses (see Appendix J for all χ^2 statistics). Five extraversion adjectives, two agreeableness adjectives, and one neuroticism adjective were identified. Table 3 shows the list of final adjectives used to calculate RRM scores.

Local dependency was not used to examine fit of the 2PL models because there was not a consensus in the literature on when two items should be considered locally dependent. For example, Yen's (1984) Q_3 statistic is a popular way to measure local independence, but there are no clear guidelines on what cutoff scores to use. Even though a simulation study has suggested to use cutoff scores below and above the Q_3 means (Christensen et al., 2017), there is no further evidence on the appropriateness of this method.

Overall, extraversion, agreeableness, conscientiousness, and neuroticism demonstrated appropriate fits to the 2PL model: They satisfied the unidimensionality and model-data fit

[3] The negatively coded adjectives were omitted instead of treated as a separate factor because the second factor appeared to be spurious. Even after reverse coding, the two factors had a negative correlation ($r = -.15$). Further, many of the negatively coded adjectives had high negative factor loadings in the single factor EFA. Therefore, it was highly possible that the second factor emerged due to the participant's lack of proper reading ability, and therefore analyzing the factor would not add additional value to the study.

assumptions (RQ1). However, openness did not show unidimensionality and was therefore excluded from further IRT-related analyses.

Response Latency Model

RQ2a asked whether the B5-RRM demonstrates an appropriate fit to the 2PL response latency model. The extraversion, agreeableness, conscientiousness, and neuroticism RRM scales were used to fit Ranger and Ortner's (2011) response latency model.

2PL and Linear Regression Model

In the first step, response data were used to fit regular 2PL models to obtain the item parameters. Item parameter estimates were the same across individuals. Second, because response latency could be modeled with the log-normal distribution (Schnipke & Scrams, 2002; van der Linden, 2006), the log response latency of each item was regressed onto the response probability from 2PL models. For each item, the standard residuals were plotted against fitted values to examine homoscedasticity and linearity, and the Q-Q plot was used to examine normality. Table 4 shows the 2PL model item parameters, regression coefficients, and R^2 values. The response probability was positively correlated with the log response latency for almost every item, and 25 out of the 55 regression coefficients were statistically significant. Response probability explained 0% to 12% of the variance in response latency (Ranger & Ortner, 2011).

Item-level fit to the response latency model was determined by examining the regression coefficient significance, linearity, normality, and homoscedasticity. The agreeableness and conscientiousness scales demonstrated appropriate fit to the response latency model (RQ2a). Six of the 13 agreeableness items and 12 of the 20 conscientiousness (three items had significant regression coefficients but were not homoscedastic) items fit the response latency model. On the other hand, only three of the 13 extraversion items and one of the nine neuroticism items fitted the response latency model. Therefore, only the agreeableness and conscientiousness scales were analyzed further. For the third and fourth steps, the scales were reconstructed by selecting items with good response latency model fits. The reconstructed agreeableness scale included: Patient, fair-natured, kind, quarrelsome, antagonistic, and vindictive; The reconstructed conscientiousness scale included: Organized, orderly, responsible, constant, efficient, unreliable, irresponsible, disorganized, inexact, inconsistent, undependable, and unsystematic.

Multilevel Regression Model

In the third step, a multilevel model was used to fit the log response latency of the selected items to decompose response latency variance into individual differences and item-specific effects. Equation (2) displays the multilevel model:

$$\begin{aligned} &\text{Level 1:} \quad \log(t_{\text{ij}}) = \beta_{0\text{ij}} + \beta_{1\text{ij}} P(x|\theta)_{\text{ij}} \\ &\text{Level 2:} \quad \beta_{0\text{i}} = \gamma^{00} + \mu_{0\text{i}} \\ &\qquad\qquad\quad \beta_{1\text{i}} = \gamma_{10} \end{aligned} \tag{2}$$

In this model, the dependent variable was log response latency, and the independent variable was the probability of response (x) given a person's latent trait level (θ). Greek letters represented the values of coefficients related to outcome variables (i.e.., intercept and slope). In level 1, the intercept represented the expected level of log response latency for person i on item j when their response probability was zero. The slope, β_{1ij}, represented the item-specific effect of response probability for person i. The individual intercepts (β_{0i}) and slopes (β_{1i}) became outcome variables in level 2, where γ_{00} represented the average time demand of an item, u_{0i} represented person i's general response speed, and γ_{10} represented the average item-specific effect of response probability across the sample. On average, the higher the response probability, the longer the log response latency was observed in both agreeableness (γ_{10}= 0.04, p = 0.63) and conscientiousness (γ_{10}= 0.11, p = 0.07), but neither of these relationships was statistically significant.

Test Information

In the fourth step, test information at given latent trait levels (θ) was calculated using the following equation and parameter estimates from (1) and (2):

$$I(\theta) = \sum_{i=1}^{n} a_i^2 \frac{\exp\left[a_i(\theta - b_i)\right]}{\left[1+\exp(a_i(\theta - b_i))\right]^2} + \sum_{i=1}^{n} \frac{a_i^2 \beta_1^2}{\sigma^2} \left(\frac{\exp\left[a_i(\theta - b_i)\right]}{\left[1+\exp(a_i(\theta - b_i))\right]^2}\right)^2 \qquad (3)$$

where n was the number of items in the scale and σ2was the variance of the log response latency distribution. The first summand was the information of an item response given latent trait level θ, and the second summand represented the information gained from considering response latency. To answer RQ2b, which asked how much incorporating response latency into the 2PL model increased test information, test information of the 2PL and the 2PL response latency models were compared. Figure 2 shows test information curves of the 2PL model and the 2PL response latency model for agreeableness and conscientiousness, respectively. The inclusion of response latency increased the test information by 0.44% for agreeableness and 3.28% for conscientiousness (RQ2b). Given such negligible increases in test information and large reductions in the number of items, 2PL models without incorporating response latency were used in further analysis.

Validity

After demonstrating the practicality of fitting the B5-RRM to 2PL models, the validity of this IRT-based scoring framework was examined and compared to that of the original RRM scoring framework. RRM scores were calculated following Meade et al.’s (2020) guidelines (referred to as the original RRM scores for the rest of the discussion), and participant latent trait levels (θ) of extraversion, agreeableness, conscientiousness, and neuroticism were estimated using maximum likelihood from the fitted 2PL models (referred to as the IRT RRM scores for the rest of the discussion). Appendix K displays the correlation table with means and standard deviations of all measures.

Construct Validity

RQ3a and RQ4a asked whether the IRT-based scoring framework demonstrated convergent and discriminant validity. Construct validity was first examined using the multitrait-multimethod matrix (MTMM; Campbell & Friske, 1959). Table 5 shows the MTMM comparing the original RRM, the IRT RRM, and the IPIP.

All monotrait-heteromethod correlations (i.e., validity diagonals) were statistically significant and practically meaningful. The observed monotrait-heteromethod correlations were also larger than any other correlations on the same row and column within each heteromethod block (Campbell & Friske, 1959), thus supporting convergent validity (RQ3a). Relatedly, the observed heterotrait correlations were generally lower than monotrait-heteromethod correlations, except those three correlations between agreeableness and conscientiousness (r = .60 for all relationships, p < .001) were larger than the correlation between IRT RRM agreeableness and IPIP agreeableness (r = .59, p < .001). Therefore, discriminant validity was only partially supported because the scale's ability to discriminate between agreeableness and conscientiousness appeared to be compromised (RQ4a).

To compare the convergent validity of the original and IRT RRM (RQ3b), the William's test (Steiger, 1980) was used to test the difference between coefficients on the two validity diagonals. The monotrait-heteromethod correlations between the original RRM and IPIP were larger than those between the IRT RRM and IPIP for extraversion (t(183) = 3.15, p = .002), agreeableness(t(183) = 0.47, p = .64), and conscientiousness (t(183) = 1.19, p = .24), but smaller for neuroticism ($t(183) = -1.36$, $p = .17$). Therefore, results partially supported the original RRM scoring framework having better convergent validity (RQ3b). Regarding discriminant validity, the original RRM scores show similar results compared to the IRT RRM scores: All observed heterotrait correlations were lower than all observed monotrait-heteromethod correlations except for the three above-mentioned relationships. Therefore, there was no evidence supporting differences in discriminant validity between the two RRM scoring frameworks (RQ4b).

In addition to using the MTMM, construct validity was also examined using GSE, CQ, and delay of gratification measures. Table 6 shows the correlations between personality measures and the above-mentioned construct validity measures. As expected, GSE was significantly related to extraversion ($r = .27$, $p = .01$), conscientiousness ($r = .37$, $p < .001$), and neuroticism ($r = -.27$, $p = .01$). However, the relationships between CQ factors and the personality traits were different from expected (see Table 1 for details). Conscientiousness did not significantly predict delay of gratification, but there was a moderate correlation in the expected direction ($r = -.20$, $p = .17$). Nevertheless, the IRT RRM scores had construct validity coefficients that were in the same directions and similar patterns as the IPIP. Furthermore, because the expected CQ-personality relationships were derived from a single study, it was reasonable that CQ did not correlate to personality in the same way. Overall, correlations with these construct validity measures further supported the construct validity of the IRT RRM (RQ3a and 4a).

The William's test (Steiger, 1980) was used to compare construct validity between the original and the IRT RRM (see Appendix L for full results). Most construct validity coefficients between the two scoring frameworks were not significantly different, except for the relationship between agreeableness and delay of gratification ($t(183) = 2.00$, $p = .04$). The relationship was larger in the IRT RRM framework (*r* = -.19, *p* = .23) compared to the original framework (*r* = -.13, *p* = 1.00). Nevertheless, because neither relationship was significant, it was difficult to interpret which scoring framework showed superior validity. Therefore, this analysis did not provide further evidence regarding differences in construct validity between the two scoring frameworks (RQ3b and 4b).

Criterion-Related Validity

Criterion-related validity was examined using task performance, contextual performance, CWB, OCB, OC, and turnover intentions. Table 7 shows the correlations between personality measures and the above-mentioned criterion-validity measures. Task performance, contextual performance, CWB, and turnover intentions were significantly correlated with the expected personality traits (see Table 1), except for task performance with neuroticism (*r* = -0.21, *p* = .12) and contextual performance with agreeableness (*r* = .07, *p* = 1.00). For OCB and OC, even though some correlations were different from the expected relationships (see Table 1), the IRT RRM scores correlated with these criteria in the same directions and similar patterns compared to the IPIP. Therefore, these unexpected relations might be specific to the sample. Overall, criterion-related validity of the IRT RRM scoring framework was generally supported (RQ5a).

The William's test (Steiger, 1980) was used to compare criterion-related validity between the original and the IRT RRM (see Appendix L for full results). Most validity coefficients between the two scoring frameworks were not significantly different, except for the correlation of conscientiousness with affective commitment ($t(183) = 2.42$, $p = 0.02$). The relationship was smaller in the IRT RRM framework (*r* = .38, *p* < .001) compared to the original framework (*r* = .45, *p* < .001). However, because the relationship between affective commitment and conscientiousness was unexpected, it was difficult to infer which RRM framework had superior criterion-related validity (RQ5b).

Discussion

The current study extends the existing scoring framework of the B5-RRM by relating latent personality traits to observed RRM scores and response latency using IRT. Out of the five Big Five personality traits, extraversion, agreeableness, conscientiousness, and neuroticism demonstrate appropriate fit to the response-only 2PL model; Agreeableness and conscientiousness demonstrate appropriate fit to the 2PL model incorporating response latency. Moreover, latent trait estimates from the IRT models demonstrate acceptable construct and criterion-related validity. However, there is inconclusive evidence comparing the validities between the original and the IRT-related scoring frameworks of the B5-RRM.

Implications for Research and Practice

Unidimensionality

In addition to extending the scoring framework, the current study also provides further item-level analysis of the B5-RRM. More specifically, the neuroticism scale appears to have two factors and the openness scale does not appear to be unidimensional. Moreover, many adjectives have poor factor loadings (i.e., factor loadings were smaller than 0.30 or negative; see Appendix J). By observation, many of these poorly loaded items are negatively worded adjectives, such as “undemanding” ($\lambda = 0.08$), “uncritical” ($\lambda = -0.30$), and “unemotional” ($\lambda = -0.21$). A plausible explanation for such loadings is that these adjectives are less common and may require longer processing times. However, given the RRM context that prompts rapid responses, it is likely that participants do not have enough time to fully comprehend these words before making their responses.

The majority of samples Meade et al. (2020) used to validate the B5-RRM were undergraduate students, but 26% of the current sample reports the highest level of education received to be below college. Such a discrepancy in education may have resulted in a lower level of average reading ability in the current sample, thus preventing some participants from fully understanding the adjectives within the short frame of time. Additionally, the possibility of having non-native English speakers in the sample can also contribute to participants' lack of ability to fully understand the adjectives. Nevertheless, the use of an internet-based crowdsourced sample helps examine the generalizability of the B5-RRM. It is suggested that future research in RRM consider examining the required reading ability of items, especially when administered to populations that are non-native English speakers or have lower levels of education.

Response Latency

The current study is the first known study to apply IRT models incorporating response latency to measure the Big Five personality traits. Response latency is significantly related to RRM responses for many adjectives. However, it appears that incorporating response latency adds little test information. Such results are not surprising because response time has been reported to have little contribution to the B5-RRM scores. Meade et al. (2020) reported that when B5-RRM scores were computed without response latency, these scores correlated highly (r > .97) with the B5-RRM scores considering response latency. Therefore, results from both the current study and previous literature suggest that response latency of the B5-RAM provides limited information regarding the test taker’s personality. According to Ranger and Ortner’s (2011) interpretation, the majority of variance in response latency is likely explained by the person’s response speed and the item’s demand of response time.

Another finding regarding response latency is the counterintuitive relationship between response latency and the probability of a test taker making a particular response (i.e., 0 or 1; referred to as response probability). In the current study, RRM response probability is generally positively related to response time. However, Ranger and Ortner (2011) found that response probability was negatively correlated with response time. In other words, these authors found that the more probable a test taker is going to endorse or reject an item, the quicker they respond. On the other hand, in the current study, higher response probability is associated with longer response times. For example, if a test taker has a high probability of answering “like me”, they tend to respond slower compared to when the probability is low. It appears that response time may be an implication of how cautious the participant is when making a response, rather than a measure of how closely an adjective is related to one’s self-schemata. The more cautious a test taker is, the longer time they may spend considering the response and the more likely they choose the response with the higher probability. It appears that the DD hypothesis (i.e., the uncertainty of an item response is negatively related to the difference between the item location and the person’s latent trait level) does not hold true under the rapid response context. Given the counterintuitive finding regarding response latency, the current study suggests future RRM research consider giving more weights to responses with longer response times or allowing for longer response times.

Validity

One of the biggest caveats this study reveals is the high correlation between the conscientiousness and agreeableness factors of the B5-RRM (r = .60), which implies the compromised discriminant validity between the two constructs. While both scales show good face validity (i.e., the adjectives appear to be measuring the appropriate constructs without overlaps), an EFA reveals that all adjectives measuring agreeableness and conscientiousness load well onto a single factor (i.e., factor loadings = .35-.73). Future research should investigate whether the RRM format is making these two constructs to correlate unusually highly.

