6-22-2011 12:00 AM

# The Utility and Feasibility of Metric Calibration for Basic Psychological Research

Etienne LeBel, *The University of Western Ontario*

Supervisor: Bertram Gawronski, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Psychology

# THE UTILITY AND FEASIBILITY OF METRIC CALIBRATION FOR BASIC PSYCHOLOGICAL RESEARCH

(Spine title: Metric Calibration in Psychological Research)

(Thesis format: Monograph)

by

Etienne P. LeBel

Graduate Program in Psychology

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO
School of Graduate and Postdoctoral Studies

## CERTIFICATE OF EXAMINATION

Supervisor

______________________________
Dr. Bertram Gawronski

Supervisory Committee

______________________________
Dr. Sampo V. Paunonen

______________________________
Dr. Richard Goffin

Examiners

______________________________
Dr. Richard Goffin

______________________________
Dr. Lorne Campbell

______________________________
Dr. James O’Brien

______________________________
Dr. Brent Donnellan

The thesis by

**Etienne Philippe LeBel**

entitled:

**The Utility and Feasibility of Metric Calibration for Basic Psychological Research**

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

______________________ Date

______________________________ Chair of the Thesis Examination Board

# Abstract

Inspired by the history of the development of instruments in the physical sciences, and by past psychology giants, the following dissertation aimed to advance basic psychological science by investigating the metric calibration of psychological instruments. The over-arching goal of the dissertation was to demonstrate that it is both *useful* and *feasible* to calibrate the metric of psychological instruments so as to render their metrics non-arbitrary. Concerning utility, a conceptual analysis was executed delineating four categories of proposed benefits of non-arbitrary metrics including (a) help in the interpretation of data, (b) facilitation of construct validity research, (c) contribution to theory development, and (d) facilitation of general accumulation of knowledge. With respect to feasibility, the metric calibration approach was successfully applied to instruments of seven distinct constructs commonly studied in psychology, across three empirical demonstration studies and re-analyses of other researchers' data. Extending past research, metric calibration was achieved in these empirical demonstration studies by finding empirical linkages between scores of the measures and specifically configured theoretically-relevant behaviors argued to reflect particular locations (i.e., ranges) of the relevant underlying psychological dimension. More generally, such configured behaviors can serve as common reference points to calibrate the scores of different instruments, rendering the metric of those instruments non-arbitrary.

Study 1 showed a meaningful metric mapping between scores of a frequently used instrument to measure need for cognition and probability of choosing to complete a cognitively effortful over a cognitively simpler task. Study 1 also found an interesting metric linkage between scores of a practically useful self-report measure of task persistence and actual persistence in an anagram persistence task. Study 2, set in the context of the debate of pan-cultural self-enhancement, found theoretically interesting metric mappings between a trait rating measure of self-enhancement often used in the debate and a specifically configured behavioral measure of self-enhancement (i.e., over-claiming of knowledge). Study 3 demonstrated the metric calibration approach for popular behavioral measures of risk-taking often used in experimental studies and found meaningful metric linkages to risky gambles in binary lottery choices involving the possibility of winning real money. Re-analyses of relevant datasets shared by other researchers also revealed meaningful metric mappings for instruments assessing extraversion, conscientiousness, and self-control. Gregariousness facet scores were empirically linked to number of social parties attended per month, Dutifulness facet scores (conscientiousness) were connected to maximum driving speed, and trait self-control scores were calibrated to GPA. In addition, to further demonstrate the utility of non-arbitrary metrics for basic psychological research, some of my preliminary metric calibration findings were applied to actual research findings from the literature. Limitations and obstacles of metric calibration and promising future directions are also discussed.

## Keywords

Metric calibration, arbitrary metrics, psychological units of measurement, psychological measurement, psychometrics.

# Acknowledgments

First and foremost, I would like to thank my advisor, Bertram Gawronski, for his guidance and knowledge, but most importantly his inspiration to think broadly, aim high, and follow one's scientific passions. I have learned so much from him and have benefited in many ways from his intellectual astuteness and openness, clarity of thought, and high standards of thoroughness and exactness in his work.

I would also like to thank my supervisory committee members, Sampo Paunonen and Rick Goffin, for their valuable input in the early stages of my dissertation. Many thanks to my other examiners, Lorne Campbell, James O'Brien, and Brent Donnellan for their helpful comments and suggestions. I also owe gratitude to Scott Leith for his assistance with data collection. Thanks also go to June Tangney, Ross O'Hara, Kenneth DeMaree, Joseph Ditre, and Sampo Paunonen for graciously sharing relevant datasets with me.

Many thanks go to graduate student colleagues – in particular Kurt Peters, Chris Wilbur, Paul Conway, and Yang Ye – who in recent years have engaged in constructive and interesting conversations about metric calibration which have ultimately clarified and improved my thinking on the subject. Many thanks also go to all my psychology friends for their support, camaraderie, and general positive energy. You have contributed greatly to this rewarding intellectual journey and I am grateful for that. Special thanks also go to my non-psychology friends who provided much needed diversions from the analytically relentless world of academic psychology.

