REVISITING THE RELATIONSHIP BETWEEN CONSCIENTIOUSNESS AND JOB PERFORMANCE: LINEARITY OR NON-LINEARITY?

Ian S. Little

A Dissertation
Submitted to the Graduate College of Bowling Green State University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2007

Committee:

Michael Zickar, Advisor

Stephen H. Cady
Graduate Faculty Representative

Anne Gordon

Scott Highhouse
ABSTRACT

Michael Zickar, Advisor

Recent research into the possibility of a nonlinear relationship between conscientiousness and job performance has been unable to show that a nonlinear model was more appropriate than a linear model. This finding may be due to a lack of psychometric information at the higher end of conscientiousness in current measures of the construct. The current study attempted to create a new measure with test information focused at the higher end of the scale by combining conscientiousness and obsessive-compulsive personality items, and then used this new measure to revisit the notion of a nonlinear relationship between conscientiousness and performance via hierarchical multiple linear regression. Results indicated that the addition of obsessive-compulsive personality items served primarily to increase test information at the low end of the scale. Results of regression analyses between the original conscientiousness scale, obsessive-compulsive personality scale, the new conscientiousness scale, and leadership performance ratings suggest that the new scale may lack construct validity, and the original scales may be more appropriately applied separately.
This work is dedicated to my wife Dana. Without your patience and support this would not have been possible.
ACKNOWLEDGMENTS

Thank you to Michael Zickar for his help and guidance throughout the dissertation process, Scott Davies for his valuable suggestions regarding data analysis, and Robert and Joyce Hogan for providing the data that made these studies possible.
TABLE OF CONTENTS

INTRODUCTION ........................................................................................................................................ 1

  Previous Research Investigating Non-Linearity ........................................................................ 4

  Precision and Psychological Measurement ............................................................................ 8

  Precision of Personality Measures ..................................................................................... 11

  The Current Study .............................................................................................................. 13

STUDY 1 ........................................................................................................................................... 17

STUDY 1 METHOD ......................................................................................................................... 19

  Participants .......................................................................................................................... 19

  Materials .............................................................................................................................. 19

  Procedure ............................................................................................................................ 20

STUDY 1 RESULTS ....................................................................................................................... 24

  Classical Test Theory Item and Scale Analysis ................................................................. 24

  Item Response Theory Item and Scale Analysis ............................................................... 39

  IRT Model Fit ....................................................................................................................... 52

STUDY 1 DISCUSSION ...................................................................................................................... 66

STUDY 2 ........................................................................................................................................... 68

STUDY 2 METHOD ......................................................................................................................... 69

  Participants .......................................................................................................................... 69

  Materials .............................................................................................................................. 69

  Procedure ............................................................................................................................ 70

STUDY 2 RESULTS ....................................................................................................................... 72
### LIST OF FIGURES/TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Prudence classical test theory item statistics</td>
</tr>
<tr>
<td>2</td>
<td>Diligent classical test theory item statistics</td>
</tr>
<tr>
<td>3</td>
<td>Combined Prudence and Diligent Classical Test Theory Item Analysis</td>
</tr>
<tr>
<td>4</td>
<td>Combined Prudence and Diligent Classical Test Theory Item Analysis</td>
</tr>
<tr>
<td>5</td>
<td>Combined Prudence and Diligent Item Statistics without Bad Items</td>
</tr>
<tr>
<td>6</td>
<td>Prudence item parameters</td>
</tr>
<tr>
<td>7</td>
<td>Diligent Item Parameters</td>
</tr>
<tr>
<td>8</td>
<td>Combined Scale Item Parameters</td>
</tr>
<tr>
<td>9</td>
<td>New Conscientiousness Scale Item Parameters</td>
</tr>
<tr>
<td>10</td>
<td>New Conscientiousness Scale Item Parameters in Hold-out Sample</td>
</tr>
<tr>
<td>11</td>
<td>EFA Factor Loadings</td>
</tr>
<tr>
<td>12</td>
<td>Comparison of Item Parameters Estimated from Original and New Scales</td>
</tr>
<tr>
<td>13</td>
<td>Frequency Table of Chi Square/DF Ratios</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Linear and non-linear regression between Prudence and Diligent scales</td>
</tr>
<tr>
<td>2</td>
<td>Prudence score distribution</td>
</tr>
<tr>
<td>3</td>
<td>Diligent score distribution</td>
</tr>
<tr>
<td>4</td>
<td>Prudence test information function and standard error</td>
</tr>
<tr>
<td>5</td>
<td>Diligent test information function and standard error</td>
</tr>
<tr>
<td>6</td>
<td>Combined scale test information function and standard error</td>
</tr>
<tr>
<td>7</td>
<td>New conscientiousness scale test information function and standard error</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>8</td>
<td>New conscientiousness scale test information function and standard error in hold-out sample</td>
</tr>
<tr>
<td>9</td>
<td>Prudence and Diligent exploratory factor analysis scree plot</td>
</tr>
<tr>
<td>10</td>
<td>New conscientiousness scale exploratory factor analysis scree plot</td>
</tr>
<tr>
<td>11</td>
<td>New conscientiousness scale item difficulty parameter estimates by male and female subgroup</td>
</tr>
<tr>
<td>12</td>
<td>New conscientiousness scale item difficulty parameter estimates by randomly equivalent group</td>
</tr>
<tr>
<td>13</td>
<td>Linear and non-linear regression of Prudence and Leadership Performance Ratings</td>
</tr>
<tr>
<td>14</td>
<td>Linear and non-linear regression of Diligent and Leadership Performance Ratings</td>
</tr>
<tr>
<td>15</td>
<td>Linear and non-linear regression of new conscientiousness scale and Leadership Performance Ratings</td>
</tr>
<tr>
<td>16</td>
<td>Prudence and Diligent interaction plot</td>
</tr>
</tbody>
</table>
INTRODUCTION

The use of personality in organizational research has seen a dramatic increase over the last fifteen years. This increase is due, at least in part, to studies that have identified conscientiousness as a generalizable predictor of job performance (e.g., Barrick & Mount, 1991; Hogan & Holland, 2003; Salgado, 1997). The low meta-analytic correlations between these two constructs (a fully corrected correlation of .18 is reported by Tett, Jackson, & Rothstein, 1991, and Barrick, Mount, & Judge, 2001 report a mean validity coefficient for conscientiousness of .12), however, has led some researchers to hypothesize that extremely high levels of conscientiousness may actually reduce job performance (Murphy, 1996; Robie & Ryan, 1999). Although the idea of a non-linear relationship seems intriguing, empirical support has been mixed. Some studies have shown empirical support for a nonlinear relationship (e.g., Cucina & Vasilopoulos, 2005; Day & Silverman, 1989; Lahuis, Martin, & Avis, 2005), but a comprehensive examination by Robie and Ryan (1999) tested this hypothesis and found no support for the idea. The current two part study will assess whether this failure to find non-linearity is truly representative of the relationship, or if the results in support of linearity found by Robie and Ryan are due to a lack of test information at the upper levels of conscientiousness scales.

Prior to the development of the Big Five taxonomy for the classification of personality traits (McCrae & Costa, 1987), personality had a tenuous reputation as a job performance predictor. In the 1960s, Guion and Gottier’s (1965) influential review all but halted research on the relationship between personality and job performance. This review concluded with three recommendations. The final recommendation stated, “Of greatest importance, it must be concluded that, taken as a whole, there is no generalizable evidence that personality measures
can be recommended as good or practical tools for employee selection. The number of significance tests resulting in acceptable statements of validity is greater than might be expected by pure chance—but not much.” (Guion & Gottier, 1965; p. 159). With the widespread acceptance of the Big Five taxonomy in industrial-organizational psychology came a resurgence of interest in the relationship between personality and job performance. In the 1990s many studies were produced that established personality, and the personality trait conscientiousness specifically, as a valid predictor of job performance (e.g. Barrick & Mount, 1991; Salgado, 1997; Tett, et al., 1991).

Despite the evidence that has been amassed showing that job performance can be predicted with personality, low meta-analytic correlations (such as that reported by Schmidt & Hunter, 1998) could be cause for concern. Many possible explanations for attenuation of the relationship between job performance and personality have been posited. These include effects stemming from the use of broad versus narrow trait measures (e.g., Hogan & Roberts, 1996; Murphy & Dzieweczynski, 2005), faking or response distortion on self-report measures (e.g., Rosse, Stecher, Miller, & Levin, 1998), and non-linear relationships between certain personality traits and the job performance criterion (e.g., Day & Silverman, 1989; Robie & Ryan, 1999). The nature of the relationship between personality and job performance (is the relationship linear or non-linear?) is probably the least understood of the explanations mentioned here.

In order to understand the issue of non-linearity between predictors and criteria, one must first address why a non-linear relationship might be expected. Most examinations of deviations from linearity follow from Murphy’s (1996) assertion that in the case of non-cognitive individual differences, more is not always better. He outlines two likely deviations from linearity: the asymptotic relationship and the inverted-U relationship. An asymptotic relationship would
indicate that having a certain characteristic or trait is beneficial to a point, but beyond that point there is little added value (the psychological equivalent of the law of diminishing returns). An inverted-U type relationship, however, indicates that as standing on a trait or characteristic increases it reaches an ideal point at which performance peaks. Beyond this point having more of the characteristic is actually detrimental to performance. Murphy posited scenarios that could result in an inverted-U relationship for several personality variables. Of interest for this dissertation are his thoughts regarding conscientiousness.

Murphy’s (1996) ideas as to why the correlations between performance and conscientiousness are low center on the notion that an individual who was extremely high in conscientiousness might actually be so bound to doing his or her job appropriately, or by the rules, that this individual would actually display lower performance than those able to circumvent small bureaucratic rules in order to facilitate the completion of a task or project. In other words, the tendency of those high in conscientiousness to be conventional, perfectionist, and rule-bound can become a hindrance in the context of many current jobs that requires flexibility, tight deadlines, and ever changing priorities. All academics, for instance, must strike a balance between quality and productivity in developing a manuscript. As is evident from the number of articles accepted for publication as submitted, there is always room for another idea, theoretical perspective, or even another study. If a person were so perfectionist and conventional that this individual was unable to let go of the manuscript without further revision, he or she would never submit the manuscript to be reviewed. Despite the intuitive appeal of this logic, evidence for the accuracy of this notion is mixed.
Previous Research Investigating Non-Linearity

Day and Silverman (1989) conducted one of the earliest studies suggesting a non-linear relationship between conscientiousness and job performance. The intended purpose of this study was to show the incremental validity of personality variables for predicting job performance beyond cognitive ability when predictors and criteria are appropriately matched, and it did not utilize the Big Five framework. Yet, in noting that the dimension akin to conscientiousness (impulse expression) showed promise as a predictor, the authors cautioned that the relationship was non-linear, and that moderate levels of the trait were most predictive of success. Their determination that the relationship between these two constructs is not linear, however, utilized a small sample (n = 43) and was based only upon the review of a scatter plot. Day and Silverman’s investigation of the scatter plot revealed an inverted-U relationship between conscientiousness and job performance. More quantitatively sophisticated methods for assessing linearity can and have been used in the literature.

LaHuis, et al. (2005) found evidence for a non-linear relationship between conscientiousness and job performance in clerical workers. This study utilized two samples of clerical workers (n = 192 and 203 respectively) and two distinct methods of personality assessment in examining the relationship between the two variables. In the first sample conscientiousness was assessed using a situational judgment test and biodata. In the second sample the conscientiousness sub-factor of the NEO Five-Factor Inventory (Costa & McCrae, 1995) was used to assess personality. Overall supervisor ratings of performance served as the criterion for both studies. Both studies used hierarchical regression to assess the relationship between conscientiousness and job performance. In these regression models, conscientiousness was entered at the first step in the model, a quadratic term used to model an inverted-U
A cubic term used to model an S-shaped relationship was entered at the second step, and a cubic term used to model an S-shaped relationship was entered in the final step. In both samples, the addition of the quadratic term added significant incremental validity ($\Delta R^2 = .02$ for both samples), but the addition of the cubic term did not. An examination of the regression plots reported by LaHuis, et al. shows that a performance decrement did occur in both samples, but the decrement was more pronounced in the first sample. Thus, the relationship between conscientiousness and job performance for clerical workers is best described by an inverted-U shaped non-linear regression model. It is important to note, however, that the more dramatic curvilinear relationship was found using a non-traditional assessment of conscientiousness. When a traditional conscientiousness measure (the NEO-FFM) was used, the relationship was closer to asymptotic. As will be discussed in greater depth later in this dissertation, evidence exists that the NEO lacks reliability at the high end of the scale (Reise & Henson, 2000). These authors suggest that the curvilinear relationship might be caused by faking. Their interpretation suggests that, were faking controlled for, the relationship would be linear. This interpretation is based on the conclusion that measurement error introduced by faking “flattened” the linear relationship. I would suggest that the situational judgment and biodata inventories are less susceptible to faking than the self-report NEO-FFM. Thus, if faking is causing increased measurement error in high conscientiousness scales, the strong curvilinear relationship found in the first sample is more indicative of the true nature of this relationship. If the reliability of high self-reported conscientiousness scores could be improved a similar relationship would likely be found.

Cucina and Vasilopoulos (2005) have shown that the relationship between conscientiousness and academic performance is non-linear. This study explored the validity of traitedness measures using a sample of 262 undergraduate psychology students. One goal of this
paper was to assess the moderating effect of traitedness on the relationship between conscientiousness and academic performance. Conscientiousness, and the other Big-Five personality dimensions, were measured using the 100-item Preliminary International Personality Item Pool (IPIP; Goldberg, 1999). Unfortunately, there is currently no information available about the measurement precision of the IPIP measure. In preparation for this analysis the authors hypothesized that the relationship between conscientiousness and academic performance, as measured by GPA, could be described by an inverted-U relationship. The authors tested this hypothesis using a hierarchical regression with a quadratic term entered at the second step similar to the procedure used by LaHuis, et al. (2005). Similar to the results of LaHuis and colleagues, the introduction of the quadratic term improved the fit of the model ($\Delta R^2 = .02$). The nature of this relationship is similar to that reported by Day and Silverman in that moderate levels of conscientiousness were ideal for success. Not all evidence, however, has been so supportive.

Robie and Ryan (1999) produced the most thorough study to date testing for nonlinearity in the relationship between conscientiousness and job performance. In their study, five diverse samples (including managers, sales representatives, truck drivers, Federal employees, and various private sector employees) with sizes ranging from 146 to 999, were used to predict job performance from conscientiousness utilizing both concurrent and predictive validation approaches. The people in these samples completed either the Revised NEO Personality Inventory (Costa & McCrae, 1992) or the Personal Characteristics Inventory (Mount & Barrick, 1995). Supervisory ratings of overall job performance were used as the criterion for this study. Across the five samples, the addition of quadratic and cubic terms to the regression equation predicting job performance did not significantly improve the fit of the regression line. This
would suggest that a generalizable relationship between these two constructs is best described as linear.