The current study also contributes to the RRM literature by further validating the B5-RRM using new construct and criteria measures. The estimated trait scores correlate with GSE, task performance, contextual performance, CWB, and turnover intentions as expected, but do not correlate with CQ, OCB, and OC in the expected directions. However, when measuring the Big Five personality with the IPIP, most of these unexpected correlations are still present and significant. Because the IPIP has been widely validated (Ehrhart et al., 2006; Goldberg, 1992, 1999; Goldberg et al., 2006) and shows good reliability in this sample, it is reasonable to suggest that these unexpected relationships are not due to scale deficiency or contamination.

Even though these unexpected relationships can be sample-specific, it is also possible that the relationship between variables has changed since the previous studies were conducted. For example, extraversion is the strongest trait that relates to the interpersonal-related OCB factor (r > .20), while previous meta-analysis suggested a smaller relationship between the two variables (Chiaburu et al., 2011). The outbreak of the coronavirus disease, which resulted in over 70% of U.S employees to be working remotely from home (Gallup Organization, 2021), may have changed the relationship between extraversion and OCB. When working remotely, extraversion may become a significant driver for employees to seek and initiate interpersonal interactions, such as lending a compassionate ear to a coworker.

IRT-Based Scoring Framework

Evidence regarding whether applying an IRT-based scoring framework improves the validity of the B5-RRM is inconclusive. Nevertheless, the successful application of 2PL models to the extraversion, agreeableness, conscientiousness, and neuroticism scales sheds light on the future development of RRM. The estimated item location and discrimination parameters allow researchers to examine the quality of items. For example, if constructing a shorter version of the scale, researchers can select items with a wide range of location parameters and large discrimination parameters. Furthermore, the test information curve allows researchers to examine the different accuracy the scale provides at different levels of the latent trait. For example, the test information curve of conscientiousness is slightly flatter compared to that of agreeableness (see Figure 2). Such results imply that the conscientiousness scale can capture a slightly wider range of information regarding the latent trait. Overall, the successful application of the IRT-based scoring framework of the B5-RRM has built a foundation for future item-level analysis and new scale construction.

Limitations and Future Directions

One of the most obvious limitations of the current study is the subsetting of items. Many items were removed from analyses because the items did not satisfy IRT assumptions (i.e., unidimensionality and model-data fit). In particular, this study follows a set of rules that are relatively strict in eliminating items for the sake of theoretical soundness and better model fit. As a result, findings from the current study may not generalize to the B5-RRM that are publicly in use (i.e., the B5-RRM using all available adjectives).

The second limitation is the use of crowdsourced data. Recent research identified declining data qualities from MTurk and suggested mitigating this decline by extensive data screening (e.g., Chmielewski & Kucker, 2020; Kennedy et al., 2020). This data quality issue was evident in the current study: Nearly 70% of the sample was not included in the final sample due to careless or invalid responses. Even though the data quality of the final sample was acceptable, the extensive identification of careless responses might have created sampling bias and restriction of range. For example, Ward et al. (2017) found removing careless responses to create more conscientious, more agreeable, and less extroverted samples.

The third limitation is that all assessments were administered under a low-stakes setting. However, when used in the workplace, personality tests are often administered under high-stakes settings such as selection. Previous research has identified the Big Five personality test to be non-invariant between high-stakes and low-stakes settings (Ion & Iliescu, 2017). Even though the RRM is more resistant to faking compared to Likert-type personality tests (Meade et al., 2020), it is unknown whether the validity and item properties of the RRM will be maintained under high-stakes settings. Therefore, it is advised that future RRM research investigate measurement invariance between low-stakes and high-stakes situations.

The fourth limitation is the use of a self-report job performance measure. Self-report measures can produce significant biases that threaten the validity of research, especially on sensitive constructs like job performance (Donaldson & Grant-Vallone, 2002). Therefore, results regarding job performance should be interpreted with caution. It is suggested that future research validates the RRM using more objective performance measures.

Conclusion

The current study demonstrates the practicality of applying an IRT-based scoring framework to the B5-RRM and is the first to relate latent Big Five personality traits to both observed test responses and observed response latency. The B5-RRM is also further analyzed at the item level and validated with additional work-related measures. Overall, the study provides additional validity evidence to the B5-RRM and paves the foundation for future RRM development.

References

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

Akrami, N., Hedlund, L. E., & Ekehammar, B. (2007). Personality scale response latencies as self-schema indicators: The inverted-U effect revisited. *Personality and Individual Differences*, *43*(3), 611-618. https://doi.org/10.1016/j.paid.2006.12.005

Allen, N. J., & Meyer, J. P. (1990). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology*, *63*(1), 1-18. https://doi.org/10.1111/j.2044-8325.1990.tb00506.x

Ang, S., Van Dyne, L., & Koh, C. (2006). Personality correlates of the four-factor model of cultural intelligence. *Group & Organization Management*, *31*(1), 100-123. https://doi.org/10.1177/1059601105275267

Ang, S., Van Dyne, L., Koh, C., Ng, K. Y., Templer, K. J., Tay, C., & Chandrasekar, N. A. (2007). Cultural intelligence: Its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. *Management and Organization Review*, *3*(3), 335-371. https://doi.org/10.1111/j.1740-8784.2007.00082.x

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology.* In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Lawrence Erlbaum Associates, Inc. https://psycnet.apa.org/record/1993-97193-001

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, *44*(1), 1-26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next?. *International Journal of Selection and assessment*, *9*(1-2), 9-30. https://doi.org/10.1111/1468-2389.00160

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317-335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability.* In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Addison-Wesley. https://ci.nii.ac.jp/naid/10011544105/

Bothma, C. F., & Roodt, G. (2013). The validation of the turnover intention scale. *SA Journal of Human Resource Management, 11*(1), 1-12. https://hdl.handle.net/10520/EJC134707

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*. https://doi.org/10.1177/1094428120947794

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage. https://us.sagepub.com/en-us/nam/methods-for-identifying-biased-test-items/book3416

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81. https://doi.org/10.1037/h0046016

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods*, *4*(1), 62-83. https://doi.org/10.1177/109442810141004

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523-562. https://doi.org/10.1207/S15327906MBR3604_03

Chiaburu, D. S., Oh, I.-S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 96*(6), 1140–1166. https://doi.org/10.1037/a0024004

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178-194. https://doi.org/10.1177/0146621616677520

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4*(1), 5–13. https://doi.org/10.1037/1040-3590.4.1.5

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt McDougal. https://eric.ed.gov/?id=ed312281

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*(1), 417-440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology, 17*(2), 245–260. https://doi.org/10.1023/A:1019637632584

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-yet-effective measures of the Big Five Factors of Personality. *Psychological Assessment*, *18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Earley, P. C., & Ang, S. (2003). Cultural intelligence: Individual interactions across cultures. Stanford University Press. https://dl.acm.org/doi/10.5555/1405871

Ebesutani, C., Regan, J., Smith, A., Reise, S., Higa-McMillan, C., & Chorpita, B. F. (2012). The 10-item Positive and Negative Affect Schedule for Children, child and parent shortened versions: Application of item response theory for more efficient assessment. *Journal of Psychopathology and Behavioral Assessment*, *34*(2), 191–203. https://doi.org/10.1007/s10862-011-9273-2

Ehrhart, K. H., Roesch, S. C., Ehrhart, M. G., & Kilian, B. (2008). A test of the factor structure equivalence of the 50-item IPIP Five-factor model measure across gender and ethnic groups. *Journal of Personality Assessment, 90*(5), 507-516. https://doi.org/10.1080/00223890802248869

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92*(2), 386–395. https://doi.org/10.1037/0021-9010.92.2.386

Embretson, S. E., & Reise, S. P. (2000). *Multivariate applications books series. Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410605269

Erdheim, J., Wang, M., & Zickar, M. J. (2006). Linking the Big Five personality constructs to organizational commitment. *Personality and Individual Differences*, *41*(5), 959-970. https://doi.org/10.1016/j.paid.2006.04.005

Fekken, G. C., & Holden, R. R. (1992). Response latency evidence for viewing personality traits as schema indicators. *Journal of Research in Personality*, *26*(2), 103-120. https://doi.org/10.1016/0092-6566(92)90047-8

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525-543. https://doi.org/10.1177/0146621606295197

Fisher, P. A., Robie, C. & Rock, L. (2020, April 23-25). Does general mental ability saturate faking behavior on the RRM? In A. W. Meade (Chair), *Life in the fast lane: Advances in rapid response measurement* [Symposium]. The Society of Industrial and Organizational Psychology 35th Annual Conference, virtual.

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. https://doi.org/10.1037/0022-3514.78.2.350

Gallup Organization. (2021). *The Gallup poll briefing.* Gallup Organization. https://news.gallup.com/poll/329501/majority-workers-continue-punch-virtually.aspx

Gibson, A. M., & Bowling, N. A. (2019). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment, 36*(2), 410-420. https://doi.org/10.1027/1015-5759/a000526.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26

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

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528. https://doi.org/10.1016/S0092-6566(03)00046-1

Grijalva, E., & Newman, D. A. (2015). Narcissism and counterproductive work behavior (CWB): Meta-analysis and consideration of collectivist culture, Big Five personality, and narcissism's facet structure. *Applied Psychology, 64*(1), 93-126. https://doi.org/10.1111/apps.12025

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Measurement methods for the social sciences series, Vol. 2. Fundamentals of item response theory.* Sage Publications, Inc. https://psycnet.apa.org/record/1991-98425-000

Hendy, N., Krammer, G., Schermer, J. A., & Biderman, M. D. (2021). Using bifactor models to identify faking on Big Five questionnaires. *International Journal of Selection and Assessment*, *29*(1), 81-99. https://doi.org/10.1111/ijsa.12316

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*(5), 1270–1285. https://doi.org/10.1037/0021-9010.92.5.1270

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist*, *51*(5), 469–477. https://doi.org/10.1037/0003-066X.51.5.469

Hogan, J., & Ones, D. S. (1997). Conscientiousness and integrity at work. In *Handbook of personality psychology* (pp. 849-870). Academic Press. https://doi.org/10.1016/B978-012134645-4/50033-0

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences*, *44*(1), 311-321. https://doi.org/10.1016/j.paid.2007.08.012

Holden, R. R., & Fekken, G. C. (1993). Can personality test item response latencies have construct validity? Issues of reliability and convergent and discriminant validity. *Personality and Individual Differences*, *15*(3), 243-248. https://doi.org/10.1016/0191-8869(93)90213-M

Holden, R. R., Fekken, G. C., & Cotton, D. H. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(1), 111–118. https://doi.org/10.1037/1040-3590.3.1.111

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. https://doi.org/10.1007/BF02289447

Hough, L. M. (1992). The 'Big Five' personality variables--construct confusion: Description versus prediction. *Human Performance, 5*(1-2), 139-155. https://doi.org/10.1080/08959285.1992.9667929

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*(2-3), 209-244. https://doi.org/10.1080/08959285.1998.9668032

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, *75*(5), 581–595. https://doi.org/10.1037/0021-9010.75.5.581

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

Ion, A., & Iliescu, D. (2017). The measurement equivalence of personality measures across high- and low-stake test taking settings. *Personality and Individual Differences, 110*, 1-6. https://doi.org/10.1016/j.paid.2017.01.008

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution?. *Human Performance*, *13*(4), 371-388. https://doi.org/10.1207/S15327043HUP1304_3

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *Big Five Inventory (BFI)* [Database record]. APA PsycTests. https://doi.org/10.1037/t07550-000

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103-129. https://doi.org/10.1016/j.jrp.2004.09.009

Johnson, J. W., & Schneider, R. J. (2013). Advancing our understanding of process in personality-performance relationship. *Handbook of Personality at Work* (1st ed., pp. 71-100). Routledge. https://doi.org/10.4324/9780203526910

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology*, *83*(3), 693–710. https://doi.org/10.1037/0022-3514.83.3.693

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*(3), 530–541. https://doi.org/10.1037/0021-9010.87.3.530

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141–151. http://dx.doi.org/10.1177/001316446002000116

Koffarnus, M. N., & Bickel, W. K. (2014). A 5-trial adjusting delay discounting task: Accurate discount rates in less than one minute. *Experimental and Clinical Psychopharmacology, 22*(3), 222–228. https://doi.org/10.1037/a0035973

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93*(1), 140–154. https://doi.org/10.1037/0021-9010.93.1.140

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., De Vet, H. C., & Van der Beek, A. J. (2014). Construct validity of the individual work performance questionnaire. *Journal of Occupational and Environmental Medicine, 56*(3), 331-337. https://doi.org/10.1097/JOM.0000000000000113

Koopmans, L., Bernaards, C., Hildebrandt, V., van Buuren, S., Van der Beek, A. J., & de Vet, H. C. (2013). Development of an individual work performance questionnaire. *International Journal of Productivity and Performance Management, 62*(1), 6-28. https://doi.org/10.1108/17410401311285273

Kuiper, N. A. (1981). Convergent evidence for the self as a prototype: The" inverted-U RT effect" for self and other judgments. *Personality and Social Psychology Bulletin*, *7*(3), 438-443. https://doi.org/10.1177/014616728173012

Kuncel, R. B., & Fiske, D. W. (1974). Stability of response process and response. *Educational and Psychological Measurement*, *34*(4), 743-755. https://doi.org/10.1177/001316447403400401

LePine, J. A., & Van Dyne, L. (2001). Voice and cooperative behavior as contrasting forms of contextual performance: Evidence of differential relationships with Big Five personality characteristics and cognitive ability. Journal of Applied Psychology, 86(2), 326–336. https://doi.org/10.1037/0021-9010.86.2.326

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*(2), 63–78. https://doi.org/10.1037/0022-3514.35.2.63

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455. https:doi.org/10.1037/a0028085

Meade, A. W., Pappalardo, G., Braddy, P. W., & Fleenor, J. W. (2020). Rapid Response Measurement: Development of a faking-resistant assessment method for personality.