Finally, I would like to thank my parents for their unwavering, emotional, financial, and intellectual support over the years. My father, for his mad-scientist genes and my mother for her detail-oriented and creativity genes. Without you, I would never have been able to achieve this goal. Sincere thanks also go to Michelle Lau and Theodore Cole for all the fun and creativity-inspiring adventures.

# Table of Contents

**CERTIFICATE OF EXAMINATION** ii

Abstract iii

Acknowledgments v

Table of Contents vi

List of Tables ix

List of Figures x

List of Appendices xii

Chapter 1 1

1 Introduction 1

1.1 The Nature of Metrics and Metric Arbitrariness 3

1.2 Reducing Metric Arbitrariness: Background 10

1.3 Main Empirical Strategies of Metric Calibration 12

1.3.1 Strategy 1: Noteworthy Differences in Behavior 12

1.3.2 Strategy 2: Gradation in Behavior 16

1.3.3 Strategy 3: Experimental Approach 17

1.4 Ideal Characteristics of Behavioral Reference Points 22

1.5 Inspirations from the History of the Development of Instruments in the Natural Sciences 24

1.6 Relevant Theorizing by Past Psychologists 29

1.7 Past Research on Arbitrary Metrics 33

1.7.1 Clinical Psychology 34

1.7.2 Forensic Psychology 37

1.7.3 Sport Psychology 39

1.7.4 Individual-Level Diagnoses 40

Chapter 2 43

2 Benefits of Metric Calibration for Basic Psychological Research 43

2.1 Help in the Interpretation of Data 44

2.1.1 Enhanced Interpretability of Statistical Effects 44

2.1.2 Allow Extraction of More Information from Data Patterns 48

2.1.3 Help Overcome Limitations of NHST 51

2.2 Facilitate Construct Validity Research 54

2.2.1 Construct Illumination ........ 55

2.2.2 Help with Conceptual Challenges ........ 56

2.2.3 Measurement Benchmark ........ 58

2.3 Contribution to Theoretical Development ........ 59

2.3.1 Aid in Theoretical Debates Involving Absolute Claims ........ 60

2.3.2 Allow More Precise Theorizing via Enhanced Scientific Language ........ 61

2.3.3 Quantitative Testing of Psychological Theories ........ 62

2.4 Facilitation of General Accumulation of Knowledge ........ 64

2.4.1 Valuable Information in its Own Right ........ 65

2.4.2 Guiding Framework for Cataloguing the Magnitude of Psychological Effects ........ 65

2.4.3 Facilitate Phenomenon-Based Research ........ 66

Chapter 3 ........ 68

3 Empirical Demonstrations ........ 68

3.1 Study 1 ........ 68

3.1.1 Theoretical Derivation of Relevant Behavioral Referents ........ 69

3.1.2 Method ........ 74

3.1.3 Results ........ 78

3.1.4 Discussion ........ 87

3.2 Study 2 ........ 88

3.2.1 Pan-cultural Self-Enhancement Debate ........ 89

3.2.2 Method ........ 92

3.2.3 Results ........ 95

3.2.4 Discussion ........ 100

3.3 Study 3 ........ 100

3.3.1 Risk-taking Measures to be Calibrated ........ 101

3.3.2 Theoretical Derivation of Behavioral Reference Points ........ 102

3.3.3 Method ........ 104

3.3.4 Results ........ 109

3.3.5 Discussion ........ 114

3.4 Other Analyses ........ 116

3.4.1 Trait Self-control ........ 116

3.4.2 Extraversion and Conscientiousness ........ 118

Chapter 4 ........ 121

4 General Discussion .......... 121

4.1 Feasibility .......... 121

4.2 Implications .......... 123

4.2.1 Help in the Interpretation of Data .......... 123

4.2.2 Facilitate Construct Validity Research .......... 128

4.2.3 Contribute to Theoretical Development .......... 131

4.2.4 Facilitate the General Accumulation of Knowledge .......... 134

4.3 Relatedness to Other Past Measurement Approaches .......... 136

4.4 Limitations and Caveats .......... 140

4.5 Future Directions .......... 142

4.6 Coda .......... 145

References .......... 146

Appendices .......... 163

Curriculum Vitae .......... 173

# List of Tables

Table 1: List of proposed benefits of non-arbitrary metrics. ............................................43

Table 2: Descriptive statistics and correlations for variables in Study 1 (*N* = 94). ...........79

Table 3: Descriptive statistics and correlations for variables in Study 2 (*N* = 97). ...........96

Table 4: Lottery options format in lottery risk task. .......................................................107

Table 5: Descriptive statistics and correlations for variables in Study 3 (*N* = 99). .........109