Why would a well-conducted examination of non-linearity show the more parsimonious linear model to be favorable when other studies have found support for curvilinearity? Larger samples, such as those used in the Robie and Ryan (1999) studies, should contain a larger number of people ranking highly on the conscientiousness continuum. This lack of range restriction should better allow for an investigation of non-linear relationships, especially when only extremely high levels of a trait are hypothesized to result in dysfunction. This is exactly the relationship described by Murphy (1996).

One difficulty in examining the conscientiousness – job performance relationship described by Murphy (1996) is measurement precision. Measurement precision might be the explanation for why the Robie and Ryan (1999) study provided different results than the other studies. Specifically, to examine the relationship between extreme levels of conscientiousness and job performance, one must have a measure that is capable of discriminating between those who are high in conscientiousness and those who are extremely high in conscientiousness. The majority of existing personality measures (including the NEO-PI used by Robie and Ryan) are most informative at average levels of conscientiousness. These tests, constructed using Classical Test Theory (CTT; see Lord & Novick, 1968 for a comprehensive review), generally utilize items with difficulties (p values in CTT jargon) that range from .3 to .7 and average .5 (a p value of .5 would indicate that half of those who were administered an item endorsed it; see McIntire & Miller, 2000). This produces a test that is most highly informative for people of average standing on a trait, but is unable to differentiate between those of high standing and those of extremely high standing on a trait (the same pertains for people low on the trait in question).
Thus, without the measurement precision to differentiate between those who are high or extremely high in conscientiousness Robie and Ryan’s regression equations could not model the performance decrement that is hypothesized to occur with extreme conscientiousness. Again, there is some evidence to be discussed later that the NEO-PI, one of the personality measures used by Robie and Ryan, lacks measurement precision at the high end of the scale. This lack of measurement range could be imposing an artificial ceiling on conscientiousness. It is impossible to model the true relationship between conscientiousness and performance if only a portion of the conscientiousness trait continuum is being measured.

*Precision and Psychological Measurement*

Tackling issues of measurement precision with CTT is difficult. This difficulty stems from the fact that CTT statistics are sample dependent (Embretson & Reise, 2000). Consider, for example, the p-value statistics mentioned previously. A p-value is a difficulty estimate for binary items, and is calculated by dividing the number of respondents who endorse the item by the total number of respondents. Thus, if 100 people endorse an item out of 200 who respond to the item the p-value equals .5. This statistic can change dramatically depending on the characteristics of the sample under examination. Newer quantitative methods, however, avoid this problem by estimating parameters that are invariant across samples within a population.

Item response theory (IRT), or modern test theory, was developed from classical test theory in the 1940’s, but failed to gather momentum until Lord and Novick published their seminal text on psychological testing in 1968. In recent years IRT has become a far more integral part of mainstream testing. Much of IRT’s popularity stems from its myriad of practical advantages over CTT such as the ability to create adaptive tests (see Wainer, 2000), distinguish item bias from true difference in cross-cultural measurement (e.g., Kim, Cohen, & Park, 1995),
and, of particular interest for the current study, the ability to create tests with maximum measurement precision at a desired range of a trait (Hulin, Drasgow, & Parsons, 1983).

IRT is a model based approach to measurement that relates the characteristics of people (person parameters) to the probability of item (or response option) endorsement using an item response function (IRF) described by the characteristics of the item (item parameters). The most commonly used set of IRT models are designed for use with binary data. Of these binary models the 1pl (or Rasch model; Rasch, 1960), 2pl (Birnbaum, 1968), and 3pl (Lord, 1980) models have gained the most popularity. The 3pl IRF is described by the following equation:

\[
P_j(\theta) \equiv P(x_j = 1|\theta; a_j, b_j, c_j) = \frac{c_j + (1-c_j)}{1 + \exp\left[-D a_j (\theta - b_j)\right]}
\]

The 3pl model is a generalized version of the 1pl and 2pl models, and the 3 parameters in this model are the \(a\) parameter (discrimination parameter), \(b\) parameter (difficulty or location parameter), and the \(c\) parameter (pseudo-guessing parameter). \(D\) is a scaling constant equal to 1.702. Theta (\(\theta\)) is the person parameter, and denotes an individuals standing on the latent trait being measured. The theta distribution follows the standard normal distribution with a mean of 0 and a standard deviation of 1. The \(a\) parameter describes the slope of the IRF, and a steeper slope indicates an item with greater discrimination. The \(b\) parameter describes the horizontal location of the IRF along the theta scale (x-axis). Positive \(b\) parameters indicate an item of above average difficulty and negative \(b\) parameters are indicative of an item with below average difficulty. The \(c\) parameter sets the lower asymptote for the IRF, and indicates the probability that an item will be endorsed solely by chance. If the \(c\) parameter is constrained to 0, then the above equation describes the 2pl model. If the \(a\) parameter is set to 1 and the \(c\) parameter
constrained to 0, the above equation then describes the 1pl model. Although the IRF varies between these models, all possess a similar monotonically increasing structure.

The IRF for any item can be used to determine the amount of psychometric information the item yields. In the case of the 3pl model, an IRF can be transformed into an item information function (IIF) with the following equation:

$$I(\theta) = \left[ \alpha_i^2 \frac{1 - P_i(\theta)}{P_i(\theta)} \right] \left[ \frac{(P_i(\theta) - c_i)^2}{(1 - c_i)^2} \right]$$

An IIC illustrates the range of the theta scale at which an item contains the most psychometric information. If an item was particularly informative at a theta range from 0 to 1.5 then the item would be most capable of discrimination for individuals that fall within this range of theta on the trait or ability of interest. This item would be highly informative for an individual with a theta of .75, but would provide little information about an individual with a theta of -1. Thus, item information can be used to hone the measurement precision of a test. If a research question or decision requires discrimination at a certain trait or ability level a test can be constructed with maximum information at that point on the theta scale.

Another desirable property of IRT models is that IICs can be combined to obtain a test information function (TIF). A TIF can be calculated using the following equation:

$$TI(\theta) = \sum_{i=1}^{j} I(\theta)$$

A TIF communicates the precision of the entire set of items across the theta scale. Thus, if one is interested in measurement at a certain trait range TIF curves can be examined to assess how well a given measure will accomplish this goal. This is particularly important given that test information is directly related to the standard error of measurement as indicated in the equation below:
The standard error of measurement can be thought of along the same lines as reliability in CTT. A low standard error of measurement suggests a lack of error in the test score in the same way that a large reliability coefficient would. The distinction with IRT, however, is that the standard error of measurement is not static across the entire range of a test whereas reliability is static. This means that one can control and manipulate the measurement precision of a test to suit one’s needs using IRT, but CTT does not allow for such sophisticated test construction.

**Precision of Personality Measures**

The use of IRT to answer questions about personality measures is becoming increasingly commonplace. The vast majority of this research was not intended to assess the precision of these measures, but general trends can be observed regarding the precision of commonly used personality measures based on the information reported in this research. In this section three studies that report item parameters, test parameters, or both from personality measures designed for a variety of populations will be reviewed.

Reise and Henson (2000) examined the potential benefits of adaptive administration of the NEO PI-R (Costa & McCrae, 1992, 1994). The NEO PI-R is a Big-Five personality measure intended for normal (as opposed to clinical) populations composed of 30 facet scales. Item and test parameters were estimated for all facets using data from 1059 undergraduates, but item parameters and test information were only reported for the anxiety facet of the neuroticism scale. The test information curve indicates that discrimination for this facet begins to drop off severely for those with a theta score of two or higher. This indicates that, as expected, the scale is not informative, or reliable, for people who are particularly high on this trait.
Zickar and Robie (1999), in a study of the effects of faking on personality items, report polytomous item parameters for the ABLE. The ABLE is a personality measure designed to predict success in army occupations. Items from three of the eleven content scales were examined in this study. The military recruits who participated in this study were divided into faking and honest conditions, but regardless of condition no item threshold parameter across the three scales reached positive two, and only three within any condition had a threshold parameter that exceeded one. Many threshold parameters, however, exceeded negative two, and some even exceeded negative three. This provides further support for the notion that it is difficult to write items that discriminate at the extreme positive end of a personality trait.

Finally, Waller, Thompson, and Wenk (2000) used IRT for an investigation of measurement bias with the MMPI (Hathaway & McKinley, 1940), a measure of personality designed for clinical populations. Using data from 1,788 youths, test characteristic curves were estimated for each of the 12 MMPI factor scales using the 2-pl model. A visual inspection of the test characteristic curves shows that only 4 of the 12 scales have appreciable discrimination for the theta range of two or higher. Taken together with the results of the two studies reviewed previously, it appears that most personality scales do in fact lack discrimination and score reliability at the high positive ranges of personality.

This lack of reliability is problematic for both linear and curvilinear regression. A lack of reliability results in a direct increase in random error. Obviously, increased random error attenuates any observable relationship between the latent variables under examination. In the case of a curvilinear relationship between conscientiousness and job performance, the performance decrement is hypothesized to occur at the point where most conscientiousness scales tend to have unacceptably large amounts of error. This may account for the mixed
findings regarding the nature of this relationship reviewed earlier. The probability of finding a significant curvilinear relationship is influenced by the reliability of the conscientiousness measure employed and the strength of the relationship in the sample under examination. Increasing the reliability of conscientiousness measures used in this line of research should help to provide more consistent findings.

The Current Study

The purpose of the current study was two fold. First, a psychometric investigation attempted to create a measure of conscientiousness capable of reliably differentiating between people who are high or extremely high in conscientiousness. Once the new measure was developed, the relationship between conscientiousness and job performance was reassessed using this measure and a large managerial sample. The first step in this process was a re-examination of personality theory in general and the nature of conscientiousness in particular in order to devise a measure that would assess the high register of the conscientiousness continuum.

Personality research in industrial-organizational psychology is typically limited to the Big-Five personality traits (cf. McCrae & Costa, 1987; John, 1990). These five dimensions of personality were uncovered via the factor analysis of self-report and peer-ratings (Goldberg, 1981) using both trait adjectives and standardized questionnaires (McCrae & Costa, 1985). The five factors in the model are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism or emotional stability. The factor of interest to the current study was conscientiousness.

In their review of conscientiousness, Hogan and Ones (1997) noted that conscientiousness was first empirically identified by Allport and Odbert (1936), and that all comprehensive lexical studies of personality have identified a conscientiousness dimension.
They defined conscientiousness as a propensity for “conformity and socially prescribed impulse control” (p. 849). The study of conscientiousness in applied psychology gained popularity after Goldberg’s (1990) factor-analytic work identified conscientiousness as a component of the “Big Five” structure of personality. Research has shown this framework to be consistent across assessments (e.g., Costa & McCrae, 1992), raters (e.g., McCrae & Costa, 1987), and cultures (e.g., Borkenau & Ostendorf, 1989).

A conscientious person is typically careful and thorough in their behavior (McCrae & Costa, 1987). Other labels that have been used to describe this dimension include superego strength (Cattell, Eber, & Tatsuoka, 1970) and will to achieve (Digman & Takemoto-Chock, 1981). The will to achieve label indicates that people high in conscientiousness always strive to do the best possible job in completing tasks. The lower pole of this dimension has been labeled undirectedness (McCrae & Costa, 1987). This label recognizes that people high in conscientiousness are highly concerned with sticking to plans, requirements, and schedules. McCrae and Costa have also used the phrase *habitually careful* to describe people high in conscientiousness. From these descriptors one can plainly see that highly conscientious people are concerned with doing things the right way, to the best of their ability, every time.

Many well established measures have subscales assessing conscientiousness (e.g., the Hogan Personality Inventory; HPI; R. Hogan & Hogan, 1995; the NEO Five-Factor Inventory; Costa & McCrae, 1995). The dilemma that must be overcome to obtain proper measurement precision for examining non-linearity in the relationship between conscientiousness and job performance lies in developing items that will only be endorsed by those who are extremely conscientious. One method for accomplishing this is to identify a different trait construct that
theoretically mirrors the high end of the conscientiousness continuum. A construct that shows promise in this respect is obsessive-compulsive personality.

A review of the diagnostic criteria for obsessive-compulsive personality disorder showed a clear conceptual link between this personality disorder and conscientiousness. Pfohl and Blum (1991) review the prototypic traits for obsessive-compulsive personality in making recommendations for the Diagnostic and Statistical Manual of Mental Disorders IV (DSM-IV). A list of these traits includes perfectionism, preoccupation with details, indecisiveness, and over-conscientiousness. Pfohl and Blum also trace the historical roots of the personality disorder back to Freud’s (1908/1959) anal character. In this classic paper Freud notes that people with this character perform small or insignificant duties with great conscientiousness. It is clear that some conceptual link exists between these two constructs, and that extremely high conscientiousness is actually part of the diagnostic criteria for obsessive-compulsive personality disorder.

In addition to a conceptual link between the two constructs, a similar link exists between obsessive-compulsive personality and the occupational maladies theorized to exist with individuals extremely high in conscientiousness. Pfohl and Blum’s (1991) review of the prototypic traits of the personality disorder highlights this connection as well. In their article they list indecisiveness from the DSM-III-R, argumentativeness from Freud, obstinacy from Lazare, and fastidiousness masking disarray, tendency to postpone action, and unproductive perseverance from Abraham as manifestations of obsessive-compulsive personality. Clearly an argumentative, indecisive, harried employee who postpones action to persevere on a trivial task is highly unlikely to be productive at work. In reviewing the vocational implications of obsessive-compulsive disorder (OCD, a more serious manifestation of the personality disorder; Garamoni & Schwartz, 1986) Miranda and Rollins (1997) conclude that OCD can result in the
loss of numerous jobs throughout a person’s working life. It is important to note that debate still rages in the clinical literature regarding the relation between obsessive-compulsive disorder and obsessive-compulsive personality disorder (Wu, Clark, & Watson, 2006). In their article Wu, et al. note that the comorbidity of these disorders is relatively low (around 18%). OCD tends to be associated with general negative affectivity, a low self-image, and a lack of conscientiousness whereas obsessive-compulsive personality disorder is characterized, and in fact defined by perfectionism and hyper-conscientiousness. Based on this conceptual and empirical evidence I expected that items from conscientiousness and obsessive-compulsive personality disorder measures could be combined to create a single scale with the measurement precision necessary to predict the performance decrement theoretically suggested at the extreme high end of the conscientiousness trait continuum.
STUDY 1

This first study developed a new measure of conscientiousness designed to have greater discrimination and reliability at the upper-levels of the conscientiousness continuum. This was accomplished by melding existing measures of conscientiousness (HPI Prudence) and obsessive-compulsive personality disorder (HDS Diligent) through the use of classical test theory (CTT) item analysis and item response theory (IRT; see Embretson & Reise, 2000 or Hambleton, Swaminathan, & Rogers, 1991 for an in-depth introduction to this measurement theory). The measure of conscientiousness used for this study was drawn from the HPI (R. Hogan & Hogan, 1995). This measure has a long history of successful use in personnel selection as well as a wealth of psychometric and validity evidence supporting the use of this instrument for selection purposes (e.g., Hogan & Holland, 2003).