Organizational Research Methods, *23*(1), 181-207. https://doi.org/10.1177/1094428118795295

Mount, M. K., & Barrick, M. R. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. *Research in Personnel and Human Resources Management*, *13*(3), 153-200. https://doi.org/10.4102/sajip.v29i1.88

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. III. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, *88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them?. *Journal of Statistical Distributions and Applications, 7*(1), 1-12. https://doi.org/10.1186/s40488-020-00108-7

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*(6), 660–679. https://doi.org/10.1037/0021-9010.81.6.660

Organ, D. W., Podsakoff, P. M., & MacKenzie, S. B. (2006). *Organizational citizenship behavior: Its nature, antecedents, and consequences.* Sage Publications. https://doi.org/10.4135/9781452231082

Oswald, F. L., & Hough, L. M. (2011). *Personality and its assessment in organizations: Theoretical and empirical developments.* In S. Zedeck (Ed.), *APA handbooks in psychology®. APA handbook of industrial and organizational psychology, Vol. 2. Selecting and developing members for the organization* (pp. 153–184). American Psychological Association. https://doi.org/10.1037/12170-005

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36*(6), 556-563. https://doi.org/10.1016/S0092-6566(02)00505-6

Ployhart, R. E., Schneider, B., & Schmitt, N. (2005). *Staffing organizations: Contemporary practice and theory* (3rd ed.). CRC Press. https://doi.org/10.1201/9781439847053

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement, 71*(2), 389-406. https://doi.org/10.1177/0013164410382895

Roberts, B. W., Lejuez, C., Krueger, R. F., Richards, J. M., & Hill, P. L. (2014). What is conscientiousness and how can it be assessed? *Developmental Psychology, 50*(5), 1315–1330. https://doi.org/10.1037/a0031109

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*(4), 634–644. https://doi.org/10.1037/0021-9010.83.4.634

Şahin, F., Gurbuz, S., & Köksal, O. (2014). Cultural intelligence (CQ) in action: The effects of personality and international assignment on the development of CQ. *International Journal of Intercultural Relations, 39*, 152-163. https://doi.org/10.1016/j.ijintrel.2013.11.002

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. https://doi.org/10.1037/0021-9010.82.1.30

Schnipke, D., & Scrams, D. (2002). *Exploring issues of examinee behavior: Insights gained from response-time analysis.* In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), C*omputer-based testing: Building the foundation for future assessments* (pp. 237-266). Lawrence Erlbaum. https://doi.org/10.4324/9781410612250

Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *Journal of Applied Psychology*, *95*(4), 781–790. https://doi.org/10.1037/a0019477

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*(1), 58-75. https://doi.org/10.1111/j.1745-3984.2000.tb01076.x

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408. https://doi.org/10.1007/BF02294363

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, 88(3), 500–517. https://doi.org/10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, 34(4), 397-423. https://doi.org/10.1006/jrpe.2000.2292

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*(4), 703–742. https://doi.org/10.1111/j.1744-6570.1991.tb00696.x

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). *Detection of differential item functioning using the parameters of item response models.* In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates, Inc. https://psycnet.apa.org/record/1993-97193-004

Valone, A. L., Meade, A. W., Archibald, J., & Ahmad, U. S. (2020, April 23-25). Development of a rapid response measure of dark personality. In A. W. Meade (Chair), *Life in the fast lane: Advances in rapid response measurement* [Symposium]. The Society of Industrial and Organizational Psychology 35th Annual Conference, virtual.

van der Linden, W. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181-204. https://doi.org/10.3102/10769986031002181

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197–210. https://doi.org/10.1177/00131649921969802

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*(5), 323-339. https://doi.org/10.1177/0146621605275984

Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior, 76*, 417-430. https://doi.org/10.1016/j.chb.2017.06.032

Wilgus, S., Meade, A. W., Priest, L., & Archibald, J. (2020, April 23-25). Job satisfaction rapid response measure. In A. W. Meade (Chair), *Life in the fast lane: Advances in rapid response measurement* [Symposium]. The Society of Industrial and Organizational Psychology 35th Annual Conference, virtual.

Wilmot, M. P., & Ones, D. S. (2019). A century of research on conscientiousness at work. Proceedings of the National Academy of Sciences, 116(46), 23004-23010. https://doi.org/10.1073/pnas.1908430116

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95-114. https://doi.org/10.1207/s15324818ame1902_2

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145. https://doi.org/10.1177/014662168400800201

Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. Personnel *Psychology, 61*(2), 309-348. https://doi.org/10.1111/j.1744-6570.2008.00115.x

Table 1

Expected Correlates of the Big Five Personality Traits.

Trait	Expected Correlates
Extraversion	GSE (+), Cognitive CQ (+), Motivational CQ (+), Behavioral CQ (+), Contextual performance (+), Affective commitment (+), Continuance commitment (-), Normative commitment (+)
Agreeableness	Behavioral CQ (+), Task performance (+), Contextual performance (+), OCB (+), CWB (-), Normative commitment (+), Turnover intentions (-)
Conscientiousness	GSE (+), Metacognitive CQ (+), Delay discounting (-), Task performance (+), Contextual performance (+), OCB (+), CWB (-), Continuance commitment (+), Turnover intentions (-)
Neuroticism	GSE (-), Behavioral CQ (+), Task performance (-), CWB (+), Continuance commitment (+), Turnover intentions (+)
Openness	Metacognitive CQ (+), Cognitive CQ (+), Motivational CQ (+), Behavioral CQ (+), Continuance commitment (-)

Note. (+) = positively correlated with trait; (-) = negatively correlated with trait.

GSE = general self-efficacy; CQ = cultural intelligence; OCB = organizational citizenship behavior; CWB = counterproductive work behavior.

Table 2

Eigenvalues and Variance Explained from Exploratory Factor Analysis.

Factor	Eigenvalue	Variance Explained
Extraversion		
1	4.09	0.48
2	0.84	0.10
Agreeableness		
1	4.09	0.49
2	0.71	0.08
Conscientiousness		
1	4.15	0.43
2	0.71	0.07
Neuroticism		
1	2.59	0.38
2	0.95	0.14
Openness		
1	1.64	0.27
2	0.82	0.14

Table 3

Final Adjectives Used to Calculate RRM Scores.

Trait	Adjectives
Extraversion (positive)	Outgoing, Talkative, Conversational, Extraverted, Vocal, Merry, Lively
Extraversion (negative)	Unenthusiastic, Quiet, Introverted, Quiet-spoken, Reserved, Shy
Agreeableness (positive)	Patient, Fair-natured, Gentle, Kind, Good-tempered
Agreeableness (negative)	Spiteful, Quarrelsome, Vengeful, Revengeful, Antagonistic, Ruthless, Harsh, Vindictive
Conscientiousness (positive)	Organized, Thorough, Orderly, Precise, Responsible, Polished, Constant, Efficient, Purposeful, Consistent
Conscientiousness (negative)	Undisciplined, Irresponsible, Unreliable, Haphazard, Disorganized, Inexact, Untidy, Inconsistent, Undependable, Unsystematic
Neuroticism (positive)	Emotional, Hypersensitive, Overemotional, Fearful, Weepy, Worrying, High-strung, Tense, Sensitive

Table 4

Item Parameters of 2PL Models and Response Latency Regression Results.

Adjective	*a*	SE *a*	*b*	SE *b*	*B*	SE *B*	R^2
Extraversion							
Conversational	1.31	0.30	-0.91	0.21	0.45**	0.14	.05
Extraverted	1.26	0.28	0.13	0.16	-0.03	0.23	.00
Introverted	1.37	0.31	0.16	0.15	0.04	0.18	.00
Lively	1.83	0.41	-0.91	0.18	0.07	0.12	.00
Merry	1.20	0.29	-1.03	0.24	0.22	0.16	.01
Outgoing	1.77	0.38	-0.52	0.15	0.31	0.16	.01
Quiet	2.09	0.49	0.67	0.14	0.53**	0.20	.03
Quiet-spoken	1.26	0.30	0.86	0.21	-0.02	0.15	.00
Reserved	1.11	0.29	1.16	0.28	-0.01	0.13	.00
Shy	1.18	0.27	0.19	0.16	0.02	0.17	.00
Talkative	2.03	0.45	-0.44	0.13	0.27	0.14	.01
Unenthusiastic	0.80	0.26	-2.17	0.62	0.47**	0.16	.04
Vocal	1.57	0.35	-0.73	0.17	0.21	0.16	.00
Agreeableness							
Antagonistic	1.61	0.35	-0.92	0.17	0.49**	0.17	.04
Fair-natured	0.98	0.27	-1.55	0.37	0.36**	0.13	.03
Gentle	1.14	0.31	-1.66	0.37	0.01	0.17	.00
Good-tempered	1.18	0.32	-1.61	0.34	0.38**	0.14	.03
Harsh	2.02	0.45	-1.01	0.16	0.29	0.17	.01
Kind	1.23	0.38	-2.06	0.48	0.46**	0.17	.04
Patient	1.09	0.31	-1.73	0.39	0.38*	0.19	.02
Quarrelsome	1.97	0.45	-1.08	0.17	0.33*	0.16	.02
Revengeful	1.81	0.39	-0.89	0.16	0.12	0.17	.00
Ruthless	4.07	1.30	-0.98	0.12	-0.03	0.19	.00
Spiteful	1.47	0.30	-0.35	0.15	0.15	0.14	.00
Vengeful	2.12	0.46	-0.86	0.14	0.21	0.15	.01
Vindictive	2.17	0.46	-0.74	0.13	0.36	0.17	.02
Conscientiousness							
Consistent	2.24	0.52	-1.21	0.17	0.50***	0.14	.06
Constant	1.30	0.33	-1.46	0.29	0.52*	0.27	.02
Disorganized	1.14	0.31	-1.66	0.36	0.33*	0.15	.02
Efficient	1.83	0.56	-1.90	0.34	0.60**	0.20	.04
Haphazard	1.14	0.30	-1.45	0.31	0.31	0.16	.01
Inconsistent	1.78	0.45	-1.50	0.24	0.86***	0.21	.08
Inexact	1.17	0.30	-1.36	0.29	0.62**	0.22	.04
Irresponsible	2.89	0.72	-1.15	0.15	0.53***	0.14	.07

Table 4 (continued).

Adjective	*a*	SE *a*	*b*	SE *b*	*B*	SE *B*	R^2
Orderly	2.04	0.48	-1.27	0.19	0.82***	0.16	.12
Organized	1.85	0.49	-1.61	0.26	0.68**	0.21	.05
Polished	0.73	0.23	-1.40	0.43	-0.04	0.19	.00
Precise	2.24	0.52	-1.19	0.17	0.14	0.13	.00
Purposeful	1.51	0.41	-1.69	0.32	0.16	0.13	.00
Responsible	3.21	1.05	-1.55	0.20	0.78**	0.24	.05
Thorough	1.53	0.39	-1.51	0.27	0.46*	0.18	.03
Undependable	1.21	0.35	-1.92	0.42	0.58**	0.16	.07
Undisciplined	2.61	0.68	-1.35	0.18	0.70***	0.15	.10
Unreliable	1.51	0.44	-1.84	0.36	0.95***	0.19	.12
Unsystematic	1.66	0.40	-1.31	0.22	0.33*	0.17	.02
Untidy	0.90	0.26	-1.58	0.40	0.16	0.17	.00
Neuroticism							
Emotional	1.31	0.32	-0.83	0.20	0.04	0.15	.00
Fearful	1.38	0.33	0.50	0.16	0.27	0.16	.01
High-strung	0.97	0.25	0.40	0.20	0.06	0.20	.00
Hypersensitive	2.02	0.49	0.54	0.14	0.32*	0.14	.02
Overemotional	1.94	0.48	0.64	0.15	-0.02	0.19	.00
Sensitive	0.64	0.22	-1.58	0.55	-0.08	0.17	.00
Tense	1.17	0.28	0.24	0.17	-0.11	0.15	.00
Weepy	1.66	0.40	0.69	0.16	0.22	0.16	.01
Worrying	1.28	0.30	0.33	0.16	0.22	0.27	.00

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

a = item discrimination; *b* = item difficulty.

Table 5

Multitrait-Multimethod Matrix Comparing RRM and IPIP.

Trait	Original RRM					IRT RRM			
	E	A	C	N	O	E	A	C	N
Original RRM									
E	(.83)								
A	.06	(.87)							
C	.15	.60***	(.84)						
N	-.04	-.47***	-.35***	(.83)					
O	.20*	.16	.24**	-.03	(.55)				
IRT RRM									
E	.94***	.08	.13	-.02	.20				
A	.01	.92***	.54***	-.47***	.12	.04			
C	.07	.60***	.90***	-.33***	.21	.07	.60***		
N	.01	-.42***	-.32***	.92***	-.02	.01	-.44***	-.30***	
IPIP									
E	.76***	-.01	.08	.02	.21	.71***	-.06	.00	.04
A	.30***	.60***	.46***	-.18	.24*	.32***	.59***	.46***	-.14
C	.13	.40***	.70***	-.26**	.16	.11	.39***	.67***	-.23*
N	-.18	-.52***	-.43***	.62***	-.14	-.16	-.52***	-.42***	.65***
O	.15	.27**	.28***	-.16	.51***	.14	.26**	.27***	-.15

Table 5 (continued).

Trait	IPIP				
	E	A	C	N	O
Original RRM					
E					
A					
C					
N					
O					
IRT RRM					
E					
A					
C					
N					
IPIP					
E	(.84)				
A	.29***	(.81)			
C	.13	.48***	(.84)		
N	-.16	-.43***	-.41***	(.88)	
O	.26***	.49***	.35***	-.25***	(.74)

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness.

p values were adjusted for multiple tests using the Holm-Bonferroni method.

Table 6

Construct Validity Results.

Measure	IPIP				Original RRM				IRT RRM			
	E	A	C	N	E	A	C	N	E	A	C	N
NGSE	.28***	.40***	.51***	-.36***	.27**	.21	.39***	-.28**	.27**	.19	.37***	-.27***
CQS												
Metacognitive	.28***	.32***	.13	-.16	.24*	.18	.14	-.12	.23*	.17	.16	-.12
Cognitive	.40***	-.15	-.20	.15	.25*	-.25*	-.22	.14	.22	-.28***	-.28***	.01
Motivational	30***	.18	.10	-.16	.18	.09	.10	-.22	.15	.06	.08	-.19
Behavioral	.36***	.09	-.01	.07	.26*	-.06	-.09	.06	.23*	-.08	-.11	.09
Delay of Gratification	.08	-.17	-.14	.15	.07	**-.13**	-.15	.06	.03	**-.19**	-.20	.08

Note. * *p* < .05, ** *p* < .01, *** *p* < .001; *p* values were adjusted for multiple tests using the Holm-Bonferroni method.

Bold values represent significant William's test results between the original and the IRT RRM scores.

E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness; NGSE = New General Self-Efficacy Scale; CQS = Cultural Intelligence Scale.

Table 7

Criterion-related Validity Results.

Measure	IPIP				Original RRM			
	E	A	C	N	E	A	C	N
IWPQ								
Task	0.22	0.47***	0.64***	-0.39***	0.14	0.30***	0.44***	-0.20
Contextual	0.35***	0.28***	0.43***	-0.18	0.35***	0.09	0.31***	-0.10
CWB	0.08	-0.44***	-0.41***	0.44***	0.06	-0.47***	-0.41***	0.28**
OCB-C								
Interpersonal	0.23	0.05	0.03	0.04	0.23	-0.17	-0.05	0.01
Organizational	0.18	-0.17	-0.17	0.28***	0.18	-0.30***	-0.25*	0.22
CWB-C	0.04	-0.58***	-0.56***	0.53***	-0.01	-0.64***	-0.59***	0.38***
OCQ								
Affective	0.33***	0.43***	0.53***	-0.41***	0.25*	0.32***	**0.45*****	-0.25*
Continuance	0.09	-0.10	-0.13	0.30***	0.02	-0.19	-0.20	0.25*
Normative	0.27**	0.14	0.15	-0.04	0.16	-0.03	0.01	0.00
TIS	-0.16	-0.35***	-0.41***	0.50***	-0.12	-0.34***	-0.35***	0.31***

Table 7 (continued).

Measure	IRT RRM			
	E	A	C	N
IWPQ				
Task	0.12	0.29***	0.43***	-0.21
Contextual	0.32***	0.07	0.25*	-0.10
CWB	0.09	-0.50***	-0.44***	0.28**
OCB-C				
Interpersonal	0.26**	-0.20	-0.07	0.00
Organizational	0.15	-0.28**	-0.25*	0.22
CWB-C	-0.01	-0.64***	-0.59***	0.36***
OCQ				
Affective	0.22	0.27**	**0.38*****	-0.22
Continuance	0.01	-0.18	-0.23	0.24*
Normative	0.14	-0.04	-0.03	0.02
TIS	-0.12	-0.34***	-0.33***	0.31***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; p values were adjusted for multiple tests using the Holm-Bonferroni method.

Bold values represent significant William's test results between the original and the IRT RRM scores.

E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness; IWPQ = Individual Work Performance Questionnaire; CWB = counterproductive work behavior; OCB-C = Organizational Citizenship Behavior Checklist; CWB-C = Counterproductive Work Behavior Checklist; OCQ = Organizational Commitment Questionnaire; TIS = Turnover Intention Scale.

