7-14-2016 12:00 AM

Item Properties and the Validity of Personality Assessment

Rachel A. Plouffe, *The University of Western Ontario*

Supervisor: Dr. Donald Saklofske, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Psychology

ABSTRACT

The aim of the current research was to evaluate psychometric properties of personality questionnaire items that can affect their validity, as measured by comparing self-ratings with roommate ratings on a series of personality questionnaires. The item properties under investigation included item content saturation, item social desirability, and mean item responses. Archival data collected between 1981 and 2004 representing various groups of same-sex undergraduate roommate dyads were used for the purpose of this research. Results demonstrated that item content saturation was the most consistent predictor of item response accuracy. Mean item responses also predicted accurate responding curvilinearly, with moderate mean item responses eliciting the most accuracy. In order to better predict outcome variables in education, clinical, and vocational contexts using scores on personality questionnaires, it is important for researchers to employ item selection procedures that take into account the specific item properties that we know can affect personality test validity.

Keywords: test construction; validity; accuracy; content saturation; social desirability; mean responses; personality; self-report.

ACKNOWLEDGEMENTS

First and foremost, I would like to extend many thanks to my family for encouraging and supporting me throughout my many years in academia. I thank my parents, Keri and Marc, for emphasizing the importance of pursuing my academic goals. Thank you for always lending a listening ear, believing in any dream I have, and celebrating the milestones with me.

I am immensely grateful for the support that I have received from my supervisor, Dr. Don Saklofske. Thank you for your guidance and supervision over the past two years. Thank you for your words of encouragement, and for the many hours of work you have put into helping me grow as a student. Thank you for always supporting my ideas, and for affording me countless opportunities to develop experience within the field.

Thank you to my committee members, Dr. Paul Tremblay, Dr. Tony Vernon, and Dr. Julie Aitken Schermer, for taking the time out of your busy schedules to assist me with the completion of this thesis. I very much appreciate the time and effort you have put into helping me to achieve my goals.

Lastly, I would like to express my sincere gratitude to my former supervisor, the late Dr. Sampo Paunonen. I am infinitely grateful to have had the privilege of working and collaborating with you for a year, and for the boundless knowledge that you shared with me. Thank you for cultivating my passion for research. Thank you for sharing with me your expertise, your words of encouragement, your generosity, your ideas, and of course, your witty humour. Without your unwavering support and your invaluable insight, I would not be where I am today in academia.

TABLE OF CONTENTS

CHAPTER 2: METHOD

CHAPTER 3: RESULTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1. Introduction

Personality psychology is predominantly concerned with the study of individual differences. The most common method for assessing personality and individual differences for more than a century has been the self-report personality questionnaire (Jackson & Paunonen, 1980; Paunonen & Hong, 2015; Paunonen & O'Neill, 2010). Traditional methods of administering such questionnaires include the completion of paper-based rating scales, in which participants indicate how representative a specific trait label, behaviour tendency, or attitude is to them (Holden & Troister, 2009). Currently, modern technology allows for these self-report questionnaires to be computerized and tailored to individual respondents, making them arguably one of the most efficient indicators of personality. These measures, regardless of how they are administered, provide researchers with an expedient way to assess an individual's attitudes and behaviours (Jackson & Paunonen, 1980; Paunonen & O'Neill, 2010).

The logic underlying self-report measures of personality is such that individuals possess enough insight into their own psychological processes and past behaviours to make accurate judgments about their personality characteristics (John & Benet-Martinez, 2000; Paunonen & O'Neill, 2010). The notion of the validity of such reports, however, has not gone unchallenged (Epstein, 1983; Epstein & O'Brien, 1985). In the 1960s, personality theory faced a paradigm crisis when a wealth of evidence introduced by critics of self-report personality testing revealed that individuals' responses to personality test items demonstrated little cross-situational consistency. That is, there appeared to be little stability of reported behaviour across time and situations (e.g., Mischel, 1968; Shrauger & Schoeneman, 1979). For instance, Shrauger and Schoeneman (1979) contended that there was little consistency in the agreement between one's

self-perceptions and others' perceptions of them. Furthermore, correlations between personality trait measures and relevant behaviours seldom surpassed a ceiling of .30 (Epstein, 1983; Jackson & Paunonen, 1985). The fundamental assumption underlying personality testing, which maintains that the characteristics and behaviours of individuals remain stable enough across diverse situations to classify them as enduring personality traits, was thus undermined (Epstein & O'Brien, 1985).

The "person-situation" debate, which has since weathered, has provided the basis for a wealth of research conducted by personality theorists concerning the improvement of traditional methods of personality test construction and assessment (Paunonen, 1984). Such improvements involve two fundamental requirements for sound personality measurement: the establishment of the measure's reliability and validity (Clark & Watson, 1995; Epstein, 1983; Loevinger, 1957). Many of the apparent inconsistencies in personality across time reported in some studies can be explained by the use of scales lacking in these psychometric properties (Jackson & Paunonen, 1985). One notable problem with past measures, for example, has been the use of self-report questionnaire items that tend to elicit socially desirable responding from people (Jackson, 1984).

Personality scales, even within the same omnibus questionnaire, typically do not have identical reliabilities or validities. The items on two scales might look similar in style and format, be written by the same item writers, and be the result of the same statistical item selection strategy, yet the scales have different validities. One reason could be differential desirability in the scales' items, as mentioned earlier. But there are other item properties that could be at work including an item's difficulty, wording, direction of keying, face validity, content saturation, and more. The primary purpose of this study is to evaluate some of these psychometric properties of items in terms of their contribution to the validity of self-report measures of personality.

1.1. Evaluating the Validity of Self-Report Personality Measures

In the area of personality test construction, there are a number of ways to evaluate the accuracy of individuals' total scores on self-report personality inventories. In classical test theory, a person's obtained scores on a personality measure is a composite of both his or her true score on the trait plus an error score. The goal of test construction is to maximize the true score variance and to eliminate any random or systematic error variance. In this context, reliability can be defined as the extent to which an obtained score on a measure reflects one's true score on the particular trait in question (Cone, 1981; Foster & Cone, 1995). Validity is the extent to which that true score reflects the construct of interest. This generally requires that new measures be compared to a criterion index that is known to reflect one's true score on the personality trait of interest (Foster & Cone, 1995). In other terms, we need to establish convergent validity for a measure, which we do by comparing different methods for assessing the same construct and looking for agreement (Campbell & Fiske, 1959; Nunnally & Bernstein, 1994).

Convergent validity typically involves testing the relationship between a new measure of a construct with an established criterion measure. For instance, in order to determine such validity, Paunonen (1985) correlated self-reports on one-item ad hoc personality adjective scales written to represent the 20 dimensions of Jackson's (1984) Personality Research Form (PRF), with the 20-item scale scores themselves. These 20 published scales represented the criterion measures used to validate the ad hoc adjective scales.

An alternate way to compute convergent validity coefficients involves having an individual who is well acquainted with the target complete the same questionnaires in a peer rating format, and then to correlate the peer responses with the target responses to items (Holden & Troister, 2009; Paunonen, 1984; Paunonen & O'Neill, 2010). This well-acquainted individual

could be a parent, friend, significant other, teacher, or sibling. High correlations between self- and peer ratings provide evidence for the measure's accuracy or validity. To reiterate, different methods of measuring the same personality construct should produce nontrivial correlations if they are indeed accurate representations of the individual (Campbell & Fiske, 1959; Foster & Cone, 1995). But, as described in the next section, test validation through peer ratings engenders certain issues not relevant to other types of test validation.

1.2. Person Perception and the Validity of Personality Measures

When using self-peer comparisons to validate a psychological measure, characteristics of the self, characteristics of the peer, and psychometric properties of test items are three primary factors that can affect the validity coefficients. Much can be learned of these factors from the voluminous literature on accuracy in person perception. Two particularly salient factors affecting person perception accuracy that have been studied are degree of self-peer acquaintanceship and level of observability of the rated behaviours.

When peer acquaintanceship with a target individual (test respondent) increases, a peer's knowledge about the target is likely to increase, and the peer will use this knowledge to make increasingly valid inferences about the target's personality and behaviours (Paunonen, 1989; Paunonen & O'Neill, 2010). Furthermore, the peer will have to rely less on heuristics such as guessing, baserate estimates, or assumed similarity in order to report target behaviours and attitudes (Paunonen & Kam, 2014). Thus, as the level of target acquaintanceship with the peer increases, self- and peer ratings of personality tend to correlate to a greater degree (e.g., Connolly, Kavanagh, & Viswesvaran, 2007; Paulhus & Bruce, 1992; Paunonen, 1989), leading to higher indices of test validity.

Studies have demonstrated that peers have a tendency to rate observable behaviours more accurately than internal cognitions and beliefs for a target individual (e.g., Cheek, 1982; Paunonen & Kam, 2014). Evidence has also demonstrated that scores on observable traits such as extraversion interact with acquaintanceship, such that observability of a trait is more important in making accurate judgments of a target when pairs are low to moderately acquainted, but less important for those pairs who are highly acquainted (Paunonen, 1989). Thus, in order to elicit peer ratings of targets that are maximally accurate, thereby demonstrating maximal test validity, having well-acquainted peers complete observable behaviour measures is advantageous.

1.3. Item Content and Personality Measures

The focus of the current study is on the psychometric properties of personality test items as determinants of test validity. In order to construct a valid measure of personality, there are four fundamental principles that some test developers consider important at the level of the item. These include (a) an emphasis on psychological theory and item content, (b) the control of response style variance in items, (c) maintaining item homogeneity within scales, and (d) fostering convergent and discriminant validity at the item writing stage (Holden, Fekken, & Jackson, 1985; Jackson, 1970; Paunonen & Hong, 2015). Improper evaluation of these item properties can lead to low reliability and low validity in the total scale scores, and to potential issues such as the excessive intrusion of method variance, in which test score variance is more attributable to the measurement method than to the construct under investigation (Coleman, 2013; Podsakoff, Mackenzie, Lee, & Podsakoff, 2003).

It is important to note that, with regard to the item properties mentioned above, not every test constructor believes that item content is important in a personality measure. Some believe that item-criterion predictive power should be the primary objective of the test developer. These

are issues that separate construct-based approaches to test construction from criterion-based approaches, and they are described in more detail in the following sections.

1.3.1. Construct-Based Personality Measures

Construct-based personality measures are rooted in psychological theory, such that the trait in question has demonstrable theoretical underpinnings, can be operationally defined, and is clearly conceptualized (Carretero-Dios, Perez, & Beula Casal, 2009; Jackson & Paunonen, 1985). Items are written so that their content is manifestly connected to the trait. Responses to items on construct-based measures should thus reflect to some degree an individual's true score on the theoretically-based dimension under investigation. If those responses accurately reflect the person's true score on the trait, then the items are said to possess construct validity, which cannot exist in the absence of a clearly articulated theory of the trait (Clark & Watson, 1995; Loevinger, 1957; Paunonen, 1984).

Because a construct-based scale is developed based on a theory of the personality trait under consideration, items tend to be homogeneous and thematic in content. Such items are said to be high in content saturation. A highly content saturated item is discernibly prototypical of, or centrally related to, the construct being measured. Scale items generally vary somewhat in their content saturation, but the higher the mean level in a scale, the higher the scale's internal consistency reliability (Holden & Troister, 2009). Scales high in content saturation can be constructed using factor analytic procedures by retaining items with high loadings on the first unrotated principal component extracted from the scale's item pool (Paunonen, 1984). Items with high loadings on the first factor are also likely to have high item-total correlations, thus indicating that the scale measures homogeneous content (Paunonen, 1987). A variation on using item-total correlations to identify content saturated items is to use Neill and Jackson's (1976)

Item Efficiency Index (IEI), where item correlations with irrelevant traits reduce the estimates of content saturation. Thus, to the extent that an item is saturated with relevant content variance, it should (a) load highly on the first factor among the scale's items, (b) have a high item-total correlation, and (c) have a high Item Efficiency Index. We would also add that content saturated items should (d) be perceived by judges as highly related to the trait being measured.

1.3.2. Criterion-Based Personality Measures

The criterion-oriented approach to personality test construction entails the development and selection of items based primarily on how well they contribute to the prediction of a criterion variable. Researchers employing criterion-based methods of scale construction might administer scale items to two groups known to differ on the characteristic being measured (e.g., a depressed group and a control group for a scale measuring depression), and then select items based on how well they differentiate between the two groups (John & Benet-Martinez, 2000). Perhaps the most commonly employed criterion-based questionnaire today is the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1940). The MMPI is a widely used psychometric tool with 10 clinical scales that were designed to distinguish between individuals who were exhibiting various psychopathologies from those who were not (Hathaway & McKinley, 1940). Note that item-criterion correlations are considered of primary importance, and not item content. As a result, one finds an item like "I enjoy detective or mystery stories" on the MMPI Hysteria scale.

Some notable advantages and disadvantages of the criterion-based approach to scale construction exist. Because items are chosen to be predictive of a specific criterion, there already exists, by definition, some inherent criterion validity to the scale (Smith, Fischer & Fister, 2003). However, criterion-oriented items may be heterogeneous in content and unrelated to any

theoretical construct they are intended to measure (see the example MMPI Hysteria item above). As a result, total scores on the scales may be psychologically difficult to interpret (Paunonen, 1984). Also, the use of criterion-oriented scales with ambiguous content and that are low in content saturation may result in self- and peer rating discrepancies due to a lack of focus by raters on the same content domain, or due to guessing (Paunonen, 1984).

Criterion-based measures are expected to do well at prediction because that is what they were designed to do. However, some comparison studies have shown that when using construct-based measures for the purposes of prediction, items that are high in content saturation can perform as well as criterion-based measures, even when predicting the same criterion. Furthermore, the use of scales with homogeneous content often allows for validity coefficients to exceed the .30 ceiling described in previous literature (Paunonen, 1984). Perhaps most important, a respondent's total score on a construct-based scale allows one to make some conclusion about the person's standing on some theoretically-based psychological continuum. It is for these reasons that the construct-oriented approach is often favoured over the largely atheoretical criterion-oriented approach (Jackson & Paunonen, 1985; Paunonen, 1984).

1.4. Item Properties Affecting Test Validity

As has already been suggested, there are a number of psychometric item properties that can have a demonstrable effect on the validity of a personality questionnaire. Such properties include, for example, item face validity, item subtlety, item content saturation, item means, item observability, item desirability, and item wording. Many of the putative shortcomings of personality assessment described in the literature, such as the absence of findings of test-behaviour predictability, can be said to be due to a lack of consideration for these item

characteristics (Jackson & Paunonen, 1980). These item properties are articulated in the separate sections that follow.

1.4.1. Face Validity and Item Subtlety

Personality test items differ in two related but distinguishable properties: face validity and subtlety (Holden & Jackson, 1979). In order for an item to be face valid, it must seem to be relevant to the respondent in the particular assessment context in which it appears. A subtle item, on the other hand, is one in which the respondent is unaware of the construct being measured. Thus, while an item's face validity refers to its apparent relevance to a given assessment context, an item's subtlety refers to the item's lack of an obvious link to the trait being measured. Face validity and item subtlety do not exist as opposite ends of one bipolar continuum. They are distinct but correlated concepts, with some estimates putting that correlation at around -.55 (Holden & Jackson, 1979).

Different strategies of constructing personality measures have placed varying emphasis on the concepts of face validity and item subtlety, each of which can have an effect on a measure's criterion validity (Holden & Jackson, 1979). With regard to criterion-based scales, because item content is considered largely irrelevant, face validity and item subtlety are not of much concern to item selection. Note, however that subtle items have been viewed as an effective means to reduce faking, even in criterion-based scales, because respondents cannot easily link their item responses to some intended personality type (Duff, 1965; Holden & Jackson, 1979). With construct-based scales, on the other hand, content is largely homogeneous and rooted in psychological theory, where item content is ostensibly related to the measured trait. Thus, face validity tends to be high and subtlety tends to be low, at least compared to criterion-based scales.

Of consequence to the present context, some research has indicated that high item face validity and low item subtlety are both associated with higher criterion validity (Holden & Jackson, 1979). Furthermore, studies investigating the effects of subtle items on the discriminatory power of the MMPI have indicated that more transparent items (less subtle) more effectively distinguish between clinical and non-clinical groups (Burgess, Campbell, & Zylberberg, 1984; Duff, 1965). These findings support the notion that criterion-oriented scales, wherein scales are constructed without regard for face validity or item subtlety, may not be as effective or as valid as scales that are constructed using methods that are heavily focused on item content.

1.4.2. Content Saturation

A goal in the construction of theoretical, construct-based measures of personality is to establish construct validity, which is defined by Loevinger (1957) as "the degree to which it measures some trait which really exists in some sense" (p. 685). An important consideration in establishing such construct validity is the extent to which items are content saturated. As already described in an earlier section, content saturated items contain trait-relevant content, and the best ones are the most prototypical representations of their content domains (Paunonen, 1984). A general assumption in conventional personality scale construction is that single scales should measure single, unitary personality constructs. Thus, a highly content saturated scale will have high scale homogeneity, with all items representing trait-relevant content.

Content saturation is not to be confused with item subtlety. Subtle items were described in an earlier section as those in which respondents are unable to infer a substantive link between the item and the trait being measured. Content saturated items, on the other hand, are trait-relevant representations of the content domain. Although item subtlety and content saturation

may, at first glance, appear to be indistinguishable attributes of personality questionnaires, they are not the same. According to Jackson (1971), a subtle item is one in which (a) there is a link between the item's content and the trait being measured by the item (i.e., the item has some level of content saturation), (b) that link is not immediately apparent to a respondent so he or she is not sure what the item measures, but (c) the item-trait connection is readily seen once the respondent is made aware of it. Jackson provided an example using the item "I think newborn babies look very much like little monkeys." This item is subtle in that it would be difficult for respondents to determine what trait is being measured. However, if a respondent is made aware that this item measures the negative pole on the Nurturance trait continuum, it becomes immediately apparent to him or her that the item indeed represents the content domain of interest. That is, even if the respondent cannot immediately delineate the connection between the trait and item content, it can be made evident to the person that this item is content saturated and prototypical of the trait of interest (Jackson, 1971). Therefore, item subtlety and content saturation are distinct but related entities.