Given that the Americans with Disabilities Act of 1990 prohibits selection on the basis of any mental disability, it was most appropriate to find a measure of normal personality built to mirror the non-clinical aspects of obsessive-compulsive personality disorder to combine with the conscientiousness scale of the HPI. The Hogan Development Survey (HDS; R. Hogan & Hogan, 1997) was created to assess the personality types associated with axis II personality disorders of the DSM IV in non-clinical populations. This measure is commonly employed in the selection and development of managers and executive in order screen for behaviors that might derail an otherwise successful manager (Hogan & Hogan, 2001). The diligent scale of the HDS was designed to assess non-clinical manifestations of obsessive-compulsive personality. In addition to their long track record of successful selection use, these measures were used for their identical response scales. Both scales use a true/false response scale to assess personality. If the two measures utilized different response scales (e.g., one using a likert scale and the other a
dichotomous scale), response categories would need to be collapsed in order to model both measures simultaneously. The true/false response format means that each item contains maximum information over a fairly specific range of the trait continuum.
STUDY 1 METHOD

Participants

This study utilized an archival sample of 16,540 working adults from various organizations throughout the United States. The majority (58%) of the people in this sample completed the measures as part of a management development and executive coaching program. The remainder of the people in this sample completed the assessments as part of a pre-employment selection battery. The positions these people were applying for represent several different occupational categories ranging from management to protective services. The majority of the sample was male (70%) and White (85%), although race/ethnicity was only reported by 49% of the sample. Other groups represented include Black (5%), Hispanic (4%), Asian/Pacific Islander (4%), and Native American (1%). The sample ranged in age from 18 to 74, with a mean age of 40 years. Age was reported by 52% of the sample.

Materials

The individuals included in this dataset have completed both the Hogan Personality Inventory (HPI; R. Hogan & Hogan, 1995) and the Hogan Development Survey (HDS; R. Hogan & Hogan, 1997). The HPI consists of seven scales, but only the prudence scale, the scale intended to measure conscientiousness, was used in this study. The HPI technical manual reports internal consistency reliability of .78 and test-retest reliability of .74 for the Prudence scale. The manual also reports positive correlations between peer ratings of conscientiousness and HPI Prudence scores. These correlations range from .54 to .79. The HPI Prudence scale is also correlated with the Minnesota Multiphasic Personality Inventory (MMPI) PROFILE scale perfectionism (r = .29). Given that the HPI is a commercially available test it is not possible to publish the Prudence items, but the HPI manual does provide sample items. Sample Prudence
items are as follows: I strive for perfection in everything I do; I do my job as well as I possibly can.

The HDS consists of 11 scales but only the dutiful scale, the scale relating to the obsessive-compulsive personality disorder from axis II of the DSM IV, was used in this study. The HDS technical manual reports internal consistency reliability of .65 and test-retest reliability of .77 for the Diligent scale. The manual also reports positive correlations between the Diligent scale and coaches ratings of the following trait adjectives: Perfectionistic, overly conscientious, resistant to change (r = .30, .42, & .15 respectively). The HDS manual also reports a correlation between the Diligent scale and the MMPI personality disorder scale Compulsion of r = .14. This is not a large correlation, but given that the HDS is intended for non-clinical populations the magnitude of the relation is understandable. No sample items are available for the HDS.

Procedure

In order to construct a new measure of conscientiousness with a large amount of test information, a small standard error of measurement, and high reliability at high levels of the trait, prudence (conscientiousness) items from the HPI and dutiful (obsessive-compulsive personality) items from the HDS were subjected to a series of analyses beginning with a CTT based item analysis. Given that test development requires an iterative process of item removal based on a series of consecutive analyses it was important to cross-validate findings to ensure that the results did not stem from capitalization on sample specific characteristics. To allow for cross-validation, the total sample was randomly divided in half. One half of the sample was used for test development, and the second half was used to cross-validate findings.

CTT item statistics were used as an initial screening to select items for more in-depth IRT analysis. Specifically, p-values (difficulty estimates) and corrected item-total correlations were
used to reduce the number of items analyzed to a more manageable set. Items with high p-values were not of particular interest for this study. These items tend to be endorsed by the majority of respondents, and thus provide little information at the high end of the conscientiousness continuum. Although some high p-value items were necessary for a thorough and construct valid scale, most were removed from further analysis. An excessive amount of high p-value items coupled with the large number of low p-value items necessary to achieve the desired measurement precision could have caused multidimensionality in the scale. To avoid this, these high p-value items were removed on the basis of corrected item-total correlations. A low corrected item-total correlation is an indication that the item in question does not correspond to the construct being measured. Eliminating items with low item total correlations helped ensure a unidimensional scale. Once this process was complete the newly created scale was expected to have an internal consistency estimate, as measured by coefficient alpha, that exceeds .70.

When the CTT analyses were completed the surviving items were subjected to an IRT analysis. The previous analytic steps primarily served to create a unidimensional scale composed of both conscientiousness and obsessive-compulsive personality items. Although steps were taken to increase test information, decrease the standard error of measurement, and increase reliability at the extreme positive end of the scale, at this point little was known about actual test information, standard error, or reliability. In order to confirm that the items chosen to compose the new broad-range measure of conscientiousness did in fact load on the scale and provide psychometric information at the extreme positive end of the scale these items were submitted to an IRT analysis.

Past research with the HPI suggests that the 3pl model was a good fit for this test (Wadlington, Davies, & Phillips, 2006). As such the 3pl model was used to estimate item
parameters, single scale factor loadings, and IRFs for the items chosen for the new scale. A single scale factor loading was estimated based on the IRT \( a \) parameter using the following equation:

\[
\text{Factor loading} = \frac{a}{\sqrt{1 + a^2}}
\]

These parameters were used to construct IIFs and TIFs. Both IIFs and TIFs, along with the standard error of measurement for the test, were examined to ensure that the test possessed the measurement precision necessary to test for nonlinearity in the relationship between conscientiousness and job performance. In particular, I focused on the range of test scores for which the test provides a sufficiently large amount of information and small standard error to be considered reliable. Green, Bock, Humphreys, Linn, and Reckase (1984) suggest that marginal reliability (defined in the following equation) is an IRT based summary of test precision.

\[
\bar{\rho} = 1 - \overline{SE^2} [\theta]
\]

This statistic, however, is not accurate when test information varies across the range of theta (Thissen, 2003). In addition, a summary reliability statistic would not provide the information necessary to determine if the new scale is capable of reliably measuring the extreme high end of conscientiousness. The equation for marginal reliability was rewritten as a point estimate of reliability:

\[
\rho(\theta) = 1 - SE^2 [\theta]
\]

This point estimate give the reliability of a score at a specific point on the theta scale. A transformation of this equation was used to calculate the standard error of measurement associated with a specific level of reliability:

\[
SE(\theta) = \sqrt{1 - \rho(\theta)}
\]
As previously mentioned, this study adopted the .70 rule of thumb for acceptable test reliability. Using the equation above, one can determine that a score with reliability of .70 is associated with a standard error of .55. Thus, a score was considered reliable if the associated standard error was less than or equal to .55.
STUDY 1 RESULTS

Classical Test Theory Item and Scale Analysis

In order to provide a baseline view of the psychometric functioning of the HPI Prudence (conscientiousness) scale and the HDS Diligent (obsessive-compulsive personality) scale, descriptive statistics, internal consistency reliability, item endorsement rates, and corrected item-total correlations were reported for these scales in Tables 1 and 2 respectively. The Prudence scale showed an acceptable level of internal consistency within this sample ($\alpha = .71$), but the Diligent scale did not ($\alpha = .58$). A closer examination of the corrected item-total correlations revealed that, on average, the items functioned similarly (average corrected item-total correlation is .22 for Prudence and .23 for Diligent). This suggested that the differences in internal consistency were due to scale length rather than item quality. A few items on both scales, however, were functioning poorly both in terms of low corrected item-total correlations and excessively high item endorsement rates. Prudence and Diligent were correlated .34 ($p < .01$; corrected for unreliability in both measures this correlation increases to .53) indicating that these scales share significant variance, but certainly would not be considered isomorphic. The relationship between these two scales was also examined using hierarchical regression with power polynomials in order to check for the existence of a curvilinear relationship. Results of this regression analysis show significant linear ($F(493) = 64.16, p < .01; R^2 = .12$) and nonlinear relationship ($F(492) = 36.66, p < .01; R^2 = .13$). The regression lines for the linear and nonlinear relationships are presented in Figure 1. The nonlinear regression line suggests that as Prudence score increase Diligent scores increase to a point, and then as Prudence scores increase beyond that point Diligent scores level off and even decrease slightly. This curvilinear relationship suggests that people who score high on the Prudence scale tend not to score high on the Diligent
Table 1.

Prudence classical test theory item statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.36</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>0.62</td>
<td>0.28</td>
</tr>
<tr>
<td>3</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>0.71</td>
<td>0.20</td>
</tr>
<tr>
<td>5</td>
<td>0.87</td>
<td>0.20</td>
</tr>
<tr>
<td>6</td>
<td>0.69</td>
<td>0.23</td>
</tr>
<tr>
<td>7</td>
<td>0.69</td>
<td>0.19</td>
</tr>
<tr>
<td>8</td>
<td>0.82</td>
<td>0.23</td>
</tr>
<tr>
<td>9</td>
<td>0.98</td>
<td>0.16</td>
</tr>
<tr>
<td>10</td>
<td>0.34</td>
<td>0.25</td>
</tr>
<tr>
<td>11</td>
<td>0.97</td>
<td>0.13</td>
</tr>
<tr>
<td>12</td>
<td>0.78</td>
<td>0.25</td>
</tr>
<tr>
<td>13</td>
<td>0.82</td>
<td>0.22</td>
</tr>
<tr>
<td>14</td>
<td>0.67</td>
<td>0.26</td>
</tr>
<tr>
<td>15</td>
<td>0.59</td>
<td>0.17</td>
</tr>
<tr>
<td>16</td>
<td>0.76</td>
<td>0.12</td>
</tr>
<tr>
<td>17</td>
<td>0.83</td>
<td>0.08</td>
</tr>
<tr>
<td>18</td>
<td>0.92</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Table 1 (continued).

**Prudence classical test theory item statistics**

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>0.92</td>
<td>0.20</td>
</tr>
<tr>
<td>20</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>21</td>
<td>0.56</td>
<td>0.15</td>
</tr>
<tr>
<td>22</td>
<td>0.81</td>
<td>0.30</td>
</tr>
<tr>
<td>23</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>24</td>
<td>0.63</td>
<td>0.32</td>
</tr>
<tr>
<td>25</td>
<td>0.63</td>
<td>0.30</td>
</tr>
<tr>
<td>26</td>
<td>0.46</td>
<td>0.42</td>
</tr>
<tr>
<td>27</td>
<td>0.93</td>
<td>0.25</td>
</tr>
<tr>
<td>28</td>
<td>0.87</td>
<td>0.28</td>
</tr>
<tr>
<td>29</td>
<td>0.73</td>
<td>0.17</td>
</tr>
<tr>
<td>30</td>
<td>0.79</td>
<td>0.25</td>
</tr>
<tr>
<td>31</td>
<td>0.81</td>
<td>0.13</td>
</tr>
<tr>
<td>Mean</td>
<td>21.50</td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>4.10</td>
<td></td>
</tr>
<tr>
<td>Coef. α</td>
<td>0.71</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Diligent classical test theory item statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.99</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.98</td>
<td>0.15</td>
</tr>
<tr>
<td>3</td>
<td>0.61</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>0.09</td>
</tr>
<tr>
<td>5</td>
<td>0.74</td>
<td>0.41</td>
</tr>
<tr>
<td>6</td>
<td>0.95</td>
<td>0.31</td>
</tr>
<tr>
<td>7</td>
<td>0.85</td>
<td>0.29</td>
</tr>
<tr>
<td>8</td>
<td>0.37</td>
<td>0.21</td>
</tr>
<tr>
<td>9</td>
<td>0.83</td>
<td>0.28</td>
</tr>
<tr>
<td>10</td>
<td>0.24</td>
<td>0.03</td>
</tr>
<tr>
<td>11</td>
<td>0.66</td>
<td>0.37</td>
</tr>
<tr>
<td>12</td>
<td>0.84</td>
<td>0.27</td>
</tr>
<tr>
<td>13</td>
<td>0.75</td>
<td>0.35</td>
</tr>
<tr>
<td>14</td>
<td>0.53</td>
<td>0.24</td>
</tr>
<tr>
<td>Mean</td>
<td>9.60</td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>2.17</td>
<td></td>
</tr>
<tr>
<td>Coef. α</td>
<td>0.58</td>
<td></td>
</tr>
</tbody>
</table>
scale, although the relatively small increase in explained variance ($\Delta R^2 = .01$) suggests that this is a weak effect.

*Figure 1.* Linear and non-linear regression between Prudence and Diligent scales.

Figures 2 and 3 present the distribution of scores for these two scales. For the purposes of this research, it would have been ideal if the score distributions for the Prudence (conscientiousness) scale were skewed left, and the distributions for the Diligent (obsessive-compulsive personality) scale skewed right. If this were the case, combining the two distributions would likely result in a relatively normal score distribution that covers a large range
of scores. Unfortunately, as one can see by examining Figures 2 and 3, the distributions for both scales were skewed slightly left.

Figure 2. Prudence score distribution.

The ultimate goal of this study was to produce a single conscientiousness scale by combining the Prudence and Diligent scales. Once the baseline psychometric functioning of the original scales had been established, an analysis of the combined scales was conducted. Descriptive statistics, internal consistency reliability, item endorsement rates, and corrected item-total correlations are reported for the combined scale in Table 3. At this point, items were removed from further analysis on the basis of corrected item-total correlations. Item-total correlations can be used to judge to construct relevance of the item. Items that correlate highly with the rest of the scale are more likely to be tapping the same construct than items that do not
correlate with the rest of the scale. In order to allow all items that may prove beneficial to be included in the more in-depth IRT analyses a relatively lenient exclusion criterion was established. Any item with a corrected item total less than .10 was excluded from further analysis. Large item total correlations tend to relate to large \( a \) parameters, and large \( a \) result in increased item information. The relationship between item total correlations and \( a \) parameters is not perfect, and this exclusion criterion was intended to remove items that were highly unlikely to prove informative while still allowing as many items as possible to be analyzed with the more powerful IRT methodology. This criterion eliminated three HPI items (items 16, 17, and 31) and three HDS items (items 3, 4, and 10) from further analysis.

*Figure 3.* Diligent score distribution.
Table 3.

