TRANSITIVITY AND INTER-DEFINITION CONSISTENCY OF NEO NEUROTICISM-DOMAIN RATINGS

By

STEFANIE I. BADZINSKI

Bachelor of Science in Psychology
Southern Nazarene University
Bethany, Oklahoma
2004

Master of Psychology
University of Dallas
Irving, Texas
2006

Master of Science in Psychology
Oklahoma State University
Stillwater, Oklahoma
2009

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2012

TRANSITIVITY AND INTER-
DEFINITION CONSISTENCY OF NEO
NEUROTICISM-DOMAIN RATINGS

Dissertation Approved:

Dr. James W. Grice

Dissertation Adviser

Dr. John Chaney

Dr. David Thomas

Dr. Dale Fuqua

Name: STEFANIE I. BADZINSKI

Date of Degree: DECEMBER 2012

Title of Study: TRANSITIVYT AND INTER-DEFINITION CONSISTENCY OF NEO NEUROTICISM CONSTRUCTS

Major Field: PSYCHOLOGY, LIFESPAN DEVELOPMENTAL OPTION

Abstract: It has been suggested that psychometrics is a pathological science on the basis that its normal processes prevent the attainment of its stated goal and conflicting interests prevent its constituents from acknowledging that this is the case. This study addressed basic measurement concerns associated with self- and other-report personality measures that are not addressed under the prevailing psychometric for establishing psychometric validity. Participants completed a series of pairwise comparisons (i.e., "more," "less," "equal") of themselves and known others with respect to 13 NEO Neuroticism-domain constructs. These judgments were examined for triplet transitivity; participants often made judgments that violated the condition of transitivity, suggesting that researchers are not justified in self- and other-ratings with respect to these attributes as ordinal or quantitative relations. Participants also compared themselves and known others in a complex ranking procedures with respect to three broad-level definitions of Neuroticism that appear in the NEO literature. Responses were examined for rank consistency across definitions. Low rates of rank consistency across definitions suggest that these definitions are not practically synonymous among lay persons. These studies do not prove that people cannot consistently order person with respect to Neuroticism-domain attributes, nor do they prove that multiple definitions of Neuroticism evoke different constructs for all people; however, in the absence of evidence that self- and other-ratings are ordinal and in the presence of evidence that they are often *not* ordinal, it is inappropriate to treat them as representations of rank or quantity. Likewise, in the absence of demonstrations that multiple definitions of Neuroticism are practically synonymous to individual raters and in the presence of evidence that they are often *not* practically synonymous, it is inappropriate to use them interchangeably. In brief, these are the kinds of measurement tasks that researchers need to engage in order to justify the measurement claims they are already making. These methods are recommended for use in conjunction with conversations with raters as a means to confront and correct the present state of psychometrics.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Psychometrics as a Pathological Science

Joel Michell (e.g., 2000, 2008a, 2008b) suggests that an enterprise should be classified as pathological when its normal processes prevent the attainment of its stated aims and conflicting interests prevent its constituents from recognizing that this is the case. He has also argued that psychometrics satisfies these criteria. In the course of evaluating this claim, at least three broad questions should be asked of psychometrics: 1) what would it take to meet the goal of psychometrics? 2) Do the prevailing methods prevent the attainment of that goal? 3) If the answer to the former is affirmative, who benefits from maintaining the illusion that psychometrics is "on track?" While all three inquiries are important, the current paper focuses primarily on the first two.

CHAPTER II

REVIEW OF THE LITERATURE

The Goal: Quantitative Measurement

Michell calls psychometrics pathological on the grounds that a) the goal of psychometrics is the measurement of psychological characteristics, yet the characteristic methods of psychometrics merely assume that psychological attributes are measurable and b) that mainstream psychometricians have accepted this assumption as true in the absence of serious attempts to question it.

In accordance with classical measurement theory, he defines measurement as "the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute," (Michell, 1997, p. 358), where an attribute is a range of properties or relations that may vary from instance to instance, and a quantitative attribute is a range of properties or relations that admits infinite degrees of variation in magnitude (Ellis, 1996; Michell, 1990; Trendler, 2009). According to this view of measurement, numbers or values are relations that obtain between real things; more specifically, numbers are faithful representations of magnitudes of quantitative attributes possessed by real things.

Quantitivity

Quantitative properties or attributes are variables that meet conditions of ordinal, additive and continuous structure, where variable may be defined as "anything relative to which objects may

vary" (Barrett, 2003, p. 423). The axiomatic conditions of quantitative structure as outlined by J.S. Mill (1843/1974) and later by Hölder (1901) are as follows:

Let X, Y and Z by any three values of variable Q. Variable Q satisfies the conditions of quantitative structure if and only if:

1) $X \geq Y$ and $Y \geq Z$ then $X \geq Z$ (transitivity; this extends to $X < Y < Z$);

2) $X \geq Y$ and $Y \leq X$ then $X = Y$ (antisymmetry);
3) either $X \geq Y$ or $Y \geq X$ (strong connexity);

Let Variable Q be any ordinal variable such that for any of its values X, Y and Z:

4) $X + (Y + Z) = (X + Y) + Z$ (associativity);
5) $X + Y = Y + X$ (commutativity);
6) $X \geq Y$ if and only if $X + Y \geq Y + Z$ (monotonicity);
7) If $X > Y$ then there exists a value of Z such that $X = Y + Z$ (solvability);
8) $X + Y > Z$ (positivity);
9) There exists a natural number n such that $nXY \geq Y$ (where $1X = X$ and $(n+1)X = nX + X$) (the Archimedian or continuity condition).

Axioms 1-3 describe the conditions of ordinal structure; an attribute with ordinal structure permits variation in magnitude (i.e., "more" or "less"). Axioms 1- 6 describe the conditions of additivity; magnitudes of an attribute with additive structure can be stated as quantities of a common unit. Magnitudes of attributes with additive structure can be meaningfully combined (e.g., 2 + 3 = 5 only where 2 = two units of an attribute and 3 = three units of the same attribute). The final condition of quantitivity is the Archimedian condition, which establishes continuity; continuous attributes are those for which there is no limit to the smallness of the units into which they may be divided. Length, for example, is a continuous attribute as length may theoretically be divided infinitely. Discrete attributes on the other hand, are indivisible units (e.g., a "quantity" of people must always be a positive integer because half of a person, for example, is not a person.)

Under the classical measurement paradigm, variable properties satisfying all nine conditions are, in principle, measurable as for any values *a* and *b* of Q, the magnitude of *a* relative to *b* may be expressed as a positive real number (i.e., *a* may be stated in terms of units *b*). The length of a single object, for example, is stated relative to an arbitrary unit of length (e.g., the

length *a* of a yardstick may be stated as a ratio of units of inches *b* in, such that $a = 36b$.) The relative magnitudes of two objects with respect to length may be stated as a ratio of their respective magnitudes of length (e.g., 36 inches = 3*12 inches, or 1 inch = 1/3 * 36 inches). Measurability extends to relations between magnitudes of two or more attributes that are quantitatively related. Density, for example, is equal to a ratio of mass, a quantitative attribute, to units of volumes, also a quantitative attribute; therefore, the density of any object may be stated in terms of the density of any other object (e.g., the gm/cm^3 of lead = .15 gm/cm^3 of magnesium or the gm/cm^3 of magnesium = 6.647 gm/cm^3 of lead).

Michell (e.g., 2008a, 2008b) is adamant that, in a scientific context, the term "measurement" properly applies solely to the assessment of quantity (i.e., continuous quantities). To date, no psychological attribute has been demonstrated to have quantitative structure; if we accept the definition of measurement as the estimation or discovery of magnitudes of quantity, without discounting their statistical elegance, any claim that psychometric modeling techniques produce measurements of psychological attributes is misleading. Michell is by no means the first to make this charge against psychometrics. In the 1930's, the British Association for the Advancement of Science established the Ferguson committee, which included noted physicist and measurement theorist Norman Robert Campbell, to investigate whether psychological attributes satisfied the conditions of quantitative measurability, and by implication the viability of a field of psychometrics. In its final report the committee determined that, in the absence of empirical demonstrations that ratings of psychological attributes could sustain concatenation operations (i.e., conditions of additivity), psychological attributes could not be considered scientifically measureable (Ferguson et al., 1940).

An Alternative Goal: Non-quantitative Measurement

Psychologist and psychometrician S. S. Stevens responded to the committee's report by challenging their definition of measurement as the estimation of quantities of units and ratios of quantities. Admitting that psychometricians had not verified the quantitivity of psychological

attributes, he famously defended psychometrics by invoking an alternate definition of measurement, writing that:

> "…in dealing with the aspects of objects we can invoke empirical operations for determining equality (the basis for classifying things), for rank ordering, and for determining when differences and ratios between the aspects of objects are equal…This isomorphism between the formal system and empirical operations performed with material things justifies the use of a formal system as a model to stand for aspects of the empirical world" (1951, p. 23).

In other words, measurement is the assignment of values to entities according to a consistent rule. Representational measurement theory, grounded in Tarski's metamathematics, particularly his general theory of models, holds that measurement is a homomorphism, or structure preserving map, between finite sets of elements in a particular relational system (Trendler, 2009; see Sher, 1999 for a review of Tarski's work). A relational system is purely formal, and the elements are abstract symbols. If the elements are applied to empirical objects, the corresponding set of objects is referred to as an empirical relational system or an empirical structure. When elements take the form of numbers, relations between the elements are expressed in mathematical terms; a relational system in which elements take the form of numbers is called a numerical relational system or a numerical structure. Provided that the relations between a numerical structures and corresponding empirical structures are maintained, empirical relations may be studied in mathematical terms. That is, mathematical operations performed on measurement values can serve as a proxy for *in vivo* operations performed on empirical systems.

Under a representational measurement paradigm, inasmuch as a procedure estimates empirical relationships, it may be considered measurement. Quantitative measurement, then, is the representation of one among many empirical systems; that is, it is a model of formal relationships among elements with respect to quantity. Many non-quantitative formal relationships holding between sets of entities may also be meaningfully represented by numerical

structures and mathematical operations performed on the values produced by non-quantitative measurement procedures can also serve as a proxy for *in vivo* operations performed on empirical systems, provided that the operations performed on the values do not disrupt their representational status.