Figure 1

Scree Plot of the Openness Scale.

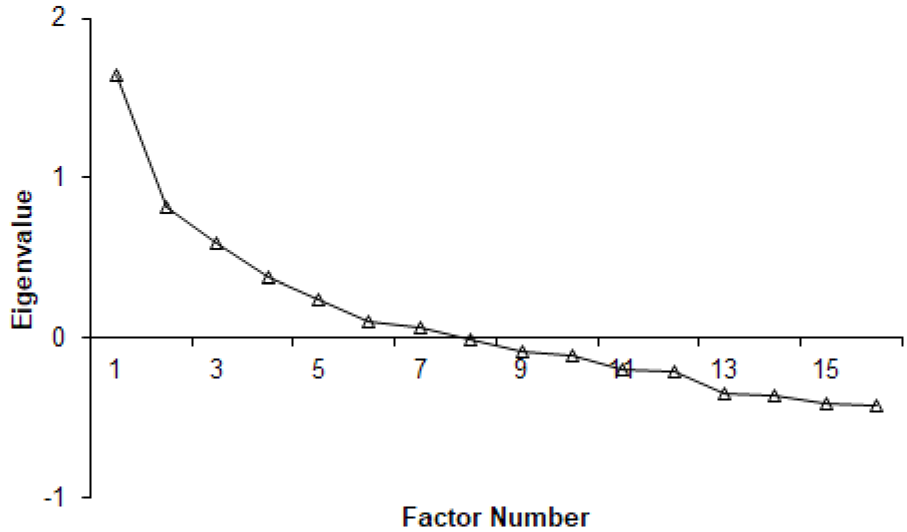
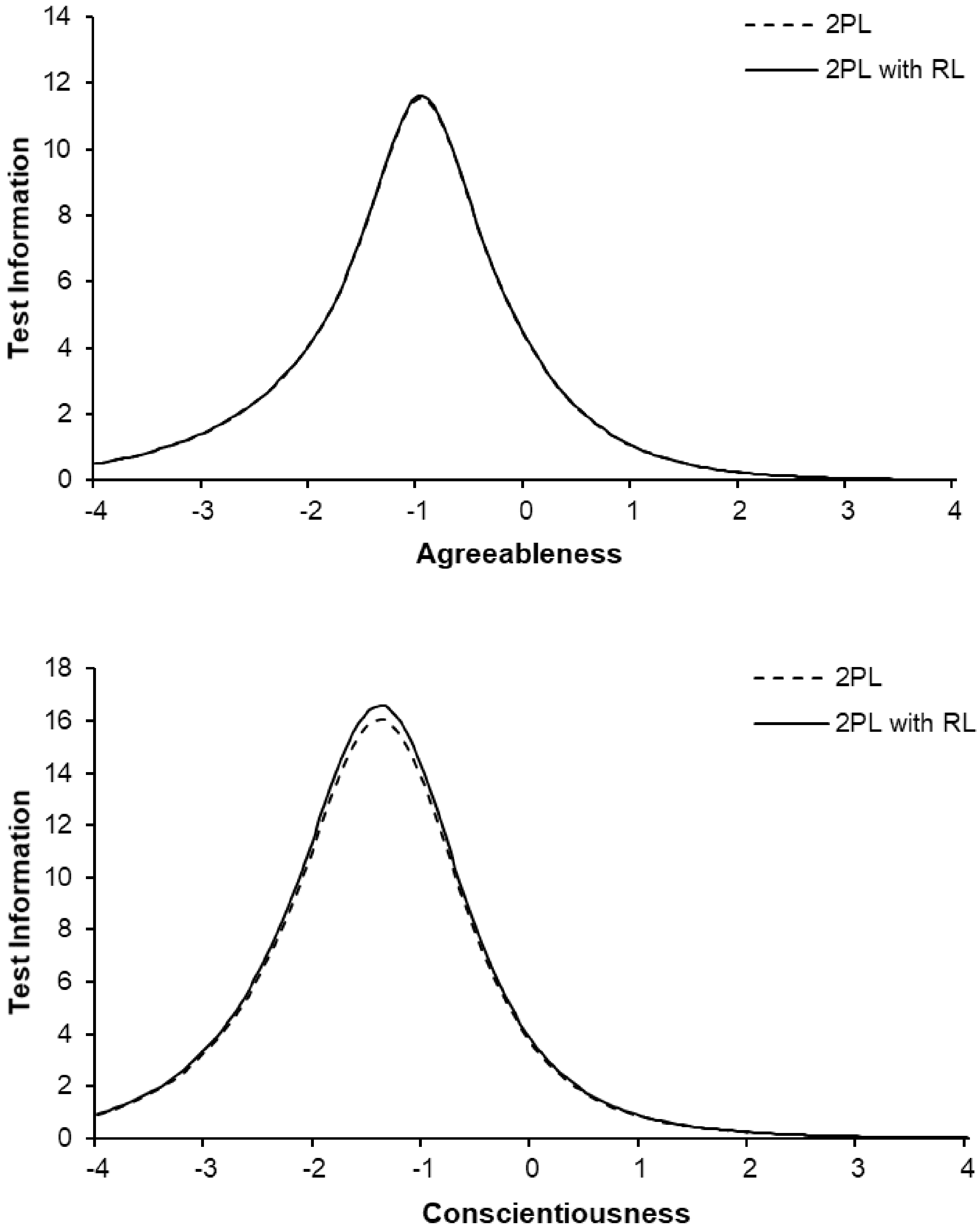


Figure 2

Test Information Curves of Agreeableness and Conscientiousness.



Note. Test Information Curves of the 2PL model and Ranger and Ortner's (2011) 2PL model with response latency. RL = response latency.

APPENDICES

Appendix A
Big Five Rapid Response Measure Adjectives (Meade et al., 2020)

Big Five Personality Trait	Adjectives
Agreeableness (positive)	Tolerant, Uncritical, Patient, Fair-natured, Gentle, Kind, Good-tempered, Undemanding
Agreeableness (negative)	Spiteful, Quarrelsome, Vengeful, Revengeful, Antagonistic, Ruthless, Harsh, Hostile, Vindictive, Rude
Conscientiousness (positive)	Organized, Thorough, Orderly, Precise, Responsible, Polished, Constant, Efficient, Purposeful, Consistent
Conscientiousness (negative)	Undisciplined, Irresponsible, Unreliable, Haphazard, Disorganized, Inexact, Untidy, Inconsistent, Undependable, Unsystematic
Extraversion (positive)	Outgoing, Jolly, Bubbly, Talkative, Conversational, Peppy, Extraverted, Vocal, Merry, Lively
Extraversion (negative)	Withdrawn, Unenthusiastic, Unsociable, Quiet, Introverted, Quiet-spoken, Reserved, Shy
Neuroticism (positive)	Emotional, Hypersensitive, Overemotional, Fearful, Weepy, Worrying, High-strung, Tense, Nervous, Sensitive
Neuroticism (negative)	Unemotional, Unfeeling, Fearless, Emotionless, Tough, Unsentimental
Openness (positive)	Rebellious, Inquisitive, Questioning, Ingenious, Unconventional, Poetic, Profound, Intense, Imaginative
Openness (negative)	Tame, Unphilosophical, Uninquisitive, Unquestioning, Shallow, Unreflective, Traditional

Appendix B
The International Personality Item Pool - Five-Factor Model (Goldberg, 1999)

Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. 1. Very Inaccurate Neither Accurate, 2. Moderately Inaccurate, 3. Neither Accurate Nor Inaccurate, 4. Moderately Accurate, or 5. Very Accurate as a description of you.

Extraversion

E1	Am the life of the party.
E2 (R)	Don't talk a lot.
E3	Feel comfortable around people.
E4 (R)	Keep in the background.
E5	Start conversations.
E6 (R)	Have little to say.
E7	Talk to a lot of different people at parties.
E8 (R)	Don't like to draw attention to myself.
E9	Don't mind being the center of attention.
E10 (R)	Am quiet around strangers.

Agreeableness

A1 (R)	Feel little concern for others.
A2	Am interested in people.
A3 (R)	Insult people.
A4	Sympathize with others' feelings.
A5 (R)	Am not interested in other people's problems.
A6	Have a soft heart.
A7 (R)	Am not really interested in others.
A8	Take time out for others.
A9	Feel others' emotions.
A10	Make people feel at ease.

Conscientiousness

C1	Am always prepared.
C2 (R)	Leave my belongings around.

C3	Pay attention to details.
C4 (R)	Make a mess of things.
C5	Get chores done right away.
C6 (R)	Often forget to put things back in their proper place.
C7	Like order.
C8 (R)	Shirk my duties.
C9	Follow a schedule.
C10	Am exacting in my work.

Neuroticism

ES1	Get stressed out easily.
ES2 (R)	Am relaxed most of the time.
ES3	Worry about things.
ES4 (R)	Seldom feel blue.
ES5	Am easily disturbed.
ES6	Get upset easily.
ES7	Change my mood a lot.
ES8	Have frequent mood swings.
ES9	Get irritated easily.
ES10	Often feel blue.

Intellect Openness

O1	Have a rich vocabulary.
O2 (R)	Have difficulty understanding abstract ideas.
O3	Have a vivid imagination.
O4 (R)	Am not interested in abstract ideas.
O5	Have excellent ideas.
O6 (R)	Do not have a good imagination.
O7	Am quick to understand things.
O8	Use difficult words.
O9	Spend time reflecting on things.
O10	Am full of ideas.

Appendix C
The New General Self-Efficacy Scale (Chen et al., 2001)

Rate 5-point Likert-type scale ranging from strongly disagree (1) to strongly agree (5).

NGSE1 I will be able to achieve most of the goals that I have set for myself.

NGSE2 When facing difficult tasks, I am certain that I will accomplish them.

NGSE3 In general, I think that I can obtain outcomes that are important to me.

NGSE4 I believe I can succeed at most any endeavor to which I set my mind.

NGSE5 I will be able to successfully overcome many challenges.

NGSE6 I am confident that I can perform effectively on many different tasks.

NGSE7 Compared to other people, I can do most tasks very well.

NGSE8 Even when things are tough, I can perform quite well.

Appendix D
The Cultural Intelligence Scale (Ang et al., 2007)

Read each statement and select the response that best describes your capabilities. Select the answer that BEST describes you AS YOU REALLY ARE (1 = strongly disagree; 5 = strongly agree)

Metacognitive CQ

MC1	I am conscious of the cultural knowledge I use when interacting with people with different cultural backgrounds.
MC2	I adjust my cultural knowledge as I interact with people from a culture that is unfamiliar to me.
MC3	I am conscious of the cultural knowledge I apply to cross-cultural interactions.
MC4	I check the accuracy of my cultural knowledge as I interact with people from different cultures.

Cognitive CQ

COG1	I know the legal and economic systems of other cultures.
COG2	I know the rules (e.g., vocabulary, grammar) of other languages.
COG3	I know the cultural values and religious beliefs of other cultures.
COG4	I know the marriage systems of other cultures.
COG5	I know the arts and crafts of other cultures.
COG6	I know the rules for expressing nonverbal behaviors in other cultures.

Motivational CQ

MOT1	I enjoy interacting with people from different cultures.
MOT2	I am confident that I can socialize with locals in a culture that is unfamiliar to me.
MOT3	I am sure I can deal with the stresses of adjusting to a culture that is new to me.
MOT4	I enjoy living in cultures that are unfamiliar to me.
MOT5	I am confident that I can get accustomed to the shopping conditions in a different culture.

Behavioral CQ

BEH1	I change my verbal behavior (e.g., accent, tone) when a cross-cultural interaction requires it.
BEH2	I use pause and silence differently to suit different cross-cultural situations.
BEH3	I vary the rate of my speaking when a cross-cultural situation requires it.
BEH4	I change my nonverbal behavior when a cross-cultural situation requires it.
BEH5	I alter my facial expressions when a cross-cultural interaction requires it.

Appendix E
The Individual Work Performance Questionnaire (Koopmans et al., 2013, 2014)

Read each statement and select the response that best describes you in the past 3 months. Select the answer that BEST describes you AS YOU REALLY ARE.
All items have a recall period of 3 months and a 5-point rating scale ("seldom" to "always" for task and contextual performance, "never" to "often" for counterproductive work behavior).

Task Performance

TP1	I managed to plan my work so that it was done on time.
TP2	My planning was optimal.
TP3	I kept in mind the results that I had to achieve in my work.
TP4	I was able to separate main issues from side issues at work.
TP5	I was able to perform my work well with minimal time and effort.

Contextual Performance

CP1	I took on extra responsibilities.
CP2	I started new tasks myself, when my old ones were finished.
CP3	I took on challenging work tasks, when available.
CP4	I worked at keeping my job knowledge up-to-date.
CP5	I worked at keeping my job skills up-to-date.
CP6	I came up with creative solutions to new problems.
CP7	I kept looking for new challenges in my job.
CP8	I actively participated in work meetings.

Counterproductive Work Behavior

CWB1	I complained about unimportant matters at work.
CWB2	I made problems greater than they were at work.
CWB3	I focused on the negative aspects of a work situation, instead of on the positive aspects.
CWB4	I spoke with colleagues about the negative aspects of my work.
CWB5	I spoke with people from outside the organization about the negative aspects of my work.

Appendix F
The Organizational Citizenship Behaviors Checklist (Spector et al., 2010)

How often have you done each of the following things on your present job? (1 = never, 2 = once or twice, 3 = once or twice per month, 4 = once or twice per week, 5 = every day)

OCB-C1	Took time to advise, coach, or mentor a co-worker.
OCB-C2	Helped co-worker learn new skills or shared job knowledge.
OCB-C3	Helped new employees get oriented to the job.
OCB-C4	Lent a compassionate ear when someone had a work problem.
OCB-C5	Offered suggestions to improve how work is done.
OCB-C6	Helped a co-worker who had too much to do.
OCB-C7	Volunteered for extra work assignments.
OCB-C8	Worked weekends or other days off to complete a project or task.
OCB-C9	Volunteered to attend meetings or work on committees on own time.
OCB-C10	Gave up meal and other breaks to complete work.

Note. Exploratory factor analysis with an oblique rotation suggests that the two factors are: individual-related OCB (OCB-C1 to OCB-C6) and organization-related OCB (OCB-C7 to OCB-C10).

Appendix G
The Counterproductive Workplace Behaviors Checklist (Spector et al., 2010)

How often have you done each of the following things on your present job? (1 = never, 2 = once or twice, 3 = once or twice per month, 4 = once or twice per week, 5 = every day)

CWB-C1	Purposely wasted your employer’s materials/supplies
CWB-C2	Complained about insignificant things at work
CWB-C3	Told people outside the job what a lousy place you work for
CWB-C4	Came to work late without permission
CWB-C5	Stayed home from work and said you were sick when you weren’t
CWB-C6	Insulted someone about their job performance
CWB-C7	Made fun of someone’s personal life
CWB-C8	Ignored someone at work
CWB-C9	Started an argument with someone at work
CWB-C10	Insulted or made fun of someone at work

Appendix H
The Organizational Commitment Questionnaire (Allen & Meyer, 1990)

Responses are made on 5-point scales ranging from strongly disagree (1) to strongly agree (5).

Affective Commitment

ACS1	I would be very happy to spend the rest of my career in this organization.
ACS2	I really feel as if this organization's problems are my own.
ACS3 (R)	I do not feel like "part of my family" at this organization.
ACS4 (R)	I do not feel "emotionally attached" to this organization.
ACS5	This organization has a great deal of personal meaning for me.
ACS6 (R)	I do not feel a strong sense of belonging to this organization.