**Combined Prudence and Diligent Classical Test Theory Item Analysis**

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 1</td>
<td>0.36</td>
<td>0.23</td>
</tr>
<tr>
<td>Prudence 2</td>
<td>0.62</td>
<td>0.29</td>
</tr>
<tr>
<td>Prudence 3</td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td>Prudence 4</td>
<td>0.71</td>
<td>0.19</td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.87</td>
<td>0.24</td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.69</td>
<td>0.26</td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.69</td>
<td>0.34</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.82</td>
<td>0.30</td>
</tr>
<tr>
<td>Prudence 9</td>
<td>0.98</td>
<td>0.20</td>
</tr>
<tr>
<td>Prudence 10</td>
<td>0.34</td>
<td>0.23</td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.97</td>
<td>0.16</td>
</tr>
<tr>
<td>Prudence 12</td>
<td>0.78</td>
<td>0.27</td>
</tr>
<tr>
<td>Prudence 13</td>
<td>0.82</td>
<td>0.19</td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.67</td>
<td>0.21</td>
</tr>
<tr>
<td>Prudence 15</td>
<td>0.59</td>
<td>0.14</td>
</tr>
<tr>
<td>Prudence 16</td>
<td>0.76</td>
<td>0.08</td>
</tr>
<tr>
<td>Prudence 17</td>
<td>0.83</td>
<td>0.05</td>
</tr>
<tr>
<td>Prudence 18</td>
<td>0.92</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Table 3 (continued).

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 19</td>
<td>0.92</td>
<td>0.22</td>
</tr>
<tr>
<td>Prudence 20</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>Prudence 21</td>
<td>0.56</td>
<td>0.17</td>
</tr>
<tr>
<td>Prudence 22</td>
<td>0.81</td>
<td>0.28</td>
</tr>
<tr>
<td>Prudence 23</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>Prudence 24</td>
<td>0.63</td>
<td>0.31</td>
</tr>
<tr>
<td>Prudence 25</td>
<td>0.63</td>
<td>0.29</td>
</tr>
<tr>
<td>Prudence 26</td>
<td>0.46</td>
<td>0.40</td>
</tr>
<tr>
<td>Prudence 27</td>
<td>0.93</td>
<td>0.22</td>
</tr>
<tr>
<td>Prudence 28</td>
<td>0.87</td>
<td>0.25</td>
</tr>
<tr>
<td>Prudence 29</td>
<td>0.73</td>
<td>0.10</td>
</tr>
<tr>
<td>Prudence 30</td>
<td>0.79</td>
<td>0.23</td>
</tr>
<tr>
<td>Prudence 31</td>
<td>0.81</td>
<td>0.07</td>
</tr>
<tr>
<td>Diligent 1</td>
<td>0.99</td>
<td>0.10</td>
</tr>
<tr>
<td>Diligent 2</td>
<td>0.98</td>
<td>0.16</td>
</tr>
<tr>
<td>Diligent 3</td>
<td>0.61</td>
<td>0.07</td>
</tr>
<tr>
<td>Diligent 4</td>
<td>0.25</td>
<td>-0.05</td>
</tr>
<tr>
<td>Diligent 5</td>
<td>0.74</td>
<td>0.33</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>0.95</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Table 3 (continued).

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diligent 7</td>
<td>0.85</td>
<td>0.30</td>
</tr>
<tr>
<td>Diligent 8</td>
<td>0.37</td>
<td>0.13</td>
</tr>
<tr>
<td>Diligent 9</td>
<td>0.83</td>
<td>0.26</td>
</tr>
<tr>
<td>Diligent 10</td>
<td>0.24</td>
<td>-0.02</td>
</tr>
<tr>
<td>Diligent 11</td>
<td>0.66</td>
<td>0.40</td>
</tr>
<tr>
<td>Diligent 12</td>
<td>0.84</td>
<td>0.32</td>
</tr>
<tr>
<td>Diligent 13</td>
<td>0.75</td>
<td>0.35</td>
</tr>
<tr>
<td>Diligent 14</td>
<td>0.53</td>
<td>0.15</td>
</tr>
<tr>
<td>Mean</td>
<td>31.00</td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>5.19</td>
<td></td>
</tr>
<tr>
<td>Coef. α</td>
<td>0.74</td>
<td></td>
</tr>
</tbody>
</table>

Descriptive statistics, internal consistency reliability, item endorsement rates, and corrected item-total correlations for the remaining items are reported in Table 4. Removal of the six poorly functioning items resulted in increased internal consistency, but also reduced the item total correlations of some of the surviving items below the .10 threshold. A total of three items (Prudence items 15 and 29 along with Diligent item 1) showed item-total correlations below the acceptable level, and were therefore removed from further analysis. At this point the new conscientiousness measure contained 36 items (26 Prudence items and 10 Diligent items).
Descriptive statistics, internal consistency reliability, item endorsement rates, and corrected item-total correlations for these items are reported in Table 5.

Table 4.

Combined Prudence and Diligent Classical Test Theory Item Analysis

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 1</td>
<td>0.36</td>
<td>0.24</td>
</tr>
<tr>
<td>Prudence 2</td>
<td>0.62</td>
<td>0.31</td>
</tr>
<tr>
<td>Prudence 3</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>Prudence 4</td>
<td>0.71</td>
<td>0.20</td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.87</td>
<td>0.24</td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.69</td>
<td>0.26</td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.69</td>
<td>0.34</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.82</td>
<td>0.31</td>
</tr>
<tr>
<td>Prudence 9</td>
<td>0.98</td>
<td>0.20</td>
</tr>
<tr>
<td>Prudence 10</td>
<td>0.34</td>
<td>0.26</td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.97</td>
<td>0.17</td>
</tr>
<tr>
<td>Prudence 12</td>
<td>0.78</td>
<td>0.29</td>
</tr>
<tr>
<td>Prudence 13</td>
<td>0.82</td>
<td>0.20</td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.67</td>
<td>0.21</td>
</tr>
<tr>
<td>Prudence 15</td>
<td>0.59</td>
<td>0.09</td>
</tr>
<tr>
<td>Prudence 18</td>
<td>0.92</td>
<td>0.14</td>
</tr>
<tr>
<td>Prudence 19</td>
<td>0.92</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Table 4 (continued).

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 20</td>
<td>0.31</td>
<td>0.26</td>
</tr>
<tr>
<td>Prudence 21</td>
<td>0.56</td>
<td>0.16</td>
</tr>
<tr>
<td>Prudence 22</td>
<td>0.81</td>
<td>0.29</td>
</tr>
<tr>
<td>Prudence 23</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>Prudence 24</td>
<td>0.63</td>
<td>0.32</td>
</tr>
<tr>
<td>Prudence 25</td>
<td>0.63</td>
<td>0.30</td>
</tr>
<tr>
<td>Prudence 26</td>
<td>0.46</td>
<td>0.42</td>
</tr>
<tr>
<td>Prudence 27</td>
<td>0.93</td>
<td>0.22</td>
</tr>
<tr>
<td>Prudence 28</td>
<td>0.87</td>
<td>0.26</td>
</tr>
<tr>
<td>Prudence 29</td>
<td>0.73</td>
<td>0.09</td>
</tr>
<tr>
<td>Prudence 30</td>
<td>0.79</td>
<td>0.23</td>
</tr>
<tr>
<td>Diligent 1</td>
<td>0.99</td>
<td>0.09</td>
</tr>
<tr>
<td>Diligent 2</td>
<td>0.98</td>
<td>0.16</td>
</tr>
<tr>
<td>Diligent 5</td>
<td>0.74</td>
<td>0.32</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>0.95</td>
<td>0.27</td>
</tr>
<tr>
<td>Diligent 7</td>
<td>0.85</td>
<td>0.32</td>
</tr>
<tr>
<td>Diligent 8</td>
<td>0.37</td>
<td>0.11</td>
</tr>
<tr>
<td>Diligent 9</td>
<td>0.83</td>
<td>0.27</td>
</tr>
<tr>
<td>Diligent 11</td>
<td>0.66</td>
<td>0.43</td>
</tr>
</tbody>
</table>
A review of the endorsement rates for the items in Table 5 was disappointing with respect to the goal of this study. Item endorsement rates ranged from .23 to .98 with a mean of .70. This suggested that few, if any items were endorsed only by individuals extremely high in conscientiousness. At this point items with high endorsement rates (e.g., a p-value greater than .90) could have been removed, but given the relatively low overall internal consistency and the relatively large item-total correlations of some high endorsement rate items, all 36 items were submitted to further analysis using IRT.

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diligent 12</td>
<td>0.84</td>
<td>0.35</td>
</tr>
<tr>
<td>Diligent 13</td>
<td>0.75</td>
<td>0.36</td>
</tr>
<tr>
<td>Diligent 14</td>
<td>0.53</td>
<td>0.14</td>
</tr>
<tr>
<td>Mean</td>
<td>27.50</td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>5.01</td>
<td></td>
</tr>
<tr>
<td>Coef. α</td>
<td>0.77</td>
<td></td>
</tr>
</tbody>
</table>
Table 5.

Combined Prudence and Diligent Item Statistics without Bad Items

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 1</td>
<td>0.36</td>
<td>0.24</td>
</tr>
<tr>
<td>Prudence 2</td>
<td>0.62</td>
<td>0.32</td>
</tr>
<tr>
<td>Prudence 3</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>Prudence 4</td>
<td>0.71</td>
<td>0.21</td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.87</td>
<td>0.25</td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.69</td>
<td>0.27</td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.69</td>
<td>0.35</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.82</td>
<td>0.32</td>
</tr>
<tr>
<td>Prudence 9</td>
<td>0.98</td>
<td>0.20</td>
</tr>
<tr>
<td>Prudence 10</td>
<td>0.34</td>
<td>0.27</td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.97</td>
<td>0.18</td>
</tr>
<tr>
<td>Prudence 12</td>
<td>0.78</td>
<td>0.30</td>
</tr>
<tr>
<td>Prudence 13</td>
<td>0.82</td>
<td>0.19</td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.67</td>
<td>0.21</td>
</tr>
<tr>
<td>Prudence 18</td>
<td>0.92</td>
<td>0.13</td>
</tr>
<tr>
<td>Prudence 19</td>
<td>0.92</td>
<td>0.22</td>
</tr>
<tr>
<td>Prudence 20</td>
<td>0.31</td>
<td>0.26</td>
</tr>
<tr>
<td>Prudence 21</td>
<td>0.56</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Table 5 (continued).

Combined Prudence and Diligent Item Statistics without Bad Items

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 22</td>
<td>0.81</td>
<td>0.28</td>
</tr>
<tr>
<td>Prudence 23</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>Prudence 24</td>
<td>0.63</td>
<td>0.31</td>
</tr>
<tr>
<td>Prudence 25</td>
<td>0.63</td>
<td>0.29</td>
</tr>
<tr>
<td>Prudence 26</td>
<td>0.46</td>
<td>0.42</td>
</tr>
<tr>
<td>Prudence 27</td>
<td>0.93</td>
<td>0.21</td>
</tr>
<tr>
<td>Prudence 28</td>
<td>0.87</td>
<td>0.25</td>
</tr>
<tr>
<td>Prudence 30</td>
<td>0.79</td>
<td>0.22</td>
</tr>
<tr>
<td>Diligent 2</td>
<td>0.98</td>
<td>0.16</td>
</tr>
<tr>
<td>Diligent 5</td>
<td>0.74</td>
<td>0.33</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>0.95</td>
<td>0.27</td>
</tr>
<tr>
<td>Diligent 7</td>
<td>0.85</td>
<td>0.33</td>
</tr>
<tr>
<td>Diligent 8</td>
<td>0.37</td>
<td>0.13</td>
</tr>
<tr>
<td>Diligent 9</td>
<td>0.83</td>
<td>0.27</td>
</tr>
<tr>
<td>Diligent 11</td>
<td>0.66</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Table 5 (continued).

Combined Prudence and Diligent Item Statistics without Bad Items

<table>
<thead>
<tr>
<th>Item</th>
<th>P-Value</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diligent 12</td>
<td>0.84</td>
<td>0.35</td>
</tr>
<tr>
<td>Diligent 13</td>
<td>0.75</td>
<td>0.36</td>
</tr>
<tr>
<td>Diligent 14</td>
<td>0.53</td>
<td>0.15</td>
</tr>
<tr>
<td>Mean</td>
<td>25.20</td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>4.88</td>
<td></td>
</tr>
<tr>
<td>Coef. α</td>
<td>0.77</td>
<td></td>
</tr>
</tbody>
</table>

*Item Response Theory Item and Scale Analysis*

As with the Classical item and scale analysis section, I will begin by presenting an IRT based item and scale analysis of the original Prudence and Diligent scales. Item parameters for the Prudence scale are reported in Table 6, and the test information function (TIF) and standard error curve are presented in Figure 4. The $a$ parameters for the Prudence items ranged from .14 to 1.25 suggesting, similar to the CTT analyses, that some of the items were not accurately tapping conscientiousness. A review of the TIF in Figure 4 indicated that, like most personality tests, the prudence scale had relatively poor reliability at the extreme positive end of the theta scale.
Table 6.

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 1</td>
<td>1.25</td>
<td>1.18</td>
<td>0.22</td>
</tr>
<tr>
<td>Prudence 2</td>
<td>1.17</td>
<td>0.54</td>
<td>0.41</td>
</tr>
<tr>
<td>Prudence 3</td>
<td>1.15</td>
<td>1.55</td>
<td>0.12</td>
</tr>
<tr>
<td>Prudence 4</td>
<td>0.58</td>
<td>0.40</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.63</td>
<td>-1.17</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.48</td>
<td>-0.53</td>
<td>0.24</td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.53</td>
<td>0.55</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.71</td>
<td>-0.58</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 9</td>
<td>0.64</td>
<td>-2.45</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 10</td>
<td>1.05</td>
<td>1.25</td>
<td>0.19</td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.52</td>
<td>-2.84</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 12</td>
<td>0.81</td>
<td>-0.26</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 13</td>
<td>0.56</td>
<td>-0.71</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.86</td>
<td>0.59</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 15</td>
<td>0.28</td>
<td>1.04</td>
<td>0.34</td>
</tr>
<tr>
<td>Prudence 16</td>
<td>0.17</td>
<td>-0.27</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 17</td>
<td>0.14</td>
<td>-2.93</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 18</td>
<td>0.38</td>
<td>-3.10</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Table 6 (continued).