# List of Figures

Figure 1: The nature of arbitrary metrics visually depicted................................................4

Figure 2: Hypothetical probability of suicide attempt as a function of depression scores. ...............................................................................................................................13

Figure 3: Hypothetical mapping between neuroticism scores and probability of reacting to an environmental stressor with anger rather than calmness. ..................................15

Figure 4: Hypothetical linear (solid line) and non-linear (dotted line) mappings between extraversion scores and average hours spent socializing............................................16

Figure 5: Schematic diagrams of the key concepts of the strong (panel A) and weak (panel B) variants of the proposed experimental metric approach. ............................19

Figure 6: Early thermometers having no metric (left) or arbitrary metric (right). Reprinted with permission of The John Hopkins University Press (© 1969) from Middleton (1969, p. 87, Figure 3.1). ..........................................................................................25

Figure 7: Santorio's early string hygrometer having a scale with arbitrary metric. Reprinted with permission of The John Hopkins University Press (© 1966) from Middleton (1966, p. 21, Figure 1.9)...........................................................................27

Figure 8: A typical moderated multiple regression model when both predictors are continuous (panel A) with an actual example from the literature (panel B, Jordan et al., 2003; reproduced with permission)......................................................................45

Figure 9: Hypothetical experimental results across two samples at different locations on the DV scale. ...........................................................................................................49

Figure 10: Predicted probability of choosing a cognitively challenging task (Task 2) over a cognitively simpler task (Task 1) given need for cognition (NFC) scores. ..............80

Figure 11: Mean time spent on unsolved near-impossible anagrams (in minutes) in the anagram persistence task (APT) plotted against self-reported task persistence scores using a linear (solid line) or cubic function (dotted line)............................................83

Figure 12: Mean time spent on unsolved near-impossible anagrams in the anagram persistence task plotted against self-reported Impulse-Control scores (Goldberg). ....87

Figure 13: Number of non-existent words claimed familiar in OCT plotted against trait rating scores using a linear (solid line) or cubic function (dotted line). ......................97

Figure 14: Number of non-existent words claimed familiar in OCT plotted against trait rating scores made separately using a linear function. ................................................98

Figure 15: Number of non-existent words claimed familiar in OCT as a function of BIDR scores using a linear function..........................................................................99

Figure 16: Predicted probabilities of choosing $10 gamble over $4 safe bet plotted against adjusted BART scores. .........................................................................111

Figure 17: Predicted probabilities of choosing $10 gamble over $4 safe bet (solid line) or over $6 safe bet (dotted line) given CCT scores......................................................112

Figure 18: Grade point average plotted against Tangney et al.’s (2004) trait self-control scores.........................................................................................................118

Figure 19: Number of parties per month given gregariousness facet scores (panel A) and maximum driving speed given dutifulness facet scores (panel B). ...........................119

Figure 20: Moderated multiple regression re-analysis of O’Hara et al. (2009) using conventional +/-1 SD (panel A) or calibrated values (panel B)................................124

# List of Appendices

Appendix A: Conscientiousness items (MPQ and NEO-FFI) used in Study 1. ..............163

Appendix B: Conscientiousness items (AB5C) used in Study 1. ...................................164

Appendix C: NFC items used in Study 1..........................................................................165

Appendix D: Anagram persistence task (APT) materials used in Study 1. .....................166

Appendix E: Words used in over-claiming technique (OCT) of Study 2. .......................167

Appendix F: Domain-specific risk-taking scale (DOSPERT) items used in Study 3......169

Appendix G: Ethics approval for Study1..........................................................................170

Appendix H: Ethics approval for Study 2. .......................................................................171

Appendix I: Ethics approval for Study 3. .........................................................................172

# Chapter 1

## 1 Introduction

It is undeniable that measurement is a cornerstone of psychology as an empirical science. Any field of scientific enquiry takes it as a given that empirical observation – rather than judgments of faith – is the primary evidentiary entity used to make claims regarding reality. The significance of measurement cannot be over-stated. The importance of measurement, however, can be further illuminated by considering the definition of the word *cornerstone*. The Oxford English dictionary defines *cornerstone* as "the first stone set in the construction of a masonry foundation, important since all other stones will be set in reference to this stone, thus determining the position of the entire structure." This analogy visually reveals that measurement is critical because all empirical findings and theoretical claims are fundamentally tied or emerge in reference to measurement. Furthermore, measurement as a first stone ultimately determines the structure and sturdiness of an entire body of scientific knowledge.