There are a number of proposed ways to construct personality questionnaires that reflect high content saturation. One such method employs factor analytic procedures in order to maximize item homogeneity and internal consistency (e.g., Briggs & Cheek, 1986; Holden & Fekken, 1990; Paunonen, 1984; Paunonen, 1987). Items that have the highest loadings on the first factor underlying the scale's item intercorrelation matrix are inferred to be most content saturated. Such items, of course, correlate more highly among themselves than do those with low loadings on the factor, which is likely due to their relation to a common theme – that is, the trait being measured (assuming irrelevant homogenizing factors such as desirability can be ruled out). Paunonen (1984) constructed ad hoc PRF subscales of varying length with varying content

saturation by retaining items with high to low loadings on the first unrotated principal component extracted from a scale's items. Results of that study demonstrated that scales constructed to reflect maximum content saturation (i.e., having the highest factor loadings) were more highly correlated with criteria such as peer ratings than were scales simply constructed to maximize items' contributions to the prediction of a relevant trait criterion.

A second method employed to maximize content saturation involves computing item-total correlations, wherein responses to individual items are correlated with total scale scores. The inference is that higher item-total correlations will be associated with more content saturation. This index is clearly linked to the above-mentioned factor loading index. Paunonen (1987) demonstrated that item loadings in a multiple group factor analysis, where each scale's items were assigned to their own factor, correlated in excess of .99 with the items' item-total correlations.

Related to the item-total correlation index, another statistical indicator of content saturation is the Item Efficiency Index (IEI; Neill & Jackson, 1976). An item's IEI is a measure of the degree to which the item relates to the content domain of interest after partialing out variance due to irrelevant content domains. If a scale item correlates too highly with irrelevant traits, this could serve to lower the measure's discriminant validity (Jackson & Paunonen, 1985). Computing an item's IEI involves subtracting the variance due to irrelevant content, based on item-irrelevant scale correlations, from the item's item-total correlation (Neill & Jackson, 1976). Items with high IEI values are preferred because there is little variance in the item that is attributable to irrelevant trait domains.

Yet another way to evaluate item content saturation involves asking a group of judges to estimate item-construct linkages of individual test items. Subjects should be provided with a

clear and comprehensive definition of the trait in question, complete with trait-defining behaviours. The negative end of the bipolar trait continuum should also be described (Paunonen & Hong, 2015). The judges would then be asked to estimate how prototypical each item is of the trait domain. It is expected that such judgments of content saturation should correlate with the other statistical indices of content saturation described above.

As mentioned earlier, construct-based scale items that are most saturated with trait relevant content can be more valid than even criterion-based scale items. Paunonen (1984) argued that such higher correlations between peer ratings and self-ratings on the more content saturated PRF items in his study were attributable to those items being most prototypical of the trait. In other words, content saturated items are highly salient and likely to be highly representative of concrete trait-relevant behaviours. Items low in content saturation, on the other hand, might measure multidimensional content or ambiguous content, which might be difficult for respondents, be they selves or peers, to interpret consensually.

1.4.3. Item Means

It is generally proposed that the optimal items to select in test construction are those with moderate means or *p*-values (popularity or probability of endorsement values). Items with moderate mean endorsement levels (e.g., around .50 on a binary true/false response scale, or 3 on a 5-point Likert scale) can demonstrate maximal observed score variance and respondent discrimination (i.e., how well the item distinguishes between respondents on the measured trait). On the other hand, items with extreme *p*-values (i.e., values close to 0.0 or 1.0 on a binary scale, or to 1 or 5 on a 5-point scale) fail to differentiate between individuals because of the restricted variance of item responses. Furthermore, items with extreme *p*-values impose limits on the

strength of the correlations that the measure can demonstrate with criterion variables, thus attenuating indices of validity (Epstein, 1983).

Holden et al. (1985) examined the relationship between absolute endorsement frequency of 80 binary PRF items and criterion validity. Their results demonstrated a significant correlation of -.29 between extreme endorsement levels and criterion validity. Thus, items that are endorsed by many respondents or by few respondents hinder the criterion validity of a measure (Nunnally, 1978). This does not mean that items with moderate means are definitely more valid; it just means that such items do not have the same statistical constraint on validity.

1.4.4. Item Observability

Research in person perception has some bearing on the type of item that might best serve personality test validity. For instance, Paunonen and Kam (2014) found that the accuracy of roommate ratings of a target individual's personality varied as a function of the observability of the personality test items used. Specifically, items describing observable behaviours that were related to the trait under investigation were more accurately rated by peers than were items describing trait-related beliefs or attitudes. Unlike behaviours, beliefs and attitudes are unobservable cognitions from an outsider's perspective, and therefore must be inferred from observable behaviours that are related to the attitude.

Other research in person perception supports the notion of better personality assessment on more observable characteristics rather than less observable. For instance, Albright, Kenny, and Malloy (1988) found significant correlations between target and rater judgments of conscientiousness and extraversion, where target and rater were complete strangers. The researchers speculated that the consensus of self-other ratings was driven by information that the judge acquired during the testing session about the unfamiliar target's highly observable

characteristics, such as talkativeness or neatness of dress (Albright et al., 1988). Similarly, Beer and Watson (2008) found that for unacquainted dyads, the correlation between target and judge ratings was .37 for extraversion, indicating some degree of accuracy on that observable trait. They contended that strangers rely on judgments of physical appearance such as facial attributes and physical fitness in order to make judgments about a target. John and Robins (1993) found the greatest discrepancies between judge and target ratings for agreeableness (r = .13), and the smallest discrepancies for extraversion (r = .29), speculating that this was due to the high observability of extraverted behaviours. Thus, when manifestations of a particular trait are highly observable as opposed to unobservable, then the observers should be able to make more valid judgments of a target (but see Funder, 1980).

The implications of person perception studies for the construction of personality measures are such that, if peers are more accurately able to report on a target person's behaviours than on their attitudes, then perhaps standardized measures of personality should be developed to reflect observable characteristics. Paunonen and Kam (2014) contended that peers judge a target's personality characteristics on the basis of a sequential model of judgment. If the peer is aware of the target's characteristics through direct knowledge, he or she will be able to accurately judge the target. But peers can use heuristics in judging the target person. If target characteristics are unknown, peers can infer these based on information from related behaviour cues, base rates, or assumed similarity to the self (Paunonen & Hong, 2013; Paunonen & Kam, 2014). Because observable trait behaviours require less subjectivity in their judgments than do unseen behaviours, there should be less error in such behaviour judgments than in attitude or belief judgments. Thus, for the purpose of test validation studies in which self-peer agreement is assessed, behaviour-based measures are more likely to succeed than are belief-based measures.

1.4.5. Item Desirability

Socially desirable responding (SDR) is a response bias in which individuals endorse response options to items on personality measures in order to present a favourable image of the self and to prevent negative perceptions from others (Paulhus, 2002; Podsakoff et al., 2003). This tendency has problematic effects on personality assessment validity. Not only might the respondent be grossly misrepresenting his or her true level of trait, which compromises the measure's construct validity, but SDR can affect mean levels of responding, thus altering relationships between the test and variables such as validation criteria (Ganster, Hennessey, & Luthans, 1983; Podsakoff et al., 2003; but cf. Paunonen & LeBel, 2012).

Personality test items differ in their tendency to elicit SDR. One method used to determine an item's level of social desirability is to compute its social desirability scale value (SDSV) by asking a group of judges to read the item and rate how socially desirable or undesirable they consider it to be as applied to others (Edwards, 1969; Paunonen, 2015). Evidence has consistently demonstrated that level of item endorsement is a strong linear function of item SDSV, such that items with high SDSVs are endorsed more frequently than items with low SDSVs (Berg, 1967; Edwards, 1969; Edwards, 1970).

Another method used to determine the social desirability of an item is to correlate respondent scores on the item with their scores on a social desirability scale (e.g., Kam, 2013). Social desirability scales are developed for the purpose of assessing whether or not an individual has a tendency to respond in an overly favourable manner (Paulhus, 2002). Such scales typically represent items that are heterogeneous with respect to trait content, but homogeneous in desirability. A strong correlation (positive or negative) between a sample's responses to a particular test item and the respondents' social desirability scale scores indicates that the item has

a preferred response option by those who are motivated to engage in SDR – in other words, it is susceptible to desirability bias (Paulhus, 2002). One technique used to ensure that items are content saturated and do not reflect a high desirability component is to select items with high Differential Reliability Index values (DRI; Jackson, 1970). An item's DRI is essentially an item's IEI (Neill & Jackson, 1976) where only one irrelevant scale that measures SDR is examined. Specifically, item DRIs measure how strongly an item relates to the content domain of interest (its item-own scale correlation) while partialing out variance due to social desirability (its item-desirability scale correlation).

Items that are neutral in desirability are generally preferred in personality assessment. This is because neutral items are likely to elicit the most accurate representations of an individual because there is no mechanism for misrepresentation by the target, even if so inclined (Paunonen & LeBel, 2012). But what about the use of such items if peer ratings on those items form the basis of a validation criterion? From one point of view, desirable items are less likely to elicit desirability biases in the peer, as the peer might not be as motivated to inflate target ratings as they are to inflate self-ratings (Paunonen & O'Neill, 2010). Consistent with this notion is research that has found a curvilinear relationship between item SDSVs and self-peer agreement on 76 unipolar personality trait adjectives (John & Robins, 1993). Traits that were rated by judges as being neutral in social desirability elicited more self-peer agreement than did traits that were rated as being high or low in desirability (John & Robins, 1993). The researchers found a significant correlation between absolute evaluativeness and self-peer agreement of r = -.53 (John & Robins, 1993). Thus, peers making personality judgments are less likely to agree with a target when the traits are highly evaluative (positive or negative), and more likely to agree with a target

when the traits are neutral, possibly because the target's self-ratings are "wrong" with the former ratings.

Other research has not found the same relationship between SDSV item ratings and self-peer agreement on personality measures. In a study using personality adjectives as items, Funder (1980) demonstrated that socially desirable items elicited greater self-peer agreement than did socially undesirable or neutral items, with a significant correlation of .30 between item SDSVs and self-peer agreement. This suggests a linear relationship between SDSVs and self-peer agreement, such that items with low SDSVs elicit less agreement between the self and the peer, and items with high SDSVs elicit greater agreement. Funder (1980) surmised that perhaps the targets and their peers discussed traits higher in social desirability with one another, and avoided any discussion of traits that appear to be undesirable. Thus, peers may have had access to a wealth of information regarding desirable traits, and much less information about undesirable traits. An alternative explanation is that traits low in desirability exhibit little variance, and this restriction of range may have attenuated correlations between self- and peer ratings of personality (Funder, 1980). Yet a third explanation is that people rating their friends might also engage in a bias to rate their targets high on desirable traits, especially if they like them (Leising, Erbs, & Fritz, 2010).

1.4.6. Item Wording

Sources of variance impairing the validity of personality measures can arise due to response sets, which are conceptualized as tendencies to respond to items in a specific manner regardless of item content (Coleman, 2013). Acquiescence is one such response set, described by Cronbach (1946) as a tendency for a respondent to choose the "true" response option rather than "false" (or "agree" rather than "disagree"), regardless of their actual level of the trait, thus

distorting their trait scores. One proposed solution to minimize the effects of acquiescence is to make use of balanced scales, in which half of the items are keyed positively and half of the items are keyed negatively (Holden et al., 1985). The use of balanced scales does not serve to eliminate acquiescent responding. Instead, the objective of balanced scales is to eliminate extreme scores for the acquiescent respondent and to transform them into summed scores that are closer to the mean (Coleman, 2013). Extreme scores are given more attention than moderate scores in most real-world personality assessment contexts. According to Holden et al. (1985), controlling acquiescence through the use of balanced scales permits for more accurate evaluations of scale intercorrelations.

Personality test items can differ in specific wording attributes that might affect validity. Negatively worded items make use of variants of the word "not", or use negative qualifiers such as "seldom" or "rarely" (Holden et al., 1985). Such items have sometimes been used to reduce the effects of response biases by requiring that respondents engage in less automatic cognitive processing (Podsakoff et al., 2003). Whereas some research has indicated that the use of negatively worded items decreases instances of acquiescence, others have indicated that such items produce confounded factor structures in which a separate, content-irrelevant factor can emerge (Coleman, 2013). Furthermore, longer and negatively worded items generally demonstrate less criterion validity than do positively worded items (Holden et al., 1985; Schriseheim & Hill, 1981). Specifically, Holden et al. (1985) demonstrated that on 80 PRF items, longer items correlated with criterion validity with coefficients ranging from -.14 to -.23. When item length was partialed out, the negatively worded items correlated with criterion validity with a coefficient of -.22. These findings are consistent with the notion that longer and more complex items tend to be less reliable and valid than short and concise items because the

former are more likely measure multiple constructs (Smith et al., 2003). By contrast, no relationship was found in the Holden et al. (1985) study between the keyed direction of the items and their criterion validity.

Additional validity issues arise from item complexity and item ambiguity. Although items should be written in a clear and concise manner, it is not uncommon for researchers to make use of double-barreled questions, words with more than one meaning, and infrequently used or unfamiliar words (Podsakoff et al., 2003). As a result, respondents often develop their own idiosyncratic interpretations of items, which will in turn either elicit a tendency to respond in a biased or unusual manner or increase rates of random responding (Podsakoff et al., 2003). Items exhibiting high levels of complexity and ambiguity might affect validity correlations based on self- and peer ratings of personality because of the differing interpretations of item content (Podsakoff et al., 2003).

1.5. Objective

The purpose of the current study is to extend findings from previous literature by formally evaluating a number of properties of personality test items that can affect their validity, and determining which properties make the most important contributions to overall scale validity. The validity, or accuracy, of these personality scales will be estimated by comparing self-ratings of personality with peer ratings of personality, with the inference being that the most accurate measures of personality are those in which self- and peer ratings of personality are highly correlated (e.g., Holden & Troister, 2009; Paunonen, 1984; Paunonen & O'Neill, 2010). Of the specific properties described earlier, those under consideration for the purpose of this study are thought to be among the most important considerations in writing and selecting

personality test items: (a) item social desirability, (b) item difficulty (i.e., mean item responses), and (c) item content saturation.

1.5.1. Hypotheses

Based on the results of previous studies described in this introduction, four hypotheses regarding item properties and test validity will be evaluated in the current investigation.

1. *Item social desirability scale values (SDSVs) will predict self-peer agreement on a series of personality measures, but curvilinearly*. We predict that to the extent that items have high or low SDSVs, this will elicit low self-peer agreement, because the self will tend to misrepresent him or herself more on these evaluative items. Items with neutral SDSVs will elicit relatively high self-peer agreement.

2. *Correlations between scores on a socially desirable responding (SDR) measure and item responses on a personality measure will linearly predict self-peer agreement.* We predict that to the extent that there are strong positive (or negative) correlations between personality items and SDR measures, this will elicit low self-peer agreement, and to the extent that there are weak correlations, this will elicit high self-peer agreement. The reasoning is the same as with item SDSVs. That is, items that are strongly correlated with SDR will tend to elicit misrepresentation in self-reports.

3. *Item difficulty will predict self-peer agreement in a curvilinear manner.* We predict that to the extent that items have extreme high and low mean endorsement values, this will elicit low self-peer agreement, and moderate mean endorsement values will elicit high self-peer agreement. This prediction is based on expected restriction of range effects at the extremes of responding.

4. *Items exhibiting higher content saturation will result in higher self-peer agreement than will items exhibiting lower content saturation.* The reasoning here is that content saturated items are more likely to be prototypical of the trait being measured and less likely to be only tangentially related to the trait. This should then result in greater consensual interpretations of item content, leading to greater internal consistency and construct validity for such items.

In summary, it is predicted that items with neutral SDSVs, weak correlations with SDR measures, moderate difficulty values, and high indices of content saturation will produce higher validity coefficients against peer ratings on a series of personality measures.

CHAPTER 2: METHOD

2. Method

2.1. Participants

Archival data representing self- and peer personality ratings were used for the purposes of the current study. The data were collected at five time points between 1981 and 2004. Specifically, the studies were conducted in 1981 (see Paunonen, 1982), 1993 (see Paunonen, 1998), 1994, 1997 (see Paunonen & Ashton, 2001), and 2004. The data represent various groups of same-sex undergraduate roommate dyads living in a dormitory at Western University. Students received cash compensation in return for their participation.

A detailed summary of demographic variables for each of the five samples is presented in Appendix A. The samples used for the current study comprised mostly undergraduate students in their first year of university. The samples' mean ages ranged from 18.79 (*SD* = .69) to 19.24 (*SD* = .84). Each of the studies was conducted in the second to last month of the academic year to ensure that participants had substantial time to become acquainted with their roommates. Sample means of participants' reported duration of time acquainted with their respective roommates ranged from 14.43 months (*SD* = 21.59) to 28.59 months (*SD* = 71.08), indicating that many participants knew their roommates prior to moving into residence. Of the samples collected in 1981 and 1994, 6.7% and 3.3% of participants, respectively, reported being acquainted with their roommate for less than three months. Although it has been documented that accuracy increases as acquaintanceship increases (e.g., Connolly et al., 2007; Paulhus & Bruce, 1992; Paunonen, 1989), it is unlikely that this proportion of the data was large enough to affect the results. Participants were also asked the question: "In terms of how well you can know anyone, how well do you know your roommate?" Responses ranged from "don't know at all" to "extremely well."

Sample means for self-reported acquaintanceship ratings ranged from 5.82 (*SD* = 1.00) on a 7-point scale to 7.37 (*SD* = 1.03) on a 9-point scale. Thus, participants generally indicated that they were well-acquainted with their roommates at the time of testing.

2.2. Procedure

Participants were recruited using posters titled "Roommate Personality Study," or variations thereof, which were posted in the lobby of a co-ed Western University dormitory. Those interested would sign up to participate, and were subsequently contacted regarding the nature of the study.