Stevens outlined four scales of measurement: nominal, ordinal, interval and ratio. From a representational measurement perspective, nominal scales are a form of measurement as they represent formal relations among a group of entities with respect to a particularly category. For example, provided that some rule exists for determining "redness," most physical objects may be classified in relation to one another using a dichotomous scale of red = 1, not red = 0. Nominal scales are the basis for all measurement scales, in part because they identify sets of entities that may be further classified with respect to a common attribute (e.g., we do not measure length in things that do not satisfy the criterion of possessing length). Estimation of ordinality also constitutes measurement as ordinal values preserve formal relations between objects with respect to rank magnitudes of a common attribute. Interval and ratio scales both preserve formal relations between objects with respect to magnitude of quantity, although interval quantities are stated with respect to an arbitrary zero and ratio quantities are stated with respect to an absolute zero. Stevens also outlined the statistical operations that could be meaningfully applied to values representing each type of relational system or "level of measurement."

The Stumbling Blocks

Psychometric "Validity"

Provided that measurement refers solely to the estimation or discovery of quantity, the field of psychometrics *as* that field whose project is the measurement of psychological attributes has failed to reach or even to strive for its goal. Several theories provide theoretical means of quantifying and demonstrating the quantitivity of attributes that cannot sustain concatenation operations (e.g., conjoint measurement theory, e.g., Luce & Tukey, 1964; item response theory, see Borsboom & Mellenbergh, 2004), but as they fail a) to identify the nature of the empirical

system targeted for quantification and b) are generally applied to observed data with unknown representational value, they may be considered distractions from rather than solutions to the fundamental problem of psychological measurement.

At the risk of offending those who believe that the term measurement applies solely to estimation or discovery of quantities, I submit that the goal of psychometrics is the accurate representation of relative relations of individual entities with respect to psychological attributes and emphasize that this includes but is not limited to the representation of quantitative relations *where they are claimed or assumed.* Though inconsistent with Michell's definition of psychological measurement, this is consistent with his concern that the term "measurement" evokes quantitative relations and is, thus, misleading when it is used in reference to the estimation of non-quantitative relations (2008b).

The charge that psychometrics is pathological is not made inapplicable by the adoption of a representational measurement paradigm. If anything, the lens of representational measurement theory brings the pathological nature of prevailing psychometric practices into focus more clearly as it emphasizes the standard of measurement validity as representational accuracy with respect to a particular empirical system; from this perspective, error is meaningful solely in reference to a target. In order, then, to attain the goal of psychometrics, psychological measurement procedures must correspond to a target empirical system and the values produced by them should be used in full knowledge of what that target is. For example, if a scale is intended to produce ordinal relations of individuals with respect to attribute A, then the validity of the scale should be determined solely by its ability to do so, and neither the operations applied to the values produced by the scale nor interpretations of analyses carried out using scale values should disrupt their representational validity. The same standards should apply to procedures designed to represent individual magnitudes in relation to units of a quantitative attribute, ordinal group relations with respect to an attribute, and so on. That is, the standard must be specified and respected. Failure to do so carries with it the potential of poor science, the abuse of the public's

trust in psychometrics and those who employ psychological scales in their research, and the very real possibility that the misinterpretation of research that relies on the validity of psychometric procedures will contribute to societal harm (e.g., misallocation of public or private funds, unjustified loss of employment/promotion, etc.)

Under the prevailing psychometric practices, the validity of psychological measurement scales is judged largely by whether the values it produces are statistically correlated in the anticipated direction and degree with values produced by other psychological measurement, by inter-item correlations, etc. This is inappropriate provided that a scale is meant, for example, to order individuals—or even individual groups- with respect to some attribute and that the values produced by scale administration are treated as if it does. As long as standard practices used to vet psychometric procedures allow us to conclude that a measurement procedure is valid (i.e., that it measures what it intends to measure) in the absence of a) a clear statement as to the empirical system it is intended to represent (e.g., ordinal relations of entities with respect to a named attribute) and b) empirical evidence that it succeeds in doing so (e.g., at the very least, values assigned to each entity should be shown to satisfy the expectation of whatever relation is claimed, be it ordinal, quantitative, etc.), one can confidently state that the prevailing psychometric paradigm does not simply fail to attain its goal, but actively stands in the way of its attainment. These statements reflect only a brief and practical assessment of the current state of psychometrics. It is certainly possible to delve deeper into the question of what the goal of psychometrics is or *should be* according to psychometricians—and Michell (e.g., 2008b) does, but as the products of psychometrics are so widely used and the results of their use so influential in our systems of government, education, industry, etc., I submit that there is no need to go deeper to conclude that something is rotten in the state of psychometrics.

Conflicting Interests

If psychometric procedures and practices do not and cannot do what we claim and/or believe them to do, why do we persist in using them? According to Michell (2000), the pressure

of scientism in the 19th century led psychologists to adopt and advance the rhetoric of measurement, and continuing pressures have prevented psychologists from acknowledging the resulting methodological shortcomings. One of the pressures he cites is the economic reliance of scientific research on grants from agencies that award financial support based upon apparent methodological rigor. Michell points out that the "new rigorism movement" in psychology emerged during the immediate post-WWII decades when government investment in scientific research was on the rise, and that our methods have remained largely unchanged since that time.

Psychology historian Kurt Danziger (1990) has presented a similar narrative of what he refers to as "the triumph of the aggregate" (p. 68). Prior to WWI, almost all of the published results in journals of experimental psychology were attributed to individual subjects; even in cases where responses were averaged across individuals, individual responses were generally provided and interpretations were usually based on these individual patterns of responses. During the period from WWI to WWII, the ratio of studies reporting individual data to aggregate data rapidly declined, particularly in journals geared towards "applied psychology." Danziger makes the case that this shift may be attributed to the higher marketability of aggregate studies relative to studies of individuals as a proposed means to assess and improve social conditions.

Both Michell (2000) and Barrett (2008) suggest that, having succeeding for so long in presenting psychology as a quantitative science, psychologists have become increasingly invested in preserving the appearance of measures of psychological phenomena. Psychologists are by no means alone in this investment. The increased focus on aggregate analyses may not at first glance appear to stand in the way of the pursuit of valid measurements, as even aggregate analyses presume the valid representation of individuals within a specified system. As with any product, however, once pseudo-measurement procedures had been marketed to the public (i.e., a client-base has been identified), their creators, marketers and users can easily find themselves colluding to defend their worth.

Widespread acknowledgment that the field of psychometrics has gone off plot would have equally widespread consequences because it would call for a reevaluation of its products, which have been disseminated into almost every aspect of society. Researchers do not want to question the foundations of years of work, government agencies do not want to hear that they spent millions of dollars on so-called scientific procedures that turned out to be no more than a "bad tip;" the list could go on and on. The debate over whether or in what sense psychological attributes are measurable has inspired countless books, journal articles, etc., many of which are confusing to initiates into psychology and/or measurement theory and most of which are likely impenetrable to members of the general public; yet a child knows that four apples is twice as many as two apples, while psychologists regularly ignore the importance of demonstrating that a person who obtains a score of 10 on a so-called quantitative measure of self-esteem should have twice as much self-esteem as a person who obtains a score of 5. Could it be that there are many and varied forces preventing educated psychologists from focusing on the most basic requirements of psychological measurement?

Confronting Our Own Pathology

Michell has proposed that psychometrics is pathological because its normal processes prevent the attainment of its stated aims and conflicting interests prevent its constituents from recognizing that this is the case. I concur with Michell, but submit that psychometrics has not failed to attain its goal by failing to demonstrate that psychological attributes are measurable in the sense that they are quantitative, but by failing to demonstrate that psychological measurement procedures are valid representations of any empirical system quantitative or otherwise. I also submit that, because psychometric practices and procedures are being applied by decreasingly specialized persons, the number of people who should be labeled constituents of the field of psychometrics goes well beyond those designated as "psychometricians" to include practitioners of its products and procedures (e.g., psychological researchers who follow accepted standards to design "measures" of psychological attributes). That is, the health of psychometrics should not

be judged solely at the level of the writings and beliefs of measurement specialists—not to say that the final diagnosis would differ-but also at the level of application.

If prevailing psychometric practices are pathological, what is the cure? A good first step is to acknowledge that psychometricians have not "filled the position" of answering basic measurement issues. The statistical procedures associated with psychometrics may be of great worth provided that they are applied to valid measurement values. Consider, for example, Bayes formula for calculating the probability of X given Y: If we know the true and false positive rates of a test for Disease X and the rate at which Disease X occurs in a population, we can calculate the conditional probability that a person who tests positive on the test actually has Disease X. Before we can apply this formula, however, we have to establish what it means to have Disease X; that is, the test helps us estimate something that is knowable in another way. While such a test may be more feasible than the path to concrete knowledge (e.g., perhaps 100% certainty requires a postmortem procedure), it does not replace the basic measurement task of clarifying the essential characteristics of the target state. The current study is offered as an example of some basic measurement tasks that must be fulfilled if we are to claim that we are measuring psychological attributes. The criticisms and methods outlined here are applicable to any number of psychological phenomena, but this particular study addresses a measurement claim that is made by personality trait theorists; namely, that we can and regularly do *measure* personality.

Measuring Personality

From the perspective of trait theory, personality is understood as a collection of observable patterns of behaviors (i.e., "traits") that are attributable to a finite set of underlying structures—also referred to as "traits"-the latter of which are characterized by the production of forces that motivate engagement in category-specific behaviors (e.g., agreeableness, extraversion). Most if not all of the personality differences we observe from one individual to another are believed to be produced by different combinations of values along quantitative trait-dimensions that are applicable to all people. As traits are considered to be quantitative attributes

possessed by individuals, one would expect a trait-measure to produce values that represent individual magnitudes of an identified trait-unit (e.g., a unit of motivational force specifically directed towards a class of trait-relevant behaviors if the target system is that of individuals with respect to underlying structures; a ratio of trait-relevant behaviors to all behaviors or to behaviors that are specific to a single trait-dimension if the target system is that of individuals with respect to traits as observable patterns of behavior); this, however, is not the case.

Trait-measures generally consist of a series of items in which people are asked to rank or rate the self and/or others with respect to typical behavior or personal attributes, also referred to as "traits" on the assumption that the language we use to describe ourselves is a valid indicator of actual behavior. For example, people are believed to be called "kind"- a natural language trait—because they engage in many "kind" behaviors and, perhaps, few "unkind" behaviors, and a person who engages in more "kind" acts than another is described as "kinder" than him/her. That is, trait scales are intended to represent empirical magnitudes or proportions of behaviors.

Trait models of personality rely heavily on the use of factor analysis, which have been embraced as a means both to uncover and measure individual relations with respect to latent quantitative structures that underlie variation in patterns of behaviors. Factor analysis is a statistical procedure that decomposes the shared variance among sets of observed measurement values into common and variable-specific portions, and distributes common variance across a smaller set of linear functions called factors (Harman, 1967). The latent variables that are purportedly revealed by factor analysis cannot be observed directly, but must be inferred from values that are observable (i.e., knowable). For factor models emerging from the lexical tradition, including the Five Factor Model, self- and other-rating/rankings with respect to natural language traits are generally used as measures of observable behavior. The resulting linear models of aggregate variance are yet another referent of the term "trait," though they do not correspond to any of the verbal definitions given for the term.