The vast majority of researchers in psychology generally agree with the importance of measurement. Indeed, within the last few decades, great advances in psychological measurement techniques have been achieved, which have led to important theoretical insights (e.g., the development of implicit measures; see Fazio & Olson, 2003). At the same time, however, some psychologists have pleaded for an increased focus on measurement in psychological research (e.g., Borsboom, 2006; Embretson, 2006; Merenda, 2007; Murphy & Deshon, 2000). In particular, Borsboom (2006), based on an analysis of the factors that have hindered the integration of psychometrics and psychology, argued that the incorporation of more advanced psychometric practices into psychological research is necessary for the true potential of psychological science to be realized. Notwithstanding these critiques, psychology researchers are often diligent in satisfying basic measurement requirements, for example, by providing evidence for the reliability and validity of their measures. However, an aspect of measurement that is distinct from reliability and validity that has received virtually no attention in basic research and scant attention in applied research, concerns the numbering system used to quantify observed scores on psychological measures: that is, the *metric* of psychological instruments.

The term metric refers to the numbering system used to quantify observed measurement scores when describing an individuals' standing on a psychological construct (Blanton & Jaccard, 2006a). For instance, Beck's Depression Inventory (BDI: Beck & Steer, 1987) has a metric that can range from 0 to 63. An interesting fact that is rarely discussed, however, is that virtually all measures in psychology have a metric that is *arbitrary* in nature. This means that any particular value produced by a measurement instrument does not necessarily have any precise meaning except when considered in relation to other values. That is, a score of "35" on the BDI does not – in and of itself – tell us much; however, relative to a score of "45", we can – all else being equal – infer that a person with a score of "35" has a depression of lesser severity than a person with a score of "45". In a formal sense, a metric is arbitrary when it is not empirically known where a given score locates an individual on the underlying psychological dimension or when it is not known how a 1-unit change in the observed scores reflects the magnitude of change on the underlying dimension (Blanton & Jaccard, 2006a, 2006b).

The focus of the current research is on the nature of arbitrary metrics in the context of basic psychological research. The over-arching goal of the dissertation is to argue that it is both *useful* and *feasible* to calibrate the metric of psychological instruments used in basic psychological research so as to render their metrics non-arbitrary. To achieve this goal, I will present a conceptual analysis of the utility of non-arbitrary metrics for basic psychological research by elaborating on four categories of proposed benefits of non-arbitrary metrics. Then, I will illustrate empirically that it is feasible to calibrate the metric of psychological instruments, by applying the metric calibration approach to seven distinct constructs commonly studied in psychological research, across three studies and re-analyses of other researchers' data. Finally, connecting the conceptual and empirical components of the dissertation, in the General Discussion, I will attempt to strengthen my case regarding the proposed benefits of non-arbitrary metrics by applying some of my preliminary metric calibration findings to actual published research findings in the literature.

## 1.1 The Nature of Metrics and Metric Arbitrariness

In this section, I will expand on what is meant by metric arbitrariness and elaborate on important issues that are relevant when considering the metrics of psychological measures. First, however, a clarification concerning terminology is in order. Following De Houwer (2006), I will define a measurement procedure (or measurement instrument) as the actual apparatus used to measure a psychological variable (e.g., the questionnaire, task instructions, etc.). For the term, *measure*, it is important to realize that this term can refer either to the measurement instrument or to the outcome of a measurement procedure (e.g., a particular score on a questionnaire). To avoid this ambiguity, I will use the term *measure* exclusively to refer to the measurement instrument. To refer to the outcome of measurement, I will use the term *measurement scores* (or *observed scores*), which can be viewed as the end product of applying a measurement procedure to a person to assess a psychological construct. The term *measurement* is defined using Stevens' (1946) widely adopted characterization as "the assignment of numerals to objects or events according to some rule" (p. 677; but see Luce, 1997). Hence, metric arbitrariness is a feature of the scores produced by a measurement procedure.

As mentioned in the introductory paragraph, Blanton and Jaccard (2006a) consider a metric as arbitrary when it is not known where a given score locates an individual on the underlying psychological dimension or when it is not known how a 1-unit change in the observed scores reflects the magnitude of change on the underlying dimension. In other words, a metric is arbitrary when the mapping between observed scores and the underlying dimension is unknown. In psychology, it is generally assumed that observed scores provide a proxy to an individual's actual standing on the latent construct of interest and that some response function relates the individual's actual standing on the construct to his or her observed score on the response metric (Lord & Novick, 1968). Hence, when a metric is arbitrary, the function describing the relation between observed scores on a measure and the underlying dimension is unknown (Blanton & Jaccard, 2006a). Of course, we never have direct access to the underlying dimension (as is also the case in the physical sciences). What we can do, however, is to observe theoretically-relevant behaviors, which can be argued to reflect different levels of the underlying dimension.

The general task then becomes to discover how scores from the to-be-calibrated measure link up to these theoretically-relevant behaviors, which researchers consensually agree reflect particular locations on the underlying dimension of the construct (I will elaborate below on the characteristics such behaviors should ideally possess to serve as useful reference points). These fundamental concepts of metric calibration can be clarified by turning to the panel A of Figure 1, which depicts three depression instruments having arbitrary but distinct metrics (Instrument A = Self-report Depression Scale [SDS; Zung, 1965], Instrument B = Major Depression Inventory [MDI; Bech et al., 2001], and Instrument C = Beck's Depression Inventory [BDI; Beck & Steer, 1987]).