The roommate pairs in each of the five samples were assessed at two time points separated by one week. In the first testing session, both roommates arrived together at a classroom on the Western University campus, where they completed a series of paper-and-pencil self-report personality questionnaires. They were seated at separate tables from their respective roommates to avoid potential biases in responding. Additionally, participants' anonymity was assured by assigning each individual a code number. In the second testing session, participants filled out the same measures, but instead they completed peer ratings of their roommates' personality characteristics. The researchers examined the questionnaires for missing or improper responses before participants exited the testing room. Upon completion of both testing sessions, participants were debriefed. The University of Western Ontario's Research Ethics Board approved the five studies used for the purpose of this research.

2.3. Personality Measures

The primary measures used in the five different roommate samples included: the Supernumerary Personality Inventory (SPI; Paunonen, 2002), the NEO Personality Inventory-Revised (NEO-PI-R; Costa & McCrae, 1992), the Jackson Personality Inventory (JPI; Jackson,

1976), the Personality Research Form (PRF; Jackson, 1974, 1984), and the Nonverbal Personality Questionnaire (NPQ; Paunonen, Ashton, & Jackson, 2004). Each of the measures are described in detail below.

2.3.1. Supernumerary Personality Inventory (SPI; Paunonen, 2002)

This 150-item inventory measures 10 personality traits that extend beyond the Big Five model of personality. The 10 SPI scales include: Conventionality, Seductiveness, Manipulativeness, Thriftiness, Humorourness, Integrity, Femininity, Religiosity, Risk-Taking, and Egotism. Participants responded to items on a 5-point Likert scale based on how much they agree or disagree with the statement (1 = *strongly disagree*, 5 = *strongly agree*). The SPI has demonstrated adequate internal consistency reliability, with coefficient alpha values ranging from .66 (Conventionality) to .95 (Religiosity), and a mean coefficient alpha value of .80 in a normative university sample (Paunonen, 2002).

Responses to items on the SPI were drawn from a study conducted in 2004. The SPI data collected in 2004 were used to compute self-peer correlations and item correlations with the true/false PRF Desirability scale. The remaining SPI computations, including mean item responses and item-total correlations, came from SPI normative data (N = 537; Paunonen, 2002).

2.3.2. NEO Personality Inventory-Revised (NEO-PI-R; Costa & McCrae, 1992)

This questionnaire comprises 240 items that assess the Five-Factor Model of personality. Trait scales include Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each subscale contains six facets. Items responses were measured on a 5-point Likert scale, where 1 = *strongly disagree* and 5 = *strongly agree*. The NEO-PI-R's psychometric data are sound, with coefficient alpha values ranging from .86 (Agreeableness) to .92

(Neuroticism) in an adult normative sample (mean coefficient alpha = .89; Costa & McCrae, 1992).

Responses to items on the NEO-PI-R were drawn from studies conducted in 1997 and 2004. The NEO-PI-R data collected in 1997 were used to compute item-total correlations and item correlations with the true/false PRF Desirability scale. The data collected in 2004 were used to compute self-peer correlations and mean item responses.

2.3.3. Jackson Personality Inventory (JPI; Jackson, 1976)

The JPI was designed to assess personality traits that predict behaviour in a number of settings. The 15 trait scales of the JPI include: Anxiety, Breadth of Interest, Complexity, Conformity, Energy Level, Innovation, Interpersonal Affect, Organization, Responsibility, Risk Taking, Self Esteem, Social Adroitness, Social Participation, Tolerance, and Value Orthodoxy. The measure also possesses an infrequency scale. The JPI consists of 320 true/false items in total. Jackson (1977) has reported adequate reliability coefficients, with coefficient alpha values ranging from .60 (Tolerance) to .87 (Innovation) in an American college sample (median coefficient alpha = .82).

Responses to items on the JPI were drawn from studies conducted in 1994 and 1997. The JPI data collected in 1994 were used to compute self-peer correlations and mean item responses on a 7-point scale. The data collected in 1997 were used to compute item correlations with the PRF Desirability scale and item-total correlations using the standard true/false scale.

2.3.4. Personality Research Form (PRF; Jackson, 1974, 1984)

The 352-item PRF is available in five forms, but the most commonly used is Form-E. The 20 trait scales of the PRF Form-E were developed based on Murray's (1938) framework of personality. These true/false item scales are: Abasement, Achievement, Affiliation, Aggression,

Autonomy, Change, Cognitive Structure, Defendance, Dominance, Endurance, Exhibition, Harmavoidance, Impulsivity, Nurturance, Order, Play, Sentience, Social Recognition, Succorance, and Understanding. Two additional scales measure infrequency and desirability. Jackson (1984) has reported strong internal consistency for the PRF, with Kuder-Richardson 20 reliability values ranging from .78 (Defendance) to .94 (Order), and a median KR-20 value of .93 in a university sample (Jackson, 1970; Valentine, 1969).

Responses to items on the PRF were drawn from studies conducted in 1981 and 1997. An earlier version of the PRF was used for the PRF data collected in 1981 (see Jackson, 1974). The PRF data collected in 1981 were used to compute self-peer correlations and mean item responses on a 9-point scale ranging from 1 = *extremely uncharacteristic* to 9 = *extremely characteristic* (Paunonen, 1982). The data collected in 1997 were used to compute item-total correlations and item correlations with the PRF Desirability scale using the standard true/false scale.

2.3.5. Nonverbal Personality Questionnaire (NPQ; Paunonen, Jackson, & Ashton, 2004)

This 136-item inventory is a structured, nonverbal measure of 16 traits. Items consist of line drawings of figures performing various trait-relevant behaviours. This nonverbal questionnaire has demonstrated its utility in cross-cultural personality assessment across a wide variety of cultures and languages (Paunonen, Jackson, & Keinonen, 1990). Items were developed to reflect 16 of the constructs described in Murray's (1938) trait system, and these traits measured using the NPQ directly correspond to those measured using the PRF. Participants are instructed to consider each nonverbal item and to estimate the likelihood that they would engage in the type of behaviour shown. Item responses are measured on a 7-point rating scale, and range from 1 = *extremely unlikely* to 7 = *extremely likely*. The 16 traits scales of the NPQ are: Achievement, Affiliation, Aggression, Autonomy, Dominance, Endurance, Exhibition,

Impulsivity, Nurturance, Order, Play, Sentience, Social Recognition, Succorance, Thrill-Seeking, and Understanding. Paunonen et al. (2004) reported adequate internal consistency, with coefficient alpha values ranging from .60 (Impulsivity) to .84 (Thrill-Seeking), and a mean coefficient alpha of .70 in a cross-cultural normative sample.

Responses to items on the NPQ were drawn from a study conducted in 1993. Although the most recent version of the NPQ was not published until 2004, no revisions to the scale have been made since 1990, when the initial item pool was revised (Paunonen et al., 2004). Thus, the version of the NPQ that the researcher had access to in 1993 was the same as the current version of the NPQ. The data collected in 1993 were used to compute self-peer correlations and item correlations with the true/false PRF Desirability scale. The remaining NPQ computations, including item-total correlations and mean item responses, were drawn from NPQ normative data (N = 304; Paunonen et al., 2004; Paunonen, Ashton, & Jackson, 2001).

2.4. Desirability Measures

The socially desirable response tendencies of the roommate raters were assessed with one questionnaire-based measure. The desirability levels of the individual personality inventory items were assessed using standard rating procedures.

2.4.1. Personality Research Form Desirability Scale (PRF Desirability; Jackson, 1984)

The PRF Desirability scale comprises 16 items of the larger 352-item questionnaire (Form-E). This scale was designed as a content-heterogeneous and internally consistent measure of one's tendency to respond desirably. This scale demonstrates adequate internal consistency, with a reported coefficient alpha of .70 (Jackson, 1984). In this study, participants' scores on the PRF Desirability scale were correlated with their item responses to the personality questionnaires

to evaluate item desirability. Responses to items on the PRF Desirability scale were drawn from studies conducted in 1993, 1997, and 2004.

2.4.2. Social Desirability Scale Values (SDSVs; Edwards, 1970)

Item SDSVs were evaluated by three different samples. Specifically, a group of 149 undergraduate students (67 males, 82 females) were presented with items of the SPI, and a second group of 27 undergraduate students (8 males, 19 females) were presented with items of the NEO-PI-R. Social desirability scale values for the PRF were drawn from a study by Helmes, Reed, and Jackson (1977; $N = 98$). Participants in all samples were asked to rate item SDSVs by considering each statement and estimating how socially desirable or undesirable the behaviour or belief would be if characterizing other people in general (Edwards, 1970, p. 89). An item's SDSV was equal to the mean of all participants' SDSV ratings for that item. Item SDSVs were assessed using a 9-point scale, where 1 = *extremely undesirable* and 9 = *extremely desirable*. In the current study, SDSVs were not collected for items on the NPQ and the JPI.

CHAPTER 3: RESULTS

3. Results

3.1. Data Analytic Strategy

The purpose of the current study was to evaluate which item properties contributed to the overall validity of personality questionnaires. The properties that were assessed included item desirability, item content saturation, and item difficulty (i.e., mean item responses).

Item desirability was computed in two ways. The first way was to correlate total scores on the PRF Desirability scale with item responses to the personality measures. The second way was to assign SDSVs to items by having students rate their social desirability, and taking the mean SDSV ratings of all students for each item. Content saturation was measured by item-total correlations (i.e., a measure of reliability); that is, correlating participants' responses to individual items with total subscale scores (cf. Paunonen, 1984, 1987). Item difficulty was evaluated by computing mean item responses.

The validity, or accuracy, of the personality questionnaires was calculated by correlating self- and peer report responses to items in the roommate rating data. Accuracy, as measured by the self-peer correlations, was regressed onto personality test item SDSVs, item correlations with the PRF Desirability scale, item-total correlations, and mean item responses, separately by personality questionnaire. Where possible, item accuracy and the predictor statistics were computed on different data sets so as to prevent these variables from being inextricably linked in the analyses. For example, if test validity and reliability are computed on the same data sets, this could spuriously inflate the relationship between the variables, as reliability is a precursor to validity. In order to assess curvilinear components of the relationships between accuracy and the other item properties, accuracy was regressed onto the squared values of these test item statistics.

3.2. Data Cleaning

Prior to conducting analyses, the data underwent standard cleaning procedures. At the time of data collection, questionnaires were examined for missing or improper responses before participants exited the testing room. In the 1981 sample, 27 (.00085%) self-rating items and 37 (.0012%) peer rating items were left blank on the PRF questionnaires. These items were replaced with the number "5," indicative of a neutral response (Paunonen, 1982). In the same sample, random or careless responding was inferred based on participants' results on the PRF Infrequency scale (i.e., a scale reflecting extremely rare forms of behaviour). The researchers concluded that no participant scored high enough on this scale to be removed from the data pool (Paunonen, 1982). A small amount of the data collected in 1993 was removed using pairwise deletion procedures (Paunonen, 1998). In the 1997 sample, one roommate failed to show up to the second testing session, and one participant showed up to the first testing session without their roommate. Aside from these two participants, there were no missing data reported for any of the personality scale responses collected in 1997 (Paunonen & Ashton, 2001). Information about data cleaning procedures was not available for samples collected in 1994 and 2004.

3.3. Preliminary Analyses

Prior to carrying out the main analyses, background analyses were conducted. These analyses included computing descriptive statistics for each self-report personality questionnaire, descriptive statistics for the item properties under investigation, and item property bivariate correlations.

3.3.1. Personality Questionnaire Descriptive Statistics

Descriptive statistics including means, standard deviations, skewness values, and kurtosis values for self-reported responses to each of the scales used in the current study are reported in

Appendix B. All means and standard deviation values were comparable to those reported in previous literature (e.g., Jackson, 1976, 1984; McCrae & Costa, 2010; Paunonen, 2002; Paunonen et al., 2004).

Indices of skewness (SI) and kurtosis (KI) for each subscale were examined to detect any existing non-normality in the data. It is suggested that absolute SI values exceeding 3.00 indicate extremely skewed data, and that absolute KI values exceeding 3.00 are problematic in terms of kurtosis (Kline, 2011). None of the data appear problematic in terms of skewness or kurtosis (see Appendix B). The JPI Infrequency subscale data collected in 1997 and the NPQ Infrequency subscale data collected in 1993 had KI values of 11.80 and 4.53, respectively, indicating that these distributions were leptokurtic (see Tables B.5 and B.8). However, responses to items on infrequency scales can be used as indicators of non-purposeful responding or comprehension issues (Jackson, 1970). Thus, one would anticipate that if participants are responding carefully, nobody should endorse these scale items. Furthermore, the mean and variance values of these scales should be low, and their distributions should be skewed. In this case, a leptokurtic distribution, in which the variance is restricted, is to be expected. For instance, the item "I have never talked to anyone by telephone," appears on the PRF infrequency subscale. Most, if not all participants should have talked via telephone at some time in their life. Thus, most, if not all respondents should select the "false" option.

It should be noted that the NPQ Affiliation subscale administered to the 1993 sample had a KI value of 6.25, also indicative of a leptokurtic distribution (see Table B.8). The mean was also slightly higher and the variance was slightly smaller than those of the other subscales (M = 42.35, SD = 7.00), indicating that on average, participants tended to endorse the positive end of the Affiliation continuum. However, these mean and variance values are not vastly different

from those reported in NPQ normative samples (e.g., *M* = 41.30, *SD* = 6.60; Paunonen et al., 2004). Thus, this distribution is not problematic for the purpose of this research.

Coefficient alpha values for the PRF and JPI subscale data collected in 1997 range from .56 (Desirability) to .88 (Order) with a mean of .73 (Paunonen & Ashton, 2001). Furthermore, coefficient alpha values for the NEO-PI-R subscale data collected in 1997 all exceeded .88, with a mean of .89 (Paunonen & Ashton, 2001). Coefficient alpha values were not available for the remaining samples. These values could not subsequently be computed because the archival data files accessible to the researcher consisted only of total subscale scores, and the individual item scores were not available[1]. However, internal consistency can be inferred using mean item-total correlations for each questionnaire (see Appendix C, D). Specifically, scores on each item were correlated with scores on the item's respective total subscale, and high item-total correlations indicate that the item is internally consistent. The values reported in Table D.1 reflect moderate internal consistency, with mean item-total correlations ranging from .42 (JPI) to .60 (NPQ). Descriptive statistics of the item-total correlations (i.e., mean, standard deviation, minimum, maximum values) were also broken down by subscale, and are summarized in Appendix C. It should be noted that overall questionnaire mean reliability coefficients for both the JPI and the PRF questionnaires could have been attenuated due to low item-total correlations on the Infrequency subscales, which were designed to reflect heterogeneous content (see Tables C.3 and C.4).

3.3.2. Item Property Descriptive Statistics

Descriptive statistics for item properties, including item accuracy coefficients (i.e., self-peer correlations on individual items), item-total correlations, item-desirability scale correlations,

[1] Aside from item-total correlations, item data were unavailable to the author due to the passing of the researcher who collected the data (Dr. S. V. Paunonen).

item SDSVs, and mean item responses for all personality measures are summarized in Appendix D. For all scales, item accuracy was moderate, ranging from a minimum self-peer response correlation of -.21 (JPI) to a maximum of .64 (SPI and PRF), with mean self-peer response correlations ranging from .13 (JPI) to .27 (SPI).

As indicated above, mean item-total correlations were adequate, ranging from .42 (JPI) to .60 (NPQ), and the use of infrequency subscales likely attenuated these values, specifically in the JPI and PRF (see Tables C.3 and C.4). For instance, the most unreliable items belong to the PRF (i.e., "I have never talked to anyone by telephone") and the JPI (i.e., "I have kept a pet monkey for years"), with item-total correlations of -.06 and -.05, respectively. These items both belong to infrequency subscales designed to detect careless or random responding. These validity subscales generally contain items reflecting extremely rare forms of behaviour and heterogeneous content (Jackson, 1970). Furthermore, infrequency subscales tend to have low mean and standard deviation values, and are typically skewed. Thus, one can expect negative correlations to emerge between items and total infrequency subscale scores because each of the items assess entirely different content areas, and possibly due to restriction of range effects.

The questionnaires under investigation underwent rigorous statistical testing in order to eliminate items that were heavily desirable in content (i.e., Jackson, 1970, 1976; Paunonen, 2002). As such, mean item correlations with the PRF Desirability scale were low for all questionnaires, ranging from -.03 (SPI) to .04 (NPQ). Similarly, mean item SDSVs were moderate-to-low for each scale, with values ranging from 5.20 (PRF) to 5.47 (NEO-PI-R), both measured on a 9-point scale. Additionally, average mean item responses were acceptable, with values ranging from 3.07 on the 5-point SPI scale to 5.19 on the 9-point PRF scale.

3.3.3. Item Property Bivariate Correlations

Bivariate correlations between item properties are presented in Appendix E. As anticipated, strong correlations emerged between accuracy (i.e., item self-peer correlations) and item-total correlations across all questionnaires. Interestingly, significant negative correlations emerged between item social desirability and accuracy in the SPI, and modestly significant positive correlations emerged between item social desirability and accuracy in the PRF. Mean item responses were not significantly correlated with accuracy, but this was likely due to existing curvilinear associations between these variables (e.g., see Figures 4 & 6).

It should also be noted that, consistent with previous literature (e.g., Edwards, 1969), strong positive correlations emerged between item social desirability levels and mean item responses. In other terms, on all questionnaires, items with high levels of social desirability tend to be endorsed more by participants than do items low in social desirability. This is in accordance with the notion that participants who are motivated to engage in SDR have a tendency to present themselves more favourably than is warranted (Paulhus, 2002). Subsequently, items that are infused with desirable content elicit more desirable response tendencies from individuals, in which more desirable items are endorsed more frequently (e.g., Edwards, 1969).

3.4. Main Analyses

A total of five multiple regression analyses were carried out using SPSS Version 23.0 (IBM Corp., 2015) to test the four main hypotheses evaluating which item properties contributed to the overall validity of the personality questionnaires. The dependent variable was validity (i.e., accuracy), as measured by self-peer correlations on individual items. The independent variables

included item SDSVs, item-PRF Desirability scale correlations, item-total correlations, and mean item responses.