<table>
<thead>
<tr>
<th>Prudence item parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
</tr>
<tr>
<td>Prudence 19</td>
</tr>
<tr>
<td>Prudence 20</td>
</tr>
<tr>
<td>Prudence 21</td>
</tr>
<tr>
<td>Prudence 22</td>
</tr>
<tr>
<td>Prudence 23</td>
</tr>
<tr>
<td>Prudence 24</td>
</tr>
<tr>
<td>Prudence 25</td>
</tr>
<tr>
<td>Prudence 26</td>
</tr>
<tr>
<td>Prudence 27</td>
</tr>
<tr>
<td>Prudence 28</td>
</tr>
<tr>
<td>Prudence 29</td>
</tr>
<tr>
<td>Prudence 30</td>
</tr>
<tr>
<td>Prudence 31</td>
</tr>
</tbody>
</table>
Item parameters for the Diligent scale are reported in Table 7, and the TIF and standard error curve are presented in Figure 5. Diligent item 10 had to be removed from analysis in order for the model to converge, indicating that this item was not functioning properly. The $a$ parameters for the Prudence items ranged from .59 to 1.38 suggesting that the items were, in general, adequately assessing the normal range of obsessive-compulsive personality. The Diligent scale, much like the Prudence scale, showed a marked deficit in test information at the positive end of the theta scale. This finding was discouraging, given that the intent of merging this scale with the Prudence scale was to increase test information at the positive end of the theta scale.
Table 7.

Diligent Item Parameters

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diligent 1</td>
<td>0.84</td>
<td>-3.42</td>
<td>0.48</td>
</tr>
<tr>
<td>Diligent 2</td>
<td>0.92</td>
<td>-2.90</td>
<td>0.42</td>
</tr>
<tr>
<td>Diligent 3</td>
<td>0.59</td>
<td>1.38</td>
<td>0.50</td>
</tr>
<tr>
<td>Diligent 4</td>
<td>1.05</td>
<td>2.75</td>
<td>0.23</td>
</tr>
<tr>
<td>Diligent 5</td>
<td>0.81</td>
<td>-0.80</td>
<td>0.14</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>1.24</td>
<td>-1.94</td>
<td>0.23</td>
</tr>
<tr>
<td>Diligent 7</td>
<td>0.81</td>
<td>-1.39</td>
<td>0.23</td>
</tr>
<tr>
<td>Diligent 8</td>
<td>0.73</td>
<td>1.75</td>
<td>0.25</td>
</tr>
<tr>
<td>Diligent 9</td>
<td>0.68</td>
<td>-1.35</td>
<td>0.24</td>
</tr>
<tr>
<td>Diligent 11</td>
<td>1.38</td>
<td>-0.33</td>
<td>0.13</td>
</tr>
<tr>
<td>Diligent 12</td>
<td>1.24</td>
<td>-0.91</td>
<td>0.32</td>
</tr>
<tr>
<td>Diligent 13</td>
<td>1.17</td>
<td>-0.51</td>
<td>0.27</td>
</tr>
<tr>
<td>Diligent 14</td>
<td>0.65</td>
<td>0.38</td>
<td>0.19</td>
</tr>
</tbody>
</table>
The 36 Prudence and Diligent items that survived the CTT analyses were submitted to an IRT analysis. The item parameters and single scale factor loadings for this new scale are reported in Table 8, and the TIF and standard error are reported in Figure 6. Two of the 36 items had factor loadings that fell below the .30 convention, and were removed from the scale. The item parameters and single scale factor loadings for the final item set of the new conscientiousness scale are reported in Table 9, and the TIF and standard error are reported in Figure 7. A comparison of the TIF for the Prudence scale reported in Figure 4 and the TIF for the new conscientiousness scale reported in Figure 7 showed that the new scale was capable of reliable measurement over a considerably wider range of scores than the Prudence scale, but that the combination of Prudence and Diligent failed to increase test precision at the positive end of the theta scale. Counter intuitively, the addition of Diligent items increased test precision toward the negative end of the theta scale more than it increased precision at the positive end.
Table 8.

Combined Scale Item Parameters

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 1</td>
<td>1.44</td>
<td>1.38</td>
<td>0.27</td>
<td>0.82</td>
</tr>
<tr>
<td>Prudence 2</td>
<td>0.92</td>
<td>0.30</td>
<td>0.34</td>
<td>0.68</td>
</tr>
<tr>
<td>Prudence 3</td>
<td>0.98</td>
<td>1.66</td>
<td>0.12</td>
<td>0.70</td>
</tr>
<tr>
<td>Prudence 4</td>
<td>0.60</td>
<td>0.40</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.84</td>
<td>-0.96</td>
<td>0.50</td>
<td>0.64</td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.52</td>
<td>-0.51</td>
<td>0.24</td>
<td>0.46</td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.78</td>
<td>-0.58</td>
<td>0.15</td>
<td>0.61</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.80</td>
<td>-0.98</td>
<td>0.31</td>
<td>0.63</td>
</tr>
<tr>
<td>Prudence 9</td>
<td>1.47</td>
<td>-2.22</td>
<td>0.50</td>
<td>0.83</td>
</tr>
<tr>
<td>Prudence 10</td>
<td>0.94</td>
<td>1.28</td>
<td>0.18</td>
<td>0.68</td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.75</td>
<td>-2.18</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>Prudence 12</td>
<td>0.94</td>
<td>-0.26</td>
<td>0.49</td>
<td>0.68</td>
</tr>
<tr>
<td>Prudence 13</td>
<td>0.56</td>
<td>-0.69</td>
<td>0.50</td>
<td>0.49</td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.61</td>
<td>0.76</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>Prudence 18</td>
<td>0.32</td>
<td>-3.31</td>
<td>0.49</td>
<td>0.31</td>
</tr>
<tr>
<td>Prudence 19</td>
<td>0.61</td>
<td>-1.99</td>
<td>0.49</td>
<td>0.52</td>
</tr>
<tr>
<td>Prudence 20</td>
<td>0.58</td>
<td>1.50</td>
<td>0.11</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 21</td>
<td>0.33</td>
<td>1.38</td>
<td>0.35</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Table 8 (continued).

**Combined Scale Item Parameters**

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 22</td>
<td>0.66</td>
<td>-0.72</td>
<td>0.44</td>
<td>0.55</td>
</tr>
<tr>
<td>Prudence 23</td>
<td>0.53</td>
<td>1.24</td>
<td>0.13</td>
<td>0.47</td>
</tr>
<tr>
<td>Prudence 24</td>
<td>0.51</td>
<td>-0.19</td>
<td>0.20</td>
<td>0.45</td>
</tr>
<tr>
<td>Prudence 25</td>
<td>0.72</td>
<td>0.32</td>
<td>0.36</td>
<td>0.59</td>
</tr>
<tr>
<td>Prudence 26</td>
<td>1.13</td>
<td>0.54</td>
<td>0.17</td>
<td>0.75</td>
</tr>
<tr>
<td>Prudence 27</td>
<td>0.65</td>
<td>-1.87</td>
<td>0.50</td>
<td>0.54</td>
</tr>
<tr>
<td>Prudence 28</td>
<td>0.66</td>
<td>-1.12</td>
<td>0.50</td>
<td>0.55</td>
</tr>
<tr>
<td>Prudence 30</td>
<td>0.46</td>
<td>-0.73</td>
<td>0.45</td>
<td>0.41</td>
</tr>
<tr>
<td>Diligent 2</td>
<td>0.80</td>
<td>-3.20</td>
<td>0.41</td>
<td>0.62</td>
</tr>
<tr>
<td>Diligent 5</td>
<td>0.71</td>
<td>-0.88</td>
<td>0.13</td>
<td>0.58</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>1.03</td>
<td>-2.06</td>
<td>0.28</td>
<td>0.72</td>
</tr>
<tr>
<td>Diligent 7</td>
<td>0.81</td>
<td>-1.41</td>
<td>0.22</td>
<td>0.63</td>
</tr>
<tr>
<td>Diligent 8</td>
<td>0.27</td>
<td>2.56</td>
<td>0.16</td>
<td>0.26</td>
</tr>
<tr>
<td>Diligent 9</td>
<td>0.58</td>
<td>-1.53</td>
<td>0.23</td>
<td>0.50</td>
</tr>
<tr>
<td>Diligent 11</td>
<td>0.92</td>
<td>-0.49</td>
<td>0.08</td>
<td>0.68</td>
</tr>
<tr>
<td>Diligent 12</td>
<td>0.93</td>
<td>-1.19</td>
<td>0.22</td>
<td>0.68</td>
</tr>
<tr>
<td>Diligent 13</td>
<td>0.69</td>
<td>-0.93</td>
<td>0.14</td>
<td>0.57</td>
</tr>
<tr>
<td>Diligent 14</td>
<td>0.27</td>
<td>0.51</td>
<td>0.16</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Figure 6. Combined scale test information function and standard error.

Table 9.

New Conscientiousness Scale Item Parameters

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 1</td>
<td>1.41</td>
<td>1.35</td>
<td>0.26</td>
<td>0.82</td>
</tr>
<tr>
<td>Prudence 2</td>
<td>0.91</td>
<td>0.27</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>Prudence 3</td>
<td>0.97</td>
<td>1.65</td>
<td>0.11</td>
<td>0.70</td>
</tr>
<tr>
<td>Prudence 4</td>
<td>0.64</td>
<td>0.38</td>
<td>0.50</td>
<td>0.54</td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.87</td>
<td>-0.94</td>
<td>0.50</td>
<td>0.66</td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.52</td>
<td>-0.53</td>
<td>0.23</td>
<td>0.46</td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.76</td>
<td>-0.58</td>
<td>0.15</td>
<td>0.60</td>
</tr>
<tr>
<td>Item</td>
<td>$a$</td>
<td>$b$</td>
<td>$c$</td>
<td>Factor Loading</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>----------------</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.80</td>
<td>-1.00</td>
<td>0.30</td>
<td>0.62</td>
</tr>
<tr>
<td>Prudence 9</td>
<td>1.40</td>
<td>-2.21</td>
<td>0.50</td>
<td>0.81</td>
</tr>
<tr>
<td>Prudence 10</td>
<td>0.92</td>
<td>1.26</td>
<td>0.18</td>
<td>0.68</td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.76</td>
<td>-2.17</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>Prudence 12</td>
<td>0.95</td>
<td>-0.28</td>
<td>0.48</td>
<td>0.69</td>
</tr>
<tr>
<td>Prudence 13</td>
<td>0.61</td>
<td>-0.65</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.66</td>
<td>0.71</td>
<td>0.50</td>
<td>0.55</td>
</tr>
<tr>
<td>Prudence 18</td>
<td>0.33</td>
<td>-3.32</td>
<td>0.48</td>
<td>0.31</td>
</tr>
<tr>
<td>Prudence 19</td>
<td>0.61</td>
<td>-2.01</td>
<td>0.48</td>
<td>0.52</td>
</tr>
<tr>
<td>Prudence 20</td>
<td>0.58</td>
<td>1.49</td>
<td>0.11</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 21</td>
<td>0.32</td>
<td>1.42</td>
<td>0.35</td>
<td>0.31</td>
</tr>
<tr>
<td>Prudence 22</td>
<td>0.66</td>
<td>-0.78</td>
<td>0.42</td>
<td>0.55</td>
</tr>
<tr>
<td>Prudence 23</td>
<td>0.52</td>
<td>1.24</td>
<td>0.13</td>
<td>0.46</td>
</tr>
<tr>
<td>Prudence 24</td>
<td>0.51</td>
<td>-0.21</td>
<td>0.20</td>
<td>0.45</td>
</tr>
<tr>
<td>Prudence 25</td>
<td>0.71</td>
<td>0.28</td>
<td>0.35</td>
<td>0.58</td>
</tr>
<tr>
<td>Prudence 26</td>
<td>1.12</td>
<td>0.52</td>
<td>0.17</td>
<td>0.75</td>
</tr>
<tr>
<td>Prudence 27</td>
<td>0.71</td>
<td>-1.75</td>
<td>0.50</td>
<td>0.58</td>
</tr>
<tr>
<td>Prudence 28</td>
<td>0.69</td>
<td>-1.09</td>
<td>0.50</td>
<td>0.57</td>
</tr>
<tr>
<td>Prudence 30</td>
<td>0.45</td>
<td>-0.79</td>
<td>0.44</td>
<td>0.41</td>
</tr>
</tbody>
</table>
Table 9 (continued).

New Conscientiousness Scale Item Parameters

<table>
<thead>
<tr>
<th>Item</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diligent 2</td>
<td>0.79</td>
<td>-3.20</td>
<td>0.41</td>
<td>0.62</td>
</tr>
<tr>
<td>Diligent 5</td>
<td>0.68</td>
<td>-0.90</td>
<td>0.14</td>
<td>0.56</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>1.01</td>
<td>-2.08</td>
<td>0.28</td>
<td>0.71</td>
</tr>
<tr>
<td>Diligent 7</td>
<td>0.80</td>
<td>-1.42</td>
<td>0.22</td>
<td>0.63</td>
</tr>
<tr>
<td>Diligent 9</td>
<td>0.57</td>
<td>-1.55</td>
<td>0.23</td>
<td>0.50</td>
</tr>
<tr>
<td>Diligent 11</td>
<td>0.89</td>
<td>-0.50</td>
<td>0.08</td>
<td>0.66</td>
</tr>
<tr>
<td>Diligent 12</td>
<td>0.93</td>
<td>-1.19</td>
<td>0.21</td>
<td>0.68</td>
</tr>
<tr>
<td>Diligent 13</td>
<td>0.66</td>
<td>-0.96</td>
<td>0.14</td>
<td>0.55</td>
</tr>
</tbody>
</table>

*Figure 7. New conscientiousness scale test information function and standard error.*
In order to cross-validate these results, the new conscientiousness scale was re-analyzed using the hold-out sample. The item parameters and single scale factor loadings for the final item set of the new conscientiousness scale are reported in Table 10, and the TIF and standard error are reported in Figure 8. These results corresponded closely to those obtained using the exploratory sample. This suggests that parameter invariance held across these randomly equivalent groups, and that the psychometric properties of the new conscientiousness scale are generalizable.

Table 10.

New Conscientiousness Scale Item Parameters in Hold-out Sample

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 1</td>
<td>1.25</td>
<td>1.30</td>
<td>0.24</td>
<td>0.78</td>
</tr>
<tr>
<td>Prudence 2</td>
<td>0.99</td>
<td>0.34</td>
<td>0.37</td>
<td>0.70</td>
</tr>
<tr>
<td>Prudence 3</td>
<td>0.95</td>
<td>1.73</td>
<td>0.12</td>
<td>0.69</td>
</tr>
<tr>
<td>Prudence 4</td>
<td>0.68</td>
<td>0.25</td>
<td>0.50</td>
<td>0.56</td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.85</td>
<td>-1.00</td>
<td>0.50</td>
<td>0.65</td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.52</td>
<td>-0.51</td>
<td>0.23</td>
<td>0.46</td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.85</td>
<td>-0.50</td>
<td>0.18</td>
<td>0.65</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.90</td>
<td>-0.87</td>
<td>0.36</td>
<td>0.67</td>
</tr>
<tr>
<td>Prudence 9</td>
<td>1.04</td>
<td>-1.86</td>
<td>0.50</td>
<td>0.72</td>
</tr>
<tr>
<td>Prudence 10</td>
<td>1.04</td>
<td>1.23</td>
<td>0.19</td>
<td>0.72</td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.81</td>
<td>-2.11</td>
<td>0.50</td>
<td>0.63</td>
</tr>
<tr>
<td>Prudence 12</td>
<td>1.19</td>
<td>-0.24</td>
<td>0.50</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Table 10 (continued).