So far, four definitions of "trait" have been outlined: 1) patterns of observable behavior displayed by individuals, 2) internal structures believed to cause observable patterns of behavior in individuals, 3) natural language terms that are believed to encode whether and to what extent a person engages in a specific class of behaviors and 4) aggregate shared variance in observable behaviors, often operationalized in terms of the previous definition, modeled as a linear function. It is in the latter sense that traits are believed to be quantitative, although trait researchers are not clear as to whether factors are intended as models of hypothesized internal structures that cause variation in observed behaviors or a collective portion of the variance of a syndrome of observed behaviors that are caused by hypothesized internal structures (Boag, 2011). Regardless of which definition is adopted, factors represent relations among a set of items— not persons- with respect to the proportion of variance a sample of ratings made on each item contribute to a pool of shared variance. As Lamiell (2003) has pointed out elsewhere, the personality trait-dimensions extracted using factor analysis do not refer to any attribute, quantitative or otherwise, possessed by individuals, and any conclusion to the contrary is based on a misunderstanding of the statistical procedure.

Trait measures constructed in accordance with factor models generally consist of a number of dispositional or behavioral items that load highly onto a single factor along which people are asked to rate themselves using a graded scale (e.g., a Likert-type scale; i.e., the same type of procedure used to obtain the raw data used to create many factor-based scales). The values assigned to the resulting responses may then be summed or averaged under the assumption that they are quantities—these summed or averaged responses may also be weighted by the proportion of variance the item contributes to the total variance of the factor--to produce a score that is then treated as an indicator of the magnitude with which the rated person possesses the target latent trait. Individuals may then be assigned a percentile score relative to scores obtained from a demographic of interest (e.g., females, Americans, etc.) These are considered to be valid quantitative measurements of traits because when the aggregate variance obtained from the

measures is modeled as a linear function, it closely reproduces traits *qua* linear factor models of aggregate variance. That is, the target relational system pertains to aggregate variance plus error, so the measures are considered to be valid because they are highly correlated with a target factor and not correlated or less correlated with non-target factors. Regardless of the ontological status individual researchers ascribe to latent factors, we must acknowledge that a precondition of a meaningful factor model is that the values analyzed using factor analytic procedures are valid representations of empirical systems that are knowable and thus subject to empirical validation.

The following self-rating item is taken from the NEO-PI-R (Costa & McCrae, 1992), a personality measure that is used for a variety of purposes including personality research and employee selection. The instructions and rating scale are typical of personality measures that have been shown to have good psychometric properties.

Instructions: Indicate the extent to which you agree with the following statement. Rank yourself in relation to others who are similar to you in age and sex.

"I worry a lot."

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
1	2	3	4	5

When the target relational system is an aggregate level factor comprised of non-specific variance, the results of this measurement procedure may appear valid; however, when each item is considered as a measurement procedure that should produce values consistent with the axioms proper to the type of relational representation that is claimed, this common format is revealed as inordinately complex. Among other shortcomings, the item reflects the inattention to the target system that is typical of measurement procedures that are validated by virtue of statistical convergence/divergence, and without which measurement validity and error lack meaning. Before we can treat ratings on this item as representations of quantity, we must first establish that a rating of "Agree" corresponds to four times the magnitude of the target attribute that

corresponds to a rating of "Strongly Agree," however it is not clear how even a single response value on this scale should be interpreted as the target unit is not clearly defined. Suppose Person A worries infrequently (i.e., not "a lot"), but is consumed by worry in those times (i.e., "a lot"), while Person B worries frequently (i.e., "a lot"), but is only mildly distressed by it (i.e., "not a lot")? Is "strongly agree" an equally appropriate description of both Person A and Person B? That is unclear. Even considered apart from the complexity of the item as a whole, the middle option (i.e., Neither Agree nor Disagree) makes no readily apparent sense. Is it intended to give the rater the opportunity to abstain from answering the question similar to a "?" option? Is it intended as a neutral point? That is also unclear. It is scored as a "3," so regardless of the meaning the test-makers intended or the rater's interpretation, it is added towards a final sum score as three times more agreement than Strongly Disagree, and 1/3 less agreement than Strongly Agree. Furthermore, treating these ratings as quantities implies that there are an infinite number of gradations of agreement/disagreement between each response value, but that the ratees exhibit only the magnitudes shown. The inclusion of the instruction for raters to consider the statement regarding his/her behavior relative to a vaguely defined comparison group adds to the difficulty of discerning the intent of the item. The belief that self-report measures provide meaningful relations among multiple persons relies on the implicit assumption that people are able to order themselves and other people in relation to one another. When raters are instructed to consider themselves in relation to a specific context group, this implicit assumption is extended to include the belief that people are able to order themselves and other people who belong to the specific group in relation to one another. Ideally, all raters are willing and able to do so, and as such raters who belong to the same specified demographic group will consider themselves in relation to one another, which could be interpreted as a basis for treating identical responses among raters as representing identical relations with respect to the attributes in question. These assumptions are untested and many procedures would be required to support them. The fact that this item was taken from what is considered to be an empirically validated personality measure

serves as clear evidence of the need to return to fundamental measurement issues in this area before we can claim to have measured traits, latent or otherwise, at the individual level.

Measuring Neuroticism

Trait Neuroticism is one of the most heavily researched personality constructs of the 20th century, but as is the case with other personality traits there is no consensus as to the referent of the term "Neuroticism" that can be identified as a unit applicable at the level of the individual. It is associated with a variety of definitions, each so vague that it is difficult where not impossible to ascertain what a single researcher means by it in a single instance. It has been identified, for example, as a "relatively enduring disposition to experience negative affect," "a dimension of psychopathology," and a "susceptibility to psychological distress." These definitions might serve to clarify the concept of Neuroticism provided that the contained terms were clarified in turn. As it is, they make use of terms that are also highly ambiguous (e.g., what is the standard that gives meaning to the phrase "relatively enduring"? how do we distinguish "susceptibility" to psychological distress from actual psychological distress? what anchors and comprises a "dimension of psychopathology"?). Depending upon how such terms are disambiguated, the construct evoked by the term "Neuroticism" may be highly consistent or highly inconsistent within and/or across researchers.

A Neuroticism or Neuroticism-like factor dimension is reliably extracted from varied data sets and is recognizable in some form within almost all theoretically or empirically based models of general personality structure (e.g., Digman, 1990; McCrae & Costa, 2003). The misinterpretation of factor analysis has resulted in the treatment of this fact as evidence that Neuroticism *qua* statistical factor is a truly general structure (i.e., present in all persons), as opposed to an aggregate structure. In Digman's (1990) enthusiastic review of the five-factor model of general personality structure, he refers to personality factors from various models representing the presence and effects of negative emotionality using the generic term "Dimension IV," a dimension generally referred to as Neuroticism "to line up with the vast work of Eysenck

over the years" (p. 422). Other theorists employ the terms "emotional stability," which often appears as the opposite of neuroticism in some models-- or simply "stability" (Guilford, 1975; Goldberg, 1993), "emotionality" (Lee & Ashton, 2004), "negative emotionality/temperament" (Watson & Tellegen, 1985; Watson & Clark, 1994), and "harm avoidance" (Cloninger, 2000), but this practice is largely considered to reflect arbitrary preference rather than construct divergence as the scales themselves produce highly convergent scores.

It is clear that Neuroticism measures that are validated on the aggregate level do not target individual relations onto Neuroticism, but what would we need to do to construct a valid representation of individual Neuroticism? Lamiell (2011) has suggested that researchers focus on individuals' raw scores instead of normalized percentile scores (i.e., make minor adjustments in the way we use current scales). It is quite harder than that because, even though the purported target system occurs at the level of individuals, our current scales are nonetheless created *as* measures of aggregate variance; that is, aggregate bias is built in to the scale. We would have to pursue a definition of Neuroticism that is applicable to individuals, but what definition should we pursue? In addition to the fact that broad definitions of Neuroticism, so or otherwise named, can be interpreted as referring to a number of types of relational systems (e.g., quantities or ratios of discrete behaviors/emotions, motivation to engage in behaviors or to experience emotions, duration of behaviors/emotions), various models and measures of Neuroticism are clearly disparate in regards to which behaviors and/or emotions are construct-relevant. Table 1 lists the facet scales of the Neuroticism scale of the NEO Personality Inventory-Revised (NEO PI-R), an assessment of personality dimensions within the five-factor model (FFM), the Negative Emotionality scale of the Positive Affect Negative Affect Schedule (PANAS; Watson, Clark & Tellegen, 1988), the Negative Affectivity scale of the Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008), the Negative Emotionality scale of the HEXACO Personality Inventory (HEXACO-PI; Lee & Ashton, 2004), the Harm Avoidance Scale of the Temperament and Character Inventory (TCI; Cloninger, Przybeck, Svrakic, &

Wetzel, 1994), an expansion of the Tridimensional Personality Questionnaire (TPQ; Cloninger, Przybeck, & Svrakic, 1991), and the neuroticism scale of the Eysenck Personality Profiler (EPP; Eysenck & Wilson, 1991). There is a readily recognizable theme both within and across these scales, but the lack of overlap between the facets of scales is striking.

Table 1. Facets of negative and/or unstable emotionality related subscales

NEO PI-R	PANAS	MPQ	HEXACO	TCI	EPP
Angry hostility	Hostility	Aggression			
Anxiety			Anxiety	Fear of uncertainty	Anxiety
Depression					
Self-CSS					
Vulnerability (to stress)		Stress Reaction			
Impulsivity					
	Irritability/Anger				
	Fear		Fear		
	Scared				
	Nervous				
	Jittery				
	Guilty				Guilt
	Ashamed				
	Upset/Distressed				
	Distressed				Unhappiness
		Alienation			
			Sentimentality		
			Dependence		Dependency
				Fatigability	
				Anticipatory worry	
				Shyness	
					Inferiority
					Hypochondriacal
					Obsessive

In short, provided that Neuroticism is an attribute, quantitative or otherwise, that is possessed by individuals and thus permits the representation of individuals relations with respect to it, existing procedures designed to measure it are inadequate to the task. The implication of the non-committal and inconsistent language associated with Neuroticism is that we cannot target a

valid representation of empirical relations, quantitative or otherwise, of individuals with respect to Neuroticism that is consistent with a single definition associated with the existing measures. Given that the present state of scientific definitions and "measures" of Neuroticism provides little foundation on which to build a serious measurement inquiry, there are a number of empirical systems that might be considered equally valid measurement targets. One approach would certainly be to disambiguate one of the many existing definitions of trait Neuroticism that appears in trait literature; it would be tedious, but possible, to create a precise, objective definition of Neuroticism that could then be imposed onto individuals as a measurement standard. The most appropriate target of any measurement procedure is dictated by theory and need. The system that is of most theoretical interest to the present study is that of subjective relations of individuals with respect to traits *qua* dispositional terms/phrases that appear in natural language. One justification for this is that this definition of a trait is the only one from trait literature that appears to be a viable target. Neither traits *qua* "patterns of observable behavior" nor traits *qua* internal structures as causes of behavior are as yet well conceptualized, and traits *qua* statistical factors do not even occur at the level of the individual. Another is that, stripping away all the confusing formatting, the trait measures that have been and continue to be used in research and industry come down to ratings with respect to traits *qua* dispositional attributes. A third is that language is a natural starting point for modeling personality if it is the aim of personality researchers to develop a scientific model of personality that is consistent with the phenomenon of personality as how we "know" one another (McAdams, 1995).