3.4.1. Supernumerary Personality Inventory

For the first main analysis, Supernumerary Personality Inventory accuracy, as measured using self-peer correlations, was regressed onto item SDSVs, item correlations with the PRF Desirability scale, item-total correlations, and mean item responses to test our four hypotheses (see Table 1). To assess curvilinear components of the relationships between accuracy and SPI item properties, accuracy was regressed onto the squared values of item SDSVs, item-PRF Desirability scale correlations, and mean item responses. Taken together, the item properties accounted for a significant amount of the variance in accuracy, $R^2 = .25$, adjusted $R^2 = .22$, $F(7, 142) = 6.89$, $p < .001$. A significant beta weight for the model was associated with item-total correlations, $\beta = .43$, $t(142) = 5.83$, $p < .001$ (see Figure 1).

Table 1

Standardized and Unstandardized Regression Coefficients: Prediction of SPI Accuracy from SPI Item Properties

Item Property	B	Standard Error	β	*t*	*p*
Item-total correlations	.48	.08	.43	5.83	.001
SDSV	-.06	.15	-.42	-.40	.69
$SDSV^2$	.01	.01	.38	.36	.72
Desirability correlation	-.14	.09	-.15	-1.53	.13
Desirability $correlation^2$	-.43	.38	-.11	-1.14	.26
Mean	.23	.30	.87	.76	.45
$Mean^2$	-.04	.05	-.89	-.77	.44

Note. Squared variables were used to test curvilinear associations.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.

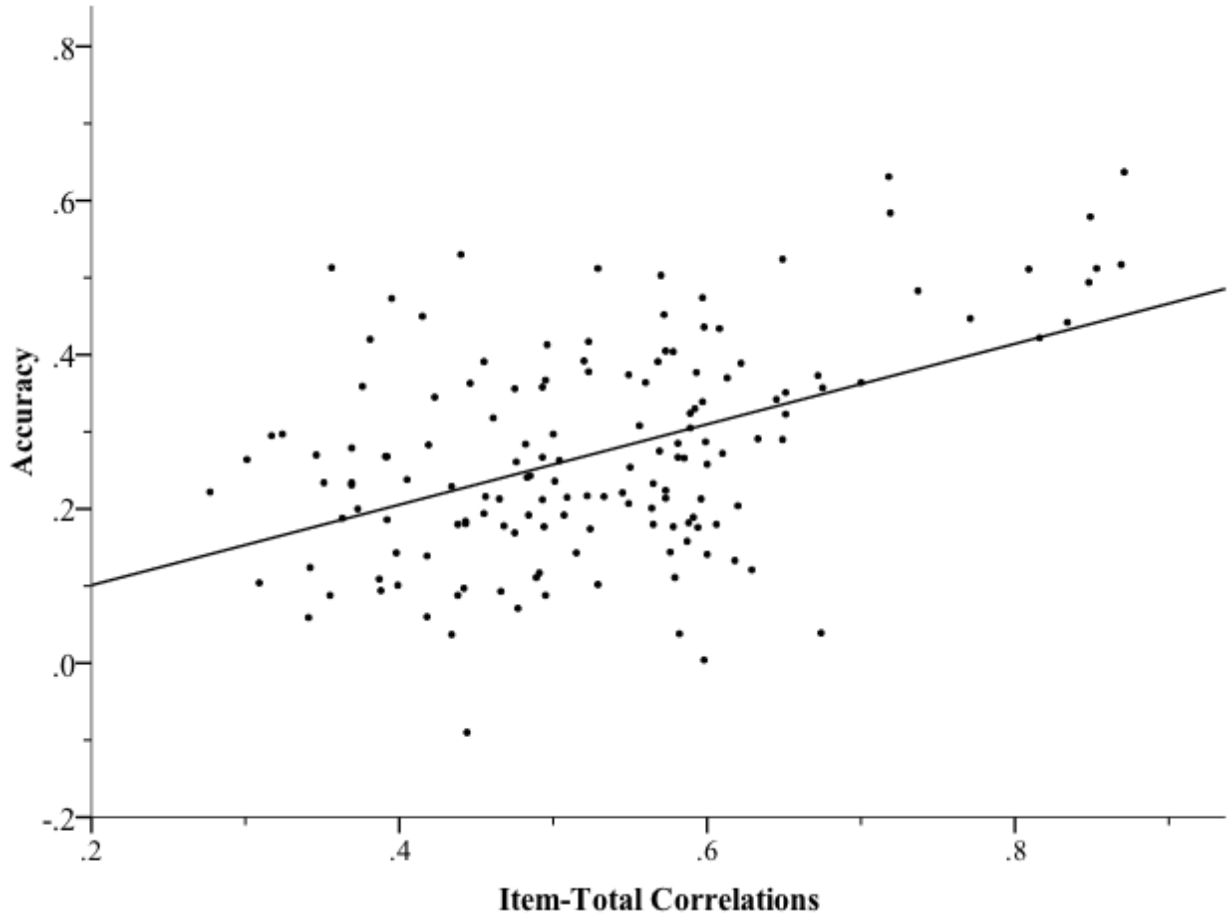

Figure 1. Scatterplot of 2004 Supernumerary Personality Inventory (SPI; *N* = 150 items) self-peer accuracy correlations (sample *N* = 124) on 2002 item-total correlations (normative sample *N* = 537).

3.4.2. NEO Personality Inventory

NEO Personality Inventory accuracy was regressed onto item SDSVs, item correlations with the PRF Desirability scale, item-total correlations, and mean item responses (see Table 2). To assess curvilinear components of the relationships between accuracy and NEO item properties, accuracy was regressed onto the squared values of item SDSVs, item-PRF Desirability scale correlations, and mean item responses. Taken together, the item properties accounted for a significant amount of the variance in accuracy, $R^2 = .12$, adjusted $R^2 = .08$, $F(7, 232) = 3.88$, $p < .001$. A significant beta weight for the model was associated with item-total correlations, $\beta = .30$, $t(232) = 4.40$, $p < .001$ (see Figure 2).

Table 2

Standardized and Unstandardized Regression Coefficients: Prediction of NEO Accuracy from NEO Item Properties

Item Property	B	Standard Error	β	*t*	*p*
Item-total correlations	.32	.07	.30	4.40	.001
SDSV	-.06	.04	-.84	-1.55	.12
$SDSV^2$	.01	.003	.78	1.48	.14
Desirability correlation	-.02	.07	-.03	-.26	.80
Desirability correlation^2	-.13	.22	-.05	-.61	.54
Mean	-.09	.14	-.43	-.62	.54
$Mean^2$	.02	.02	.54	.78	.44

Note. Squared variables were used to test curvilinear components.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.

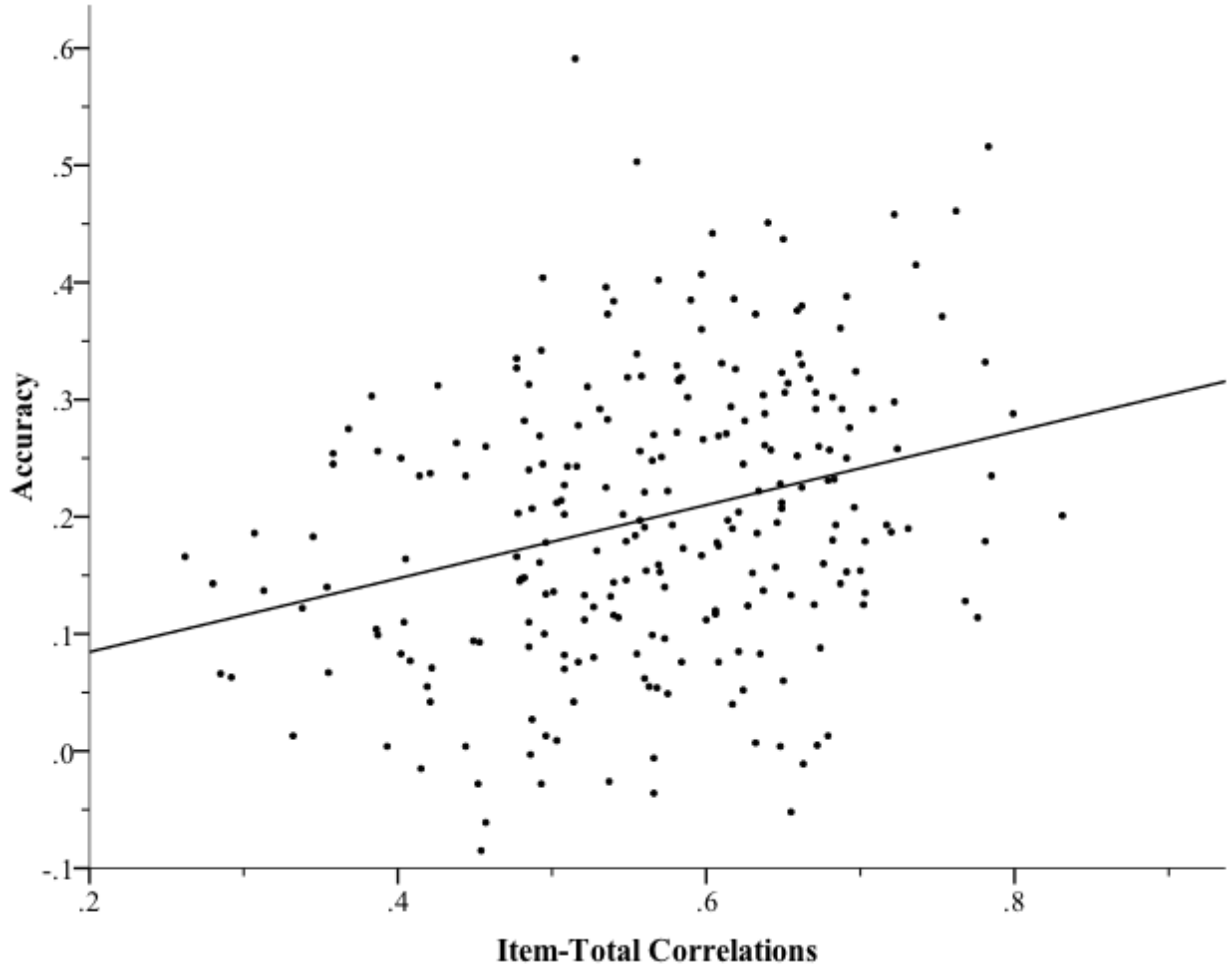

Figure 2. Scatterplot of 2004 NEO Personality Inventory - Revised (NEO-PI-R; *N* = 240 items) self-peer accuracy correlations (sample *N* = 124) on 1997 item-total correlations (sample *N* = 141).

3.4.3. Jackson Personality Inventory

Jackson Personality Inventory accuracy was regressed onto item correlations with the PRF Desirability scale, item-total correlations, and mean item responses (see Table 3). To assess curvilinear components of the relationships between accuracy and JPI item properties, accuracy was regressed onto the squared values of the item-PRF Desirability scale correlations and mean item responses. Item SDSVs were not available for the JPI. Taken together, the item properties accounted for a significant amount of the variance in accuracy, $R^2 = .11$, adjusted $R^2 = .09$, $F(5, 314) = 7.60$, $p < .001$. Significant beta weights for the model were associated with item-total correlations, $\beta = .21$, $t(314) = 3.74$, $p < .001$ (see Figure 3), and squared values of the mean item responses, $\beta = -1.14$, $t(314) = -3.45$, $p < .001$. The latter result suggests that there was a curvilinear relationship between mean item responses and self-peer correlations, such that moderate means elicited higher self-peer correlations than did low or high mean item responses (see Figure 4).

Table 3

Standardized and Unstandardized Regression Coefficients: Prediction of JPI Accuracy from JPI Item Properties

Item Property	B	Standard Error	β	*t*	*p*
Item-total correlations	.20	.05	.21	3.74	.001
Desirability correlation	.03	.06	.03	.43	.67
Desirability correlation^2	-.11	.35	-.02	-.32	.75
Mean	.17	.05	1.11	3.38	.001
Mean^2	-.02	.01	-1.14	-3.45	.001

Note. Squared variables were used to test curvilinear components.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.

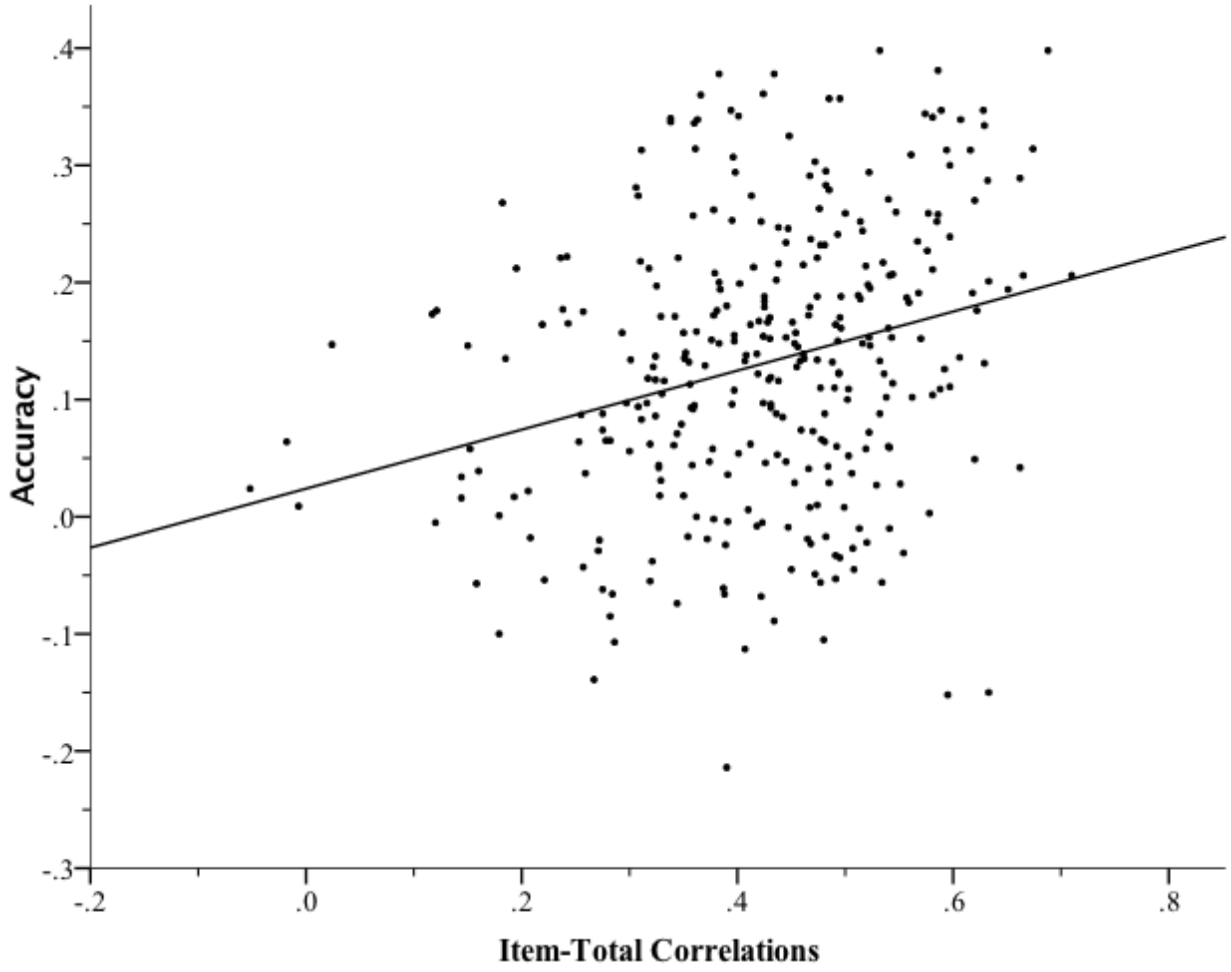

Figure 3. Scatterplot of 1994 Jackson Personality Inventory (JPI; *N* = 320 items) self-peer accuracy correlations (sample *N* = 92) regressed on 1997 item-total correlations (sample *N* = 141).

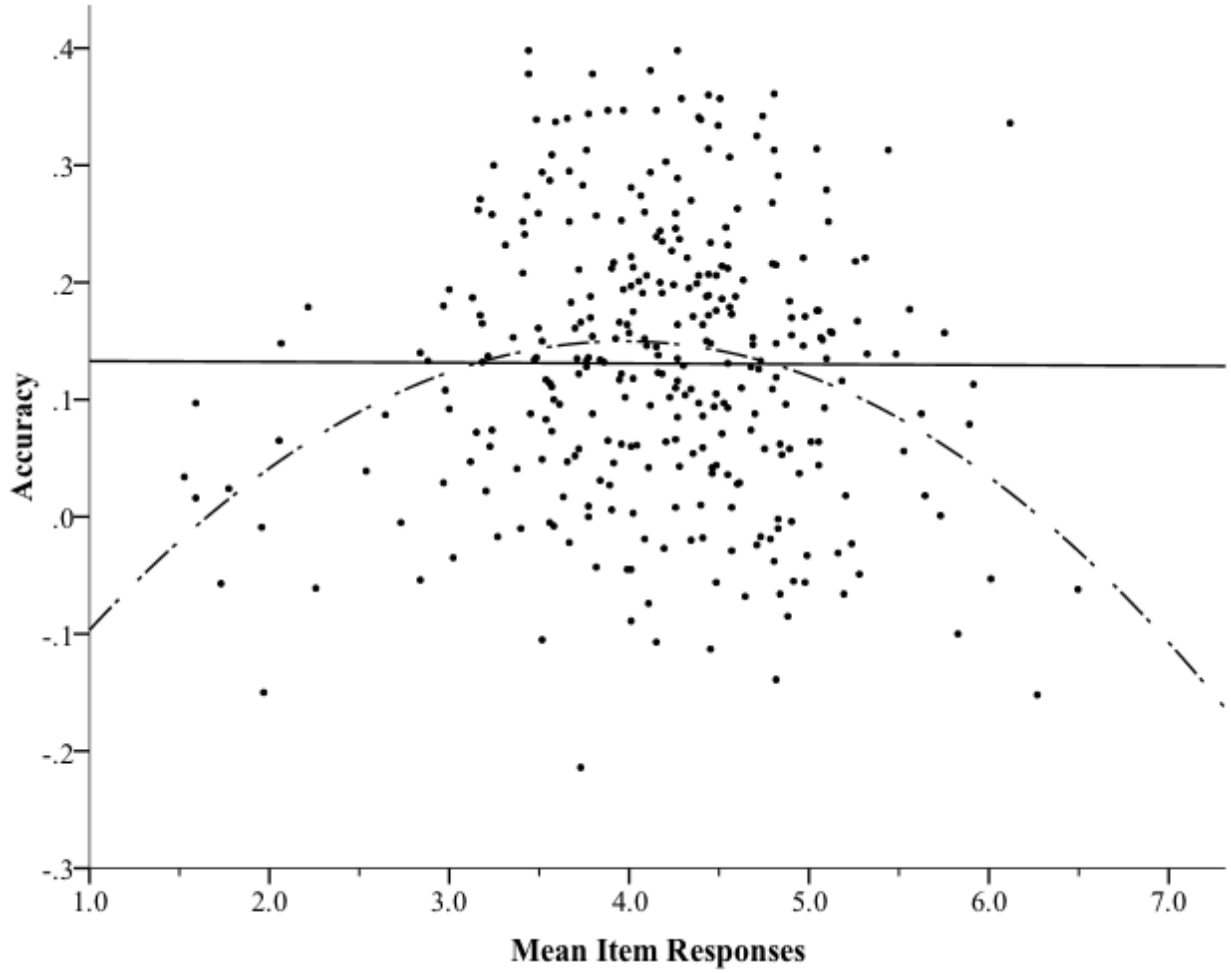

Figure 4. Scatterplot of 1994 Jackson Personality Inventory (JPI; *N* = 320 items) self-peer accuracy correlations (sample *N* = 92) on 1994 7-point scale mean item responses (sample *N* = 92).