New Conscientiousness Scale Item Parameters in Hold-out Sample

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 13</td>
<td>0.64</td>
<td>-0.59</td>
<td>0.50</td>
<td>0.54</td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.67</td>
<td>0.71</td>
<td>0.50</td>
<td>0.56</td>
</tr>
<tr>
<td>Prudence 18</td>
<td>0.36</td>
<td>-3.40</td>
<td>0.43</td>
<td>0.34</td>
</tr>
<tr>
<td>Prudence 19</td>
<td>0.63</td>
<td>-2.27</td>
<td>0.36</td>
<td>0.53</td>
</tr>
<tr>
<td>Prudence 20</td>
<td>0.57</td>
<td>1.62</td>
<td>0.13</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 21</td>
<td>0.26</td>
<td>1.10</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>Prudence 22</td>
<td>0.57</td>
<td>-1.17</td>
<td>0.30</td>
<td>0.50</td>
</tr>
<tr>
<td>Prudence 23</td>
<td>0.46</td>
<td>1.44</td>
<td>0.12</td>
<td>0.42</td>
</tr>
<tr>
<td>Prudence 24</td>
<td>0.49</td>
<td>-0.22</td>
<td>0.21</td>
<td>0.44</td>
</tr>
<tr>
<td>Prudence 25</td>
<td>0.56</td>
<td>-0.10</td>
<td>0.25</td>
<td>0.49</td>
</tr>
<tr>
<td>Prudence 26</td>
<td>1.05</td>
<td>0.50</td>
<td>0.17</td>
<td>0.73</td>
</tr>
<tr>
<td>Prudence 27</td>
<td>0.67</td>
<td>-1.84</td>
<td>0.50</td>
<td>0.56</td>
</tr>
<tr>
<td>Prudence 28</td>
<td>0.68</td>
<td>-1.15</td>
<td>0.50</td>
<td>0.56</td>
</tr>
<tr>
<td>Prudence 30</td>
<td>0.40</td>
<td>-1.10</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>Diligent 2</td>
<td>0.67</td>
<td>-3.74</td>
<td>0.46</td>
<td>0.55</td>
</tr>
<tr>
<td>Diligent 5</td>
<td>0.72</td>
<td>-0.84</td>
<td>0.13</td>
<td>0.58</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>1.17</td>
<td>-1.82</td>
<td>0.33</td>
<td>0.76</td>
</tr>
<tr>
<td>Diligent 7</td>
<td>0.84</td>
<td>-1.35</td>
<td>0.21</td>
<td>0.65</td>
</tr>
<tr>
<td>Diligent 9</td>
<td>0.56</td>
<td>-1.59</td>
<td>0.23</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Table 10 (continued).

New Conscientiousness Scale Item Parameters in Hold-out Sample

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diligent 11</td>
<td>0.88</td>
<td>-0.44</td>
<td>0.08</td>
<td>0.66</td>
</tr>
<tr>
<td>Diligent 12</td>
<td>0.90</td>
<td>-1.19</td>
<td>0.27</td>
<td>0.67</td>
</tr>
<tr>
<td>Diligent 13</td>
<td>0.66</td>
<td>-0.89</td>
<td>0.16</td>
<td>0.55</td>
</tr>
</tbody>
</table>

*Figure 8. New conscientiousness scale test information function and standard error in hold-out sample.*

*IRT Model Fit*

IRT is a model based approach to psychological measurement. As such, it is important to assess model fit when interpreting results. If the model is shown to be a poor fit for the data, results should be interpreted with caution. Hambleton and Swaminathan (1985) suggested three
approaches to assess model-data fit: Demonstrate that the assumptions of the model have been met, demonstrate that the advantages of the model (e.g., parameter invariance) have been realized, and demonstrate fit between model-based predictions and data-based observations. I began my assessment of model fit by demonstrating that the HPI and HDS data met the assumptions of the 3pl model. There are two testable assumptions for the 3pl model: The test administration should be nonspeeded and the data should be unidimensional. The HPI and HDS were decidedly nonspeeded. No time constraints were placed upon test administration. The assumption of unidimensionality, however, had to be assessed empirically, especially given that two distinct scales were combined in this study.

Factor analysis is the most appropriate method for assessing the dimensionality of a set of test items. A traditional item level factor analysis was not possible for this data. As mentioned previously, the HPI and HDS utilize a dichotomous true/false response format. Because Pearson correlations require continuous data the item level correlation matrix had to be calculated using tetrachoric correlations. Tetrachoric correlation matrices are often not positive definite, as was the case here. A not positive definite correlation matrix cannot be factor analyzed. To overcome this limitation the HPI Prudence and HDS Diligent items were bundled to create continuous variables that could be used to compute a Pearson correlation matrix. Egan, Sireci, Swaminathan, and Sweeney (1998) conducted a study looking at the impact of item bundles on the assessment of scale dimensionality. This study examined how bundling procedure, bundle length, and number of bundles affect examinations of dimensionality. This study examined three methods of bundling items: Bundling based on content, item difficulty, and random assignment to bundle. This study found that bundling procedure had essentially no impact on examinations of dimensionality. This study did find, however, that bundle length and number of bundles have
a substantial impact of the assessment of dimensionality. Larger bundles (e.g., bundles consisting of a large number of items) resulted in a smaller number of detected dimensions, and an increased number of bundles resulted in a larger number of detected dimensions. Thus, it can be concluded that a large number of small bundles is most sensitive to multidimensionality. Given that the point of this analysis was to investigate whether the Prudence and Diligent scales meet the assumption of unidimensionality, it was important that the bundling method be sensitive to multidimensionality. In order to create a large number of small bundles, items were combined in groups of two or three. In the case of the 31 item Prudence scale 15 item bundles were created (14 two item bundles and 1 three item bundle). Seven item bundles were created for the 14 item Diligent scale. Care was taken to ensure that the paired items in each bundle were somewhat diverse (e.g., items judged to have highly similar content, such as items belonging to the same HPI homogeneous item composites, were not bundled together). If the items within each bundle were too similar the shared variance could have resulted in spurious factors and artificial cross-loadings, although the findings of Egan, et al. suggest that such effects are minimal.

These 22 item bundles were submitted to a maximum likelihood exploratory factor analysis with oblique rotation using the initial Study 1 data set. The results of this analysis showed six factors with eigenvalues greater than one. A review of the scree plot (Figure 9) revealed a strong first factor. This factor was clearly interpretable as a Prudence factor (see factor loadings in Table 11), and accounted for nearly twice the variance of any other factor (First factor eigenvalue = 3.6, explained variance = 16.3%; Second factor eigenvalue = 1.9, explained variance = 8.7%, Third factor eigenvalue = 1.4, explained variance = 6.3%; Fourth factor eigenvalue = 1.4, explained variance = 6.1%). The other five factors were not easily interpretable with the notable exception of factor five. This factor appeared to be a Diligent
factor. The correlation between the Prudence and Diligent factors was .30. This correlation was nearly identical to the correlation previously reported between these two scales, thus providing further evidence for the factor interpretations. This evidence supports the validity of these scales as measuring distinct construct, but also suggests that the two related constructs could be scaled together. The Diligent factor eigenvalue was only slightly higher than one (eigenvalue = 1.2), and this factor accounted for approximately one third of the variance accounted for by the Prudence factor (explained variance = 5.5). According to Reckase (1979) this is indicative of sufficient unidimensionality for appropriate use in latent trait models.

*Figure 9.* Prudence and Diligent exploratory factor analysis scree plot.
Table 11.

EFA Factor Loadings

<table>
<thead>
<tr>
<th>Item Bundle</th>
<th>Factor 1</th>
<th>Factor 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diligent 1</td>
<td></td>
<td>0.47</td>
</tr>
<tr>
<td>Diligent 2</td>
<td>0.26</td>
<td>0.60</td>
</tr>
<tr>
<td>Diligent 3</td>
<td></td>
<td>0.36</td>
</tr>
<tr>
<td>Diligent 4</td>
<td>0.26</td>
<td>0.42</td>
</tr>
<tr>
<td>Diligent 5</td>
<td></td>
<td>0.30</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>0.21</td>
<td>0.26</td>
</tr>
<tr>
<td>Diligent 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prudence 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prudence 2</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>Prudence 3</td>
<td>0.30</td>
<td>0.20</td>
</tr>
<tr>
<td>Prudence 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.49</td>
<td>0.29</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.46</td>
<td>0.22</td>
</tr>
<tr>
<td>Prudence 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prudence 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>
Table 11 (continued).

**EFA Factor Loadings**

<table>
<thead>
<tr>
<th>Item Bundle</th>
<th>Factor 1</th>
<th>Factor 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence 12</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>Prudence 13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Prudence 15</td>
<td>0.26</td>
<td></td>
</tr>
</tbody>
</table>

*Note: For readability factor loadings less than .20 are not reported in this table.*

In order to gauge how well these items would function scaled together, this factor analysis was repeated with factor extraction constrained to one factor rather than eigenvalues greater than one. Six of the seven Diligent item bundles showed factor loadings of .20 or higher, and four of the bundles loaded on the factor at .30 or higher. Thirteen of the 15 Prudence showed factor loading of .20 or higher, and 10 of the items loaded higher than .30. Although one could not recommend operational use of an item with such factor loadings, these results do suggest that the first factor was strong enough to assume unidimensionality for the purposes of IRT analysis.

These results provided initial support for the appropriateness of these scales for IRT analysis, but it was most important to demonstrate the unidimensionality of the final item set for the new conscientiousness measure. To accomplish this item bundles were created for the final set of Prudence and Diligent items. The new conscientiousness scale contains eight Diligent items that were combined into four item bundles and 26 Prudence items that were combined into 13 item bundles. The 17 item bundles were submitted to a maximum likelihood factor analysis.
with oblique rotation using the Study 1 initial sample. The results of this analysis showed four factors with eigenvalues greater than one. An examination of the scree plot (Figure 10), however, indicated that the data represent only one primary factor. To estimate factor loadings for this factor the analysis was repeated with factor extraction constrained to one factor rather than eigenvalues greater than one. Results of this analysis showed that all of the items loaded on this factor at .25 or higher, and only two item bundles fell below .30 (both bundles were constructed from Prudence items). These results provided strong support for the unidimensionality of the new conscientiousness scale.

*Figure 10. New conscientiousness scale exploratory factor analysis scree plot.*

In order to provide one final check on the dimensionality of the new conscientiousness scale the 17 item bundles were submitted to a maximum likelihood confirmatory factor analysis using the Study 1 confirmatory sample. Two factor models were tested. The first model
specified the four Diligent item bundles as indicators of a latent Diligent factor and the 13 Prudence item bundles as indicators of a latent Prudence factor. The latent Diligent and Prudence factors were allowed to correlate. This model failed to minimize, which is indicative of extremely poor model fit. The second model specified all 17 item bundles as indicators of a single latent factor. This model showed acceptable fit as indicated by an RMSEA of .074 (an RMSEA value between .05 and .08 indicates a reasonable error of approximation; Browne & Cudeck, 1993, p. 144). These results showed that a multi-dimensional factor model was not appropriate for the new conscientiousness scale. Taken as a whole, this evidence provided strong support for the unidimensionality assumption.

To ensure that features of the model were performing as expected I next examined the sample invariance of the parameters. Hambleton and Swaminathan (1985) recommend comparing $b$ parameter estimates between subgroups of the larger test population. These plots can then be compared to baseline plots of item parameters estimated using randomly equivalent groups. The only subgroups with sufficient data for accurate estimation in this sample were men and women. Item parameters were estimated for the new conscientiousness scale using all of the men in the initial Study 1 sample ($n = 3896$), and then again using all of the women in the initial Study 1 sample ($n = 1697$). The $b$ parameters estimated in these two samples were then plotted in Figure 11. The parameter estimates for the male sample were more stable and had a smaller standard error by virtue of the fact that this sample was nearly twice as large as the female sample. As such, the standard errors associated with the parameter estimates derived from the male sample were used to construct a 95% confidence interval around the parameters. Based on these confidence intervals one can see that the majority of parameter estimates from these two samples were not significantly different. This analysis was repeated using the parameter
estimates from the randomly equivalent Study 1 initial and hold-out samples. The b parameters estimated in these two samples are presented in Figure 12. These parameter estimates converge slightly better than those derived from the male and female samples, but overall the two convergence plots were highly similar. Taken together these results suggest that parameter invariance is present in these models.

*Figure 11.* New conscientiousness scale item difficulty parameter estimates by male and female subgroup.
Figure 12. New conscientiousness scale item difficulty parameter estimates by randomly equivalent group.

The parameter estimates for these items are also fairly stable when estimated as part of the original scales and the new conscientiousness scale (the correlation between original scale and new scale $a$ parameters is .63 and $b$ parameters is .96). The parameter estimates for the new conscientiousness items are reported by scale in Table 12. In general, the item parameter estimates varied slightly when estimated as part of two different scales, but this is to be expected given the differences in item content across the scales. This finding suggests that the item parameters are invariant across scales and that items from the two original scales could be appropriately scaled together.
Table 12.

Comparison of Item Parameters Estimated from Original and New Scales

<table>
<thead>
<tr>
<th>Item</th>
<th>Estimated as Part of Original Scale</th>
<th>Estimated as Part of New Conscientiousness Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>Prudence 1</td>
<td>1.25</td>
<td>1.18</td>
</tr>
<tr>
<td>Prudence 2</td>
<td>1.17</td>
<td>0.54</td>
</tr>
<tr>
<td>Prudence 3</td>
<td>1.15</td>
<td>1.55</td>
</tr>
<tr>
<td>Prudence 4</td>
<td>0.58</td>
<td>0.40</td>
</tr>
<tr>
<td>Prudence 5</td>
<td>0.63</td>
<td>-1.17</td>
</tr>
<tr>
<td>Prudence 6</td>
<td>0.48</td>
<td>-0.53</td>
</tr>
<tr>
<td>Prudence 7</td>
<td>0.53</td>
<td>0.55</td>
</tr>
<tr>
<td>Prudence 8</td>
<td>0.71</td>
<td>-0.58</td>
</tr>
<tr>
<td>Prudence 9</td>
<td>0.64</td>
<td>-2.45</td>
</tr>
<tr>
<td>Prudence 10</td>
<td>1.05</td>
<td>1.25</td>
</tr>
<tr>
<td>Prudence 11</td>
<td>0.52</td>
<td>-2.85</td>
</tr>
<tr>
<td>Prudence 12</td>
<td>0.81</td>
<td>-0.26</td>
</tr>
<tr>
<td>Prudence 13</td>
<td>0.56</td>
<td>-0.71</td>
</tr>
<tr>
<td>Prudence 14</td>
<td>0.86</td>
<td>0.59</td>
</tr>
<tr>
<td>Prudence 18</td>
<td>0.39</td>
<td>-3.11</td>
</tr>
<tr>
<td>Prudence 19</td>
<td>0.62</td>
<td>-2.10</td>
</tr>
</tbody>
</table>
Table 12 (continued).