Continuance Commitment

CCS1	It would be very hard for me to leave my job at this organization right now even if I wanted to.
CCS2	Too much of my life would be disrupted if I leave my organization.
CCS3	Right now, staying with my job at this organization is a matter of necessity as much as desire.
CCS4	I believe I have too few options to consider leaving this organization.
CCS5	One of the few negative consequences of leaving my job at this organization would be the scarcity of available alternatives elsewhere.
CCS6	One of the major reasons I continue to work for this organization is that leaving would require considerable personal sacrifice.

Normative Commitment

NCS1 (R)	I do not feel any obligation to remain with my organization.
NCS2	Even if it were to my advantage, I do not feel it would be right to leave.
NCS3	I would feel guilty if I left this organization now.
NCS4	This organization deserves my loyalty.
NCS5	I would not leave my organization right now because of my sense of obligation to it.
NCS6	I owe a great deal to this organization.

Appendix I
The Turnover Intentions Scale (Bothma & Roodt, 2013)

All items have a recall period of 3 months and a five-point rating scale.

TIS1	How often have you considered leaving your job? 1 = never, 5 = always
TIS2	How satisfying is your job in fulfilling your personal needs? 1 = very satisfying, 5 = totally dissatisfying
TIS3	How often are you frustrated when not given the opportunity at work to achieve your personal work-related goals? 1 = never, 5 = always
TIS4	How often do you dream about getting another job that will better suit your personal needs? 1 = never, 5 = always
TIS5	How likely are you to accept another job at the same compensation level should it be offered to you? 1 = Highly unlikely, 5 = highly likely
TIS6	How often do you look forward to another day at work? 1 = always, 5 = never

Appendix J
Factor Loadings and Item Fit Statistic

Extraversion			Agreeableness			Conscientiousness			Neuroticism		
Adjective	Factor Loading	χ^2	Adjective	Factor Loading	χ^2	Adjective	Factor Loading	χ^2	Adjective	Factor Loading	χ^2
Bubbly	0.60	5.02*	Antagonistic	0.65	0.38	Consistent	0.72	1.30	Emotional	0.59	0.10
Conversational	0.63	2.50	Fair natured	0.47	2.94	Constant	0.61	0.92	Emotionless	-0.41	0.81
Extraverted	0.59	2.75	Gentle	0.47	0.53	Disorganized	0.47	2.35	Fearful	0.63	0.08
Introverted	0.65	1.61	Good tempered	0.50	0.66	Efficient	0.59	2.82	Fearless	0.09	0.13
Jolly	0.75	3.07*	Harsh	0.76	0.19	Haphazard	0.49	2.51	High strung	0.42	1.76
Lively	0.77	0.85	Hostile	0.70	4.42*	Inconsistent	0.63	2.12	Hypersensitive	0.71	0.40
Merry	0.67	1.45	Kind	0.48	1.71	Inexact	0.54	0.48	Nervous	0.60	2.28*
Outgoing	0.69	0.92	Patient	0.49	0.86	Irresponsible	0.79	2.18	Overemotional	0.70	0.62
Peppy	0.52	4.07*	Quarrelsome	0.72	1.67	Orderly	0.73	0.39	Sensitive	0.31	0.46
Quiet	0.72	1.63	Revengeful	0.68	2.42	Organized	0.62	0.80	Tense	0.54	0.44
Quiet spoken	0.56	2.67	Rude	0.78	3.89*	Polished	0.42	1.48	Tough	-0.39	1.02
Reserved	0.51	2.26	Ruthless	0.86	0.53	Precise	0.72	1.75	Unemotional	-0.21	0.53
Shy	0.55	2.83	Spiteful	0.68	1.97	Purposeful	0.57	1.41	Unfeeling	-0.46	1.78
Talkative	0.73	0.88	Tolerant	0.29	0.80	Responsible	0.76	0.34	Unsentimental	0.00	1.25
Unenthusiastic	0.47	2.55	Uncritical	-0.30	1.92	Thorough	0.66	1.61	Weepy	0.77	0.46
Unsociable	0.52	3.03*	Undemanding	0.08	2.31	Undependable	0.52	0.13	Worrying	0.59	1.24
Vocal	0.68	0.85	Vengeful	0.70	2.15	Undisciplined	0.76	0.63			
Withdrawn	0.47	6.51*	Vindictive	0.78	0.31	Unreliable	0.57	1.05			
						Unsystematic	0.60	2.37			
						Untidy	0.43	0.66			

Note. * $p < .05$.

Openness	
Adjective	Factor Loading
Imaginative	0.21
Ingenious	0.03
Inquisitive	0.18
Intense	-0.06
Poetic	0.20
Profound	0.27
Questioning	0.41
Rebellious	-0.23
Shallow	0.60
Tame	0.02
Traditional	-0.01
Unconventional	-0.10
Uninquisitive	0.84
Unphilosophical	0.39
Unquestioning	0.78
Unreflective	0.80

Neuroticism Two-factor EFA Results.

	Factor 1	Factor 2
Emotional	**0.84**	0.07
Emotionless	-0.44	**0.63**
Fearful	**0.87**	-0.24
Fearless	0.01	0.09
High.strung	**0.66**	-0.10
Hypersensitive	**0.86**	-0.09
Nervous	**0.72**	0.02
Overemotional	**0.84**	0.04
Sensitive	**0.73**	**0.46**
Tense	**0.67**	0.11
Tough	-0.32	0.21
Unemotional	0.02	**0.96**
Unfeeling	-0.51	**0.61**
Unsentimental	0.18	**0.74**
Weepy	**0.84**	-0.19
Worrying	**0.82**	0.27

Note. Bold values are factor loadings larger than .30.

Appendix K
Correlation Table of All Measures

Adjective	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 IPIP E	3.00	0.81														
2 IPIP A	3.78	0.67	.29**													
3 IPIP C	3.77	0.73	.13	.48**												
4 IPIP N	2.71	0.89	-.16*	-.43**	-.41**											
5 IPIP O	3.75	0.58	.26**	.49**	.35**	-.25**										
6 Orig RRM E	189.31	730.49	.76**	.30**	.13	-.18*	.15*									
7 Orig RRM A	791.14	730.69	-.01	.60	.40**	-.52**	.27**	.06								
8 Orig RRM C	982.21	552.49	.09	.46**	.70**	-.43**	.28	.15*								
9 Orig RRM N	-83.86	823.55	.02	-.18*	-.26**	.62**	-.16*	-.04	-.47**	-.35**						
10 Orig RRM O	322.15	416.87	.21**	.24**	.06*	-.14	-51**	.20**	.16*	.24**	-.03					
11 IRT RRM E	-0.02	1.14	.71**	.32**	.11	-.16	.14	.94**	.08	.13	-.02	.20**				
12 IRT RRM A	-0.13	0.87	-.06	.59**	.39**	-.52**	.26**	.01	.92**	.54**	-.47**	.12	.04			
13 IRT RRM C	-0.16	0.88	.00	.46**	.67**	-.42**	.27**	.07	.60**	.90**	-.33**	.21**	.07	.60**		
14 IRT RRM N	-0.01	1.16	.04	-.14	-.23**	.65**	-.15*	.01	-.42**	-.32**	.92**	-.02	.01	-.44**	-.30**	
15 NGSE	4.15	0.61	.28**	.40**	.51**	-.36**	.47**	.27**	.21**	.39**	-.28**	.16*	.27**	.19*	.37**	-.27**
16 CQS Metacognitive	4.04	0.63	.28**	.32**	.13	-.16*	.40**	.24**	.18*	.14	-.12	.18*	.23**	.17*	.16*	-.12
17 CQS Cognitive	3.38	1.03	.40**	-.15*	-.20**	.15*	.13	.25**	-.25**	-.22**	.14	.08	.22**	-.28**	-.28**	.14
18 CQS Motivational	3.88	0.72	.30**	.18*	.10	-.16*	.40**	.18*	.09	.10	-.22**	.23**	.15*	.06	.08	-.19*
19 CQS Behavioral	3.69	0.82	.36**	.09	-.01	.07	.20**	.26**	-.06	-.09	.06	.08	.23**	-.08	-.11	.09
20 Delay of Gratification	1.76	5.47	.08	-.17*	-.14*	.15*	-.05	.07	-.13	-.15*	.06	-.09	.03	-.19*	-.20**	.08
21 IWPQ Task	4.01	0.59	.22**	.47**	.64**	-.39**	.37**	.14	.30**	.44**	-.20**	.06	.12	.29**	.43**	-.21**
22 IWPQ Contextual	3.87	0.64	.35**	.28**	.43**	-.18*	.30**	.35**	.09	.31**	-.10	.10	.32**	.07	.25**	-.10
23 IWPQ CWB	2.65	1.17	.08	-.44**	-.41**	.44**	-.30**	.06	-.47**	-.41**	.28**	-.19**	.09	-.50**	-.44**	.28**
24 OCB-C I	3.47	0.79	.23**	.05	.03	.04	.00	.23**	-.17*	-.05	.01	-.09	.26**	-.20**	-.07	.00
25 OCB-C O	3.00	0.98	.18*	-.17*	-17*	.28**	-.07	.18*	-.30**	-.25**	.22**	-.03	.15*	-.28**	-.25**	.22**
26 CWB-C	2.07	1.14	.04	-.58**	-.56**	.53**	-.40**	-.01	-.64**	-.59**	.38**	-.26**	-.01	-.64**	-.59**	.36**

Adjective	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
27 OC Affective	3.42	0.89	.33**	.43**	.53**	-.41**	.31**	.25**	.32**	.45**	-.25**	.08	.22**	.27**	.38**	-.22**
28 OC Continuance	3.45	0.80	.09	-.10	-.13	.30**	-.07	.02	-.19**	-.20**	.25**	.04	.01	-.18*	-.23**	.24**
29 OC Normative	3.33	0.89	.27**	.14	.15*	-.04	.10	.16*	-.03	.01	.00	-.16*	.14	-.04	-.03	.02
30 TIS	2.60	0.73	-.16*	-.35**	-.41**	.50**	-.13	-.12	-.34**	-.35**	.31**	.06	-.12	-.34**	-.33**	.31**

Adjective	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1 IPIP E															
2 IPIP A															
3 IPIP C															
4 IPIP N															
5 IPIP O															
6 Orig RRM E															
7 Orig RRM A															
8 Orig RRM C															
9 Orig RRM N															
10 Orig RRM O															
11 IRT RRM E															
12 IRT RRM A															
13 IRT RRM C															
14 IRT RRM N															
15 NGSE															
16 CQS Metacognitive	.38**														
17 CQS Cognitive	.13	.45**													
18 CQS Motivational	.43**	.58**	.21**												
19 CQS Behavioral	25**	.53**	.67**	.57**											
20 Delay of Gratification	-.00	.06	.23**	.04	.13										
21 IWPQ Task	.59**	.28**	.04	.28**	.15*	-.02									
22 IWPQ Contextual	.47**	.33**	.33**	.41**	.34**	.09	.59**								
23 IWPQ CWB	-.21**	.02	.42**	.11	.27**	.21**	-.19**	.08							
24 OCB-C I	.27**	.20**	.37**	.29**	.34**	.16*	.31**	.53**	.31**						
25 OCB-C O	.03	.10	.48**	.15*	.38**	.18*	.02	.35**	.43**	.59**					
26 CWB-C	.029**	-.06	.44**	.02	.23**	.23**	-.36**	-.07	.71**	.30**	.52**				
27 OC Affective	.44**	.21**	.11	.22**	.13	-.03	.44**	.46**	-.26**	.19*	.06	-.38**			
28 OC Continuance	-.04	.19**	.34**	-.00	.28**	.14*	-.03	.06	.33**	.25**	.37**	.36**	-.02		
29 OC Normative	.24**	.20**	.34**	.20**	.30**	.12	.28**	.46**	.11	.41**	.38**	.13	.60**	.27**	

Adjective	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
30 TIS	-.26**	-.06	.03	-.08	.03	.09	-.43**	-.30**	.34**	-.10	.14	.45**	-.62**	.29**	-.42**

Note. * $p < .05$, ** $p < .01$.

E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness; Orig RRM = Original RRM; I = Interpersonal-Related; O = Organizational-Related

Appendix L
William's Test Results

William's Test Results Comparing Construct Validity Coefficients between RRM and IPIP.

Measure	IPIP vs. Original RRM				IPIP vs. IRT RRM				Original vs. IRT RRM			
	E	A	C	N	E	A	C	N	E	A	C	N
NGSE	0.24	3.29***	2.38*	-1.24	0.13	3.51	2.64***	-1.53	-0.20	0.58	0.63	-0.51
CQS												
Metacognitive	0.81	2.37*	-0.11	-0.62	0.80	2.45	-0.50*	-0.67	0.15	0.24	-0.71	-0.04
Cognitive	3.10**	1.62	0.28	0.18	2.09***	2.09	1.32*	0.17	1.41	1.09	1.91	-0.04
Motivational	2.35**	1.46	0.04	0.82	1.81*	1.81	0.31	0.40	1.53	0.84	0.49	-0.96
Behavioral	2.15*	2.17*	1.31	0.13	2.59*	2.59	1.67**	-0.35	1.09	0.99	0.75	-1.03
Delay of Gratification	0.23	-0.54	0.07	1.34	0.34	0.34	0.89	1.13	1.22	2.00*	1.50	-0.56

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness.

p values were adjusted for multiple tests using the Holm-Bonferroni method.

William's Test Results Comparing Criterion-related Validity Coefficients between RRM and IPIP.

Measure	IPIP vs. Original RRM				IPIP vs. IRT RRM				Original vs. IRT RRM			
	E	A	C	N	E	A	C	N	E	A	C	N
IWPQ												
Task	1.50	2.83**	4.44***	-3.13**	1.83	2.97**	4.39***	-3.05**	1.04	0.38	0.24	0.43
Contextual	-0.14	3.09**	2.18*	-1.33	0.44	3.30**	3.21**	-1.37	1.27	0.53	1.95	0.03
CWB	0.34	0.58	0.04	2.83	-0.16	1.08	0.49	2.88**	-1.05	1.12	0.81	-0.15
OCB-C												
Interpersonal	0.05	3.44***	1.54	0.55	-0.56	3.93***	1.74	0.71	-1.14	1.13	0.47	0.29
Organizational	-0.02	1.95	1.52	0.98	0.44	1.62	1.42	1.01	1.03	-0.68	-0.06	-0.02
CWB-C	0.97	1.14	0.57	2.73**	0.73	1.09	0.72	3.28**	-0.33	-0.09	0.31	0.85
OCQ												
Affective	1.79	1.81	1.48	-2.68**	2.01*	2.55*	2.79**	-3.27**	0.85	1.68	2.41*	-0.95
Continuance	1.38	1.49	1.19	0.86	1.42	1.23	1.61	0.96	0.38	-0.55	0.85	0.13
Normative	2.09*	2.63**	2.49*	-0.56	2.37*	2.77**	3.09**	-0.85	1.03	0.36	1.23	-0.59
TIS	-0.78	-0.11	-1.22	3.35***	-0.67	-0.07	-1.46	3.50***	0.07	0.08	-0.51	0.01

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness.

p values were adjusted for multiple tests using the Holm-Bonferroni method.