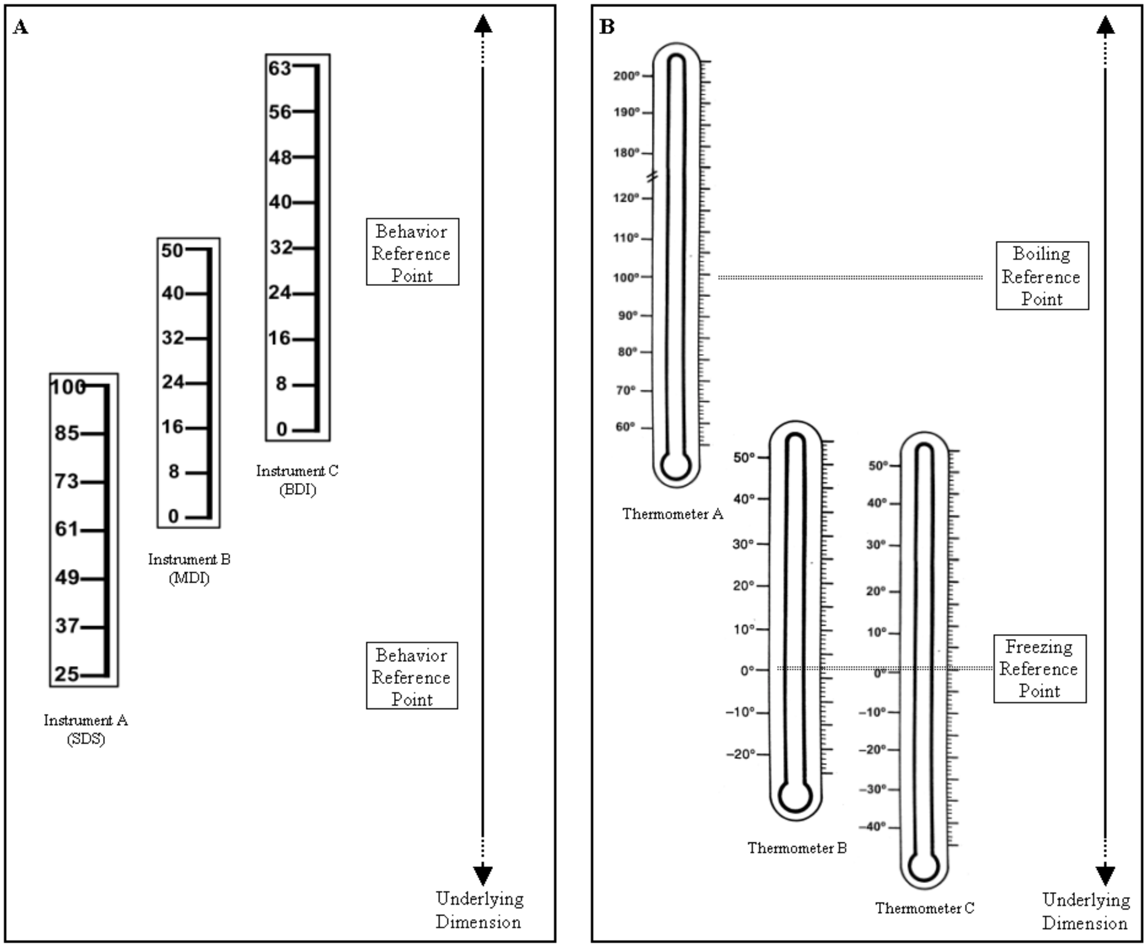


**Figure 1: The nature of arbitrary metrics visually depicted.**

The metric arbitrariness of these instruments becomes apparent in the figure given that the instruments have not been linked in any way to relevant behavioral reference points of depression. Hence, it is unclear what a high score on any of the depression instruments means with respect to the different ranges of the underlying dimension. For instance, it is possible that a high score on one instrument (e.g., instrument A) reflects a lower level of depression than a low score on another instrument (e.g., instrument C).

These issues can be further clarified by considering the metric calibration of different thermometer instruments, depicted in panel B of Figure 1. Here, the non-arbitrary metric of the different thermometers becomes apparent given they have all been empirically linked to the boiling and freezing point of water, which are reference points indicative of particular locations on the underlying dimension of temperature. As should be apparent, it is clear that the only way to know that these different thermometers are tapping into different ranges of the underlying dimension of temperature is through their empirical linkages to the relevant reference points. For instance, the cooking thermometer (thermometer A) taps into a higher and much wider range of the underlying dimension whereas the more general purpose thermometers (thermometers B and C) cover a narrower and lower range.