Comparison of Item Parameters Estimated from Original and New Scales

<table>
<thead>
<tr>
<th>Item</th>
<th>Estimated as Part of Original Scale</th>
<th>Estimated as Part of New Conscientiousness Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>Prudence 20</td>
<td>0.56</td>
<td>1.29</td>
</tr>
<tr>
<td>Prudence 21</td>
<td>0.30</td>
<td>0.89</td>
</tr>
<tr>
<td>Prudence 22</td>
<td>0.73</td>
<td>-1.11</td>
</tr>
<tr>
<td>Prudence 23</td>
<td>0.53</td>
<td>0.97</td>
</tr>
<tr>
<td>Prudence 24</td>
<td>0.58</td>
<td>-0.40</td>
</tr>
<tr>
<td>Prudence 25</td>
<td>0.63</td>
<td>-0.28</td>
</tr>
<tr>
<td>Prudence 26</td>
<td>1.01</td>
<td>0.29</td>
</tr>
<tr>
<td>Prudence 27</td>
<td>1.01</td>
<td>-1.76</td>
</tr>
<tr>
<td>Prudence 28</td>
<td>0.89</td>
<td>-1.49</td>
</tr>
<tr>
<td>Prudence 30</td>
<td>0.54</td>
<td>-1.29</td>
</tr>
<tr>
<td>Diligent 2</td>
<td>0.93</td>
<td>-2.90</td>
</tr>
<tr>
<td>Diligent 5</td>
<td>0.81</td>
<td>-0.80</td>
</tr>
<tr>
<td>Diligent 6</td>
<td>1.24</td>
<td>-1.94</td>
</tr>
<tr>
<td>Diligent 7</td>
<td>0.81</td>
<td>-1.39</td>
</tr>
<tr>
<td>Diligent 9</td>
<td>0.68</td>
<td>-1.35</td>
</tr>
<tr>
<td>Diligent 11</td>
<td>1.38</td>
<td>-0.33</td>
</tr>
<tr>
<td>Diligent 12</td>
<td>1.24</td>
<td>-0.91</td>
</tr>
<tr>
<td>Diligent 13</td>
<td>1.17</td>
<td>-0.51</td>
</tr>
</tbody>
</table>
The final step in the model fit process was to test model predictions against actual test results. This comparison should be conducted using both graphical and statistical methods (Drasgow, Levine, Tsien, Williams, & Mead, 1995). Stark’s (2001) MODFIT program was used to conduct both the statistical and graphical comparisons. This analysis compared the item parameters calibrated in the initial sample to actual item responses from the hold-out sample. The program is limited to 3,000 cases of actual item responses. The 3,000 cases used in this analysis were drawn randomly from the hold-out sample. The results of this analysis were encouraging. A visual inspection of the graphical fit plots showed that, in general, the predicted (or theoretical) IRF’s and observed (or empirical) response functions were nearly identical (all item fit plots are presented in Appendix A). This suggests that the model is a good fit for the data. The statistical results generally confirmed this interpretation. Items with Chi-Square/Degrees of Freedom Ratios of three or less are indicative of good fit. The majority of the items (28 of the 34) met this criteria. MODFIT also groups items into “doublets” and “triplets” (bundles of two and three items) to avoid over-estimating model-data fit based on non-uniform response function differences. If differences are present in more than one item the chi-square statistic will increase and highlight the fit problem. Results based on doublets and triplets were less encouraging (see Table 13 for a summary of chi-square results). Previous studies of personality test items have found similar results (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Zickar, Gibby, & Robie, 2004; Zickar & Robie, 1999).
Table 13.

<table>
<thead>
<tr>
<th>Frequency Table of Chi Square/DF Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Singlets</td>
</tr>
<tr>
<td>Doublets</td>
</tr>
<tr>
<td>Triplets</td>
</tr>
</tbody>
</table>

Taken holistically, the results of these investigations indicate that the 3pl model is appropriate for this data. Assumptions about dimensionality and speededness were justifiable, the expected benefits of employing this IRT model were realized, and the IRT model item response predictions were confirmed with empirical response data. Good model-data fit indicates that the results of Study1 could be interpreted with confidence, and it is appropriate to draw inferences based on these results.
STUDY 1 DISCUSSION

The goal of Study 1 was to create a new conscientiousness scale with increased measurement precision at the high end of the trait continuum by combining items from an existing conscientiousness scale and a measure intended to assess the normal range of obsessive-compulsive personality disorder. The results of this study show that it was possible to combine these two existing measures, and that the combination resulted in a conscientiousness measure with an increased range of measurement precision. Unexpectedly, the addition of the obsessive-compulsive personality items served primarily to increase measurement precision at the low end of the trait continuum. The curvilinear relationship found between the original scales also suggests that obsessive-compulsive personality does not constitute the high end of conscientiousness. This finding shows that as conscientiousness increases past moderate levels obsessive-compulsive personality is not more prevalent, and may actually be less prevalent at very high levels of conscientiousness.

These results do not support Murphy’s (1996) proposition that the extreme high end of conscientiousness becomes obsessive-compulsive personality. These findings suggest that those who possess obsessive-compulsive tendencies are actually relatively low or moderate in conscientiousness. It may be the case that the rigidity and preoccupation with details associated with obsessive-compulsive personality, although conceptually similar to the detail orientation and socially prescribed impulse control expressed by conscientious people, results in a lack of reliability and inability to adhere to deadlines thereby making these individuals decidedly unconscientious.

Alternatively, it may be the case that all of the items that differentiate obsessive-compulsive personality and conscientiousness were removed in creating the new
Conscientiousness and Performance

Conscientiousness scale. When the original scales were factor analyzed an obsessive-compulsive personality factor appeared. This factor was not highly correlated with the conscientiousness factor suggesting that the two constructs are distinct but related. After IRT-based scale construction the surviving items no longer showed any obsessive-compulsive personality factor. This suggests that this study may have “extracted” the conscientiousness from the obsessive-compulsiveness scale and combined only those items with the conscientiousness scale. Again, the construct validity of this scaling effort is still in question.

This study did little, however, to answer any questions about the nature of extreme conscientiousness (high or low). If obsessive-compulsive personality is not the high end of conscientiousness, then what is? Is obsessive-compulsive personality really the low end of conscientiousness? Conceptually this makes little sense. Not all people who are low in conscientiousness suffer from obsessive-compulsive personality. The answer may be that two measures assessing conceptually similar constructs can be combined into one reliable measure with little, if any construct validity. Although a thorough construct validation was beyond the scope of this paper, an examination of the predictive validity of this new scale may offer some information about the validity of the new measure. If the new conscientiousness scale related to job performance as was theoretically expected, then this would provide some evidence for the construct validity of the measure. If not, then the validity and usefulness of this new scale remains in question. This effort also allowed for an investigation of curvilinearity in both the original conscientiousness measure and the new measure with increased reliability.
STUDY 2

With the exception of a lack of measurement precision at the extreme positive end of the scale, the attempt to find a curvilinear relationship between conscientiousness and job performance conducted by Robie and Ryan (1999) was methodologically sound. Study 2 replicated the analyses used in the Robie and Ryan study utilizing the new conscientiousness scale created in study 1. It was expected that, when utilizing a measure with appropriate measurement precision, the relationship between conscientiousness and job performance would best be described by a curvilinear regression line indicating a performance decrement at high levels of conscientiousness.
STUDY 2 METHOD

Participants

An archival sample of 509 senior managers working at two multi-national Fortune 100 organizations within the U.S. (286 managers from a logistics organization and 223 from a consumer products organization) was used in this study. This sample was drawn from the validation archives of the HPI and HDS maintained by Hogan Assessment Systems. This sample was predominantly Male (74%) and White (77%). The remainder of the sample was 26% Female, 9% Black, 5% Asian/Pacific Islander, 7% Hispanic, 1% Native American, and 1% Non-US. A total of 35 people did not report sex and 48 people did not report ethnicity. Age information was not available in this database. The two organizations differ slightly in terms of demographics and significantly on the measures of interest. Demographically, the logistics organization is more homogeneous than the consumer products company. The logistics company is 88% Male and 85% White as compared to the consumer products organization that is 59% Male and 68% White. Managers with the logistics organization scored significantly higher on the Prudence, Diligent, and new conscientiousness scales. Given the large sample sizes involved in these mean comparisons it is important to consider the effect size of these differences as well as their statistical significance. The Prudence mean difference was a relatively small effect (D = 0.31), the Diligent mean difference was a medium effect (D = 0.47), and the new conscientiousness scale mean difference was a large effect (D = 1.71).

Materials

All individuals included in this dataset had taken both the HPI and the HDS, and performance data is included for each person. The Prudence and Diligent items identified as markers for a unidimensional broad-range scale of conscientiousness in Study 1 were summed to
compute a conscientiousness score for each individual. This new conscientiousness scale and the
two scales from which it was created were used in this regression analysis. Internal consistency
estimates for the three scales in this sample were as follows: New conscientiousness scale, $\alpha = .75$; Prudence, $\alpha = .77$; Diligent, $\alpha = .58$. The performance criterion used for this study was
subjective ratings of overall leadership performance. These ratings are global performance
ratings of managers or leaders from their respective organizations. These ratings are an
assessment of the target’s managerial skill as opposed to technical skill. As noted in the
participants section, these ratings came from two different organizations. These organizations
utilized different response formats for their ratings, so in order to standardize the scales the
ratings were converted to z-scores prior to aggregation. Meta-analyses suggest that
conscientiousness is predictive of both managerial and leadership performance (e.g., Bono &
Judge, 2004; Hogan & Holland, 2003).

Procedure

To test for the presence of a curvilinear relationship between conscientiousness and job
performance the new broad-range measure of conscientiousness and the job performance
criterion were submitted to a hierarchical multiple linear regression analysis. This hierarchical
multiple linear regression tested the effect of a quadratic regression term above and beyond the
effect of a linear regression term. According to Cohen and Cohen (1975), curvilinear
relationships can be examined within a multiple linear regression framework through the use of
power polynomials. In their discussion of power polynomials Cohen and Cohen note that the
majority of data analyzed in the behavioral sciences can be adequately described by linear,
quadratic, or cubic coding. Based upon the theoretical justification and empirical evidence
discussed previously this investigation was limited to linear and quadratic coding. In linear
coding, the variable is entered into the regression model as collected with no alteration or transformation. In quadratic coding, the variable (in this case conscientiousness), is raised to the second power before being entered into the regression model. In order to test for the existence of a curvilinear relationship between variables Cohen and Cohen recommend a hierarchical regression model in which the linear term is entered at the first step in the model and the quadratic term is entered at the second step. If the quadratic term represents the relationship significantly better the overall model will be statistically significant and the quadratic term will also result in a significant change in variance accounted for (R²). Given the size of the available sample, any concerns regarding statistical power were not warranted. The possibility of finding statistically significant but practically insignificant effects did exist. Thus, in the event of significant findings, effect sizes were reported for both the linear and quadratic models. Changes in effect size across model were compared in order to gage the practical implications of the findings. To ensure that significant findings were attributable to the new broad-range measure of conscientiousness these analyses were first be conducted using the original Prudence (conscientiousness) scale from the HPI and Diligent (obsessive-compulsive personality) scale from the HDS. If the addition of a quadratic term explained a significantly greater proportion of variance in the relationship between the broad-ranged measure of conscientiousness and job performance than in the relationship between prudence and job performance one could attribute the change to the increased discrimination power of the new instrument.
STUDY 2 RESULTS

Hierarchical multiple regression was used to establish the nature of the relationship between the original HPI Prudence scale and leadership performance ratings. Results indicated no significant linear relationship between the two variables (F(507) = 1.25, n.s.; $R^2 = .00$). The addition of a quadratic regression term to the model did not explain a significant amount of variance between the two variables (F(506) = 2.05, n.s.; $R^2 = .01$). The resulting regression lines are presented in Figure 13.

*Figure 13.* Linear and non-linear regression of Prudence and Leadership Performance Ratings.
The same procedure was used to establish the relationship between the original HDS Diligent scale and leadership performance ratings. Results indicated no significant linear relationship between the two variables ($F(493) = 1.77, \text{n.s.}; R^2 = .00$). The addition of a quadratic regression term to the model did not explain a significant amount of variance between the two variables ($F(492) = 1.01, \text{n.s.}; R^2 = .00$). The resulting regression lines are presented in Figure 14.

*Figure 14.* Linear and non-linear regression of Diligent and Leadership Performance Ratings.

These results partially support the findings reported by Robie and Ryan (1999) suggesting that a curvilinear relationship between conscientiousness and job performance does
not exist. Unlike Robie and Ryan, however, a significant linear relationship was not present. The lack of a linear relationship is unexpected, and will be addressed later in this section. In order to determine if the new, broader range measure of conscientiousness is more sensitive to a curvilinear relationship the regression procedure was repeated again using this new scale. Similar to the previous analyses, the results do not support a linear relationship between conscientiousness and leadership performance ratings ($F(493) = 0.14$, n.s.; $R^2 = .00$). The results of this analysis do, however, support the hypothesized curvilinear relationship between the new conscientiousness scale and leadership performance ratings ($F(492) = 6.71$, $p < .01$; $R^2 = .03$), but the nature of this relationship was not as predicted. Instead of an inverted U-shaped relationship, these results suggest that a U-shaped regression line best describes the nature of this relationship. The resulting regression lines are shown in Figure 15.

A U-shaped relationship suggests that performance decreases as conscientiousness increases to a point, and that after that point performance begins to increase as conscientiousness increases. This would mean that high and low levels of conscientiousness are beneficial for job performance, but that moderate levels of conscientiousness are detrimental. This relationship seems unlikely and counterintuitive, but an alternative explanation does exist. Reviewing the results of Study 1, one realizes that mostly Diligent items populate the low end of the new conscientiousness scale and mostly Prudence items populate the high end of this scale. Keeping this in mind, the quadratic regression results may show that low diligence is good for performance and high prudence is good for performance. This suggests that, although neither scale is particularly predictive in isolation, an interaction may be present.
In order to test for the presence of an interaction between the Prudence and Diligent scales in predicting job performance hierarchical multiple regression was used again. Prudence was entered in the first step, Diligent in the second, and an interaction term created using the product of the Prudence and Diligent scores was entered in the final step. Results of this analysis showed a significant interaction between Prudence and Diligent ($F(491) = 2.84, \ P < .05; \ \Delta R^2 = .01$). The interaction plot is shown in Figure 16. This interaction suggests that performance will be greatest when one is high in Prudence and low in Diligence.
The lack of a linear relationship between either of the original scales and the leadership performance ratings was cause for concern, as previous research on the constructs these scales measure and these specific scales suggest that such a relationship exists. The two samples were chosen based on similar occupations from organizations of a similar size that utilize a similar performance dimension. Despite this there were notable differences between the two samples, as indicated in the method section. In order to determine if the unexpected results reported previously were due to the aggregation of these distinct samples the analyses were repeated on the data from each organization independently.