3.4.4. Personality Research Form

Personality Research Form accuracy was regressed onto item SDSVs, item correlations with the PRF Desirability scale, item-total correlations, and mean item responses (see Table 4). To assess curvilinear components of the relationships between accuracy and PRF item properties, accuracy was regressed onto the squared values of item SDSVs, item-PRF Desirability scale correlations, and mean item responses. Taken together, the item properties accounted for a significant amount of the variance in accuracy, $R^2 = .18$, adjusted $R^2 = .16$, $F(7, 344) = 10.40$, $p < .001$. Significant beta weights for the model were associated with item-total

correlations, β = .29, $t(344) = 5.35$, $p < .001$ (see Figure 5), and squared values of the mean item responses, β = -.90, $t(344) = -2.95$, $p < .003$. The latter result suggests that there was a curvilinear relationship between mean item responses and self-peer correlations, such that moderate means elicited higher self-peer correlations than did low or high mean item responses (see Figure 6).

Table 4

Standardized and Unstandardized Regression Coefficients: Prediction of PRF Accuracy from PRF Item Properties

Item Property	B	Standard Error	β	*t*	*p*
Item-total correlations	.32	.06	.29	5.35	.001
SDSV	-.04	.05	-.29	-.68	.50
$SDSV^2$	.01	.01	.48	1.09	.28
Desirability correlation	.10	.07	.11	1.52	.13
Desirability correlation^2	.39	.22	.10	1.73	.09
Mean	.09	.04	.75	2.45	.02
$Mean^2$	-.01	.01	-.90	-2.95	.003

Note. Squared variables were used to test curvilinear components.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.

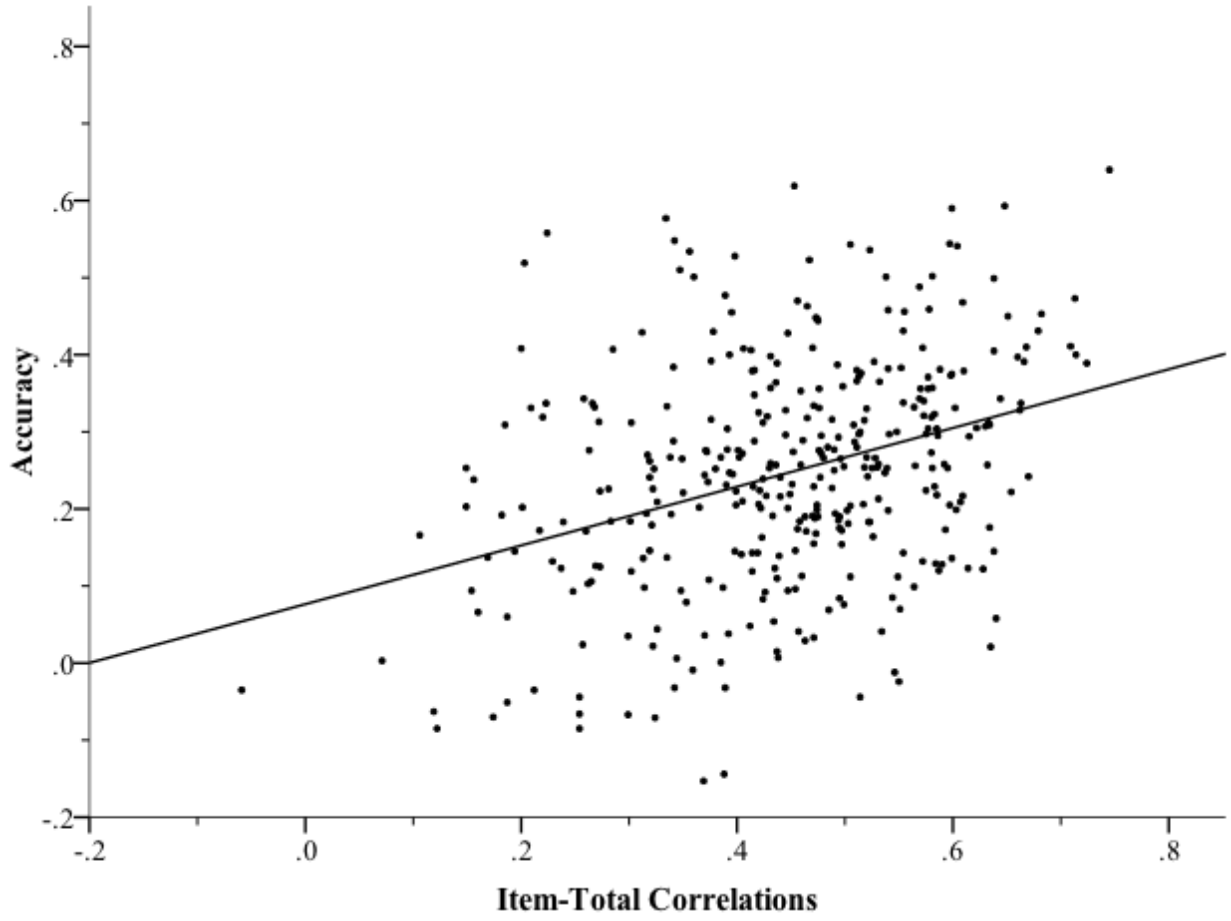

Figure 5. Scatterplot of 1981 Personality Research Form (PRF; *N* = 352 items) self-peer accuracy correlations (sample *N* = 90) on 1997 item-total correlations (sample *N* = 141).

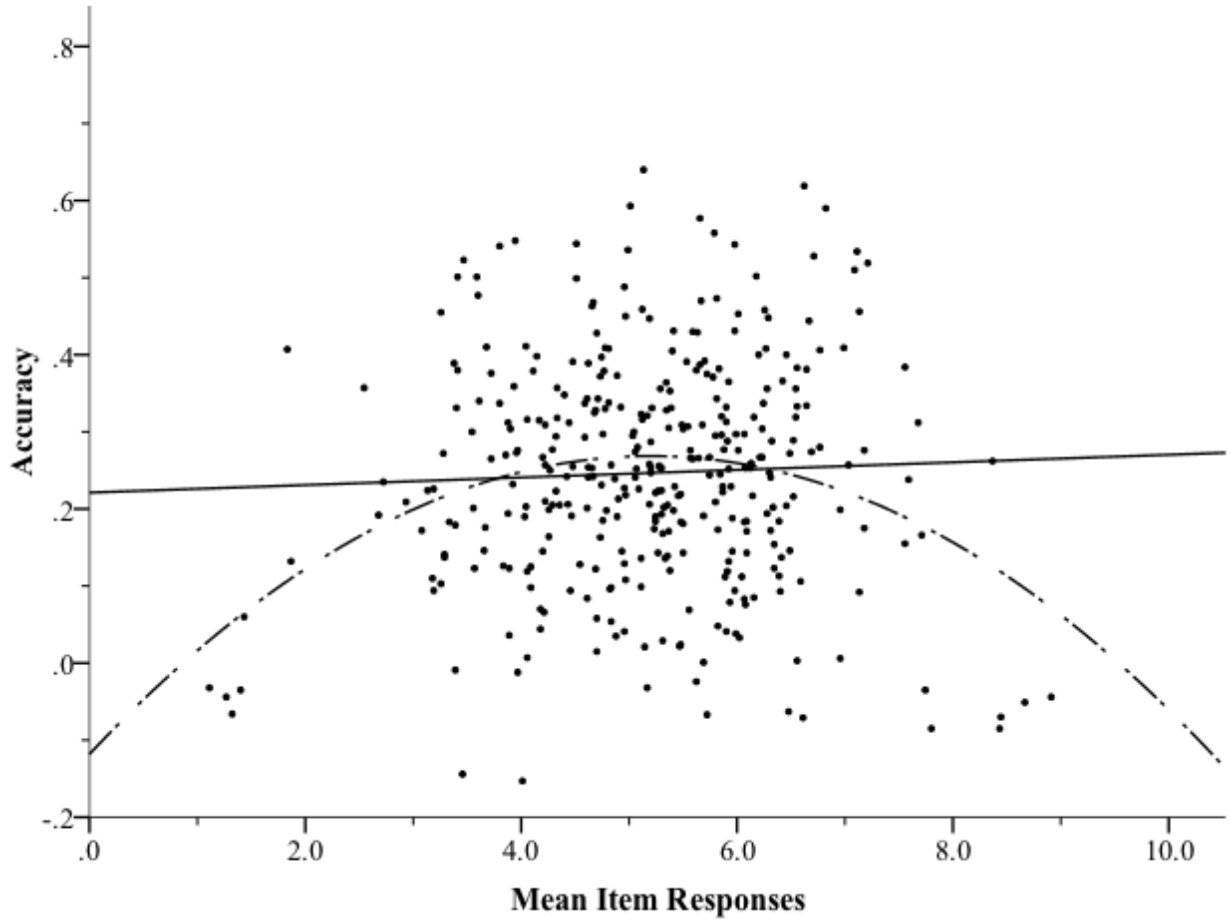

Figure 6. Scatterplot of 1981 Personality Research Form (PRF; N = 352 items) self-peer accuracy correlations (sample N = 90) on 1981 9-point scale mean item responses (sample N = 90).

3.4.5. Nonverbal Personality Questionnaire

Nonverbal Personality Questionnaire accuracy was regressed onto item correlations with the PRF Desirability scale, item-total correlations, and mean item responses (see Table 5). To assess curvilinear components of the relationships between accuracy and NPQ item properties, accuracy was regressed onto the squared values of the item-PRF Desirability scale correlations and mean item responses. Item SDSVs were not available for the NPQ. Taken together, the item properties accounted for a significant amount of the variance in accuracy, R^2 = .14, adjusted R^2 =

.10, $F(5, 130) = 4.05$, $p < .002$. A significant beta weight for the model was associated with item-total correlations, $\beta = .29$, $t(130) = 3.50$, $p < .001$ (see Figure 7).

Table 5

Standardized and Unstandardized Regression Coefficients: Prediction of NPQ Accuracy from NPQ Item Properties

Item Property	B	Standard Error	β	*t*	*p*
Item-total correlations	.41	.12	.29	3.50	.001
Desirability correlation	.15	.10	.16	1.56	.12
Desirability correlation^2	-.05	.52	-.01	-.09	.93
Mean	.09	.06	.69	1.39	.17
Mean^2	-.01	.01	-.71	-1.46	.15

Note. Squared variables were used to test curvilinear components.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.

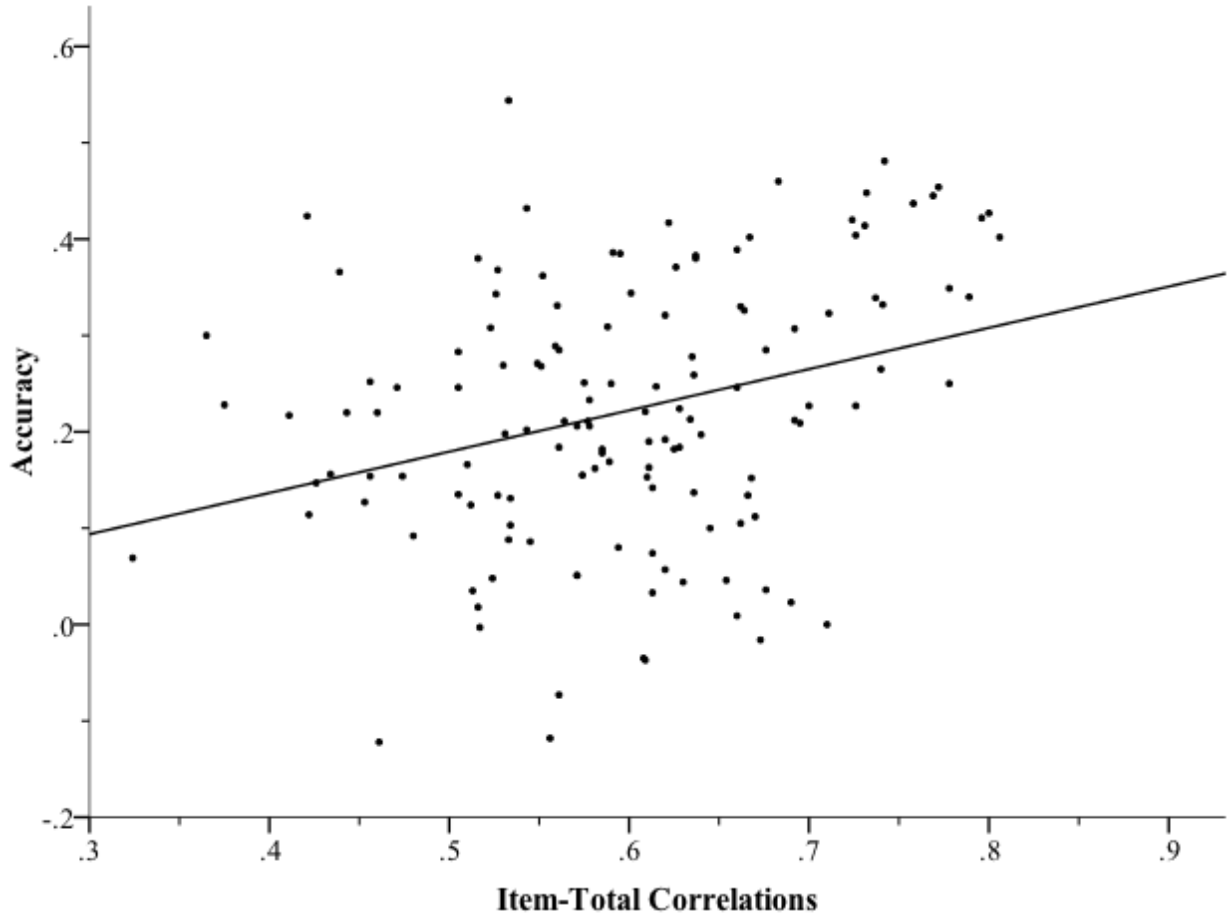

Figure 7. Scatterplot of 1993 Nonverbal Personality Questionnaire (NPQ; *N* = 136 items) self-peer accuracy correlations (*N* = 94) on 2004 item-total correlations (normative sample *N* = 304).

CHAPTER 4: DISCUSSION

4. Discussion

The current study sought to evaluate several key properties of personality test items that can affect their accuracy (i.e., validity), and to determine their impact on overall scale validity. The properties under investigation in the current study included: (a) item social desirability, (b) item difficulty (i.e., mean item responses), and (c) item content saturation. Overall scale validity, or accuracy, was estimated by correlating self-ratings on personality questionnaire items with peer ratings on the same items, with the hypothesis that the most accurate items are those in which self- and peer ratings of personality are highly correlated (e.g., Holden & Troister, 2009; Paunonen, 1984; Paunonen & O'Neill, 2010).

First, it was hypothesized that item social desirability scale values (SDSVs) would predict accuracy on a series of personality measures in a curvilinear manner. Specifically, we predicted that to the extent that items had high or low SDSVs, response accuracy would be low, whereas items with moderate SDSVs would be relatively more accurate. This view was based on the premise that items with high or low SDSVs are highly evaluative, and thus, are more likely to elicit more socially desirable responding by the target (Berg, 1967; Edwards, 1969; Edwards, 1970). Similarly, it was argued that correlations between scores on a socially desirable responding (SDR) measure and item responses on a personality measure would linearly predict self-peer agreement. Specifically, to the extent that there were strong positive (or negative) correlations between personality test items and scores on an SDR measure, it was anticipated that this would elicit low self-peer agreement and, in contrast, weak correlations would result in high self-peer agreement. Again, this is because strongly desirable or undesirable personality test items tend to elicit misrepresentation in self-report questionnaires due to their evaluative nature

(Paulhus, 2002). Additionally, it was hypothesized that item difficulty (i.e., mean item responses) would predict accuracy in a curvilinear manner. Specifically, to the extent that items had extreme high and low mean endorsement values, we anticipated that this would elicit low accuracy, whereas moderate mean endorsement values would elicit high accuracy. This hypothesis was based on expected restriction of range effects at the extremes of responding. Finally, we hypothesized that items exhibiting higher content saturation would result in higher accuracy than would items exhibiting lower content saturation. This was based on the premise that content saturated items are more likely to be homogeneous, prototypical representations of the trait being measured (Paunonen, 1987).

The current study used five different personality questionnaires to provide robust tests of the hypotheses: the Supernumerary Personality Inventory (SPI; Paunonen, 2002), the NEO Personality Inventory-Revised (NEO-PI-R; Costa & McCrae, 1992), the Jackson Personality Inventory (JPI; Jackson, 1976), the Personality Research Form (PRF; Jackson, 1984), and the Nonverbal Personality Questionnaire (NPQ; Paunonen et al., 2004). The socially desirable response tendencies of the roommate raters were assessed using one self-report questionnaire-based measure: the Personality Research Form Desirability Scale (PRF Desirability; Jackson, 1984).