Present Study

A study was designed to address two measurement questions concerning Neuroticism-domain ratings: 1) Do self- and other-ratings satisfy the transitivity, which is the first axiomatic condition of both ordinal and quantitative structure? 2) Do multiple broad definitions of Neuroticism as a general personality dimension used interchangeably in personality literature communicate the same construct to lay raters? The traits selected for use in this study represent

the facets of the NEO model of Neuroticism that are assessed in the NEO-PI-R (Costa & McCrae, 1992; i.e., Angry Hostility, Anxiety, Depression, Self-Consciousness, Vulnerability to Stress, and Impulsivity.) Three broad definition of Neuroticism were also taken from the NEO literature: 1) Emotionally unstable, 2) Sensitive, emotional and prone to experience feelings that are upsetting and 3) Prone to psychological distress. Given the lack of consistency in the trait literature, it can be argued that any set of traits and definitions appearing in a single Neuroticism trait scale/model would have been equally appropriate for use in this study. NEO Neuroticism-domain constructs were selected primarily on the basis of the model's popularity.

Question One: Transitivity

Transitivity states that, for any three values X, Y and Z of Variable Q, if $X \geq Y$ and $Y \geq Z$, then $X \geq Z$, and by implication, if $X < Y$ and $Y < Z$, then $X < Z$. In other words, provided that three entities X, Y and Z each possess some magnitude of Variable Q, if at a given point in time entity X possesses more Q than entity Y, and entity Y possesses more Q than entity Z, entity X must posses more Q than entity Z. Provided that the magnitude of Variable Q possessed by the entities is stable over time, transitivity should be satisfied even if the measurements are taken at different times. If the values produced by a procedure designed to measure the magnitudes of Variable Q violate transitivity for any three entities, it suggests that a) the measurement procedure is invalid (i.e., it fails to represent actual ordinal or quantitative relations onto Variable Q), b) Variable Q does not permit ordinal structure (e.g., it is not a unidimensional attribute) or c) if the measurements were not obtained simultaneously, the magnitude of Variable Q possessed by entities X, Y and Z is not stable over time. In any of these events, the values produced by such a measurement procedure cannot be treated as quantities of a common Variable Q.

Consider for example, a case in which a ruler (i.e., a standard of length) is used to conduct pairwise ratings of the height of three persons, Tom, Dick and Harry. When the ruler is used to compare Tom and Dick, the procedure indicates that Tom is taller than Dick; when the ruler is used to compare Dick and Harry, the procedure indicates that Dick is taller than Harry;

when the ruler is used to compare Tom and Harry, the procedure indicates that Tom is *shorter* than Harry, a value that violations the expectation of transitivity. This suggests that a) the ruler has changed over time (i.e., the measurement standard is inconsistent across ratings), or b) the height of one or more of the three persons shifted from the first pairwise comparison to the second (i.e., height is not stable over time). Whatever the reason, we cannot treat the values obtained using this procedure as quantities of height (i.e., length) without further qualification.

In the case of psychometric procedures in which people are asked to make self- and/or other-ratings on a scale with respect to psychological attributes, like personality "traits", we are essentially handing them a ruler (i.e., the scale itself) and providing training as to how the ruler should be applied (i.e., the instructions). If we are to treat these ratings as representations of quantitative relations with respect to the attribute in question, we *must* justify the following assumptions: 1) the attribute, as stated in the item, evokes a unidimensional attribute possessed by self and others and 2) that raters are able to rank themselves and others consistently with respect to said attribute. We can demonstrate grounds for these assumptions at the level of the individual by showing that individual raters can make transitive ordinal judgments (i.e., "more," "less," or "equal") of the self and others with respect to psychological attributes. Examining the transitivity of trait-ratings strips away the unnecessary and premature complexity of existing trait measurement procedures that were never designed to represent individual trait relations in the first place to expose the answer to the very basic question of whether people are able to produce trait rankings that are logically consistent with the assumptions of quantitative measurement.

If, as trait theorists believe, a person is called "impulsive" because he or she is objectively impulsive (e.g., he/she performs many impulsive acts and/or few acts that are not impulsive), and one person is called "more impulsive" than another because the former actually performs more impulsive acts and fewer non-impulsive acts than the latter, then subjective judgments of the relative "impulsivity" of the self and known others conducted in a single sitting should satisfy the assumption of transitivity provided that the rater does not observe the ratees

engaging in any "impulsive" or notably "non-impulsive" behavior during the actual rating task (i.e., objective "impulsivity" does not change.) If, however, subjective judgments of relative impulsivity violate the assumption of transitivity, it suggests that the rater's standard of "impulsivity" changes during the sitting; in either case, subjective ratings of "impulsivity" given by the rater cannot be treated as ordinal relations with respect to impulsivity and they certainly cannot be treated as quantitative relations with respect to impulsivity. Transitive trait ratings are equally anticipated by the slightly different assumption that people can order self and others with respect to traits *qua* dispositions to engage in certain behaviors rather than actual engagement in certain behaviors.

In order to test the implicit assumption of transitivity behind the treatment of NEO Neuroticism-domain ratings as quantitative relations, participants were asked to make a series of pair-wise rankings of themselves and known others with respect to both representative items from each NEO Neuroticism sub-domain and each of three broad definitions of Neuroticism as a general personality dimension. The transitive consistency of their pairwise rankings was then examined in the context of triplet relationships. A conditional hypothesis was associated with this task: If trait ratings are a valid representation of ordinal relations, then pairwise comparisons would satisfy the assumption of transitivity.

Question Two: Inter-definition Rank Consistency

Several definitions of Neuroticism as a general personality dimension appear in trait literature. At face value, these definitions are ambiguous. Not only does this make it difficult to ascertain the intended referent of a single definition, it makes the practice of using multiple definitions interchangeably highly questionable. It may be the case that these seemingly ambiguous definitions are *practically* synonymous, but this is a question that can and must be addressed empirically. If multiple definitions communicate the same referent, a person who is asked to rank or rate people with respect to each definition should produce identical relations along each definition. If a single rater ranks or rates people differently with respect to multiple

descriptions of what is purportedly a single construct, it suggests that the descriptions do not have identical meaning for that rater. This simple, straightforward test of inter-item consistency exposes entity-level relations that are obscured by measures of statistical convergence.

Participants in this study ranked all known others along with themselves on each of the three broad Neuroticism definitions simultaneously (i.e., all persons at the same time with respect to three different constructs in turn). Individual proportions of rank matches across all possible pairings and across all definitions were calculated. The conditional hypothesis associated with this task was as follows: If the broad Neuroticism definitions evoke the same construct in participants regardless of their apparent ambiguity, the rank positions of persons should be consistent across all of them.

Summary

The procedures used in the present study were design a) to test explicit and implicit measurement claims that have been made regarding self- and other-report ratings with respect to NEO Neuroticism-domain attributes and b) to test whether multiple broad-level definitions of Neuroticism taken from the NEO literature can be used interchangeably on the basic of the as-yet untested assumption that they are practically synonymous. In addition to addressing these specific questions, these procedures used here are offered as methods that may be valuable as components of a systematic effort towards rehabilitating the pathological science of psychometrics.

CHAPTER III

METHODOLOGY

Participants

A total of 147 participated in this study. All were undergraduate students at Oklahoma State University and received course credit in exchange for their participation. One-hundred and sixteen participants were female, and the remaining 31 participants were male. The majority of participants (67.3%) reported their ethnicity as Caucasian, 15.6% as African American, 2.7% as Asian, 4.15% as Hispanic, 6.8% as Native American, and 3.4% as Other. Age ranged from 18 to 50, with the majority of participants between 18 and 21 (M = 18.27, SD = 3.28).

Instruments and Methods

Name Elicitation Procedure

A Name Elicitation Procedure was used to generate a list of names to be used in the pairwise and complex ranking procedures (Appendix A). Participants were asked to provide the first names or nicknames of 7 people known personally to him/her. Participants were also asked to indicate a) the sex and age of each person and b) how well he/she knows the person using the following categories: 1) "I know this person extremely well," 2) "I do not know this person extremely well, but he/she is more than an acquaintance," and 3) "This person is an acquaintance." Participants were instructed to indicate each person's exact age when it was known and to estimate his/her age as closely as possible when the exact age was unknown.

Pairwise Ranking Task

Pairwise rankings were generated using a 364-item ranking task in which every pairwise combination of the 7 people listed on the Naming Elicitation Procedure along with "Myself" were compared with respect to 10 Neuroticism-domain Attributes including 5 positive markers (e.g., panics easily) and 5 negative markers (e.g., rarely feels irritated) taken from the International Personality Items Pool (IPIP) 20-item NEO Neuroticism-domain scale, and 3 general descriptions of Neuroticism (e.g., prone to psychological distress) for a total of 13 NEO Neuroticism-domain attributes (Appendix B).

Each item consisted of the presentation of two names drawn randomly from a pool of names along with one of the 13 Neuroticism-domain attributes. The ranking task was computerized, and the items appeared one at a time on a monitor in the following format:

Consider the attribute: **Prone to psychological distress**
Who possesses **MORE** of this attribute, Person 1 or Person 2 **(**If they exhibit the attribute equally, select Equal.)

Prone to psychological distress

□ Person 1 □ Person 2 □ Equal

Enter

Use "A," "S," "D" and Enter keys to make ratings faster.