A hierarchical regression model was applied to the HPI Prudence scale and leadership performance ratings in the logistics company sample. Results indicated a significant linear (F(284) = 6.29, P < .05; R^2 = .02) and curvilinear (F(283) = 3.67, P < .05; R^2 = .03) relationship between these two variables, but the addition of a curvilinear term did not explain a significant
amount of additional variance. The same analyses were conducted using the sample of consumer goods company managers. Results showed no significant linear (F(221) = 1.35, n.s.; R² = .01) or curvilinear (F(220) = 0.98, n.s.; R² = .01) relationship between the two variables.

This analysis was then applied to the HDS Diligent scale and leadership performance ratings in the logistics company sample. Results indicated no significant linear relationship (F(273) = 0.75, n.s.; R² = .00), but a significant curvilinear relationship (F(272) = 3.72, P < .05; R² = .03) between these variables. The curvilinear relationship took the same inexplicable U-shape as the curvilinear relationship between the new conscientiousness measure and leadership performance in the aggregated sample. This analysis was repeated using the consumer goods company sample. Results indicated significant negative linear relationship (F(218) = 6.60, P < .05; R² = .03) and a curvilinear relationship (F(217) = 5.59, P < .05; R² = .05) with leadership performance. In this sample, however, the curvilinear relationship showed an inverted U-shape.

Next this set of analyses was conducted using the new conscientiousness measure and the logistics organization managers. Results indicated no significant linear relationship (F(273) = 3.83, n.s.; R² = .01), but a significant curvilinear relationship (F(272) = 6.63, P < .01; R² = .05) with leadership performance. As in the aggregate sample results, the curvilinear relationship showed a U-shape. The analyses were repeated using the consumer goods company managers. Results indicated no significant linear (F(218) = 3.71, n.s.; R² = .02) or curvilinear relationship (F(217) = 2.00, n.s.; R² = .02).

Finally the hierarchical regression including a Prudence/Diligent interaction term was replicated using the logistics and consumer goods samples. Results of the analyses using the logistics managers showed a significant interaction (F(271) = 3.95, P < .05; ∆R² = .03). The
results of the analyses using consumer goods managers did not show a significant interaction

(F(216) = 2.23, n.s.; ΔR² = .00).
STUDY 2 DISCUSSION

The primary purpose of this study was to determine whether the increased reliability of the new conscientiousness scale constructed in Study 1 would result in the curvilinear relationship with job performance hypothesized in the literature (e.g., Murphy, 1996). A significant curvilinear relationship was found for this scale and not the original scales from which it was constructed, but this relationship was nothing like that proposed by Murphy. The curvilinear relationship was expected to take the shape of an inverted U, but instead this relationship was U-shaped. Such a relationship is both counterintuitive and difficult to interpret. The empirical results suggest that moderate levels of conscientiousness are bad for job performance, but that low and high levels of conscientiousness are good for job performance. There is little theoretical rationale one could use to explain why high and low conscientiousness would both result in good job performance, and such a relationship is at odds with all of the meta-analytic findings on the subject.

The secondary purpose of this study was to begin examination of the construct validity of the scale. The fact that the results of this study are not in line with either theory or previous research on the topic calls the construct validity of the new conscientiousness scale into question. Both theory and previous research on the relationship between conscientiousness and job performance suggest a positive linear relationship or inverted U-shaped relationship between these constructs. The U-shaped relationship shown in this research suggests that combining scales of two distinct constructs to create the new conscientiousness scale may have actually created a scale that does not measure the construct it is intended to assess.

The significant interaction between the original Prudence and Diligent scales shown in this study supports the notion that these constructs are more powerful (and appropriate) when
used individually than when combined statistically. This interaction has shown that high
performance results from high conscientiousness and a lack of obsessive-compulsive personality
traits. This interaction, unlike the curvilinear regression findings, is easily interpreted and fits
with previous research and theory suggesting that conscientiousness is beneficial for job
performance and obsessive-compulsive personality is detrimental. These results are also
informative for practical applications, whereas the quadratic regression results serve only to
confuse those interested in applying personality tests for employee selection or development
purposes.

Although the results from the larger combined sample are more appropriate for drawing
generalizable inferences, the findings from the analyses using the company specific samples
provide further evidence to suggest that these constructs are best kept separate. The results of
these analyses were not terribly consistent across samples with one notable exception. The new
conscientiousness scale showed results that run contrary to theory and previous research across
both independent samples and the aggregated sample. Each of the original scales showed results
that are consistent with both theory and previous research in one of the two samples, and in the
interaction results of the aggregated sample. This evidenced shows that the unexpected results
obtained with this new scale are not solely due to the characteristics of the two samples.
GENERAL DISCUSSION

This research had two primary objectives: Determine whether or not obsessive-compulsive personality constitutes the extreme high end of conscientiousness and examine the nature of the relationship between conscientiousness and job performance. To accomplish the first objective, items from a measure of conscientiousness and a measure of the normal range of obsessive-compulsive personality were combined in an attempt to create a measure of conscientiousness with increased test information and reliability at the high end of the scale range. The results of this analysis showed that, contrary to expectations, the addition of obsessive-compulsive personality items to a conscientiousness scale actually increases test information and reliability at the low end of the scale range. These results led to the conclusion that these two constructs, although conceptually linked, may not constitute different ranges of a single continuum, and the new conscientiousness scale may lack construct validity.

A second study was undertaken to examine both the nature of the relationship between conscientiousness and job performance and gather information about the construct validity of the new conscientiousness scale. This study initially compared both the linear and non-linear relationships between the original conscientiousness and obsessive-compulsive personality scales used in the first study and job performance and the new conscientiousness scale and job performance. Despite the unexpected properties of the new scale, this scale’s psychometrics are, in general, superior to either of the scales used to create it. This measure offers an increased range of reliable measurement that should produce better relationships with job performance if such a relationship exists.

With respect to linear relationships none of the three scales were significant predictors of job performance in the aggregated sample. This finding is inconsistent with both theory and past
research on this topic. One would have expected positive relationships between both conscientiousness scales and performance, and a negative relationship between the obsessive-compulsive personality scale and performance. The HPI Prudence (conscientiousness) scale and HDS Diligent (obsessive-compulsive personality) scale both show truncated ranges of score reliability, and the Diligent scale shows unacceptably low internal consistency. This may account for the lack of significant findings in the aggregated sample. Despite this, each scale showed a hypothesized linear relationship in one of the independent samples. This suggests that these scales possess some degree of construct validity. The new conscientiousness scale, despite its acceptable internal consistency and increased range of reliable scores, also failed to show a positive relationship with job performance in either the aggregated or company specific samples. This finding does not support the construct validity of the scale.

The Prudence and Diligent scales also failed to show significant curvilinear relationships with job performance in the aggregate sample. This is not unexpected given previous research in this area, although the Diligent scale did show an inverted U-shaped relationship in the consumer goods company sample. The new conscientiousness measure did show a significant curvilinear relationship with job performance, but the nature of this relationship was not in keeping with the theoretical literature on this subject. The literature suggests that conscientiousness would be predictive of job performance to a point, and after this point the excessive conscientiousness becomes maladaptive and performance decreases. The results of this study indicate that as conscientiousness increases performance decreases to a point, at which time increased conscientiousness results in increased performance. These findings again call the construct validity of the new measure into question.
Given that the addition of the Diligent items to the Prudence scale served primarily to increase test information and reliability at the low end of the conscientiousness scale, it is more likely that the two scales actually represent different constructs (one negatively related to performance and one positively related to performance). The factor analytic results from Study 1 along with the significant interaction shown between these two variables and job performance in both the aggregated and logistics company samples and the linear regression results found in the two independent samples in Study 2 support this conclusion. The initial factor analysis of the Prudence and Diligent scales showed that the Diligent items do form a distinct factor. The linear regression results support the construct validity of the original scales. Finally, the interaction suggests that performance is greatest for individuals who are high in conscientiousness and lack obsessive-compulsive personality characteristics. This makes conceptual sense. One who is dependable and detail oriented, but lacks any maladaptive compulsion toward such behaviors will generally be seen as high performers. All of these findings support the use of the original scales separately, not the use of a combined conscientiousness/obsessive-compulsive personality scale.

Limitations

The primary limitation of this study is that the construct validity of the newly created conscientiousness scale cannot be established. Without this information it is difficult to determine whether or not conscientiousness and obsessive-compulsive personality are two distinct constructs or part of a single continuum of personality. The factor analysis of the two original scales, although dominated by a single factor, did reveal an obsessive-compulsive personality factor, and the zero-order correlation between the two scales is relatively low for two scales that measure the same construct. The current study, however, has shown that measures of
Conscientiousness and obsessive-compulsive personality items can be scaled together, but the appropriateness of this scaling is not clear. The exploratory and confirmatory factor analysis of the new conscientiousness scale provided strong support that only one construct was being measured, but the regression results suggest that the original scales may function in a more explicable and predictable manner than the new conscientiousness scale. Without additional information regarding the construct validity of this scale such results are difficult to interpret.

The scale construction portion of this study could have benefited from the use of additional obsessive-compulsive personality items. The HDS Diligent scale used in this study contained a limited number of items, some of which functioned poorly on the original scale. The inclusion of additional items from a more diverse set of obsessive-compulsive personality scales could have altered the psychometric properties of the end product. It is possible that the scale used in this study, a measure designed to assess the “normal” range of obsessive-compulsive personality, may not fully represent the extreme behaviors expressed by individuals with this trait. If a more traditional, clinically oriented obsessive-compulsive personality measure was used it may have been possible to increase information at the high end of the scale.

A final limitation is that the regression analyses were conducted on a fairly homogeneous set of individuals. All of the people contained in this sample held managerial positions at large organizations. Although conscientiousness has been shown to predict performance across occupations (e.g., Hogan & Holland, 2003), occupational moderators may exist for curvilinear relationships. For example, a job that requires close attention to detail (e.g., accounting) may be more likely to cause difficulties for individuals with extremely high conscientiousness than a job dealing with strategic planning (e.g., senior management). In both cases an extremely
conscientious person may have difficulty with extremely detailed projects, but the strategic managerial position allows less opportunity for this than the accounting position.

Future Research

Additional research in this area should initially focus on three main areas. First, future studies should seek to definitively answer the question of whether or not obsessive-compulsive personality and conscientiousness are two parts of a single continuum. This study has shown that measures of conscientiousness and obsessive-compulsive personality can be scaled together, but it has also indicated that the validity of this scaling is suspect. It is important both theoretically and practically to understand the behaviors associated with the upper (and lower) bound of conscientiousness. I would suggest the use of a diverse set of obsessive-compulsive personality disorder measures including scales oriented toward clinical populations. It may be the case that only items from this type of assessment will yield the appropriate psychometric properties. It may also be necessary to seek out other constructs that may constitute the upper bound of conscientiousness. The term perfectionism came up frequently in reviewing the construct validity evidence for both the Prudence and Diligent scales. This may indicate that a common bond between conscientiousness and obsessive-compulsive personality is perfectionism (see Enns & Cox, 2002 for a review of the perfectionism construct).

Regardless of the success of these efforts, further research should be dedicated to increasing the range of information provided by conscientiousness measures. Without this work any attempts to investigate curvilinear relationships between conscientiousness and work outcomes will be susceptible to criticism based on a lack of score reliability at the score range where performance decrements are expected to occur. Until this is remedied I suspect that future
investigations into the nature of the relationship between conscientiousness and job performance will result in mixed results similar to those currently present in the literature. Once measures with adequate psychometric properties are available, a series of regression-based studies will need to be conducted using several different jobs and performance criteria in order to establish reliable findings about the nature of the relationship between conscientiousness and job performance. The use of multiple distinct samples and multiple criteria will allow researchers to determine when a curvilinear relationship should be expected and what job characteristics are typically associated with this relationship. This type of information will be useful to inform both theory and practice.

Practical Implications

The main findings of this study provide little information of practical benefit. The interaction between conscientiousness and obsessive-compulsive personality, however, may prove quite useful for practitioners. The benefits of including personality variables such as conscientiousness in selection batteries are well established (e.g., Schmidt & Hunter, 1998), but the results of this study suggest that, at least in the case of conscientiousness, high performance results from an interaction between positive personality traits and negative personality traits. This line of thinking is not new (see Hogan & Hogan, 2001), but this study provides further evidence that this approach may be beneficial for I-O practitioners designing selection systems.

The application of these findings to selection system design should, however, be undertaken with caution. As previously discussed, the Americans with Disabilities Act prohibits selection on the basis of mental illness. The HDS was specifically designed to assess the normal range of personality disorders, but care must be taken to ensure those suffering from mental illness are not unfairly denied opportunities without reasonable accommodation. If measures
such as the HDS are to be used to screen-out applicants studies should be undertaken to ensure that the measure does not differentiate between normal and clinical populations.
REFERENCES


Costa, P. T., Jr. & McCrae, R. R. (1994). *Bibliography for the revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.


data and recommendations for DSM-IV. *Journal of Personality Disorders, 5*, 363-375.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago:
University of Chicago Press.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and


Robie, C., & Ryan, A. M. (1999). Effects of nonlinearity and heteroscedasticity on the validity of
conscientiousness in predicting overall job performance. *International Journal of
Selection and Assessment, 7*, 157-169.

distortion on preemployment personality testing and hiring decisions. *Journal of Applied
Psychology, 83*, 634-644.


psychology: Practical and theoretical implications of 85 years of research findings.

*Psychological Bulletin, 124*, 262-274.


APPENDIX A. ITEM LEVEL FIT PLOTS

Fit Plot for Prudence 1

Fit Plot for Prudence 2
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

**Fit Plot for Prudence 5**

![Fit Plot for Prudence 5](image1)

**Fit Plot for Prudence 6**

![Fit Plot for Prudence 6](image2)
Appendix A. Item Level Fit Plots (continued)

Fit Plot for Prudence 7

Fit Plot for Prudence 8
Fit Plot for Prudence 9

Fit Plot for Prudence 10

APPENDIX A. ITEM LEVEL FIT PLOTS (continued)
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Prudence 11

Fit Plot for Prudence 12
Fit Plot for Prudence 18

Fit Plot for Prudence 19
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Prudence 20

Fit Plot for Prudence 21
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Prudence 22

Fit Plot for Prudence 23
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Prudence 24

Fit Plot for Prudence 25
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Prudence 26

Fit Plot for Prudence 27
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Prudence 28

Fit Plot for Prudence 30
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Diligent 2

Fit Plot for Diligent 5
Conscientiousness and Performance

APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Diligent 6

Fit Plot for Diligent 7
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Diligent 9

Fit Plot for Diligent 11
APPENDIX A. ITEM LEVEL FIT PLOTS (continued)

Fit Plot for Diligent 12

Fit Plot for Diligent 13