Results revealed that item content saturation, measured using item-total correlations, was the most salient linear predictor of item response accuracy (i.e., self-peer agreement) across all five personality questionnaires. Results also revealed that mean item responses curvilinearly predicted response accuracy in two of the five questionnaires: the JPI and the PRF. Neither item SDSVs, nor personality questionnaire item correlations with an SDR measure predicted response accuracy on any of the five personality questionnaires.

Consistent with prediction and with past literature (e.g., Ashton & Goldberg, 1973; Hase & Goldberg, 1967; Jackson, 1975; Paunonen, 1984), the present findings indicated that participants responded to content saturated items infused with trait-relevant content more accurately than items that were only tangentially related to the trait in question. In other words, more content saturated items on all five personality questionnaires elicited greater convergence on self- and peer responses than did less content saturated items. This finding is in accordance with past research that has empirically compared construct-oriented scales, on which items are developed to be representative and salient exemplars of the trait (Holden et al., 1985; Paunonen, 1984; Paunonen & Jackson, 1985), to criterion-keyed scales, on which items are selected on the basis of their ability to predict a particular criterion (John & Benet-Martinez, 2000). For instance, Paunonen (1984) found that ad hoc PRF scales constructed to reflect maximum content saturation elicited higher correlations with predictive criteria (i.e., peer ratings, adjective trait ratings, and nonverbal stimuli) than did criterion-oriented scales that were constructed to optimize the prediction of mean roommate ratings for a PRF trait measure.

Ashton and Goldberg (1973) found that the PRF trait scales developed based on psychological theory (i.e., more content saturated) outperformed the empirically-derived California Psychological Inventory (CPI; Gough, 1957) on accuracy measured by correlating self- and peer item responses. Similarly, Jackson (1975) reported that the construct-oriented trait scales of the Jackson Personality Inventory were more highly correlated with both self-ratings and peer ratings on adjective scales representing the traits in question than were the criterion-oriented scales of the CPI. Researchers have contended that construct-based items, developed to be more internally consistent, prototypical, and content saturated, contribute more to personality questionnaire validity than criterion-based items because they are highly salient and

representative of the trait. To the contrary, criterion-oriented scales developed without consideration of psychological theory appear to allow for greater ambiguity in responses, guessing in interpretations, and a varying focus on differential aspects of item content (Paunonen, 1984). Thus, a theory-based method of test construction, in which more salient and content saturated items are selected, is generally preferred over the criterion-based method in order to elicit accurate responding.

Consistent with prediction, mean item endorsement levels curvilinearly predicted item response accuracy in two of five personality questionnaires (i.e., the JPI and the PRF). Specifically, in these two questionnaires, items with extreme low or high mean endorsement levels (e.g., values close to 0.0 or 1.0 on a binary scale, or 1 or 5 on a 5-point Likert scale) elicited less accurate responding than did items with moderate mean endorsement levels (e.g., values close to .50 on a binary scale, or 3 on a 5-point Likert scale). This prediction was derived on the basis of restriction of range effects. Based on past literature, it was predicted that extreme mean item responses would result in a restricted variance of item responses that would attenuate the relationship between two variables, or criterion validity (e.g., Epstein, 1983; Nunnally & Bernstein, 1994; Sackett, Lievens, Berry, & Landers, 2007). To elaborate, if everybody endorsed the same response to an item, the item would be ineffective in assessing individual differences. Additionally, this restricted variance would place upper limits on the strength of the correlations that could exist between the item and criterion variables, such as self-peer response correlations (Epstein, 1983). Items with moderate mean responses, on the other hand, maximize observed score variance, allowing for maximal correlations between item responses and a criterion (e.g., Feldt, 1993). The same curvilinear relationship between mean item responses and accuracy did not emerge for the SPI, the NEO-PI-R, and the NPQ. That is, on these questionnaires, mean item

endorsement levels had no effect on the correlations between self- and peer ratings of personality.

One possible, albeit weak explanation for the inconsistent findings across personality questionnaires is that items with moderate mean endorsement levels (i.e., items with larger variances) are not always more valid than items with restricted variances, and as a result, these items may not reflect reliable individual differences (e.g., Epstein, 1983). Instead, the larger variances for items with moderate means on the SPI, NEO-PI-R, and the NPQ may have reflected error variance as opposed to true score variance. For instance, perhaps the items with moderate means were ambiguous or unclear in nature, and elicited guessing from individuals. These items would not be any more valid than those items extreme in mean endorsement levels, thus attenuating relationships between mean response levels and accuracy. However, this explanation is less plausible, given the extensive and rigorous validation procedures that each questionnaire underwent during item analysis phases of test construction.

An alternative explanation is such that items with extreme endorsement levels had been eliminated from the questionnaires following the rigorous validation procedures, and the relationship between mean item responses and accuracy was subsequently attenuated. In line with this explanation, it was observed that the majority of the mean item endorsement levels were moderate for each of the questionnaires (see Table D.1). Jackson (1970) eliminated items from the PRF that only a small percentage, or almost all of the individuals endorsed using a computer program that would classify such items as “unreliable.” Likewise, throughout the test construction process and preliminary item analysis procedures for both the SPI and the NPQ, Paunonen (2002) and Paunonen et al. (2004) discarded items representing extremely rare and extremely popular behaviours, and items for which variance values were not acceptable.

Therefore, the lack of items reflecting extreme levels of responding may have minimized potential larger correlations between mean item responses and accuracy.

Contrary to prediction, neither item SDSVs, nor personality questionnaire item correlations with SDR scale scores predicted item response accuracy across any of the five personality questionnaires. This original prediction was based on the premise that items infused with desirability bias have the potential to elicit misrepresentation from the respondent on his or her true level of trait, which, in turn, may compromise the measure's construct validity. Specifically, the respondent may endorse response options to items on personality measures in order to present a more favourable image of the self than is warranted (Paulhus, 2002; Paunonen & LeBel, 2012; Podsakoff et al., 2003). Unlike maximal performance measures, the typical performance personality questionnaires do not have inherently correct or incorrect answers (Paunonen & LeBel, 2012). Thus, test examiners cannot be certain whether or not the respondent has, in fact, selected response options representing more desirable characteristics than warranted. Not only might the construct validity of these scales be compromised, but items eliciting SDR can also affect mean levels of responding (Ganster et al., 1983; Podsakoff et al., 2003), which may alter relationships between the test and variables such as validation criteria. However, items infused with desirable content are less likely to elicit the same desirability biases in the peer, as the peer might not be as motivated to inflate target ratings as they are to inflate self-ratings (John & Robins, 1993; Paunonen & O'Neill, 2010). We hypothesized that this discrepancy between self- and peer responses to extremely (un)desirable items would subsequently attenuate validity coefficients.

Our findings are in contrast with past research that has identified a curvilinear relationship between item social desirability levels and self-peer agreement (e.g., John & Robins,

1993). Specifically, traits that were rated by college students as being neutral in social desirability elicited more self-peer agreement than did traits that were rated as being high or low in desirability. These results were replicated in a community sample wherein peers had known the target judges for an average of 18 years (John & Robins, 1993). However, these findings have been inconsistent across studies. For instance, Paunonen and Kam (2014) did not find significant relationships between item SDSVs and item accuracy coefficients in either the SPI items (r = .04, p > .05), nor the NEO-PI-R items (r = -.01, p > .05). Similarly, Holden et al. (1985) found no relationship between absolute item desirability and item criterion validity (r = -.11, p > .05). Interestingly, in a study using personality adjectives, Funder (1980) revealed that more socially desirable items elicited greater self-peer agreement than did socially undesirable or neutral items, with a significant correlation of .30 between item SDSVs and self-peer agreement.

One possible explanation for these inconsistent findings is that some personality questionnaire items may possess both high levels of social desirability and high levels of content saturation. In fact, it is near impossible to solely select items high in content saturation that are entirely neutral in desirability (Jackson, 1970). If items reflecting any (un)desirable content were removed, this would also result in removing valid content variance from the scale, which would compromise its construct and content validity. Thus, items that have some desirability in their content may elicit accurate responding, so long as the levels of item content saturation remain higher than levels of item desirability. In other words, if an item has moderate desirability variance as well as moderate-to-high content saturation (e.g., as measured using a Differential Reliability Index), the item may still elicit accurate responding (Jackson, 1970). Thus, perhaps the items in question demonstrated enough content saturation, and were prototypical enough of the trait under investigation that any desirability variance did not offset their responses.

An additional explanation for the current results involves the link between item desirability levels and construct validity (cf. criterion validity). Specifically, Paunonen and LeBel (2012) investigated the effect of social desirability bias on the criterion validity of simulated responses to bipolar personality trait adjectives using Monte Carlo procedures. These bipolar trait adjectives represented relatively desirable and undesirable traits (e.g., honesty-dishonesty poles; Paunonen & LeBel, 2012). The results revealed that although adding large components of social desirability to test scores altered observed trait scores drastically from their true level of the trait, criterion validity (i.e., test score-criterion correlations) remained relatively unaffected by the intrusion of desirability variance. The authors speculated that the reason why SDR did not act as a moderator of criterion validity (i.e., the relationship between the predictor and the criterion did not vary as a function of SDR) was because their linear data transformations did not systematically change the rankings of respondents' simulated personality trait scores depending on their levels of SDR. However, even if criterion validity had not been compromised, the intrusion of extremely desirable content into test items is still incredibly problematic for the construct validity of the measures because observed scores were drastically altered from true scores on the traits (Paunonen & LeBel, 2012). Perhaps in the current study, although extreme desirable or undesirable items did not alter self-peer personality test item response correlations, these items still may have elicited SDR tendencies from respondents, thus causing discrepancies between individuals' obtained and true scores on the traits in question. Thus, researchers constructing personality tests should take into consideration that highly desirable and undesirable items have an effect on the construct validity of their scales, and it is most likely prudent to avoid items that are extremely evaluative in nature.

Finally, another possible explanation for the findings of this study concerns the nature of the items' levels of social desirability. It could be the case that the items were not infused with enough desirable content to have had a substantial effect on the accuracy of the measures. For instance, as stated previously, the questionnaires had already undergone extensive validation procedures that eliminated extreme desirable or undesirable items from their respective item pools. During initial item analysis phases of test construction for the JPI and the PRF, Jackson (1970, 1976) used the Differential Reliability Index (DRI), a statistical indicator of content saturation, to ensure that items selected for the scales did not reflect high desirability components. Item DRIs measure how strongly an item relates to the content domain of interest while partialing out variance due to social desirability. The other questionnaires underwent similar validation procedures. For example, original SPI and NPQ items that correlated more highly with a measure of desirability than with the other items in the scale were discarded during scale construction (Paunonen, 2002; Paunonen et al., 2004). Additionally, the authors of the NEO-PI-R have contended that socially desirable responding does not pose a threat to the validity of their scale (Costa & McCrae, 1988; Costa & McCrae, 1992; McCrae & Costa, 1983) based on scale correlations with social desirability. Our findings corroborated these assertions. Social desirability scale values (SDSVs) in our samples ranged from 1.39 (PRF) to 8.61 (NEO-PI-R) on a 9-point scale, with means all approximating 5.00. Similarly, our personality questionnaire item-PRF Desirability scale correlations ranged from -.52 (PRF) to .41 (PRF), with all means approximating zero. Therefore, it is evident that the personality questionnaire items under investigation were largely free of content extreme in desirability, which may have potentially minimized the effects of item desirability levels on accuracy.

4.1. Limitations and Future Directions

Some limitations of the current study should be noted and addressed in future studies. First, the five samples used in the present research were convenience samples comprising groups of undergraduate students (see Table A.1 for demographic information). Specifically, the groups were predominantly first-year students with a mean age of 19 years, and a large number of the recruits were female. It is unclear whether the current results would generalize to a wider demographic. However, five different undergraduate samples collected over a time span of 23 years were used for the current research. Thus, it is probable that the results likely would generalize, at the very least, to undergraduate samples across time.

Another limitation of the current research is that participants varied widely in the duration of time that they were acquainted with their roommates. Specifically, sample means of participants' reported duration of time acquainted with their respective roommates ranged from 14.43 months (SD = 21.59) to 28.59 months (SD = 71.08). Additionally, although each of the five studies were carried out during the second to last month of the academic year to ensure that participants had substantial time to become familiar with their roommates, of the samples collected in 1981 and 1994, 6.7% and 3.3% of participants, respectively, reported having known their roommate for less than three months. It has been well-established in the person perception literature that accuracy of personality questionnaire responses increases as a function of acquaintanceship (e.g., Funder & Colvin, 1988; Norman & Goldberg, 1966; Paunonen, 1989). Therefore, it is a possibility that length of time of roommate-peer acquaintanceship may have influenced our results, such that relevant cues regarding behaviours available to the peer varied as a function of the number of interpersonal encounters (Paunonen, 1989). As such, the responses to personality questionnaire items may have correlated to a greater degree with their

roommates' when they were better acquainted. However, the number of participants with less than three months of acquaintanceship was small, with the majority (more than 93%) reporting high degrees of familiarity with their roommates (see Table A.1). Future research, however, should ensure that participants have been sufficiently acquainted with their roommates for a minimum duration of time so as to avoid subsequent biases in the results. As an added precaution, a check on measurement accuracy could be done by correlating self-report responses to personality questionnaire items with responses from other well-acquainted individuals such as a parent, significant other, or sibling, or with alternative behavioural criteria. This may serve to eliminate the potential bias of lack of target familiarity.

An additional methodological limitation of the current study is that item SDSVs were not collected for the JPI or the NPQ. If item SDSVs were included, they may have had an effect on self- and peer item response correlations. This is an unlikely possibility, as item social desirability was unrelated to self- and peer response correlations for all of the questionnaires investigated in the current study. However, it would still be beneficial in future research to collect item SDSVs for the JPI and the NPQ, and examine the effects of these SDSVs on accuracy.

A fourth limitation of the current study is that we only used scales that have undergone extensive and rigorous validation procedures. Each of the scales used for the current analyses are widely-used measures of general personality traits, and have been validated across many cultures and communities. Thus, items that are extremely low in content saturation, high in desirability, and extreme in mean responses were most likely eliminated from their item pools in the early scale construction process. It may be beneficial in future research to investigate the effects of these item properties on response accuracy using self-report personality questionnaires that have

not undergone the same exhaustive validation procedures (e.g., a newly developed measure). This way, one could further delineate the effects of item properties on accuracy using a wider range of item-total correlations, item desirability levels, and mean item responses. This type of study could also serve as a validation tool for new self-report measures of personality.

Due to the limited scope of this thesis, it was not tenable to investigate the effects of each of the item properties mentioned in the introduction section on accuracy. However, future research could evaluate the effects of item observability, item wording, face validity, and item subtlety on correlations between self- and peer reports of personality. Furthermore, content saturation could be assessed by asking a group of judges to estimate item-construct linkages of individual test items. Participants would be provided with a clear and comprehensive definition of the trait in question, complete with trait-defining behaviours (Paunonen & Hong, 2015). The judges would then be asked to estimate how prototypical each item is of the trait domain. Aside from using item-total correlations and content saturation ratings, one could identify content saturated items using, for instance, Neill and Jackson's (1976) Item Efficiency Index (IEI), an estimation of content saturation whereby trait-irrelevant content is subtracted from item-total correlations. A higher IEI value indicates higher content saturation. Another option would be to evaluate item social desirability using the Differential Reliability Index (DRI; Jackson, 1970). An item's DRI is essentially an item's IEI (Neill & Jackson, 1976) aimed specifically at reducing desirability variance from items.

Finally, future research should assess the effect of person perception on accuracy. Indices of person perception could include the degree of self-peer acquaintanceship and level of observability of rated behaviours. It may be expected that well-acquainted peers completing measures of observable behaviours would elicit maximally accurate peer ratings of targets

(Cheek, 1982; Paunonen, 1989; Paunonen & Kam, 2014; Paunonen & O'Neill, 2010). It is also possible that acquaintanceship would moderate the relationship between item observability and accuracy, such that observability of a trait is important in making accurate judgments of a target when pairs are low to moderately acquainted, but less important for those pairs who are highly acquainted (Paunonen, 1989).

4.2. Concluding Remarks

The primary purpose of the current study was to evaluate which item properties contribute to the overall convergent validity, or the accuracy of a measure. The current study is the first, to the author's knowledge, that has investigated the effects of the specific item properties (cf. overall scale properties) on the convergent validity of five widely-used self-report personality questionnaires. The samples constituted a diverse group of individuals tested over a 23-year time span, which provides some strong support for the generalizability of the reported results. Furthermore, diverse means of measuring the item properties investigated were employed in the current study. For example, item social desirability was evaluated in two different ways (i.e., item SDSVs and item-PRF Desirability scale correlations), and item-total correlations were used to measure content saturation as an alternative to strategies previously employed, such as factor analytic procedures (e.g., Paunonen, 1984).

This study has enhanced our current understanding of some the observed inconsistencies in personality testing over the years. The field of personality psychology faced a great deal of scrutiny in the 1960s when critics of self-report personality testing (e.g., Fiske, 1978; Mischel, 1968) argued that personality questionnaire score correlations with objective measures of behaviour did not exceed a ceiling of .30 (Epstein & O'Brien, 1985). This led researchers to conclude that scores on self-report personality questionnaires were not attributable to the

individual's stable, enduring traits, but instead, were attributable only to the particular situation at hand. These critics of self-report personality testing asserted that observable behaviours are largely invariant across time and situations, and that these behavioural tendencies are best explained by, for instance, social learning processes and operant conditioning (Mischel, 1968; Mischel & Peake, 1982).

This debate spurned a wealth of research conducted by personality theorists concerning the improvement of traditional methods of personality test construction and assessment (Epstein & O'Brien, 1985; Jackson & Paunonen, 1985; Paunonen, 1984; Paunonen & Jackson, 1985). At present, it is generally agreed upon that many of the apparent inconsistencies in personality across time and situations reported in these studies can be explained by the use of scales lacking in such basic, yet fundamental psychometric principles as reliability and validity (Epstein & O'Brien, 1985; Jackson & Paunonen, 1985). Based on the results of the current study, it is evident that in order to maximize the validity, or accuracy, of a self-report personality measure, it is necessary to write items high in content saturation (i.e., homogeneous and thematic item content), and to write items that will elicit moderate mean responses in order to achieve item-criterion correlations in excess of .30.