Returning to the depression instruments, these considerations make apparent that to get a sense of what range of the underlying dimension the different depression instruments are tapping into, it is necessary to empirically link scores from those instruments to specific depression behaviors which could be argued to reflect different locations on the underlying dimension of depression (for e.g., suicide attempt in the last six months). Then, the metric of those instruments would gain meaning and become non-arbitrary. More generally, achieving non-arbitrary metrics for psychological instruments requires that observed scores are linked to particular behaviors argued to reflect different locations on the underlying psychological dimension. Then, and only then, will the metric of psychological instruments start to gain meaning and hence shed light on what location of the underlying dimension one's instrument is tapping into. Below, I will elaborate on the specific details of this kind of metric calibration approach and outline the critical features ideal behavioral reference points should possess.

Arbitrary metrics may not be a problem per se for basic researchers who typically seek to test general theories rather than make absolute judgments about a person's standing on a construct (however, we will see later how this can be a problem for certain types of claims made even in basic psychology). Arbitrary metrics do, however, become an issue when researchers attempt to make individual-level diagnoses based solely on the scores produced by an instrument with an arbitrary metric. This is the case when a researcher attempts to characterize an individual (or a group of individuals) as "high" or "low" on the underlying dimension: that is, making a statement regarding a person's absolute level on the underlying dimension.

Blanton and Jaccard described (2006a) two inappropriate strategies researchers sometimes use to make absolute judgments from scores produced by measures with arbitrary metrics: *meter reading* and *norming*. Meter reading refers to the act of simply using the score on the observed metric to infer the standing of the person on the underlying dimension. Hence, someone with a score at the high end of the metric would be considered as being "high" on the underlying dimension whereas someone with a score at the low end of the metric would be considered as being "low" on the underlying dimension (see below for an example). Norming refers to the process of transforming raw scores into standardized scores (e.g., z-scores or percentiles) based on the distribution of data from a target population and then making inferences of location on the underlying dimension based on this new metric. Blanton and Jaccard argue that both of these strategies are unfounded and that systematic metric research linking measurement scores to meaningful psychological events is the only sufficient strategy to permit inferences regarding someone's standing on the underlying dimension.

Meter reading, the first inappropriate strategy reviewed by Blanton and Jaccard (2006a), involves inferring the standing of a person on the underlying dimension by simply examining where, on the metric range, that person's observed score lies. For example, inferring that someone with a score of "6" on a self-esteem inventory with a metric ranging from 1 to 7 is "high" on the underlying dimension of self-esteem would be an example of meter reading. Although meter reading may not be that common for situations similar to this simple example, researchers often engage in meter reading in the context of bipolar constructs, which are quite common in psychology. For bipolar constructs, the two ends of the dimension are assumed to be polar opposites, with the midpoint of the scale sometimes labeled as “neutral” or “unsure” or “neither agree nor disagree”. In this context, researchers may simply assume (based on faith) that the scale midpoint maps onto the midpoint of the underlying dimension.

The assumption that the scale midpoint maps onto the midpoint of the underlying dimension is pervasive, for example, in research on egocentric preferences for the self relative to others (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Weinstein, 1980) and in research on knowledge overconfidence (Erev, Wallsten, & Bedescu, 1994). In the research on egocentric preferences, for example, researchers infer “better-than-average” effects by testing whether individuals’ mean ratings of how they view themselves in comparison to others on certain traits (e.g., “intelligent”, “friendly”) are statistically greater than the scale midpoint (5 on a 9-point scale, where 1 = *much less than the average college student*; 5 = *about the same as the average college student*; 9 = *much more than the average college student*). There are numerous logical and empirical reasons why assuming that the scale midpoint reflects the midpoint on the underlying dimension is not warranted. Logically, and returning to my earlier analogy to thermometers (see Figure 1, panel B), it should be clear that if one was working with thermometers that have *not* been calibrated to relevant fixed points, it follows that one should not simply conclude that the numerical midpoint on one’s thermometer reflects “neutral” temperature. Similarly, it is clear from Figure 1 (panel B) that a thermometer reading near the maximum (or minimum) of the range of the instrument should not be used to infer the temperature in an absolute sense (i.e., meter reading) if the thermometer has not been empirically calibrated to relevant fixed points. Transporting these considerations into the psychological arena clearly implies that one should not engage in this type of meter reading if the psychological instruments have not been empirically linked to relevant reference points.

Above and beyond logical considerations against meter reading, there are also various empirical findings that indicate we should not engage in meter reading with psychological instruments. For example, ratings given to questionnaire items have been shown to be influenced by all of the following: the number of categories on the rating scale (e.g., 6- vs. 9-point scales; Parducci & Wedell, 1986), the extremity of previously judged items (Rotter & Tinkleman, 1970), the adjective labels or format of the scale anchors (French-Lazovik & Gibson, 1984; Schwarz, Hippler, Deutsch, & Strack, 1985), whether the intermediate categories are labeled with adverbs or not (e.g., *slightly*, *moderately*; Lam & Klockars, 1982), sheer frequency with which stimuli occur in the real world (Wedell & Parducci, 1988), and category activation processes related to the scale anchors ("very honest" vs. "not at all dishonest" activates conceptually distinct knowledge structures; Gannon & Ostrom, 1996).