Appendix M
Proposal

An Item Response Theory Approach to Rapid Response Measurement of Personality

Over the past few decades, there has been an increased optimism in the utility of personality assessments in organizational settings, especially for employee selection (Goldberg, 1993; Hogan et al., 1996; Hogan & Ones, 1997). Such optimism is partially a result of the growing consensus around the Big Five model of personality (Digman, 1990; Goldberg, 1990), which has helped remove one of the largest limitations in personality research - the lack of a common framework for organizing traits (Barrick & Mount, 1991; Goldberg, 1993; Hough, 1992; Hurtz & Donovan, 2000; Mount & Barrick, 1995). Early meta-analyses were able to identify somewhat stable relationships between personality traits and job performance (Barrick & Mount, 1991; Hurtz & Donovan, 2000; Tett et al., 1991). With such growing confidence in the utility of personality measurements, there is a growing need for accurate and efficient personality measurements. Self-report Likert-type scales have been one of the most popular ways to measure personality (e.g., Costa & McCrae, 1992; Goldberg et al., 2006; Gosling et al., 2003; John et al., 1991). Yet, there are challenges for typical Likert-type scale personality measures: Lengthy tests can increase boredom and irritation for participants (Donnellan et al., 2006) and faking in personality measurements, which meta-analyses have demonstrated possible (Viswesvaran & Ones, 1999), can be prevalent in selection settings (Birkeland et al., 2006).

Meade et al. (2020) introduced a new method called rapid response measurement (RRM) to address such challenges. RRM refers to a general measurement method that can be applied to multiple constructs. RRM presents stimuli in rapid succession to test-takers in a computer-based format. When applied to measure the Big Five personality traits, RRM has been shown to be reliable, valid, more time-efficient, and less prone to faking compared to traditional Likert-type scales (Meade et al., 2020). The current study aims to extend the existing scoring framework of RRM by relating latent personality traits to observed RRM scores. More specifically, the study applies item response theory (IRT; Embretson and Reise 2000; Lord, 1980) to score the Big Five Rapid Response Measure (B5-RRM; Meade et al., 2020). Additionally, no known studies have applied IRT models that account for response latencies in assessing Big Five personality traits. Therefore, this study also seeks to showcase IRT's use to model both response latencies and the responses in the B5-RRM.

Personality in Organizational Research

Personality assessment is widely used in organizational research today, given its association with many salient workplace outcomes (Ployhart et al., 2005). Yet, historically, there has been a major challenge in measuring personality: The lack of a common framework to organize traits (Barrick & Mount, 1991; Goldberg, 1993; Hurtz & Donovan, 2000; Hough, 1992; Mount & Barrick, 1995). Such a problem made it difficult for early meta-analytic research to establish stable relationships between personality and other important workplace variables (Hurtz & Donovan, 2000). Because there was no stable and well-accepted taxonomy of personality traits, it was almost impossible to aggregate across studies and obtain meaningful personality correlates. As a result, early personality research was often unsuccessful in showcasing the validity and value of utilizing personality in the workplace (Barrick & Mount, 1991; Hurtz & Donovan, 2000).

Nevertheless, as the consensus around the Big Five as a common trait-organizing framework grew (Barrick & Mount, 1991; Digman, 1990; Goldberg, 1990), personality gained more popularity in organizational research and practices. The Big Five framework (Digman, 1990; Goldberg, 1990) organizes personality traits into five major traits: Openness to experience

(O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N; also referred to as emotional stability). Since then, the Big Five personality traits have become a ubiquitous model in modern personality assessments (Barrick et al., 2001; Oswald & Hough, 2011). In the following sections, the literature on the personality-performance relationship will be reviewed, which will be followed by a discussion of some of the remaining challenges in measuring personality.

Personality and Job Performance

Meta-analyses have found personality to be a crucial predictor of many workplace outcomes such as job satisfaction (Judge et al., 2002), organizational citizenship (Chiaburu et al., 2011), and turnover (Barrick & Mount, 1991; Zimmerman, 2008). Among these criteria, job performance is the most widely studied variable. Research concerning personality and job performance ranges from criterion development to clarifying the context-specific nature of job performance relationships with different traits (Ployhart et al., 2005; Johnson & Schneider, 2013). Many meta-analyses have attempted to explicate the personality-performance relationships (e.g., Barrick & Mount, 1991; Hurtz & Donovan, 2000; Tett et al., 1991). Such studies examined both the overall correlations and occupation-specific relationships.

In terms of the overall personality-performance relationship, the meta-analytic studies averaged correlations across occupational groups. Early meta-analytic studies revealed discrepancies in findings regarding the personality-performance relationship. One meta-analysis found that conscientiousness was the strongest predictor of job performance (Barrick & Mount, 1991). Conversely, meta-analytic findings by Tett et al. (1991) found agreeableness and extraversion to be the strongest and weakest predictor of job performance, respectively. Still, other meta-analyses found moderate yet slightly different relationships between personality traits and job performance (Mount & Barrick, 1995; Salgado, 1997). These early meta-analyses were later criticized for including studies that did not use Big Five measurements of personality, which explained the mixed results (Hurtz & Donovan, 2000). In fact, Hurtz and Donovan (2000) argued that results obtained from using different trait taxonomies should not be aggregated. Therefore, including only studies that used explicit Big Five factors, Hertz and Donovan (2000) conducted a meta-analysis and found conscientiousness to be the most relevant personality predictor of job performance, followed by emotional stability, agreeableness, and extraversion.

In addition to investigating the personality-performance relationships cross-occupationally, the above-mentioned meta-analyses also investigated the relationship across different occupational groups. Results revealed that personality traits correlated to different occupational groups differentially. For example, extraversion was strongly associated with performance in sales, customer service, and managerial jobs, but not in skilled occupations (e.g., accountants) and professional occupations (e.g., engineers; Barrick & Mount, 1991; Hurtz & Donovan, 2000).

Trait activation theory (TAT; Tett & Burnett, 2003; Tett & Guterman, 2000) helps track the complexity of how personality traits impact job performance and workplace behaviors in general. The TAT posits that the extent to which personality traits influence behaviors is moderated by situation trait relevance (i.e., the opportunity for trait expression; Tett & Guterman, 2000). A TAT-based model of personality-performance states that the extent to which personality influences work behavior and job performance is determined by trait-relevant cues (Tett & Burnett, 2003). Accordingly, the nature of the occupation (e.g., tasks, responsibilities) serves as a cue to determine whether a personality trait is relevant to a job. In fact, different behaviors are required to perform well in different occupations, and such behaviors are associated with different personality traits. For example, a salesperson needs to talk to customers, which is a relevant cue that is associated with trait extraversion, but an accountant does not frequently experience the same situational cue. As a result, extraversion is more predictive of job performance among salespeople but not accountants. When meta-analyses aggregate criterion-related correlations for extraversion across these two jobs, the result is an estimate that underestimates the criterion-related validity for sales positions and overestimates the criterion-related validity for accountants.

Overall, meta-analyses support the utility of personality in employee selection due to its recognizable relationship with job performance (Barrick & Mount, 1991; Hurtz & Donovan, 2000; Tett et al., 1991). While conscientiousness appears to be a notable predictor of performance across occupations, the other Big Five factors can also serve as important predictors for specific occupational groups.

Challenges in Measuring Personality

Challenges in personality measurements can still hinder researchers from accurately assessing personality. First, personality measures can become lengthy: Traditional Likert-type personality scales can contain more than 200 items and take up to 40 minutes to complete (e.g., Costa & McCrae, 1992). Such lengthy scales can be especially problematic in organizational research because participants are often required to complete multiple surveys at once, making the full battery of assessment more lengthy and boring (Donnellan et al., 2006). In fact, under low-stakes testing situations, such as participating in a research study, careless responding increases as participants progress further into assessments (Bowling et al., 2020; Gibson & Bowling, 2019; Wise, 2010). The longer an assessment becomes, the higher the possibility participants will start responding carelessly. Furthermore, low-stakes testing is associated with lower motivation levels of participants (Wise, 2010). As a result, demotivated examinees are more likely to experience increasing fatigue and decreasing efforts as they progress into surveys (Ackerman & Kanfer, 2009). Given such concerns, there appears to be a need for more efficient and innovative personality measurements to overcome the problems associated with lengthy assessments.

The potential of applicant faking is another challenge in personality assessments. Meta-analytic findings have demonstrated that, when instructed, test takers are capable of distorting their responses in socially desirable ways (Viswesvaran & Ones, 1999). While some studies suggested negligible faking behaviors in job applicants (e.g., Hough et al., 1990; Ellingson et al., 2007), a meta-analysis found that job applicants do distort their response to personality assessments to present themselves in ways that are more desirable to the employer (Birkeland et al., 2006). Furthermore, there has been a debate as to whether faking will hurt personality assessments' criterion-related validity (Birkeland et al., 2006). While several studies concluded that faking does not threaten the validity of personality measurement in real-world settings (e.g., Hogan et al., 2007; Hough et al., 1990; Ones et al., 1996), other studies suggested that the impact of faking on validity should not be underestimated (e.g., Holden, 2008; Komar et al., 2008; Rosse et al., 1998). Nevertheless, there has been a consensus in the literature that faking on personality measurement can impact hiring decisions (Hough, 1998; Rosse et al., 1998), which can be harmful to organizations when the selection ratio is small (Mueller-Hanson et al., 2003). Thus, there have been several attempts to resist faking in personality tests: For example, imposing a time limit (e.g., Komar et al., 2010), statistically correcting for social desirability or faking (e.g., Ellingson et al., 1999; Hendy et al., 2020), and the use of a forced-choice format (e.g., Jackson et al., 2000).

Overview of Rapid Response Measurement Method

RRM is a general measurement method that attempts to overcome the above-mentioned hurdles in personality measurement (Meade et al., 2020). In RRM, each item is a single adjective that is related to the construct being measured. For example, in the B5-RRM, adjectives describe one of the Big Five personality traits (e.g., adjectives like “social”, “expressive”, and “talkative” describe extraversion). During an RRM session, respondents read adjectives on a computer screen successively and make responses using two keys to indicate if an adjective is “like me” or “not like me,” respectively (i.e., the “i” and “e” key in the desktop version; two virtual buttons in the mobile version). For each item, three variables are recorded: The adjective, the respondent’s response, and the response latency. Upon answering each item, the next adjective will appear on the screen immediately. To promote rapid response, RRM assessments prompt respondents to respond to items as quickly as possible at the beginning of sessions. Respondents will see a warning message if any of their responses take more than 2.5 seconds. Only responses made within 2.5 seconds are considered valid responses and are considered in trait scores calculation. Under such a setting, respondents will be able to answer items fairly quickly. In fact, Meade et al. (2020) reported that for the B5-RRM, 100 adjectives per trait (i.e., 500 items) only took seven to eight minutes to complete. Furthermore, the reliability and validity of B5-RRM were not compromised despite the quick administration, making it an efficient tool for measuring personality.

In response to obstacles related to faking in personality measurements, the B5-RRM is also more faking resistant than traditional Likert-type personality scales. Meade et al. (2020) compared honest and faking conditions within subjects. They found that even though participants in the faking condition could still increase conscientiousness, extraversion, and decrease neuroticism, there was no significant difference between the two groups in agreeableness and openness to experience. More importantly, the effect size of faking was significantly smaller in RRM compared to that in Likert-type personality measurements. The B5-RRM also had a significantly lower correlation with social desirability compared to Likert-type personality scales. Fisher et al. (2020) revealed that the “multisaturation” perspective on faking (Tett & Simonet) could be used to explain why RRM reduces faking. According to the “multisaturation” perspective, faking behavior is determined by the opportunity, motivation, and ability to fake. Fisher et al. (2020) argued that while the opportunity and motivation to fake are dispositional and relatively constant, the ability to fake can be greatly reduced by RRM.

RRM is a measurement method that can be used for constructs that seek measurement efficiency and faking resistance. In addition to the Big Five personality traits, Meade et al. (2020) also validated the Workplace RRM (W-RRM), which measures more specific work-relevant personality traits, such as assertiveness and dependability. RRM has also been successfully applied to measure the dark triad personality (Valone et al., 2020) and job satisfaction (Wilgus et al., 2020). The dark triad personality traits - psychopathy, narcissism, and Machiavellianism (Paulhus & Williams, 2002) - are socially undesirable traits that test-takers often wish to hide. Therefore, applying a faking resistant measurement to the dark triad personality traits can be crucial in many settings.

Scoring Framework

A linearly weighted scoring framework has been used in all known applications of RRM. After reversing reverse coded items, each item score is computed by multiplying the response and the response latency weight, which gives more weight to faster responses. The final score is the average of all item scores. Response latency has been noted as an appropriate indicator of how closely an item is related to a person's self-schema (Kuiper, 1981; Markus, 1977). A self-schema is a cognitive generalization of past experiences that individuals use to organize and process self-related information, such as personality (Markus, 1977). Personality measurements like RRM involve comparing items to one's self-schemata, and the more central an item is to one's network of self-schemata, the faster the individual will be able to compare and therefore respond. Similarly, if an item is extremely unlike the individual, the individual will be able to reject the item quickly (Akrami et al., 2007; Fekken & Holden, 1992; Holden & Fekken, 1993; Holden et al., 1991; Kuiper, 1981).

Item Response Theory and RRM Scoring

The current scoring framework that RRM follows is closely related to the classic test theory (CCT). CCT assumes that any measured trait has an underlying true score, and the observed score comprises an individual's true score and error (Crocker & Algina, 1986). According to the CCT, ideally, measurement errors are random and can be canceled out when the number of items is appropriately large. However, more recently, scholars have recognized limitations in CCT and have begun to use IRT models in scale construction and item analysis (see Embretson & Reise, 2000; Hambleton et al., 1991 for full discussions on differences between IRT and CCT). In fact, many psychometricians have utilized IRT to shorten (e.g., Ebesutani et al., 2012) or improve the psychometric properties of existing scales (e.g., Fraley et al., 2000). The current study does not seek to compare CCT and IRT scoring frameworks but to extend the current scoring framework using IRT models for the sake of future scale construction and item analysis. In the following sections, we will review some basic concepts of IRT, as well as discuss how using an IRT-based scoring framework can serve as a bridge to future RRM research.

Basic Concepts of IRT. IRT is a model-based measurement that links a person's latent trait to their observed response (Embretson & Reise, 2000). An item's observed response depends on both the person's latent trait and the properties of items. For the purposes of the current study, we will focus only on dichotomous IRT models, in which items are either endorsed or rejected.

The two-parameter logistic (2PL) item response model (Birnbaum, 1968) is one of the most widely-used dichotomous models (Embretson & Reise, 2000). In a 2PL model, the relationship between an item response and a person's latent trait is represented with an item characteristic curve (ICC). ICCs are monotonically increasing curves that take an "S" shape in 2PL models. ICC represents the probability of endorsing an item as a function of the person's latent trait:

$$P(x = 1|\theta) = \frac{\exp[\mathrm{a}(\theta - b)]}{1 + \exp\,[a(\theta - b)]} \tag{1}$$

where x represents the observed response, represents latent trait level, a represents item discrimination, and b represents item difficulty. In this equation, the probability of item endorsement is defined by one person parameter (θ) and two item parameters: item difficulty (a) and item discrimination (b). Item difficulty refers to the level of the latent trait that has a probability of endorsement of 0.50. For example, if the difficulty of an item is 1, then a person with a latent trait level of 1 will have a 50 percent chance of endorsing the item. Item difficulty can be viewed as the intercept of the ICC with a horizontal line of endorsement probability equalling 0.50.