Participants were always asked to indicate which of the persons shown possessed/exhibited *more* of the attribute; that is, they were never asked to indicate which of the person possessed/exhibited *less* of the attribute. This was done to minimize judgment errors that might occur as the result of misreading the item. A progress bar also appeared at the bottom of the screen to allow participants to track their progress on the task. Participants were instructed that they could use the computer mouse or the keyboard to make ratings. Participants were also informed they could change their response as many times as they wanted before proceeding to the next item, but that they would not be able to return to a previous item.

Dynamic Analog Scale (DAS): Complex Ranking Procedure

The DAS is a technique developed by Grice, Jackson and Badzinski (2011) for generating single-item measures for various personality traits. A single DAS scale is comprised of an analog-type scale anchored by trait definitions onto which participants use a computer mouse to "drag and drop" names, creating simultaneously ratings of all rated persons. Three DAS scales were devised for the purpose of this study each anchored by the three broad Neuroticism Definitions that appeared in the pairwise ranking procedure. Each vertical scale was anchored at the top by the phrase "The MOST 'Neuroticism Definition' person imaginable" and at the bottom by the phrase "The LEAST 'Neuroticism Definition' person imaginable." This phrasing was selected to avoid altering in the meaning of the phrases. For example, it is unclear whether the phrase "the least emotionally unstable person imaginable" is semantically equivalent to the phrase "the most emotionally stable person imaginable." The scale midpoint was also indicated. Participants were instructed to "drag and drop" each name onto the point on the scale that he/she believed best represented the extent to which each person exhibited the relevant attribute. For the purposes of this study, scale positions were converted into ranks.

Procedure

After providing informed consent, each participant completed a brief demographic questionnaire followed by the Name Elicitation Procedure. When participants completed the Name Elicitation Procedure, a researcher entered the seven elicited names along with "Myself" into the pairwise ranking program and the DAS complex rating program. Participants then completed the Pairwise Ranking Procedure, followed by the DAS Complex Ranking Procedure in private workspaces separated by dividers. Up to 4 participants completed the study at any given time during data collection. Following completion of all tasks, each participant was given a written debriefing statement explaining the purpose of the study.

Planned Analyses

Identifying Transitivity Violations

The pairs associated with each NEO Neuroticism-domain attribute were organized into all possible triplets (total 56) and examined for transitivity violations, or relationships that are inconsistent with the assumption that pairwise rankings are valid representations of real relationships with respect to a single, stable attribute permitting at least ordinal relations. The process of coding and classifying triplet patterns as transitive or intransitive was programmed into the Pairwise Ranking Procedure.

Transitivity Violation Indices

A transitivity violation index (i.e., TVI = # Intransitive Triplets/ (# Transitive Triplets + # Intransitive Triplets) was computed for 1) each Neuroticism-domain Attribute and broad Neuroticism Definition for all participants and 2) all Neuroticism-domain Attributes and broad Neuroticism Definitions for all participants. Participant by Attribute and Definition Transitivity Indices were also computed and summarized. While a valid ranking procedure on a supposedly ordinal/quantitative measurement procedure should produce values that satisfy transitivity across all rated entities, calculating the TVIs of various subgroups allows us to explore whether characteristics of triplets and/or constructs might facilitate ordinal consistency.

DAS Complex Ranking Procedure

The consistency of rankings across the three broad-level Neuroticism definitions were summarized in terms of 1) the proportion of names, including "Myself," that appear in the same rank position on all possible pairs of complex ranking tasks and 2) the proportion of names, including "Myself," that appear in the same rank position on all three complex ranking tasks. Only identical ranks across scales (i.e., zero difference) were considered as "matches."

Post hoc Analyses

A number of post hoc analyses were conducted to address three questions:

1) Are Neuroticism ratings any more or less consistent when triplets contain "Myself" than when triplets do not contain "Myself"?

The belief that self-report measures provide meaningful relations among multiple persons relies implicitly on the assumption that people are able to consider themselves in relation to others. It may be, however, that people take different information into account when comparing themselves with another person as opposed to comparing two other people. By definition, each triplet ranking pattern will include at least one pairwise judgment between two non-self persons. For this reason, it is unclear if we should anticipate greater or fewer transitivity violations associated with self-included triplets than with triplets that do not include the self. Nonetheless, separate TVIs were calculated for both of these groups as the results may be of interest.

2) Do triplets in which all ratees are matched with respect to sex and age yield lower rates of intransitivity than triplets in which ratees are unmatched with respect to sex and/or age?

It may be the case that certain attribute phrases evoke a different standard in the context of pairs that are similar with respect to sex and age than in the context of pairs that are dissimilar with respect to sex and/or age. For example, gender and/or age-related expectations may prompt a person to employ a different standard of "Not easily frustrated" when comparing a sex-congruent pair than comparing a sex-incongruent pair, or draw on different behavioral information in determining whether two adults are equally described by "remains calm under pressure" than in determining whether an adult and a child are equally described by the same phrase. If this is the case, shifting standards may contribute to higher levels of inconsistency among triplets that are unmatched with respect to sex and/or age.

3) Are transitivity violations rarer among triplets representing the most ideal rating scenarios appearing in this study?

Personality scales sometimes instruct raters to consider themselves in relations to others who are similar to them in sex and age. Inasmuch as this is intended to increase the validity and inter-person comparability of self-ratings, sets of pairwise ratings between "Myself" and similar

others may be expected to facilitate transitive consistency. For this reason, triplets with these characteristics were considered as possible "best case" ratings scenarios (i.e., "Best Case Scenario I" triplets). The context group intended by these instructions is somewhat vague; there may, for example, be observable differences between self-ratings made in the context of similar others who are well-known to the rater as opposed to similar others who are not well-known to the rater. In anticipation of potentially observable differences between these groups, self-included triplets in which the two non-self ratees were both similar to the rater with respect to sex and age and "extremely well known" to the rater were also considered as a potential "best case" ratings scenario (i.e., "Best Case Scenario II" triplets). Separate TVI's were calculated for "Best Case Scenario I" triplets and all other triplets as well as "Best Case Scenario II" triplets and all other triplets.

All post hoc analyses were conducted using the Proportional Analysis feature of Observation Orienting Modeling (OOM) software, a system developed by Grice (2010), to assess the degree of conformity between the deep structure matrices of observations on multiple orderings. Proportional Analysis is a feature of Observation Oriented Modeling (OOM) software (Grice, 2011) that compares a predicted pattern of proportions to an observed pattern of proportions. Proportional analysis is similar to a chi-squared goodness-of-fit test in that it calculates the difference between the expected and predicted proportions of each variable level OOM also calculates the probability that the proportion of correctly classified observations (PCC) is due to chance by comparing it to the proportions of correctly classified observations in randomized versions of the data. Through a repeated (e.g., 500 to 1000 times) two-step process of 1) generating randomized data and 2) calculating the PCC between the predicted pattern and the randomized data, OOM produces a chance value, or c-value, which is the proportion of times in which the PCC in a random data set exceeds the PCC in the observed data.

CHAPTER IV

FINDINGS

Pairwise Ranking Procedure

The Transitivity Violation Index for all triplets for all participants on all Neuroticism constructs (i.e., both the 10 Neuroticism-domain Attributes and the 3 broad Neuroticism definitions) was .22. The TVIs for all triplets for all participants on each Neuroticism-domain Attributes ranged from .20 to .24, and the TVIs for all participants on each of the broad Neuroticism definitions ranged from .21 to .23 (Table 2).

The mean Individual TVIs (i.e., participant x construct) on each of the Neuroticism-domain Attributes ranged from .21 (Attributes 5 and 7; SDs = .14 and .15, respectively) to .25 (Attribute 10; SD = .16); and the Individual TVIs on each of the broad Neuroticism definitions ranged from .21 (Definition 3; SD = .16) to .24 (Definition 1; SD = .15). On all Neuroticism constructs, the minimum TVI was .00, while the maximum TI was .66. The mean of the overall Individual TVIs (i.e., each person on all Neuroticism constructs) was .23 (SD = .11, Range = .05 to .55; Table 3).

(Attribute 10; SD = .16); and the Individual TVIs on each of the broad Neuroticism definitions ranged from .21 (Definition 3; SD = .16) to .24 (Definition 1; SD = .15). On all Neuroticism constructs, the minimum TVI was .00, while the maximum TI was .66. The mean of the overall Individual TVIs (i.e., each person on all Neuroticism constructs) was .23 (SD = .11, Range = .05 to .55; Table 3).

A summary of Individual TVIs was also prepared separately for Males (N = 26) and Females (N = 108). The mean TVIs for Females were comparable to the means for all subjects ranging from .20 (Attribute 5; SD = .14) to .24 (Attribute 6; SD = .14; see Table 4). The mean TVIs for Males ranged from .18 (Attribute 7; SD = .15) to .29 (Attribute 2; SD = .18), and the highest minimum TVIs for males were .36, on Attributes 4 and 5 and Definition 1 (see Table 5).

Post Hoc Analyses of Transitivity Violation Indices

Proportional analyses were conducted to ascertain the extent to which the TVIs associated with triplets including "Myself" could be predicted using the TVIs of the remaining triplets (e.g., all triplets not including "Myself"). The proportional analyses for transitivity violations on each of the 10 NEO Neuroticism Domain Attributes and each of the 3 Broad Neuroticism Definitions yielded a high match rate and low chance value (all PCCs ≥ .97and all c-values = .00 based on 500 randomizations of the Myself-included dataset) revealing that triplet patterns were equally inconsistent in triplets that included "Myself" and that that did not include "Myself" (see Table 6).

Proportional analyses were conducted to ascertain the extent to which the TVIs associated with triplets that were matched on both sex and age (i.e., Sex and Age Congruent Triplets) could be predicted using the TVIs of all triplets that were incongruent with respect to either sex or age. Pairs were said to be matched with respect to age if the age differences was equal to or less than 5 years; triplets were said to be matched with respect to age if all involved pairs were matched in age. The proportional analyses for transitivity violations on each of the 10 NEO Neuroticism Domain Attributes and each of the 3 Broad Neuroticism Definitions yielded a high match rate and low chance value (all PCCs ≥ .95 and all c-values = .00 based on 500 randomizations of the Sex and Age Congruent dataset) revealing that triplet patterns were equally inconsistent in triplets that were matched and unmatched with respect to Sex and Age (see Table 7).

Proportional analyses were conducted to ascertain to extent to which the TVI's associated with "Best Case Scenario I" triplets (i.e., self-included, matched with respect to sex and age) could be predicted using the TVIs of all triplets not meeting these criteria. The proportional analyses for transitivity violations on each of the 10 NEO Neuroticism Domain Attributes and each of the 3 Broad Neuroticism Definitions yielded a high match rate and low chance value (all PCCs ≥ .91 and all c-values = .00 based on 500 randomizations of the "Best Case Scenario I" dataset; Table 8) indicated that triplets satisfying these criteria were no more consistent than others.