Yet another empirical reason that supports the inappropriateness of meter reading is the issue that individuals sometimes use different standards of reference when making judgments about different targets (Biernat & Manis, 1994). For example, it has been found that for judgments of competence, individuals use different standards of comparison when judging men versus women, such that individuals – based on gender stereotypes – give higher ratings of competence to women than to men for exhibiting the same level of competence (Biernat & Manis, 1994). Hence, equal ratings of perceived competence across gender targets do not reflect equal amounts of perceived competence. In other words, a difference score of "0" on observed ratings of competence for men minus women would not indicate equal perceived competence but in actuality reflect higher perceived competence in women than in men. In a broader sense, however, one can argue that individuals' reliance on different judgment standards poses psychometric problems even for judgments made about the same target (e.g., personal attitude judgments: Olson, Goffin, & Haynes, 2007; employee performance judgments: Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996; see also Goffin & Olson, 2011). Taken together, all of these empirical findings imply that it is questionable to engage in meter reading and hence assume based on faith alone that the neutral point of a numerical scale maps onto the neutral point of the underlying dimension for bipolar constructs. Ultimately, however, these empirical reasons are not strictly required to make the case that meter reading is unwarranted given the aforementioned logic regarding the more obvious flaw of meter reading when considering, for instance, thermometers which have not been empirically calibrated to relevant fixed points (see Figure 1, panel B).

Another example of this issue – worth mentioning due to its notoriety – is the race Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), for which some researchers interpret scores of “0” as lack of implicit bias, when in fact certain features of the IAT stimuli could render a difference score of “0” ambiguous in meaning at best (Blanton & Jaccard, 2006a, 2006b; Blanton et al., 2009). This could be the case if positive words used in the IAT are more positive in character than the negative words are negative and/or if the Black faces are more prototypically Black than the White faces are prototypically White (Bluemke & Friese, 2006; hence, again shifting scores in either direction away from the theoretical midpoint). To interpret the numerical midpoint of “0” in the IAT as the neutral midpoint on the underlying dimension, one has to move beyond meter reading and actually gather evidence of empirical linkages between particular measurement scores and “meaningful and conceptually relevant behaviors” (p. 63) argued to reflect particular locations on the underlying dimension (Blanton & Jaccard, 2006b). It is important to note that the metric of *any* measure, regardless of the measure type (i.e., self-report, indirect, unobtrusive, behavioral task), is initially inherently arbitrary. Only after scores of a measure are linked to specifically chosen theoretically-relevant behaviors consensually agreed-upon to reflect particular locations on the underlying psychological dimension, do measurement scores start to gain meaning.

A second strategy identified by Blanton and Jaccard (2006a) sometimes used to inappropriately infer metric meaning, is when a researcher “norms” a distribution of scores such that it conforms to the properties of a standardized population of individuals (e.g., a population of healthy individuals). Norming typically involves the transformation of raw measure scores into standardized scores, such as z-scores, and then these scores are interpreted relative to the mean of the standardized population. For example, a raw score of “6” could be transformed to a z-score of “1.5”, which would mean that a person’s score is 1.5 standard deviations higher than the mean in the normative target population. Although it is true that standardizing scores may give important information about the *relative standing* of individuals within a particular target population, standardization alone in no way conveys information about a person’s standing on the underlying dimension in an absolute sense or in terms of the behavioral implications of a particular score. Consequently, standardizing scores from an arbitrary metric does not render the metric non-arbitrary. The only way to convey information about a person's standing on the underlying dimension in an absolute sense is to link observed scores to behavioral implications of the relevant construct, which can serve as reference points (Blanton & Jaccard, 2006a, 2006b; Sechrest, McKnight, & McKnight, 1996).[1]

## 1.2 Reducing Metric Arbitrariness: Background

To glean information about a person's standing on an underlying dimension, one has to go beyond meter reading and norming and acquire empirical evidence for making more nuanced interpretations of measurement scores. Empirical research must therefore be executed providing behavioral evidence for score interpretation rather than deciding metric meaning based on faith, by fiat, or as a measurement assumption. Although sparse, the literature contains some theorizing about different research strategies that can be used to reduce metric arbitrariness. A valuable starting point is a framework provided by Sechrest et al. (1996) who stated that metric meaning can be increased in one of three ways: (1) by estimating the degree of internal coherence of a measure, (2) by calibrating a measure with another measure, and (3) by calibrating a measure against external criteria or behavioral implications. Sechrest et al. emphasized that the third strategy is likely the most fruitful strategy and hence focus their discussion almost exclusively on this strategy. This resonates well with Blanton and Jaccard's (2006a, 2006b) position who also emphasized a strategy involving the calibration of measures by finding empirical linkages between measurement scores and meaningful behavioral referents external to the to-be-calibrated measure.