Item discrimination represents the ability of an item to differentiate between people with different levels of the latent trait. If the discrimination of an item is 0.00, then the item is incapable of differentiating latent trait levels, and the ICC will therefore be flat. A flat ICC implies no relationship between the trait and the item response. Contrastingly, when item discrimination is large, people with slightly different levels of the latent trait will have a large difference in the probability of item endorsement, resulting in a relatively steep ICC around the person parameter (θ). Item discrimination has been shown to be related to item factor loadings (Takane & de Leeuw, 1987) and item-total correlation under CCT. The stronger the item relates to the construct being measured, the higher the item discrimination parameter.

Bridging Future RRM Research. Scoring the B5-RRM using an IRT-based framework can serve as a bridge to aid future research in both new scale construction and item analysis. IRT allows for standard errors to vary across scores (Embretson & Reise, 2000; Hambleton et al., 1991). An item that is precise for individuals with a latent trait at one level might not be accurate for people with latent traits at different levels. In fact, an item might provide more information for people at one trait level than for people at another. For example, a calculus item is highly unlikely to discriminate the mathematical abilities among a group of 4th graders because most 4th graders will likely miss the item. Such an item is more appropriate for differentiating the mathematical abilities among college students because the difficulty of the item better aligns with the test-taker ability (θ). To gain appropriate test information, IRT-based assessments must match the item to the test-taker's latent trait level. For personality measurements like the B5-RRM, the test must generate desirable test information across a wide range of the personality trait levels (i.e., contain items with various difficulties). Therefore, if an IRT-based scoring framework demonstrates an appropriate fit to the model, future scale constructions and refinements using RRM can pick items with the goal of maximizing test information instead of ensuring reliability.

Second, IRT models allow for item-level analysis. For example, IRT models are often used to detect differential item functioning (DIF). Angoff (1993) defines DIF as “the simple observation that an item displays different statistical properties in different group settings” (p. 4). If a scale is to be used in selection, DIF analysis can be a critical step for identifying potentially biased items to ensure legal defensibility. DIF can be tested by fitting IRT models to different subgroups and comparing item parameters (Camilli & Shepard, 1994; Embretson & Reise, 2000; Thissen et al., 1993). If an item’s parameters are significantly different across subgroups, then that item should be further examined to determine if it is biased.

Lastly, IRT scoring frameworks allow for the possibility of computer adaptive testing (CAT). In CAT, test items are continuously chosen from an item bank for each individual according to the individual’s estimated latent trait at the moment. CAT is more efficient than traditional testing in that it can obtain the same level of test information with a smaller set of items (Embretson & Reise, 2000).

In sum, an IRT-based scoring framework paves the way for future work in RRM, which include, but are not limited to, DIF studies, the use of CAT, and more nuanced methods of scale development and refinement. As such, introducing an IRT-based measurement into RRM appears to be a logical and meaningful expansion of the RRM literature, which can shed light on future advancements in RRM.

An IRT Approach to Score RRM

The B5-RRM consists of dichotomous items. Therefore, 2PL models will be appropriate choices to fit and score the B5-RRM. In fact, many studies have successfully fitted 2PL models to personality inventories (e.g., Chernyshenko et al., 2001; Ferrando & Lorenzo-Seva 2007; Ranger & Ortner, 2011). In the case of B5-RRM, indicating an adjective that is positively related to the trait is “like me,” or a negatively related adjective is “not like me,” will be considered item endorsements. The opposites will be considered item rejections. Item difficulty will represent the latent personality trait level such that the probability of endorsing the adjective is 50 percent. And item discrimination will represent the adjective’s ability to differentiate between people with different personality trait levels.

To ensure IRT model fit, Hambleton et al. (1991) recommended examining three aspects of data: Unidimensionality, local independence, and model-data fit. To assess unidimensionality, factor analyses are often conducted (Embretson & Reise, 2000). If the eigenvalues support a one-factor solution and items load well onto a one-factor model, unidimensionality may be inferred. One way to detect local dependence is by examining Yen’s (1984) Q3 statistic, which measures the correlation of items after controlling for people’s latent traits (θ; Embretson & Reise, 2000). Finally, a popular way to assess model-data fit is through the chi-square (X^2) statistic, which measures the goodness-of-fit between the model and data (Embretson & Reise, 2000).

IRT Models with Response Latencies

The current study seeks to fill a gap in the literature by applying an IRT model that incorporates response latencies to measure the Big Five personality traits. Several studies have revealed that incorporating response latencies increases test information of dichotomous personality measures (e.g., Ferrando & Lorenzo-Seva 2007; Ranger & Ortner, 2011; Wang & Hanson 2005). However, such IRT models have not been applied to measures that assess Big Five personality traits. As the RRM currently incorporates response latency into the scoring of traits, an IRT model that introduces latency is a natural extension of the current RRM.

The tenets of the distance-difficulty (DD) hypothesis (Ferrando & Lorenzo-Seva, 2007; Kuncel & Fiske, 1974) can explain how including response latencies increases test information of IRT models. The DD hypothesis states that the uncertainty of responding to an item is the highest when the item difficulty level is close to the respondent's level of the latent trait. Response latency can be a good measure of uncertainty, such that the longer a respondent takes to respond, the more uncertain they are regarding the item. Relatedly, Ferrando and Lorenzo-Seva (2007) proposed one of the earliest IRT models that incorporates response latencies to assess personality. They used 2PL parameters (i.e., item difficulty and item discrimination) to estimate the person-item distance. Then, the person-item distance was used to estimate a latency submodel, such that the larger the distance, the higher the response latency. However, in a later article, Ranger and Otner (2011) argued that this model does not differentiate between the endorsement and rejection of an item, and is, therefore, unsatisfactory. They contend that when a person with a high latent trait level is presented with an item located at a low trait level, a situation in which this person endorses the low difficulty item should be considered differently from the improbable case that this person rejects the item. Ranger and Otner (2011) then proposed a model that connects the probability of responses to response latencies, such that the more probable a response is, the faster the response latency will be. The current study will follow Ranger and Otner's (2011) model by incorporating response latencies into the 2PL model in assessing the B5-RRM.

Research Questions

The current study investigates the appropriateness of applying IRT-based scoring frameworks to the B5-RRM. First, the study evaluates whether B5-RRM responses fit appropriately to the 2PL model. Thus, research question (RQ) 1 asks that

RQ1. Does the B5-RRM demonstrate an appropriate fit to the 2PL model?

In addition to the 2PL model, the study is also interested in fitting the data to Ranger and Otner's (2011) model, which will allow for the investigation of whether response latencies provide information beyond examinee responses. Therefore, RQ2 asks

RQ2a. Does the B5-RRM demonstrate an appropriate fit to the 2PL response latency model?

RQ2b. How much does incorporating response latencies increase test information of the 2PL model?

The proposed study will also examine whether IRT-based measurement models for RRM demonstrate appropriate construct validities. Furthermore, the study seeks to test whether the construct validities are comparable to those of the original RRM scoring model. The following RQs guide the investigation on construct validities.

RQ3a. Does IRT-based scoring frameworks of the B5-RRM demonstrate convergent validity?

RQ3b. Is the convergent validity of the IRT model different from the original RRM?

RQ4a. Does IRT-based scoring frameworks of the B5-RRM demonstrate discriminant validity?

RQ4b. Is the discriminant validity of the IRT model different from the original RRM?

Finally, the current study will investigate the criterion-related validity of the IRT-based RRM scoring framework. Job performance, which has been shown to strongly relate to conscientiousness and emotional stability (Barrick & Mount, 1995; Hurtz & Donovan, 2000), will be included as a criterion. In addition, general self-efficacy (GSE), or the self-judgement of how an individual performs across different situations (Chen et al., 2001), has been shown by a meta-analysis to relate to the Big Five personality traits (Judge et al., 2002). Judge et al. (2002)

found that the GSE construct has communality with neuroticism. In fact, GSE is strongly related to neuroticism, conscientiousness, and extraversion (Judge et al., 2002). Furthermore, cultural intelligence (CQ), defined as an individual's ability to effectively handle culturally diverse situations (Earley & Ang, 2003), has also been related to personality (Ang et al. 2006; Sahin et al. 2012). Personality traits such as extraversion and openness can influence how an individual perceives international experiences and help develop CQ (Sahin et al., 2012). Moreover, Ang et al. (2006) found that agreeableness is significantly associated with the behavioral aspect of CQ. Given such expected relationships, the proposed study will use job performance, GSE, and CQ to examine the criterion-related validity of the IRT-based B5-RRM. The following RQs guide the investigation on criterion-related validity.

RQ5a. Does IRT-based scoring frameworks of the B5-RRM demonstrate creation-related validity?

RQ5b. Is the creation-related validity of the IRT model different from the original RRM?

Method

Sample

Participants will be recruited from Amazon's Mechanical Turk (MTurk). Even though recent research has identified declining data qualities from MTurk, studies have suggested that this decline can be mitigated by extensive data quality screening (e.g., Chmielewski & Kucker, 2020; Kennedy et al., 2020). The study expects to recruit about 500 participants. Because there is no consensus regarding the appropriate sample size for dichotomous models (Morizot et al., 2007), the participant number is estimated from former empirical studies that used similar IRT models. Ferrando et al. (2007) recruited 750 participants in their empirical study to showcase their model. However, Ranger and Ortner (2011, 2012) demonstrated an appropriate fit to their

2PL response latency model with less than 170 participants. Therefore, in consideration of the fact that some participants will be removed due to data quality screening, 500 participants appear to be able to provide a sufficient sample for accurately estimating 2PLM parameters.

To be included in the study, participants must be over 18 years old, employed (i.e., work at least 20 hours per week), and work within the United States. The current study will include only employed participants because most criteria of interest are work-related.

Measures

B5-RRM

The B5-RRM assessment will be distributed following Meade et al.'s (2020) guidelines for RRM administration. Furthermore, each Big Five factor will present 30 adjectives to the participants. Appendix A shows the list of adjectives used in the B5-RRM.

Likert-Type Personality

A Likert-type personality measurement will be included to examine the convergent and divergent validity of the B5-RRM. The current study will use the 50-item version of the International Personality Item Pool - Five-Factor Model (IPIP-FFM; Goldberg, 1992, 1999; Goldberg et al., 2006). Even though longer IPIP scales exist, this shorter version will be less likely to cause fatigue and careless responding in participants (Bowling et al., 2020; Gibson & Bowling, 2019; Wise, 2010). Furthermore, studies have shown that the 50-item IPIP is not any less valid than longer versions of the IPIP (Cucina et al., 2005; Lim & Ployhart, 2006). Appendix B shows the IPIP-FFM 50-item version. The scale uses a 5-point Likert scale ranging from *very inaccurate* (1) to *very accurate* (5).

Criteria

Three self-report scales that measure job performance, GSE, and CQ, respectively, will be used to assess the criterion-related validity of the B5-RRM. Appendices C, D, and E show items that appear in the following scales, respectively.

Job Performance. Job performance will be measured with the Individual Work Performance Questionnaire (IWPQ; Koopmans et al., 2013, 2014). The IWPQ consists of three factors: task performance (5 items), contextual performance (8 items), and counterproductive work behavior (5 items). Items are rated on a 5-point Likert scale ranging from *seldom/never* (1) to *always/often* (5). Job performance is expected to positively correlate with conscientiousness and emotional stability (Barrick & Mount, 1995; Hurtz & Donovan, 2000).

General Self-Efficacy. The New General Self-Efficacy Scale (NGSE; Chen et al., 2001) will be used to measure GSE. The NGSE contains 8 items and is rated on a 5-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5). Self-efficacy is expected to be positively associated with conscientiousness, extraversion, and emotional stability (Judge et al., 2002).

Cultural Intelligence. CQ will be measured with the Cultural Intelligence Scale (CQS; Ang et al., 2007). This scale consists of four factors: metacognitive CQ (4 items), cognitive CQ (6 items), motivational CQ (5 items), and behavioral CQ (5 items). The CQS is measured on a 7-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5). CQ is expected to positively relate to openness, agreeableness, and extraversion (Ang et al. 2006; Sahin et al. 2012).

Procedure

Upon recruitment, participants will first be prompted to sign a consent form. They will complete the B5-RRM and the IPIP-FFM in random order. Then, they will complete the IWPQ, the NGSE, and the CQS. Lastly, demographic variables will be recorded (i.e., gender, race, age, work experience, and industry). Moreover, following Meade and Craig's (2012) recommendations in screening for careless responses, three instructed response questions (e.g., "Respond with 'disagree' for this item") will be included. Participants who do not pass these screening questions will not be included in the final sample. After completing all measures and passing the instructed response screening questions, they will be compensated with $0.50. Participants will complete all assessments through Qualtrics. RRM assessments will be administered through an assessment tool called PerSignt, which can be embedded in Qualtrics.

Analysis Plan

Data Quality Screening

The current study will follow Meade and Craig's (2012) guidelines to screen for careless responses. First, participants who failed to correctly answer the instructed response questions will be identified and rejected from the final sample. Second, participants who answered a large number of consecutive items with the same response pattern will be identified. Third, Mahalanobis distance, which measures how strongly a participant's response pattern deviates from the rest of the sample, will be calculated to identify participants who deviate from sample norms. Lastly, the "psychometric antonyms" index (Goldberg, 2000, cited in Johnson, 2005) will be used to identify careless responses. "Psychometric antonyms" are the 30 pairs of distinct items with the highest negative correlations. If a participant's responses to these antonyms are not in the opposite direction, then there is a high likelihood that they are responding carelessly.

The exact threshold for removal using these three indices will depend on the sample. As Johnson (2005) recommends, frequency curves should be used to decide cutoff points. All careless response screening will be done with the “careless” package in R (Yentes & Wilhelm, 2021).

2PL Model Fit

RQ1 asks to evaluate the fit of data to the 2PL model. Each Big Five personality trait will be examined separately. First, the “mirt” package in R (Chalmers, 2012) will be used to fit the data to a 2PL model. Only the RRM responses (i.e., 0’s and 1’s) will be used to fit this model. The item difficulty and item discrimination of each adjective will be reported. Analysis of model fit will follow Hambleton et al.’s (1991) suggestions. First, unidimensionality will be investigated through exploratory factor analysis. Parallel analysis, the number of eigenvalues bigger than 1, and factor loadings to a single factor model will be used to evaluate unidimensionality. Second, Yen’s (1984) Q3 statistic will be used to examine local independence. Lastly, the chi-square (X^2) statistic will be used to examine the fit between the 2PL model and the data.

Response Latency Model

RQ2a asks whether the B5-RRM demonstrates an appropriate fit to the 2PL response latency model. Each Big Five personality trait will again be examined separately. This study will follow the four steps outlined in Ranger and Ortner (2011) to model response latency. In the first step, response data will be used to fit a regular 2PL model. The second step involves linear regression analysis. Because response latencies can be modeled with the log-normal distribution (Schnipke & Scrams, 2002; van der Linden, 2006), the log response latencies of each item will be regressed onto the response probability obtained from the 2PL model. The appropriateness of the fit of these linear regression models answers RQ2a. Fit will be evaluated with residual plots:

The standard residuals plotted against fitted values for variance homogeneity and linearity, and Q-Q plots for normality. Only items that show appropriate model fit and a significant relationship between its log response latencies and response probability will be analyzed in steps 3 and 4.