REFERENCES

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55*(3), 387-395. http://dx.doi.org/10.1037/0022-3514.55.3.387.

Ashton, S. G., & Goldberg, L. R. (1973). In response to Jackson's challenge: the comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices, and laymen. *Journal of Research in Personality, 7*(1), 1-20.

Beer, A. & Watson, D. (2008). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment, 90*(3), 250-260. doi:10.1080/00223890701884970.

Berg, I. A. (1967). *Response set in personality assessment*. Chicago: Aldine.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*(1), 106-148. doi:10.1111/j.1467-6494.1986.tb00391.x.

Burgess, P. M., Campbell, I. M., & Zylberberg, A. (1984). Face validity vs. item subtlety in the MMPI D scale. *Journal of Clinical Psychology*, *40*(2), 499-504.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105. http://dx.doi.org/10.1037/h0046016.

Carretero-Dios, H., Perez, C., & Beula Casal, G. (2009). Content validity and metric properties of a pool of items developed to assess humor appreciation. *The Spanish Journal of Psychology, 12*(2), 773-787.

Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer-rating study. *Journal of Personality and Social Psychology, 43*(6), 1254-1269. doi:10.1037/0022-3514.43.6.1254.

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319. http://dx.doi.org/10.1037/1040-3590.7.3.309.

Coleman, C. M. (2013*). Effects of negative keying and wording in attitude measures: A mixed methods study* (Doctoral dissertation). Retrieved from ProQuest Theses and Dissertations Database. (Accession No. 3560664).

Cone, J. D. (1981). Psychometric considerations. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (2nd ed., pp. 38-68). NewYork: Pergamon.

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment, 15*(1), 110-117. http://dx.doi.org/10.1111/j.1468 2389.2007.00371.x.

Costa, P. T., Jr., & McCrae, R. R. (1988). From catalog to classification: Murray's needs and the five-factor model. *Journal of Personality and Social Psychology, 55*, 258-265. doi:http://dx.doi.org/10.1037/0022-3514.55.2.258.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 475-494. doi: 10.1177/001316444600600405.

Duff, F. L. (1965). Item subtlety in personality inventory scales. *Journal of Consulting Psychology, 29,* 565-570. http://dx.doi.org/10.1037/h0022753.

Edwards, A. L. (1969). Trait and evaluative consistency in self-description. *Educational and Psychological Measurement, 29,* 737-752. doi:10.1177/001316446902900401.

Edwards, A. L. (1970). *The measurement of personality traits by scales and inventories.* New York: Holt, Rinehart, & Winston.

Epstein, S. (1983) Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51*(3), 360-392. doi: 0.1111/j.1467 6494.1983.tb00338.x.

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

Feldt, L. S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education, 6*(1), 37-48. doi:http://dx.doi.org/10.1207/s15324818ame0601_3.

Fiske, D. W. (1973). Can a personality construct be validated empirically? *Psychological Bulletin, 80*(2), 89-92.

Foster, S. L. & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7*(3), 248-260. doi: 10.1037//1040-3590.7.3.248.

Funder, D. C. (1980). On seeing ourselves as others see us: Self-other agreement and discrepancy in personality ratings. *Journal of Personality, 48*(4), 473-493. doi: 10.1111/j.1467-6494.1980.tb02380.x.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55*(1), 149-158. doi:http://dx.doi.org/10.1037/0022-3514.55.1.149

Ganster, D.C., Hennessey, H., & Luthans, F. (1983). Social desirability response effects: Three alternative models. *Academy of Management Journal, 26*, 321-331. doi: 10.2307/255979.

Gough. H. G. (1957). *Manual for the California Psychological Inventory*. Palo Alto, CA: Consulting Psychologist Press.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, *67*(4), 231-248. doi:http://dx.doi.org/10.1037/h0024421.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10*, 249-254. doi:10.1093/OBO/9780199828340-0118.

Helmes, E., Reed, P. L., & Jackson, D. N. (1977). Desirability and frequency scale values and endorsement proportions for items of Personality Research Form-E. *Psychological Reports, 41*(2), 435-444. http://dx.doi.org.proxy1.lib.uwo.ca/10.2466/pr0.1977.41.2.435.

Holden, R. R., Fekken, G. C., & Jackson, D. N. (1985). Structured personality test item characteristics and validity. *Journal of Research in Personality, 19*, 386-394. doi: 10.1016/0092-6566(85)90007-8.

Holden, R. R., & Jackson, D. N. (1979). Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology, 47*(3), 459-468. doi: 10.1037/0022-006X.47.3.459.

Holden, R. R. & Troister, T. (2009). Developments in the self-report assessment of personality and psychopathology in adults. *Canadian Psychology, 50*(3), 120-130. doi: http://dx.doi.org/10.1037/a0015959.

IBM Corp. Released 2015. *IBM SPSS Statistics for Macintosh*, Version 23.0. Armonk, New York: IBM Corp.

Jackson, D. N. (1970). A sequential system for personality scale development, In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (pp. 61-96). New York: Academic Press.

Jackson, D. N. (1971). The dynamics of structured personality tests. *Psychological Review, 78*(3), 229-248.

Jackson, D. N. (1974). *Personality Research Form manual.* Port Huron, MI: Research Psychology Press.

Jackson, D. N. (1975). The relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction. *Educational and Psychological Measurement, 35*(2), 361-370.

Jackson, D. N. (1976). *Jackson Personality Inventory manual.* Port Huron, MI: Research Psychologists Press.

Jackson, D. N. (1977). Reliability of the Jackson Personality Inventory. *Psychological Reports, 40*(2), 613-614.

Jackson, D. N. (1984). *Personality Research Form manual.* Port Huron, MI: Research Psychology Press.

Jackson, D. N. & Paunonen, S. V. (1980). Personality structure and assessment. *Annual Review of Psychology, 31,* 503 – 551. doi: 10.1146/annurev.ps.31.020180.002443.

Jackson, D. N. & Paunonen, S. V. (1985). Construct validity and the predictability of behavior. *Journal of Personality and Social Psychology, 49*(2), 554-570. doi: http://dx.doi.org/10.1037/h0092818.

John, O. P. & Benet-Martinez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H.T. Reis & C.M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 339-369). New York, NY: Cambridge University Press.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61*(4), 521-551. doi:10.1111/j.1467-6494.1993.tb00781.x.

Kam, C. (2013). Probing item social desirability by correlating personality items with Balanced Inventory of Desirable Responding (BIDR): A validity examination. *Personality and Individual Differences, 54*, 513-518. doi:10.1016/j.paid.2012.10.017.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. New York: Methuen.

Kline, R. B. (2011). *Principles and practice of structural equation modeling, 3rd ed*ition. New York: The Guilford Press.

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology, 98*(4), 668-682. http://dx.doi.org/10.1037/a0018771.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694. doi:10.2466/pr0.1957.3.3.635.

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51,* 882-888. doi:http://dx.doi.org/10.1037/0022-006X.51.6.882

McCrae, R. R., & Costa, P. T., Jr. (2010). *NEO Inventories: Professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*(6), 730-755. doi:http://dx.doi.org/10.1037/0033-295X.89.6.730.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford Press.

Neill, J. A., & Jackson, D. N. (1976). Minimum redundancy item analysis. *Educational and Psychological Measurement, 36,* 123-134. doi:10.1177/001316447603600111.

Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4*(6), 681-691. doi:http://dx.doi.org/10.1037/h0024002

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

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory: Third edition.* New York: McGraw-Hill.

Paulhus, D. L. (1981). Control of social desirability in personality inventories: Principal factor deletion. *Journal of Research in Personality, 15,* 383-388. doi:10.1016/00926566(85)90036-4.

Paulhus, D. L., & Bruce, M. N. (1992). The effect of acquaintanceship on the validity of personality impressions: A longitudinal Study. *Journal of Personality and Social Psychology, 63,* 816-824. doi:10.1037/0022-3514.63.5.816.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 67-88). Hillsdale, NJ: Erlbaum.

Paunonen, S. V. (1982). Behavioral consistency and individual differences in predictive structure (Doctoral dissertation). Retrieved from ProQuest Dissertations Publishing. (Accession No. NK54116).

Paunonen, S. V. (1984). Optimizing the validity of personality assessments: The importance of aggregation and item content. *Journal of Research in Personality, 18,* 411-431. doi:10.1016/0092-6566(84)90001-1.

Paunonen, S. V. (1987). Test construction and targeted factor solutions derived by multiple group and procrustes methods. *Multivariate Behavioral Research, 22,* 437-455. http://dx.doi.org/10.1207/s15327906mbr2204_4.

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target-rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology, 56,* 823-833. doi:10.1037/0022-3514.56.5.823.

Paunonen, S. V. (1994). *Personality questionnaire self- and peer report responses.* Unpublished raw data.

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology, 74*(2), 538-556. doi:10.1.1.463.4362.

Paunonen, S. V. (2002). *Design and construction of the Supernumerary Personality Inventory (Research Bulletin 763).* London, Ontario: University of Western Ontario.

Paunonen, S. V. (2004). *Personality questionnaire self- and peer report responses.* Unpublished raw data.

Paunonen, S. V. (2015). Sex differences in judgments of social desirability. *Journal of Personality.* doi:10.1111/jopy.12169.

Paunonen, S. V. & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81*(3), 524-539. doi:10.1037//0022-3514.81.3.524.

Paunonen, S. V., Ashton, M. C., & Jackson, D. N. (2001). Nonverbal assessment of the Big Five personality factors. *European Journal of Personality, 15,* 3-18. doi: 10.1002/per.385.

Paunonen, S. V., & Hong, R. Y. (2013). The many faces of assumed similarity in perceptions of personality. *Journal of Research in Personality, 47*, 800-815. doi:10.1016/j.jrp.2013.08.007.

Paunonen, S. V., & Hong, R. Y. (2015). On the properties of personality traits. In M. Mikulincer & P. R. Shaver (Eds.), *APA handbook of personality and social psychology: Vol. 4. Personality processes and individual differences* (pp. 233-259). Washington, DC: American Psychological Association.

Paunonen, S. V., & Jackson, D. N. (1985). The validity of formal and informal personality assessments. *Journal of Research in Personality, 19,* 331-342. http://dx.doi.org/10.1016/0092-6566(85)90001-7.

Paunonen, S. V., Jackson, D. N., & Ashton, M. C. (2004). *NPQ manual: Nonverbal Personality Questionnaire (NPQ) and Five-Factor Nonverbal Personality Questionnaire (FF-NPQ).* Port Huron, MI: Sigma Assessment Systems.

Paunonen, S. V., Jackson, D. N., & Keinonen, M. (1990). The structured nonverbal assessment of personality. *Journal of Personality, 58,* 481-502.

Paunonen, S. V. & Kam, C. (2014). The accuracy of roommate ratings of behaviors versus beliefs. *Journal of Research in Personality, 52,* 55-67. doi:10.1016/j.jrp.2014.07.006.

Paunonen, S. V., & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology, 103*(1), 158-175. http://dx.doi.org/10.1037/a0028165.

Paunonen, S. V. & O'Neill, T. A. (2010). Self-reports, peer ratings and construct validity. *European Journal of Personality, 24,* 189 – 206. doi:10.1002/per.751.

Podsakoff, P., MacKenzie, S., Lee, J., & Podsakoff, N. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology,* 88, 879-903. doi:10.1037/00219010.88.5.879.

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology, 92*(2), 538-544. doi:http://dx.doi.org/10.1037/0021-9010.92.2.538

Schriesheim, C., & Hill, K. (1981). Controlling acquiescence response bias by item Reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*, 1101-1114. doi:10.1177/001316448104100420.

Scott, W. A. (1960). Measures of test homogeneity. *Educational and Psychological Measurement, 20*, 751-757. http://dx.doi.org/10.1177/001316446002000411.

Shrauger, J. S. & Schoeneman, T. J. (1979). Symbolic interactionist view of self-concept: Through the looking glass darkly. *Psychological Bulletin, 86*(3), 549-573. doi:10.1037/0033-2909.86.3.549.

Smith, G. T., Fischer, S., & Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment, 15*(4), 467-477. doi:10.1037/10403590.15.4.467.

Valentine, L. D. (1969). Review of Personality Research Form, *Professional Psychology, 1*(1), 82-83.

APPENDICES

APPENDIX A

Demographic Variables by Sample

Table A.1

Demographic Variables by Sample

Variable	1981 (*N* = 90)	1993 (*N* = 94)	1994 (*N* = 92)	1997 (*N* = 141)	2004 (*N* = 124)
Age (years)					
Mean	19.18	19.24	19.23	19.20	18.79
SD	.84	.84	.85	.66	.69
Minimum	18	18	17	17	17
Maximum	22	21	24	22	20
Gender (*N*)					
Male	20 (22%)	34 (36%)	28 (30%)	46 (33%)	42 (34%)
Female	70 (78%)	60 (64%)	64 (70%)	95 (67%)	82 (66%)
Number of months acquainted with roommate					
Mean	14.43	19.80	28.11	28.59	18.72
SD	21.59	32.27	56.57	71.08	26.59
Minimum	1	6	1	3	4
Maximum	96	156	340	720	120
Roommate acquaintanceship ratings[a]					
Mean	5.82	7.37	7.24	7.12	6.94
SD	1.00	1.03	1.28	1.39	1.12
Minimum	3	4	3	3	3
Maximum	7	9	9	9	9
Ethnicity (*N*)					
Caucasian	81	*NA*	*NA*	120	*NA*
Asian	*NA*	*NA*	*NA*	11	*NA*
Black	*NA*	*NA*	*NA*	10	*NA*
Other	9	*NA*	*NA*	*NA*	*NA*

Table A.1 (Continued)

Variable	1981 (*N* = 90)	1993 (*N* = 94)	1994 (*N* = 92)	1997 (*N* = 141)	2004 (*N* = 124)
Year of study					
1	*NA*	86 (91%)	83 (90%)	126 (89%)	122 (98%)
2	*NA*	8 (9%)	7 (8%)	13 (9%)	1 (1%)
3	*NA*	0	0	0	1 (1%)
4	*NA*	0	0	1 (1%)	0

Note. *N*s refer to total number of subjects in each sample (i.e., *N* = 90 indicates 45 pairs and 90 total participants).
NA indicates that no data were available.
Two missing cases for 1994 "year of study." One missing case for 1997 "year of study."
[a]Roommate acquaintanceship ratings in 1981 measured 7-point scale. Remaining roommate acquaintanceship ratings on 9-point scale.

APPENDIX B

Personality Questionnaire Descriptive Statistics

Table B.1
Descriptive Statistics for the 2004 Supernumerary Personality Inventory Self-Report Ratings (N = 124; Paunonen, 2004)

Scale	Mean	Standard Deviation	Skewness	Kurtosis
Conventionality	47.53	6.60	.17	.89
Seductiveness	51.61	8.01	-.20	-.24
Manipulativeness	46.21	6.91	.10	-.23
Thriftiness	41.52	8.15	.20	-.51
Humorousness	51.97	8.96	-.36	.53
Integrity	48.51	9.09	-.01	-.55
Femininity	46.70	9.06	-.04	-.74
Religiosity	41.67	15.91	.02	-1.06
Risk Taking	48.89	9.46	-.11	-.51
Egotism	51.63	6.37	-.22	-.26

Note. Self-report ratings on a 5-point scale.

Table B.2

Descriptive Statistics for the 1997 NEO Personality Inventory-Revised Self-Report Ratings (N = 141; Paunonen & Ashton, 2001)

Scale	Mean	Standard Deviation	Skewness	Kurtosis
Neuroticism	145.49	20.60	-.11	.40
Extraversion	173.49	17.46	.05	.07
Openness to Experience	172.33	18.53	.11	-.09
Agreeableness	159.84	19.29	-.36	1.07
Conscientiousness	150.77	17.92	.07	-.40

Note. Self-report ratings on a 5-point scale.

Table B.3

Descriptive Statistics for the 2004 NEO Personality Inventory-Revised Self-Report Ratings (N = 124; Paunonen, 2004)

Scale	Mean	Standard Deviation	Skewness	Kurtosis
Neuroticism	141.41	21.46	.20	.46
Extraversion	172.98	16.87	-.17	1.51
Openness to Experience	169.19	19.23	-.11	.18
Agreeableness	159.85	15.31	-.32	-.01
Conscientiousness	153.71	18.98	.05	-.18

Note. Self-report ratings on a 5-point scale.

Table B.4

Descriptive Statistics for the 1994 Jackson Personality Inventory Self-Report Ratings (N = 88; Paunonen, 1994)

Scale	Mean	Standard Deviation	Skewness	Kurtosis
Anxiety	88.06	15.99	.10	-.32
Breadth of Interest	80.16	17.73	-.05	-.62
Complexity	81.59	12.70	-.14	.56
Conformity	74.94	20.58	-.23	.27
Energy Level	82.99	16.41	.15	1.02
Innovation	87.91	17.78	.26	-.33
Interpersonal Affect	92.56	15.18	-.15	.33
Organization	78.72	14.40	-.32	.19
Responsibility	84.23	14.20	.24	-.17
Risk Taking	80.65	18.32	.10	-.34
Self Esteem	90.31	18.92	-.24	-.45
Social Adroitness	81.08	10.48	.16	.51
Social Participation	88.10	18.02	-.29	-.34
Tolerance	83.83	12.30	-.32	-.15
Value Orthodoxy	71.43	15.72	.65	.96
Infrequency	34.93	9.38	1.09	1.48

Note. Self-report ratings on a 7-point scale.