Finally, a series of proportional analyses were conducted to ascertain the extent to which the TVIs associated with "Best Case Scenario II" triplets (i.e., self-included triplets, all persons matched with respect to sex and age, both non-self ratees are "extremely well known" by the rater) could be predicted using the TVIs of all triplets not meeting these criteria. The proportional analyses for transitivity violations on each of the 10 NEO Neuroticism Domain Attributes and each of the 3 Broad Neuroticism Definitions yielded a high match rate and low chance value (all PCCs ≥ .91 and all c-values = .00 based on 500 randomizations of the Best Case Scenario dataset) revealing that triplet patterns were equally inconsistent in these triplets as in the remaining triplets (see Table 9). The 91% PCC-values, the lowest frequency match rates found in all of the proportional analyses, reflect a slightly greater incidence of intransitivity among the "Best Case Scenario I and II" triplets than among the remaining triplets; while a small effect, the direction of this difference is contrary to what we might expect if these were, in fact, representative of ideal ratings scenarios.

Inter-definition Rank Consistency

The overall match rate for ranks on the broad Neuroticism Definitions 1 and 2 (i.e., "Emotionally unstable" and "Sensitive, emotional and prone to feelings that are upsetting, respectively) ranged from .00 to .75 (i.e., zero matches to six out of eight matches; M = .29, SD = .17). The overall rank match rate on broad Neuroticism Definitions 1 and 3 (i.e., "Emotionally

unstable" and "Prone to psychological distress," respectively) ranged from .00 to 1.00 (i.e., zero to eight out of eight matches; M = .32, SD = .20). The overall rank match rate on Definitions 2 and 3 ranged from .00 to .875 (i.e., zero matches to seven out of eight matches; M = .29, SD = .19). Finally, the overall rank match rate on all three broad Neuroticism definitions ranged from .00 to .50 (M = .15, SD = .14; Table 10). An examination of rank matches by sex revealed similar patterns across Males and Females (Table 11), and a post hoc proportional analysis revealed no notable differences across sex (all PCCs $\geq$ .79, all c-values = .00; Table 12).

CHAPTER V

CONCLUSION

It has been argued that psychometrics qualifies as a pathological science on the basis that the prevailing psychometric paradigm prevents the attainment of its stated goals and conflicting interests prevent its constituents from recognizing that this is the case. Michell (1997) has made this argument from the perspective that the goal of psychometrics is the measurement of psychological attributes, where measurement is properly defined as the "the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute," p. 358, and where a quantitative attribute is a range of properties or relations that admits infinite variation in degrees of magnitude (Ellis, 1996; Michell, 1990; Trendler, 2009). Having proposed that measurement is the representation of an empirical system, I concur with Michell that psychometrics is pathological, but do so on the grounds that psychometricians and people who employ psychometric procedures have failed to demonstrate that psychological measures produces valid representations of any empirical system, quantitative or otherwise. There are several ways in which the prevailing psychometric paradigm keeps us in the dark with respect to what the values produces by psychometric procedures represent. Among them is the tendency to use ambiguous language (e.g., does a person who worries frequently, but functions well "worry a lot" in the same way as a person who worries infrequently but is, at those times, consumed with worry?) and the tendency to overlook or disregard the implications of scale labels (e.g., if a person "agrees somewhat" with a statement does that person also "disagree somewhat"

” with that statement? If so, how do we justify assigning different values to these responses?) This inattention to item/scale construction makes it difficult to determine what if any relational system is targeted by a measurement procedure. Additionally, psychometric procedures in which a single person rates multiple ratees is either done using multiple items (i.e., separate ratings for each person) or on a single item that may obscure inconsistencies in judgment (i.e., a forced ranking item). This prevents us from obtaining values that we might test against the appropriate axioms of measurement. The obfuscation does not stop with these. Proponents of the use of item response theory in the construction of personality scales present a baffling argument that peoples' ability to use a pencil to circle a response on a questionnaire are limited to some extent with respect to which responses they are able to circle by some internal quantitative property (e.g., Neuroticism). Perhaps the most self-defeating psychometric practice is that of "establishing psychometric validity" through the demonstration of statistical convergence and divergence of the values produced by measurement procedures. This might be adequate if the goal of measurement is to represent generic aggregate variance onto generic statistical functions, but if the goal of measurement is the representation of the relations of real entities with respect to real qualities this paradigm is inadequate as real qualities are defined by their essential characteristics.

Personality is one among many areas in which researchers claim to measure psychological attributes at the level of the individual, but have not addressed basic measurement concerns. This study was designed to test two necessary but insufficient conditions that must be met before self- and/or other-report ratings with respect to NEO Neuroticism-domain attributes can be treated as quantities. Specifically, it tested the assumption that people are willing and able to make pairwise ordinal comparisons that are consistent with the assumption of transitivity at the triplet level. It also tested the assumption that apparently ambiguous definitions of Neuroticism that are used interchangeably in the NEO literature are practically synonymous among raters (i.e., lay raters). In addition to addressing these specific questions, these procedures are offered as

methods that may be valuable as components of a systematic effort towards rehabilitating the pathological science of psychometrics.

Triplet Transitivity

Transitivity is a basic assumption of quantitative measurement. That is, if a measurement procedure is successful in its aim to represent ordinal or quantitative relations, comparative judgments of pairs with respect to the target variable will be consistent when considered at the triplet level. In the current study, participants were asked to conduct head-to-head comparisons (e.g., greater than, less than) of all possible pairwise combinations of people they know personally plus "Myself" with respect to 10 NEO Neuroticism-domain Attributes and 3 broad Neuroticism definitions found in the NEO literature. These pairwise judgments were considered in the context of all possible triplet combinations and examined for transitivity. Roughly 20% to 25% of the triplets across all subjects were found to be intransitive. The data were divided into several subgroups in turn based upon characteristics of the triplets that we believed might make consistency more/less likely. Triplets including "Myself" were intransitive almost exactly as often as triplets not including "Myself," sex and age congruent triplets were intransitive almost exactly as often as sex and/or age incongruent triplets. Finally, triplets a) involving "Myself" and two other people of the same sex and similar in age and b) involving "Myself" and two other people of the same sex and similar age whom the rater indicated were "extremely well known" to him/her--we believed these rating scenario were most likely to facilitate consistency--were intransitive almost exactly as often as all other triplets.

These results indicate that individuals are often inconsistent when making pairwise comparisons of people they know personally with respect to Neuroticism-domain constructs. Some participants were consistent on some of the triads with respect to some of the Neuroticism attributes, but no participant made a completely consistent set of ratings. This suggests that 1) the meaning of the attributes sometimes or always changed from one pairwise comparison to another or 2) the extent to which the known persons possessed or exhibited the attributes sometimes or

always changed from one pairwise comparison to another or that a combination of these two events occurred either simultaneously or in turn. The objective body of behaviors produced by the rated people did not change as the participants did not observe them engaging in any new behaviors during the course of the study, the exception being that the participants could monitor their own feelings and behaviors as they completed the procedures.

Implications

What is the significance of these findings for personality research that relies on Neuroticism scores taken from self-report procedures? At least one implication of intransitivity seems clear: if people cannot make consistent judgments on a simplified task like this, we would be hard pressed to produce justification for treating values produced by any procedure that assumes people are both able and motivated to do so as representations of quantitative relations. In this study, 22% of pairwise ratings made by participants were inconsistent at the triplet level, meaning that 78% *did* satisfy transitivity. Some might argue that this is an acceptable error rate, thus, good news for studies that treat trait measures as quantities, but this argument does not hold up. First, if a single pairwise relationship is altered in an inconsistent triplet, it could potentially alter the relationships in many other triplets some of which were previously consistent. Second, transitivity is a *necessary but insufficient* condition of quantitative structure. By definition, attributes that satisfy the conditions of quantitative structure also satisfy the conditions of ordinal structure, but we cannot conclude that a measurement procedure produces quantitative values simply because it produces values that satisfy the condition of transitivity.

Limitations

It is possible that the inconsistencies found in this study are attributable, at least in part, to particularities of the procedures rather than the inability of the participants to produce consistent ordinal relations with respect to the relevant constructs. The order of the items on the pairwise rating task was randomized in order to minimize the participants' ability to make consistent judgments by "keeping track" of their rating patterns instead of evaluating each pair with respect

to the attributes. If participants had been asked to make all possible pairwise ratings with respect to a single attribute before moving onto the next attribute, they may have been less likely to switch constructs from comparison to comparison. Participants were able to change their answers on each pairwise judgment an unlimited number of times, but were unable to return to a previous item once they selected "Enter." Again, this was done to minimize the participants' ability to produce artificially consistent judgments, but if participants pressed "Enter" too quickly and were unable to correct the mistake, it may have resulted in inconsistent triplet patterns even when participants were able to make consistent judgments. In the future, it may be advisable to have participants confirm their response to an item one additional time before moving to the next judgment. The language used in the task may also have affected the results. Trait phrases like "easily frustrated" read smoothly in the context of a Likert-type scale, but can be awkward in a pairwise ranking procedure. From a trait-theory perspective, it is appropriate to talk about people exhibiting or possessing trait-related behaviors and/or dispositions, but that is not necessarily how people think about trait phrases as they appear in natural language. For the purposes of this study, the trait phrases were taken from an existing measure and intentionally left unaltered in order to avoid disambiguating the language, but researchers using this method in the future may choose to change the language. For example, if the intent is to evaluate the transitivity of the relative ratings of belief or agreement that an attribute phrase describes ratees, the items may be adapted to reflect this intent (e.g., "Which of these people do you think is best described by the phrase: 'easily frustrated'?"). This format does not require researchers to alter the trait phrases that are currently used in Likert-type trait scales, but the responses are not readily interpretable as relations of rated persons or entities with respect to properties possessed or exhibited by them. If the target empirical system is the actual degree to which ratees exhibit a trait-related behavior, it might be more appropriate to adapt phrases as follows: "Who is more easily frustrated?" or "Who is more likely to remain calm under pressure?" If the target empirical system is the degree to which ratees possess the disposition to engage in a certain behavior, it might be more

appropriate to adapt the phrases in a manner similar to this: "Which of the following people is most likely to become frustrated" or "Which of the following people is most likely to remain calm under pressure?" In any event, phrasing adaptations should be carefully chosen to reflect the target system. One benefit of needing to attend carefully to item phrasing is that it requires researchers to clarify the constructs they wish to measure.