In the third step, a multi-level model will be used to fit log response latencies of the selected items to decompose response latency variance into individual differences and item-specific effects. Equation (2) displays the multilevel model:

$$\text{Level 1:} \quad \log(t_{\mathrm{ij}}) = \beta_{0\mathrm{ij}} + \beta_{1\mathrm{ij}} P(x|\theta)_{\mathrm{ij}}$$

$$\text{Level 2:} \quad \beta_{0\mathrm{i}} = \gamma^{00} + \mu_{0\mathrm{i}}$$

$$\beta_{1\mathrm{i}} = \gamma_{10} \qquad (2)$$

In this model, the dependent variable is log response latency, and the independent variable is the probability of response (x) given a person's latent trait level (). The probability of responses can be estimated using (1). Greek letters represent the values of coefficients related to outcome variables (i.e.., intercept and slope). In level 1, the intercept represents the expected level of log response latency for person i on item j when their response probability is zero. The slope, 1ij, represents the item-specific effect of response probability for person i. The individual intercepts (0i) and slopes (1i) become outcome variables in level 2, where ɣ00 represents the average time demand of an item, u0i represents person i's general response speed, and ɣ10 represents the average item-specific effect of response probability across the sample. In the fourth step, maximum likelihood estimations of a person's latent trait level and general response speed u0ican be made given the person's observed response and response latencies. In this step, test information at a given latent trait level will also be calculated using the following equation and parameter estimates from (1) and (2):

$$I(\theta) = \sum_{i=1}^{n} a_i^2 \frac{\exp\left[a_i(\theta - b_i)\right]}{\left[1+\exp(a_i(\theta - b_i))\right]^2} + \sum_{i=1}^{n} \frac{a_i^2 \beta_1^2}{\sigma^2} \left(\frac{\exp\left[a_i(\theta - b_i)\right]}{\left[1+\exp(a_i(\theta - b_i))\right]^2}\right)^2 \qquad (3)$$

where *n* is the number of items in the scale and 2is the variance of the log response latency distribution. The first summand is the information of an item response given latent trait level , and the second summand represents the information gained from considering response latencies. To answer RQ2b, which asks how much incorporating response latencies into the 2PL model increases test information, test information of the 2PL and the 2PL response latency model will be compared. Test information curves will be plotted to examine differences in test information between the 2PL model and the 2PL response latency model. Furthermore, because (3) shows that it is impossible for the response latency model to have less information than the 2PL model, the percentage increase in test information will also be reported.

Construct Validity

RQ3a and RQ4a ask whether the IRT-based scoring framework demonstrates convergent and discriminant validity. First, participants' latent trait levels () will be estimated and used as final RRM scores (referred to as the IRT RRM scores for the rest of the discussion). Then, the IPIP, IRT RRM scores, and the RRM scores calculated using the original scoring framework (referred to as the original RRM scores for the rest of the discussion) will be compared using the multitrait-multimethod matrix (MTMM; Campbell and Fiske, 1959). If correlation coefficients on the validity diagonal are statistically significant and larger than any other correlations in the same row and column, convergent validity will be supported. The MTMM showing the same pattern of correlations in all heterotrait triangles will also support convergent validity. Relatedly, if all heterotrait correlations are generally lower than monotrait-heteromethod correlations, discriminant validity will be supported.

RQ3b and RQ4b ask whether the IRT and original RRM scores show different construct validities. Several statistics can be compared to answer these RQs. First, the monotrait correlations between the IRT RRM scores and the IPIP will be compared to that between the original RRM scores and the IPIP. If correlations are larger for one RRM scoring method, that scoring method will be considered to have better convergent validity. Second, the percent of heterotrait correlations that are lower than monotrait-heteromethod correlations can also be compared between each RRM scoring method and the IPIP. The scoring method that has a higher percentage will be considered to have better discriminant validity.

Criterion-Related Validity

RQ5a asks if the IRT RRM scores demonstrate criterion-related validity. To examine criterion-related validity, the IRT RRM scores will be correlated with all criteria (i.e., the IWPQ, the NGSE, and the CQS). If these correlations are significant, then we can infer that the IRT RRM scores demonstrate acceptable criterion-related validity. Furthermore, RQ5b asks to compare criterion-related validities of the IRT and the original RRM scores. To answer RQ5b, the same correlation analysis will be applied to the original RRM scores. Then, the William's test (as described by Steiger, 1980) will be used to test for any statistically significant differences between the criterion-related validity coefficients.

References

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

Akrami, N., Hedlund, L. E., & Ekehammar, B. (2007). Personality scale response latencies as self-schema indicators: The inverted-U effect revisited. *Personality and Individual Differences*, *43*(3), 611-618. https://doi.org/10.1016/j.paid.2006.12.005

Ang, S., Van Dyne, L., & Koh, C. (2006). Personality correlates of the four-factor model of cultural intelligence. *Group & Organization Management, 31*(1), 100-123. https://doi.org/10.1177/1059601105275267

Ang, S., Van Dyne, L., Koh, C., Ng, K. Y., Templer, K. J., Tay, C., & Chandrasekar, N. A. (2007). Cultural intelligence: Its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. *Management and Organization Review, 3*(3), 335-371. https://doi.org/10.1111/j.1740-8784.2007.00082.x

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology*. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 3–23). Lawrence Erlbaum Associates, Inc. https://psycnet.apa.org/record/1993-97193-001

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, *44*(1), 1-26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next?. *International Journal of Selection and assessment*, *9*(1-2), 9-30. https://doi.org/10.1111/1468-2389.00160

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International*

Journal of Selection and Assessment, *14*(4), 317-335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397-472). Addison-Wesley. https://ci.nii.ac.jp/naid/10011544105/

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*. https://doi.org/10.1177/1094428120947794

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage. https://us.sagepub.com/en-us/nam/methods-for-identifying-biased-test-items/book3416

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81. https://doi.org/10.1037/h0046016

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods, 4*(1), 62-83. https://doi.org/10.1177/109442810141004

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://dio.org/10.18637/jss.v048.i06.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523-562. https://doi.org/10.1207/S15327906MBR3604_03

Chiaburu, D. S., Oh, I.-S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology*, *96*(6), 1140–1166. https://doi.org/10.1037/a0024004

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science, 11*(4), 464-473. https://doi.org/10.1177/1948550619875149

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*(1), 5–13. https://doi.org/10.1037/1040-3590.4.1.5

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt McDougal. https://eric.ed.gov/?id=ed312281

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417-440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-yet-effective measures of the Big Five Factors of Personality. *Psychological Assessment*, *18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Earley, P. C., & Ang, S. (2003). Cultural intelligence: Individual interactions across cultures. Stanford University Press. https://dl.acm.org/doi/10.5555/1405871

Ebesutani, C., Regan, J., Smith, A., Reise, S., Higa-McMillan, C., & Chorpita, B. F. (2012). The 10-item Positive and Negative Affect Schedule for Children, child and parent shortened versions: Application of item response theory for more efficient assessment. *Journal of Psychopathology and Behavioral Assessment, 34*(2), 191–203. https://doi.org/10.1007/s10862-011-9273-2

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology*, *92*(2), 386–395. https://doi.org/10.1037/0021-9010.92.2.386

Embretson, S. E., & Reise, S. P. (2000). *Multivariate Applications Books Series. Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410605269

Fekken, G. C., & Holden, R. R. (1992). Response latency evidence for viewing personality traits as schema indicators. *Journal of Research in Personality*, *26*(2), 103-120. https://doi.org/10.1016/0092-6566(92)90047-8

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525-543. https://doi.org/10.1177/0146621606295197

Fisher, P. A., Robie, C. & Rock, L. (2020, April 23-25). Does general mental ability saturate faking behavior on the RRM? In A. W. Meade (Chair), *Life in the fast lane: Advances in rapid response measurement* [Symposium]. The Society of Industrial and Organizational Psychology 35th Annual Conference, virtual.

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. https://doi.org/10.1037/0022-3514.78.2.350

Gibson, A. M., & Bowling, N. A. (2019). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, *36*(2), 410-420. https://doi.org/10.1027/1015-5759/a000526.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment,* 4(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26

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

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528. https://doi.org/10.1016/S0092-6566(03)00046-1

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Measurement methods for the social sciences series, Vol. 2. Fundamentals of item response theory.* Sage Publications, Inc. https://psycnet.apa.org/record/1991-98425-000

Hendy, N., Krammer, G., Schermer, J. A., & Biderman, M. D. (2021). Using bifactor models to identify faking on Big Five questionnaires. *International Journal of Selection and Assessment*, *29*(1), 81-99. https://doi.org/10.1111/ijsa.12316

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*(5), 1270–1285. https://doi.org/10.1037/0021-9010.92.5.1270

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist*, *51*(5), 469–477. https://doi.org/10.1037/0003-066X.51.5.469

Hogan, J., & Ones, D. S. (1997). Conscientiousness and integrity at work. In *Handbook of personality psychology* (pp. 849-870). Academic Press. https://doi.org/10.1016/B978-012134645-4/50033-0

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences*, *44*(1), 311-321. https://doi.org/10.1016/j.paid.2007.08.012

Holden, R. R., & Fekken, G. C. (1993). Can personality test item response latencies have construct validity? Issues of reliability and convergent and discriminant validity. *Personality and Individual Differences*, *15*(3), 243-248. https://doi.org/10.1016/0191-8869(93)90213-M

Holden, R. R., Fekken, G. C., & Cotton, D. H. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*(1), 111–118. https://doi.org/10.1037/1040-3590.3.1.111

Hough, L. M. (1992). The 'Big Five' personality variables--construct confusion: Description versus prediction. *Human Performance, 5*(1-2), 139-155. https://doi.org/10.1080/08959285.1992.9667929

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*(2-3), 209-244. https://doi.org/10.1080/08959285.1998.9668032

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*(5), 581–595. https://doi.org/10.1037/0021-9010.75.5.581

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

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution?. *Human Performance, 13*(4), 371-388. https://doi.org/10.1207/S15327043HUP1304_3

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *Big Five Inventory (BFI)* [Database record]. APA PsycTests. https://doi.org/10.1037/t07550-000

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1), 103-129. https://doi.org/10.1016/j.jrp.2004.09.009

Johnson, J. W., & Schneider, R. J. (2013). Advancing our understanding of process in personality-performance relationship. *Handbook of Personality at Work* (1st ed., pp. 71-100). Routledge. https://doi.org/10.4324/9780203526910

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology, 83*(3), 693–710. https://doi.org/10.1037/0022-3514.83.3.693

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, *87*(3), 530–541. https://doi.org/10.1037/0021-9010.87.3.530

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods, 8*(4), 614-629. https://doi.org/10.1017/psrm.2020.6

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology*, *93*(1), 140–154. https://doi.org/10.1037/0021-9010.93.1.140

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., De Vet, H. C., & Van der Beek, A. J. (2014). Construct validity of the individual work performance questionnaire. *Journal of Occupational and Environmental Medicine, 56*(3), 331-337. https://doi.org/10.1097/JOM.0000000000000113

Koopmans, L., Bernaards, C., Hildebrandt, V., van Buuren, S., Van der Beek, A. J., & de Vet, H. C. (2013). Development of an individual work performance questionnaire. *International Journal of Productivity and Performance Management, 62*(1), 6-28. https://doi.org/10.1108/17410401311285273

Kuiper, N. A. (1981). Convergent evidence for the self as a prototype: The" inverted-U RT effect" for self and other judgments. *Personality and Social Psychology Bulletin*, *7*(3), 438-443. https://doi.org/10.1177/014616728173012

Kuncel, R. B., & Fiske, D. W. (1974). Stability of response process and response. *Educational and Psychological Measurement*, *34*(4), 743-755. https://doi.org/10.1177/001316447403400401

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*(2), 63–78. https://doi.org/10.1037/0022-3514.35.2.63

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455. https:doi.org/10.1037/a0028085

Meade, A. W., Pappalardo, G., Braddy, P. W., & Fleenor, J. W. (2020). Rapid Response Measurement: Development of a faking-resistant assessment method for personality. *Organizational Research Methods*, *23*(1), 181-207. https://doi.org/10.1177/1094428118795295

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). *Toward modern psychometrics: Application of item response theory models in personality research*. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (p. 407–423). The Guilford Press. https://doi.org/10.1111/j.1744-6570.2008.01133_6.x

Mount, M. K., & Barrick, M. R. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. *Research in Personnel and Human Resources Management*, *13*(3), 153-200. https://doi.org/10.4102/sajip.v29i1.88

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. III. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660–679. https://doi.org/10.1037/0021-9010.81.6.660

Oswald, F. L., & Hough, L. M. (2011). *Personality and its assessment in organizations: Theoretical and empirical developments.* In S. Zedeck (Ed.), *APA handbooks in psychology®. APA handbook of industrial and organizational psychology, Vol. 2. Selecting and developing members for the organization* (p. 153–184). American Psychological Association. https://doi.org/10.1037/12170-005

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36*(6), 556-563. https://doi.org/10.1016/S0092-6566(02)00505-6

Ployhart, R. E., Schneider, B., & Schmitt, N. (2005). *Staffing organizations: Contemporary practice and theory* (3rd ed.). CRC Press. https://doi.org/10.1201/9781439847053

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*(2), 389-406. https://doi.org/10.1177/0013164410382895

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, *83*(4), 634–644. https://doi.org/10.1037/0021-9010.83.4.634

Şahin, F., Gurbuz, S., & Köksal, O. (2014). Cultural intelligence (CQ) in action: The effects of personality and international assignment on the development of CQ. *International Journal of Intercultural Relations, 39*, 152-163. https://doi.org/10.1016/j.ijintrel.2013.11.002

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. https://doi.org/10.1037/0021-9010.82.1.30

Schnipke, D., & Scrams, D. (2002). *Exploring issues of examinee behavior: Insights gained from response-time analysis.* In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), C*omputer-based testing: Building the foundation for future assessments* (p. 237-266). Lawrence Erlbaum. https://doi.org/10.4324/9781410612250

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408. https://doi.org/10.1007/BF02294363

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, 88(3), 500–517. https://doi.org/10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, 34(4), 397-423. https://doi.org/10.1006/jrpe.2000.2292

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*(4), 703–742. https://doi.org/10.1111/j.1744-6570.1991.tb00696.x

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). *Detection of differential item functioning using the parameters of item response models.* In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 67–113). Lawrence Erlbaum Associates, Inc. https://psycnet.apa.org/record/1993-97193-004

Valone, A. L., Meade, A. W., Archibald, J., & Ahmad, U. S. (2020, April 23-25). Development of a rapid response measure of dark personality. In A. W. Meade (Chair), *Life in the fast lane: Advances in rapid response measurement* [Symposium]. The Society of Industrial and Organizational Psychology 35th Annual Conference, virtual.

van der Linden, W. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181-204. https://doi.org/10.3102/10769986031002181

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197–210. https://doi.org/10.1177/00131649921969802

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*(5), 323-339. https://doi.org/10.1177/0146621605275984

Wilgus, S., Meade, A. W., Priest, L., & Archibald, J. (2020, April 23-25). Job satisfaction rapid response measure. In A. W. Meade (Chair), *Life in the fast lane: Advances in rapid* *response measurement* [Symposium]. The Society of Industrial and Organizational Psychology 35th Annual Conference, virtual.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95-114. https://doi.org/10.1207/s15324818ame1902_2

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145. https://doi.org/10.1177/014662168400800201

Yentes R. D., & Wilhelm F. (2021). *careless: Procedures for computing indices of careless responding*. R package version 1.2.1. https://cran.r-project.org/web/packages/careless/index.html

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, *45*(1), 166-183. https://doi.org/10.3102/0002831207312909