Table B.5

Descriptive Statistics for the 1997 Jackson Personality Inventory Self-Report Ratings (N = 139; Paunonen & Ashton, 2001)

Scale	Mean	Standard Deviation	Skewness	Kurtosis
Anxiety	12.25	3.80	-.28	-.49
Breadth of Interest	10.40	4.31	.05	-.66
Complexity	10.09	3.45	.14	-.94
Conformity	7.60	4.64	.18	-.79
Energy Level	10.36	3.90	.09	-.45
Innovation	12.91	4.76	-.69	-.29
Interpersonal Affect	13.19	3.90	-.33	-.75
Organization	9.25	4.00	.22	-.22
Responsibility	10.69	3.29	-.24	-.07
Risk Taking	10.87	5.02	-.08	-.82
Self Esteem	13.23	4.52	-.44	-.63
Social Adroitness	10.73	3.35	-.06	-.30
Social Participation	11.65	4.43	-.37	-.32
Tolerance	11.77	3.03	-.22	-.37
Value Orthodoxy	7.99	3.44	.44	-.23
Infrequency	.80	1.30	2.81	11.80

Note. Self-report ratings on true/false scale.

Table B.6

Descriptive Statistics for the 1981 Personality Research Form Self-Report Ratings (N = 90; Paunonen, 1982)

Scale	Mean	Standard Deviation	Skewness	Kurtosis
Abasement	74.67	13.55	-.44	1.13
Achievement	90.99	17.06	-.47	.22
Affiliation	100.76	18.54	-.54	.41
Aggression	76.11	19.15	.14	-.57
Autonomy	75.86	15.33	.82	1.18
Change	93.17	14.77	-.19	-.26
Cognitive Structure	88.08	16.11	-.22	.47
Defendance	74.33	16.47	.62	.89
Dominance	90.04	22.09	-.30	-.37
Endurance	85.78	17.42	-.32	.24
Exhibition	91.12	24.20	-.22	-.38
Harm Avoidance	81.54	23.41	-.29	-.50
Impulsivity	72.31	20.82	.28	.07
Nurturance	98.19	18.02	-.50	.14
Order	82.81	24.75	-.35	-.61
Play	90.82	16.58	-.20	-.12
Sentience	93.94	15.96	-.35	.04
Social Recognition	89.56	16.49	-.41	.55
Succorance	79.22	19.37	-.35	.10
Understanding	81.99	15.76	.24	1.36
Infrequency	26.86	9.70	1.125	1.01
Desirability	100.92	16.22	-.27	-.18

Note. Self-report ratings on a 9-point scale.

Table B.7

Descriptive Statistics for the 1997 Personality Research Form Self-Report Ratings (N = 141; Paunonen & Ashton, 2001)

Scale	Mean	Standard Deviation	Skewness	Kurtosis
Abasement	6.96	3.01	.10	-.59
Achievement	9.37	3.41	-.23	-.45
Affiliation	11.16	2.98	-.91	.75
Aggression	8.37	3.68	-.08	-.68
Autonomy	7.89	3.22	.32	-.50
Change	9.35	3.00	-.30	-.67
Cognitive Structure	7.52	2.70	.36	-.27
Defendance	7.18	3.46	.29	-.43
Dominance	9.76	4.22	-.34	-.87
Endurance	8.80	3.27	-.23	-.31
Exhibition	9.71	4.44	-.43	-.81
Harm Avoidance	7.49	4.50	.07	-1.12
Impulsivity	8.41	3.52	-.11	-.78
Nurturance	10.87	3.12	-.45	-.31
Order	6.58	4.61	.42	-1.04
Play	11.51	2.58	-.75	.35
Sentience	10.23	2.95	-.30	-.77
Social Recognition	8.27	3.36	-.01	-.74
Succorance	8.09	3.64	.03	-.71
Understanding	7.69	2.97	.25	-.30
Infrequency	.62	.89	1.70	2.84
Desirability	10.54	2.56	-.48	.21

Note. Self-report ratings on a true/false scale.

Table B.8

Descriptive Statistics for the 1993 Nonverbal Personality Questionnaire Self-Report Ratings (N = 94; Paunonen, 1998)

Scale	Mean	Standard Deviation	Skewness	Kurtosis
Achievement	35.04	6.40	-.07	-.46
Affiliation	42.35	7.00	-1.77	6.25
Aggression	23.86	9.50	1.12	1.53
Autonomy	32.64	9.63	-.07	-.71
Dominance	32.46	6.65	-.30	.44
Endurance	32.15	6.86	-.21	-.17
Exhibition	32.24	9.54	-.11	-.63
Harm Avoidance	30.68	12.53	.09	-1.12
Impulsivity	32.34	6.05	.05	-.04
Nurturance	41.01	8.28	-1.15	1.84
Order	35.54	9.27	-.01	-.60
Play	37.02	7.85	-.73	1.69
Sentience	39.27	7.52	-.65	1.16
Social Recognition	30.39	7.29	.02	.11
Succorance	34.40	8.96	-.36	-.16
Understanding	34.21	7.87	-.34	-.57
Infrequency	19.10	7.53	1.84	4.53

Note. Self-report ratings on a 7-point scale.

APPENDIX C

Personality Questionnaire Item-Total Correlations by Subscale

Table C.1

Item-Total Correlations by Subscale for Normative Supernumerary Personality Inventory Self-Report Ratings (N = 537; Paunonen, 2002)

Scale	Mean	Standard Deviation	Minimum	Maximum
Conventionality	.40	.06	.23	.52
Seductiveness	.55	.08	.39	.65
Manipulativeness	.45	.11	.30	.65
Thriftiness	.49	.09	.34	.62
Humorousness	.56	.07	.39	.63
Integrity	.52	.09	.37	.67
Femininity	.54	.08	.38	.72
Religiosity	.77	.09	.58	.87
Risk Taking	.49	.09	.32	.61
Egotism	.51	.06	.44	.67

Note. Self-report ratings on a 5-point scale.
N = 537 total participants, N = 15 items per subscale.

Table C.2

Item-Total Correlations by Subscale for the 1997 NEO Personality Inventory-Revised Self-Report Ratings (N = 141; Paunonen & Ashton, 2001)

Scale	Mean	Standard Deviation	Minimum	Maximum
Neuroticism	.58	.11	.29	.83
Extraversion	.56	.09	.34	.75
Openness to Experience	.57	.10	.36	.80
Agreeableness	.59	.12	.31	.79
Conscientiousness	.54	.13	.26	.78

Note. Self-report ratings on a 5-point scale.
N = 141 total participants, *N* = 48 items per subscale.

Table C.3

Item-Total Correlations by Subscale for the 1997 Jackson Personality Inventory Self-Report Ratings (N = 139; Paunonen & Ashton, 2001)

Scale	Mean	Standard Deviation	Minimum	Maximum
Anxiety	.42	.11	.12	.60
Breadth of Interest	.44	.09	.28	.59
Complexity	.38	.14	.15	.58
Conformity	.48	.10	.35	.65
Energy Level	.43	.10	.16	.55
Innovation	.53	.09	.38	.71
Interpersonal Affect	.43	.11	.15	.62
Organization	.41	.08	.27	.60
Responsibility	.36	.11	.21	.56
Risk Taking	.52	.10	.32	.67
Self Esteem	.50	.09	.31	.66
Social Adroitness	.35	.12	.02	.51
Social Participation	.47	.12	.19	.63
Tolerance	.34	.14	-.01	.54
Value Orthodoxy	.37	.13	-.02	.51
Infrequency	.33	.16	-.05	.63

Note. Self-report ratings are on true/false scale.
N = 139 total participants, *N* = 20 items per subscale.

Table C.4

Item-Total Correlations by Subscale for the 1997 Personality Research Form Self-Report Ratings (N = 141; Paunonen & Ashton, 2001)

Scale	Mean	Standard Deviation	Minimum	Maximum
Abasement	.41	.09	.21	.53
Achievement	.44	.09	.26	.55
Affiliation	.45	.10	.22	.60
Aggression	.49	.09	.25	.62
Autonomy	.42	.09	.26	.56
Change	.40	.13	.15	.58
Cognitive Structure	.37	.15	.12	.58
Defendance	.45	.09	.24	.60
Dominance	.56	.12	.30	.71
Endurance	.43	.13	.15	.63
Exhibition	.58	.07	.47	.71
Harm Avoidance	.59	.06	.47	.67
Impulsivity	.45	.12	.19	.64
Nurturance	.44	.09	.32	.60
Order	.60	.12	.32	.75
Play	.39	.10	.22	.58
Sentience	.41	.17	.07	.64
Social Recognition	.44	.15	.11	.61
Succorance	.48	.10	.27	.63
Understanding	.41	.11	.16	.55
Infrequency	.27	.16	-.06	.58
Desirability	.36	.11	.20	.52

Note. Self-report ratings on a true/false scale.
N = 141 total participants, N = 16 items per subscale.

Table C.5

Item-Total Correlations by Subscale for Normative Nonverbal Personality Questionnaire Self-Report Ratings (N = 304; Paunonen et al., 2001; Paunonen et al., 2004)

Scale	Mean	Standard Deviation	Minimum	Maximum
Achievement	.53	.08	.41	.63
Affiliation	.57	.07	.46	.64
Aggression	.64	.08	.52	.78
Autonomy	.62	.12	.43	.74
Dominance	.53	.10	.32	.61
Endurance	.54	.07	.44	.67
Exhibition	.67	.14	.42	.81
Harm Avoidance	.72	.09	.53	.80
Impulsivity	.51	.11	.37	.64
Nurturance	.62	.06	.51	.72
Order	.61	.11	.46	.76
Play	.57	.06	.47	.63
Sentience	.62	.11	.45	.74
Social Recognition	.62	.07	.52	.71
Succorance	.63	.08	.52	.73
Understanding	.57	.04	.53	.65
Infrequency	.59	.07	.51	.69

Note. Self-report ratings on a 7-point scale.
N = 304 total participants, N = 8 items per subscale.

APPENDIX D

Descriptive Statistics of Item Properties by Personality Questionnaire

Table D.1

Descriptive Statistics of Item Properties by Personality Questionnaire

Item Property by Scale	*N*	Minimum	Maximum	Mean	Standard Deviation
SPI					
Accuracy	150	-.09	.64	.27	.14
Item-total correlations	150	.28	.87	.53	.12
SDSV	150	2.68	7.33	5.29	.97
$SDSV^2$	150	7.17	53.71	28.86	10.22
Desirability correlation	150	-.42	.38	-.03	.15
Desirability $correlation^2$	150	.00	.18	.02	.03
Mean	150	1.91	4.12	3.07	.52
$Mean^2$	150	3.66	17.00	9.71	3.20
NEO-PI-R					
Accuracy	240	-.09	.59	.20	.12
Item-total correlations	240	.26	.83	.57	.11
SDSV	240	2.07	8.61	5.47	1.69
$SDSV^2$	240	4.29	74.11	32.78	18.47
Desirability correlation	240	-.47	.41	.00	.19
Desirability $correlation^2$	240	.00	.22	.04	.04
Mean	240	1.76	4.36	3.20	.58
$Mean^2$	240	3.09	18.97	10.56	3.70
JPI					
Accuracy	320	-.21	.40	.13	.12
Item-total correlations	320	-.05	.71	.42	.13
Desirability correlation	320	-.38	.30	-.02	.12
Desirability $correlation^2$	320	.00	.14	.01	.02
Mean	320	1.53	6.50	4.16	.80
$Mean^2$	320	2.33	42.19	17.98	6.48
PRF					
Accuracy	352	-.15	.64	.25	.15
Item-total correlations	352	-.06	.75	.45	.13
SDSV	352	1.39	7.76	5.20	1.25
$SDSV^2$	352	1.93	60.22	28.60	12.87
Desirability correlation	352	-.52	.41	-.02	.15
Desirability $correlation^2$	352	.00	.27	.02	.04

Table D.1 (Continued)

Item Property by Scale	*N*	Minimum	Maximum	Mean	Standard Deviation
PRF					
Mean	352	1.11	8.91	5.19	1.24
Mean^2	352	1.23	79.41	28.49	12.73
NPQ					
Accuracy	136	-.12	.54	.22	.14
Item-total correlations	136	.32	.81	.60	.10
Desirability correlation	136	-.31	.30	.04	.14
Desirability correlation^2	136	.00	.10	.02	.02
Mean	136	1.28	5.93	4.07	1.07
Mean^2	136	1.64	35.21	17.70	8.31

Note. Squared variables were used to test curvilinear associations.
*N*s refer to total number of items.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.

APPENDIX E

Item Property Bivariate Correlations

Table E.1

Supernumerary Personality Inventory Item Property Bivariate Correlations (N=150; Paunonen, 2002, 2004)

Item Property	1	2	3	4
1. Accuracy				
2. Item-total correlations	.47**			
3. SDSV	-.18*	-.13		
4. Desirability correlation	-.20*	-.12	.59**	
5. Mean	-.15	-.11	.83**	.39**

Note. N refers to total number of items.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.
*$p < .01$, **$p < 001$.

Table E.2

NEO Personality Inventory Item Property Bivariate Correlations (N=240; Paunonen, 2004; Paunonen & Ashton, 2001)

Item Property	1	2	3	4
1. Accuracy				
2. Item-total correlations	.29**			
3. SDSV	-.03	-.07		
4. Desirability correlation	-.02	.02	.83**	
5. Mean	.01	-.14*	.70**	.50**

Note. N refers to total number of items.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.
*$p < .01$, **$p < 001$.

Table E.3

Jackson Personality Inventory Item Property Bivariate Correlations (N=320; Paunonen, 1994; Paunonen & Ashton, 2001)

Item Property	1	2	3
1. Accuracy			
2. Item-total correlations	.27**		
3. Desirability correlation	.03	.03	
4. Mean	-.01	-.01	.23**

Note. N refers to total number of items.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.
*p < .01, **p < 001.

Table E.4

Personality Research Form Item Property Bivariate Correlations (N=352; Paunonen, 1982; Paunonen & Ashton, 2001)

Item Property	1	2	3	4
1. Accuracy				
2. Item-total correlations	.35**			
3. SDSV	.11*	-.03		
4. Desirability correlation	.11*	.00	.66**	
5. Mean	.04	-.02	.78**	.49**

Note. N refers to total number of items.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.
*$p < .01$, **$p < 001$.

Table E.5

Nonverbal Personality Inventory Item Property Bivariate Correlations (N=136; Paunonen, 1998; Paunonen et al., 2004)

Item Property	1	2	3
1. Accuracy			
2. Item-total correlations	.31**		
3. Desirability correlation	.16	-.004	
4. Mean	.03	-.16	.56**

Note. N refers to total number of items.
Desirability correlation refers to personality questionnaire item correlation with PRF Desirability scale.
*p < .01, **p < 001.

CURRICULUM VITAE

Name: Rachel A. Plouffe

Post-secondary Education and Degrees:

Western University
London, Ontario, Canada
2014 – 2016 M.Sc.

Queen's University
Kingston, Ontario, Canada
2009 – 2013 B.A.H.

Honours and Awards:

Queen's University
Honours Degree Awarded with Distinction
2013

Queen's University
Dean's Honour List
2011 – 2013

Queen's University
Dean's Honour List
2009 – 2010

Queen's University
Excellence Scholarship ($1,500)
2009

Related Work Experience:

Western University
Graduate Teaching Assistant
2014 – Present

Western University
Graduate Research Assistant
2014 – 2015

Research Contributions

Peer-Reviewed Journal Articles:

Plouffe, R. A., Paunonen, S. V., & Saklofske, D. H. (2016). *Item properties and the validity of personality assessment.* Manuscript submitted for publication.

Plouffe, R. A., Saklofske, D. H., & Smith, M. M. (2016). *The Subclinical Sadism Scale: Preliminary psychometric evidence for a new measure.* Manuscript submitted for publication.

Balakrishnan, A., **Plouffe, R. A.,** & Saklofske, D. H. (2016). *What do sadists value? Is honesty humility an intermediary? Extending findings on the link between values and "dark" personalities.* Manuscript submitted for publication.

Other Refereed Publications:

Plouffe, R. A., Wilson, C. A. & Smith, M. M. (2015). The Dark Triad. In B. Carducci (Ed.), *The Wiley Encyclopedia of Personality and Individual Differences* (Vol. 3). Hoboken, NJ: Wiley-Blackwell. (Invited chapter submitted for publication).

Chen, S. **& Plouffe, R. A.** (2015). Psychopathy. In B. Carducci (Ed.), *The Wiley Encyclopedia of Personality and Individual Differences* (Vol. 3). Hoboken, NJ: Wiley-Blackwell. (Invited chapter submitted for publication).

Conference Presentations:

Plouffe, R. A., & Tremblay, P. F. (submitted). *The effect of income on life satisfaction: Does religiosity play a role?* Poster to be presented at the Society for Personality and Social Psychology conference, San Antonio, TX.

Balakrishnan, A., **Plouffe, R. A.,** & Saklofske, D. H. (2016, July). *Extending findings on the link between values and "dark" personalities: What do sadists value? Are gender and honesty humility intermediaries?* Poster presented at the International Association for Relationship Research conference, Toronto, ON.

Plouffe, R. A., Saklofske, D. H., & Smith, M. M. (2016, June). *Validation of a new subclinical sadism measure.* Poster presented at the annual Canadian Psychological Association conference, Victoria, BC.

Plouffe, R. A., Paunonen, S. V., & Saklofske, D. H. (2016, May). *Item properties and the validity of personality assessment.* Poster presented at the Association for Psychological Science conference, Chicago, IL.

Plouffe, R. A. & Paunonen, S. V. (2015, July). *Personality traits underlying socially desirable responding in men versus women.* Poster presented at the International Society for the Study of Individual Differences conference, London, ON.

Plouffe, R. A., Saklofske, D. H., & Smith, M. M. (2015, July). *The development and validation of a "Dark Tetrad" measure of personality.* Poster presented at the International Society for the Study of Individual Differences conference, London, ON.

Plouffe, R. A. & Fekken, G. C. (2014, August). *Academic dishonesty: Who cares what other people think?* Poster presented at the annual American Psychological Association conference, Washington, D.C.

Plouffe, R. A. & Fekken, G. C. (2014, June). *Effects of sanctions on academic dishonesty*. Poster presented at the annual Canadian Psychological Association conference, Vancouver, BC.

Stead, R., **Plouffe, R. A.**, Kay, A., & Fekken, G. C. (2014, June). *The Dark Triad of personality and social desirability: lying to oneself or lying to other people?* Poster presented at the annual Canadian Psychological Association conference, Vancouver, BC.