Inter-definition Rank Consistency

In addition to the pairwise rating procedure, participants were asked to rank themselves and seven known others with respect to three broad definitions of Neuroticism that appear in the NEO literature: 1) emotional instability, 2) sensitivity, emotionality and proneness to experience feeling that are upsetting, and 3) proneness to psychological distress. Rank matches were calculated for each pair of descriptions and for all three descriptions. The mean rank match rate for all pairs ranged from 29% to 32%, and the overall mean rank match rate was 15%, suggesting that, for the majority of participants, these definitions do not evoke the same construct.

Implications

These results suggests that the language used by Neuroticism researchers are not practically synonymous among lay people (i.e., people who are not familiar with trait jargon), which in turn suggests that trait Neuroticism researchers have not been successful in capturing important differences in the language that people use to describe the personalities of themselves and others along this "general dimension."

Limitations

The DAS procedure that was used to obtain the ranks was not a forced rank procedure. Participants "dragged and dropped" names onto an analog-type scale, and the scale positions were then converted into ranks. While it is highly unlikely that participants intended to place participants at different scale points but placed them close enough together on the scale that they were assigned the same rank, it is possible that participants intended to place one or more names at the same scale point, but placed them far enough apart that they were assigned different ranks.

Researchers using this general method (i.e., inter-definition rank consistency) in the future may choose to alter the scale precision or to use another method to obtain rank-values.

Conclusion

When we ask participants to make self-ratings on a scale with respect to a psychological attribute, we attempt to communicate a standard (i.e. the scale and attribute in combination) and we assume that their answers are meaningful. We do not have full faith in the judgment of participants, as we assume that all measurement procedures are subject to error, but we believe that statistical procedures allows us to separate the wheat from the chaff (i.e., variance attributable to true departures from the mean from variance attributable to error). Unfortunately, all statistical analyses presume that the numbers we feed into them represent certain relations and it is our knowledge of the kinds of relations they represent that guides our selection of statistical procedures. A chi-squared test, for example, presumes that all people who indicate that they are "depressed" share an understanding with one another and with the researchers of what it means to be "depressed." Various ordinal procedures assume that people who indicate they "Strongly Agree" with this or that statement a) share a common standard of what it means to strongly agree, b) share a common understanding of what it is with which they strongly agree, and c) agree more strongly with the statement than people who merely "Agree" with it. That is, in order to distill measurement error from the values produced by psychometric procedures, we have to begin with valid measurements. Quantitative procedures presume even more knowledge on the part of researchers than many ordinal procedures as they are meaningful only when applied to values representing quantities. Procedures that treat ratings as ordinal relations comparable across raters (e.g., all people who are assigned a rank of "4" are equal with respect to the target) require additions assumptions beyond those that assume only intra-rater ordinality.

The current study addressed some basic concerns regarding the measurement of Neuroticism that cannot be addressed using the prevailing psychometric vetting procedures. When asked to make pairwise comparisons of themselves and others with respect to Neuroticism-

domain attributes, participants often made intransitive judgments. These results may be attributable to some degree to particularities of the format of the pairwise ranking procedure; nonetheless, this is the kind of task that is required to justify the assumption that self- and/or other-ratings may be treated as representations of ordinal values. We are, therefore, unjustified in continuing to treat self- and/or other-ratings with respect to these or any attributes as ordinal or quantitative relations based on any real or perceived limitations of this study. Self- and other-rankings with respect to three broad-level definitions of Neuroticism were also found to be highly inconsistent across definitions suggesting that these definitions are not practically synonymous—at least among lay populations, and should not be treated as such. Again, these results may be attributable, in whole or in part, to particularities of the ranking tasks that were used or to particularities of the sample; however, this is the kind of task that is required to justify using multiple terms interchangeably. If it is only one's intent to use terms interchangeably when communicating with a specific audience, then the practical equivalence of the terms need only be demonstrated with that audience, but it should be demonstrated.

The methods used in this study are not just useful for providing evidence for or against assumptions about existing measures. They may also be used to help build valid measures and increase the conceptual clarity of any number of states or attributes. If we want to count things, we need to begin with some idea of what constitutes a "thing." If a person is asked to compare the number of "things" in a set of baskets and we find that the task produces values that are intransitive, we are presented with an opportunity to explore that persons' understand of a "thing." We can then begin to adapt our understanding of thing-ness to conform to their understanding and/or adapt their understanding to conform to ours until the procedure does produce consistent responses. That is, by using tasks like these in conjunction with conversation with raters, we stand to improve the precision of our measures by increasing our conceptual precision (i.e., our understanding of target systems). By shifting the focus of measurement efforts away from the reduction of non-specific aggregate error and onto the reduction of conceptual

ambiguity, we may acknowledge the intimate relationship between quality and quantity to our great benefit. One potential outcome of this paradigm shift is that researchers may begin to take more ownership of the measures they use; that is, they may be encouraged to develop measures and procedures that answer the questions they want to ask rather than relying on psychometric formats and procedures that work against them. Statistical procedures provide probabilistic answers; a premature reliance on statistics may have led psychologists into a trap of believing that the answers to all of their questions are ultimately unknowable. If psychological scientists adopt this attitude, the science of psychology is defeated before it begins. We may not be able to say with certainty whether a person will or will not be in a specific state at a specific time, but it is certainly within our reach to define and recognize the states that concern us and this is ultimately what the prevailing psychometric paradigm prevents us from doing.

REFERENCES

Barrett, P.T. (2003) Beyond Psychometrics: Measurement, non-quantitative structure, and applied numerics. *Journal of Managerial Psychology*, *3*(18), 421-439.

Barrett, P. (2008). The consequence of sustaining a pathology: Scientific stagnation. A commentary on the target article 'Is psychometrics a pathological science?' by Joel Michell. *Measurement*, *6*, 78–123.

Boag, S. (April 2011). Explanation in personality psychology: Verbal magic and the Five-Factor Model, *Philosophical Psychology, 24*(2), 223-243.

Cloninger, C. R., Przybeck, R. & Svrakic, D. (1991). The Tridimensional Personality Questionnaire: U.S. normative data. *Psychological Reports*, *69*, 1047–1057.

Cloninger, C. R. (2000). A practical way to diagnose personality disorder: A proposal. *Journal of Personality Disorders, 14*, 98-108.

Cloninger, C.R., Svrakic, D.M., Pryzbeck, T.R., & Wetzel, T.R. (1994). The Temperament and Character Inventory (TCI): A Guide to its Development and Use. Center for Psychobiology of Personality, St. Louis, MO.

Costa, P. T., Jr. & McCrae, R. R. (1992). *The NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological* research. *New York: Cambridge University Press.*

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

of Psychology, 41, 417-470.

Ellis, B. (1966). *Basic concepts of measurement*. Cambridge, UK: Cambridge University Press.

Eysenck, H.J., & Wilson, G. (1991). *The Eysenck Personality Profiler,* 1st ed. Guildford: Psi-Press.

Ferguson, A., Myers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown,W., et al. (1940). Quantitative estimates of sensory events: Final report of the Committee Appointed to Consider and Report upon the Possibility of Quantitative Estimates of Sensory Events. *Advancement of Science, 1*, 331–349.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Grice, J. W. (2011). Observation Oriented Modeling: Analysis of cause in the behavioral sciences. Elsevier.

Guilford, J.P. (1975). Factors and factors of personality. *Psychological Bulletin, 82*, 802-814.

Harman, H.H., 1967, Modern factor analysis, University of Chicago Press: Chicago, IL.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass [The axioms of quantity and the theory of measurement]. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Classe*, *53*, 1–64.

International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (http://ipip.ori.org/). Internet Web Site.

Lamiell, J. T. (2003). *Beyond individual and group differences: Human individuality, scientific psychology and William Stern's critical personalism,* Sage Publications: Thousand Oaks, CA.

Lamiell, J. T. (*In Press*). Statisticism in personality psychologists' use of trait constructs; What is it? How was it contracted? Is there a cure? *New Ideas in Psychology* (2011), doi:10.1016/j.newideapsych.2011.02.009.

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

Luce, R.D., & Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.

McAdams, D. P. (1995). What do we know when we know a person? *Journal of Personality, 63*, 3, 365-396.

McCrae, R.R., & Costa, P.T. (2003). *Personality in adulthood: a five-factor theory perspective*. New York: Guilford.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology*, *10*, 639–667.

Michell, J. (2008a). Is Psychometrics Pathological Science? *Measurement, 6*, 7-24.

Michell, J. (2008b). Rejoinder. *Measurement, 6,* 125-133.

Mill, J.S. (1974). *A system of logic: Ratiocinative and inductive*. Toronto, ON: University of Toronto Press. (Original work published 1843).

Sher, G. (1999). Is there a place for philosophy in Quine's theory?, *The Journal of Philosophy 96,* 491-524.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Tellegen, A., & Waller, N. G. (2008). *Exploring personality through test construction: development of the Multidimensional Personality Questionnaire.* In G. J. Boyle,

G. Matthews, & D. H. Saklofske (Eds.), The Sage handbook of personality theory and assessment: Vol. II. Personality measurement and testing (pp. 261–292). London: Sage.

Watson, D., & Clark, L.A. (1994). *Manual for the Positive and Negative Affect Schedule: Expanded form*. Iowa City, Iowa: University of Iowa.

Watson, D. & Tellegen, A. (1985). Towards a consensual structure of mood. *Psychological Bulletin*, 98, 219–235.

Watson D, Clark LA, Tellegen A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070.

APPENDICES

Appendix A

Instructions: In the spaces below, please provide the first name, nickname, or initials of 7 people you know personally. For confidentiality purposes, please, do not include both first and last names. Beside each name, please indicate each person's sex (e.g., M = Male and F = Female) and approximate age. Below each name, please place a check next to the option that best describes how well you know the person.

	Name/Nickname	**Sex**	**Age**
1.	_____	_____	_____

___ I know this person extremely well.
___ I do not know this person extremely well, but he/she is more than an "acquaintance."
___ This person is an acquaintance.

	Name/Nickname	**Sex**	**Age**
2.	_____	_____	_____

___ I know this person extremely well.
___ I do not know this person extremely well, but he/she is more than an "acquaintance."
___ This person is an acquaintance.

	Name/Nickname	**Sex**	**Age**
3.	_____	_____	_____

___ I know this person extremely well.
___ I do not know this person extremely well, but he/she is more than an "acquaintance."
___ This person is an acquaintance.

Name/Nickname **Sex** **Age**

4. _____ _____ _____

___ I know this person extremely well.
___ I do not know this person extremely well, but he/she is more than an "acquaintance."
___ This person is an acquaintance.

Name/Nickname **Sex** **Age**

5. _____ _____ _____

___ I know this person extremely well.
___ I do not know this person extremely well, but he/she is more than an "acquaintance."
___ This person is an acquaintance.

Name/Nickname **Sex** **Age**

6. _____ _____ _____

___ I know this person extremely well.
___ I do not know this person extremely well, but he/she is more than an "acquaintance."
___ This person is an acquaintance.

Name/Nickname **Sex** **Age**

7. _____ _____ _____

___ I know this person extremely well.
___ I do not know this person extremely well, but he/she is more than an "acquaintance."
___ This person is an acquaintance.

Appendix B

Neuroticism-domain Attributes

Items taken from the 20-Item International Personality Item Pool (IPIP) NEO Neuroticism Domain Scale

1. Often feels “blue” (positive indicator)
2. Seldom gets mad (negative indicator)
3. Not easily frustrated (negative indicator)
4. Has frequent mood swings (positive indicator)
5. Panics easily (positive indicator)
6. Rarely gets irritated (negative indicator)
7. Remains calm under pressure (negative indicator)
8. Feels comfortable with himself/herself (negative indicator)
9. Fears for the worst (positive indicator)
10. Feels threatened easily (positive indicator)

General Definitions

11. Emotionally unstable
12. Sensitive, emotional and prone to experience feelings that are upsetting
13. Prone to psychological distress

Table 2
Overall Transitivity Violations for Each Attribute and Definitions

Attribute	Total Violations	Possible Violations	Transitivity Index
1	1650	7672	.21
2	1767	7672	.23
3	1764	7672	.23
4	1640	7672	.21
5	1561	7672	.20
6	1836	7672	.24
7	1559	7672	.20
8	1795	7672	.23
9	1657	7672	.21
10	1842	7672	.24
Definition			
1	1776	7672	.23
2	1763	7672	.23
3	1618	7672	.21

Table 3
Summary of Individual Transitivity Indices: All Subjects

Attribute	Minimum	Maximum	Median	Mode	Mean	SD
1	.00	.61	.20	.21	.22	.15
2	.00	.61	.21	.71	.24	.16
3	.00	.61	.20	.14	.23	.15
4	.00	.63	.20	.16	.22	.15
5	.00	.64	.18	.14	.21	.14
6	.00	.64	.24	.18	.24	.14
7	.00	.66	.18	.89	.21	.15
8	.00	.59	.21	.00	.24	.16
9	.00	.63	.18	.14	.22	.15
10	.00	.66	.23	.71	.25	.16
Definition						
1	.00	.57	.21	.11	.24	.15
2	.00	.64	.20	.18	.23	.16
3	.00	.64	.18	.89	.21	.16
Total	.50	.55	.22	.17	.23	.11

*Note: N = 134

Table 4
Summary of Individual Transitivity Indices: Males

Attribute	Minimum	Maximum	Median	Mode	Mean	SD
1	.00	.48	.18	.18	.22	.15
2	.00	.59	.19	.59	.26	.18
3	.00	.59	.27	.27	.29	.14
4	.36	.63	.16	.16	.21	.15
5	.36	.54	.16	.11	.23	.15
6	.00	.54	.27	.29	.26	.15
7	.18	.61	.13	.11	.18	.15
8	.00	.57	.27	.30	.26	.17
9	.18	.63	.18	.23	.22	.16
10	.18	.54	.24	.43	.25	.15
Definition						
1	.36	.52	.21	.11	.23	.13
2	.00	.57	.21	.34	.25	.16
3	.00	.59	.20	.23	.23	.15
Total	.06	.52	.20	.20	.24	.11

*Note: N = 26

Table 5
Summary of Individual Transitivity Indices: Females

Attribute	Minimum	Maximum	Median	Mode	Mean	SD
1	0	.61	.20	.21	.22	.15
2	.18	.61	.21	.71	.23	.15
3	0	.61	.18	.14	.22	.14
4	0	.61	.21	.71	.22	.14
5	0	.64	.18	.25	.20	.14
6	.18	.64	.22	.18	.24	.14
7	0	.66	.18	.23	.21	.14
8	0	.59	.21	.11	.23	.15
9	0	.61	.18	.18	.22	.15
10	0	.66	.23	.25	.24	.15
Definition						
1	0	.57	.21	.11	.24	.15
2	0	.64	.20	.18	.23	.16
3	0	.64	.18	.98	.21	.17
Total	.05	.64	.22	.17	.23	.11

*Note: N = 108

Table 6
Proportional Analysis: Self Included Triplets vs. Self Not Included Triplets

	Transitivity Index		Proportion	
Attribute	No Self	Self	No Self	Self
1	.21	.23	1.00	.98
2	.22	.25	1.00	.97
3	.23	.25	1.00	.98
4	.21	.23	1.00	.98
5	.20	.23	1.00	.97
6	.23	.27	1.00	.96
7	.20	.21	1.00	.99
8	.22	.26	1.00	.96
9	.21	.24	1.00	.97
10	.23	.27	1.00	.96
Definition				
1	.25	.25	1.00	1.00
2	.25	.25	1.00	1.00
3	.22	.22	1.00	1.00

*Note: No Self = triplets in which "Myself" was not included (N = 4704); Self = triplets in which "Myself" was included (N = 2835); All c-values = .00 based on 500 randomizations.

Table 7

Proportional Analysis: Sex and Age Congruent Triplets vs. Sex and/or Age Incongruent Triplets

	Transitivity Index		Proportion	
Attribute	Incon	Con	Incon	Con
1	.23	.23	1.00	1.00
2	.24	.24	1.00	1.00
3	.25	.25	1.00	1.00
4	.27	.26	1.00	.99
5	.22	.22	1.00	1.00
6	.25	.25	1.00	1.00
7	.20	.22	1.00	.98
8	.22	.27	1.00	.95
9	.27	.25	1.00	.98
10	.25	.28	1.00	.97
Definition				
1	.28	.27	1.00	.99
2	.27	.27	1.00	1.00
3	.23	.23	1.00	1.00

*Note: Incon = sex and/or age incongruent triplets (N = 6430); Con = sex and age incongruent triplets (N = 1109); All c-values = .00 based on 500 randomizations.

Table 8

Proportional Analysis: "Best Case Scenario I" Triplets vs. All Others

	Transitivity Index		Proportion	
Attribute	Not Best Case	Best Case I	Not Best Case	Best Case I
1	.22	.22	1.00	1.00
2	.23	.26	1.00	.97
3	.23	.30	1.00	.93
4	.21	.30	1.00	.91
5	.21	.22	1.00	.99
6	.24	.29	1.00	.95
7	.21	.23	1.00	.98
8	.24	.29	1.00	.95
9	.22	.26	1.00	.96
10	.24	.29	1.00	.95
Definition				
1	.23	.26	1.00	.97
2	.23	.27	1.00	.96
3	.21	.26	1.00	.95

*Note: Best Case I = self-included triplets, matched with respect to sex and age (N = 594); Not Best Case = triplets not satisfying the Best Case I criteria (N = 6946); All c-values = .00 based on 500 randomizations.

Table 9

Proportional Analysis: "Best Case Scenario II" Triplets vs. All Other Triplets

	Transitivity Index		Proportion	
Attribute	Not Best	Best Case II	Not Best	Best Case II
1	.22	.22	1.00	1.00
2	.23	.26	1.00	.97
3	.23	.30	1.00	.93
4	.21	.30	1.00	.91
5	.21	.22	1.00	.99
6	.21	.23	1.00	.98
7	.24	.29	1.00	.95
8	.24	.29	1.00	.95
9	.22	.26	1.00	.96
10	.24	.29	1.00	.95
Definition				
1	.23	.26	1.00	.97
2	.23	.27	1.00	.96
3	.21	.26	1.00	.95

*Note: Best Case II= self-included triplets in which all people are matched for age and sex and the two non-self ratees are extremely well-known (N = 389); Not Best = all remaining triplets (N = 7150); All c-values = .00 based on 500 randomizations.

Table 10
Proportion of Rank Matches on Broad Neuroticism Definitions: All Subjects

	Minimum	Maximum	Median	Mode	Mean	N
Definitions 1 & 2	0.00	.75	.25	.25	.29	.17
.Definitions 1 &3	0.00	1.00	.25	.125	.32	.20
Definitions 2 & 3	0.00	.875	.25	.25	.29	.19
All	0.00	.50	.125	.125	.15	.14

*Note: N = 143

Table 11

Males	Minimum	Maximum	Median	Mode	Mean	SD
Definitions 1 & 2	.00	.625	.25	.25	.29	.14
Definitions 1 & 3	.00	1.00	.25	.25	.30	.20
Definitions 2 & 3	.00	.875	.25	.375	.29	.20
All	.00	.375	.125	.125	.14	.12
Females						
Definitions 1 & 2	.00	.75	.25	.125	.29	.18
Definitions 1 & 3	.00	.875	.31	.125	.33	.20
Definitions 2 & 3	.00	.75	.25	.25	.29	.18
All	.00	.50	.125	.125	.16	.14

Proportions of Rank Matches on Broad Neuroticism Definitions by Sex

*Note: N for Males = 31

Table 12
Proportional Analysis of Match Count by Sex

	Definitions 1 & 2		Definitions 1 & 3		Definitions 2 & 3		All Definitions	
Matches	Male	Female	Male	Female	Male	Female	Male	Female
0	.03	.06	.07	.07	.13	.09	.32	.29
1	.20	.28	.19	.24	.19	.24	.32	.35
2	.35	.23	.35	.19	.26	.29	.26	.21
3	.26	.20	.23	.18	.26	.15	.10	.10
4	.10	.14	.10	.20	.10	.17	.00	.04
5	.03	.08	.00	.07	.00	.04	.00	.00
6	.00	.01	.03	.04	.03	.03	.00	.00
7	.00	.00	.00	.01	.03	.00	.00	.00
8	.00	.00	.03	.00	.00	.00	.00	.00
Proportion	1.00	.79	1.00	.76	1.00	.82	1.00	.93
C-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

*Note: C-values based on 500 randomizatons.

VITA

Stefanie Ilene Badzinski

Candidate for the Degree of

Doctor of Philosophy

Thesis: IMPLICATIVE DILEMMAS AND GENERAL PSYCHOLOGICAL WELL-BEING: PREDICTIVE VALUES OF THREE PROPOSED SUBTYPES

Major Field: Psychology, Lifespan Human Development option

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Psychology at Oklahoma State University, Stillwater, Oklahoma in December, 2012.

Completed the requirements for the Master of Science in Psychology at Oklahoma State University, Stillwater, Oklahoma in 2009.

Completed the requirements for the Master of Psychology at University of Dallas, Irving, Texas in 2006.

Completed the requirements for the Bachelor of Science in Psychology at Southern Nazarene University, Bethany, Oklahoma in 2004.