Sechrest et al. (1996) elaborated on five strategies to increase the meaning of score metrics via calibration of a measure against external criteria: direct personal experience, empirically established behavioral implications, cross-experiential equivalence

[1] Some may find Blanton and Jaccard's (2006a) position that normative data cannot speak to the metric issue too strong, given that normative IQ data have led to an intuitive metric of IQ scores. Though it is true that the IQ metric in some sense has gained an intuitive metric via normative data, strictly speaking metric calibration requires systematic empirical research linking test scores to behavioral reference points rather than metric meaning based on informal data regarding the kind of behaviors one can expect from individuals with particular IQ scores.

(formulating unfamiliar phenomena in more familiar terms), cross-modal representation (representing psychological states in terms of experiences in other modalities; e.g., loudness), and the method of just noticeable differences (minimum difference in scores required to observe a difference in theoretically-relevant behavior). I will focus my attention on the first two strategies as they reflect most closely the conceptualization of metric arbitrariness taken in this dissertation. In terms of direct personal experience, clinicians working with patients may sometimes have extensive experience in the use of psychological instruments and hence have an intuitive sense of the kinds of behaviors that correspond to particular scores on a measure. For example, a clinician who regularly uses the BDI potentially could have an intuitive understanding of how certain BDI scores map onto different types of depression-related behaviors (i.e., frequency of crying, suicide ideation) exhibited by his or her patients. It can be argued, however, that these types of mappings need to be established more systematically and precisely. Indeed, as Sechrest et al. mention, these kinds of personal experiences may not be of much help to the majority of the researchers in the field, who lack dual contact with clients and a psychological measure, unless this information can be captured and organized in some systematic fashion. The second strategy proposed by Sechrest et al. involves using empirically established behavioral implications to calibrate a measure's scores. For example, given that a sad demeanor is one of the most salient features of depression, one could examine the specific mapping between a 1-point increase in BDI score and reduction in probability of being found smiling. More generally, Sechrest et al. argue that using behavioral implications of any particular measure (not just clinical measures) can be a fairly direct method of imbuing measurement scores with more meaning.

In a similar vein, Blanton and Jaccard (2006a, 2006b) provided valuable information on strategies to reduce the metric arbitrariness of psychological measures. Indeed, they went beyond Sechrest et al. (1996) by providing a more nuanced conceptual analysis of the tricky issues surrounding the metric calibration of psychological measures and also by elaborating more concretely on the actual steps required to carry out empirical research aimed at reducing metric arbitrariness. According to Blanton and Jaccard's conceptual analysis, an important preliminary step, which can be seen as a pre-condition to the metric calibration of a measure, is to develop consensus among researchers as to which particular behaviors (or symptoms, or manifestations of a certain state) likely places an individual at the high or low end of the underlying psychological dimension. If this particular behavior empirically corresponds to a certain observed score on the measure, then the *approximate* location of the score on the underlying dimension can be inferred. Granted, this inference nonetheless caries some degree of uncertainty given the number of complexities inherent in the metric calibration process (e.g., reaching consensus among experts on the most appropriate behavioral reference points, assessing the particular behavior, modeling the metric mapping). Indeed, Sechrest et al. mention that the calibration of measures in psychology may never be as “tight” as in the physical sciences, but that this should not detract psychologists from engaging in metric calibration research. Moreover, the fact that the constructs of interest in psychology are unobservable should also not detract psychologists from calibrating their measures. Although this fact undoubtedly renders the calibration task quite challenging, it is important to keep in mind that most constructs in the physical sciences are also unobservable (e.g., temperature, electricity, magnetism), but that this did not prevent natural scientists from calibrating measures of unobservable constructs.

## 1.3 Main Empirical Strategies of Metric Calibration

Three primary strategies can be followed to reduce metric arbitrariness, two roughly following from Blanton and Jaccard’s (2006a, 2006b) and Sechrest et al.’s (1996) analyses and one stemming from my own ideas derived from an analysis of how thermometers and hygrometers are experimentally calibrated to theoretically-relevant reference points. The main empirical strategies are: (1) mapping observed scores to noteworthy differences in behavior tied to the phenomenon in question, (2) mapping observed scores to the gradation of theoretically-relevant behaviors, and (3) using an experimental approach to experimentally manipulate the construct to increasingly extreme levels. I will elaborate on each one of these strategies in turn.

### 1.3.1 Strategy 1: Noteworthy Differences in Behavior

In a first sense, a metric can be made more meaningful by finding an empirical mapping between observed scores and noteworthy differences in behavior tied to the construct of interest. This can be implemented, for example, by finding an empirical mapping between observed scores and the probability of performing a theoretically-relevant behavior. In this context, the presence or absence of a behavior can be seen as a clear noteworthy difference in behavior. Hence, one seeks to document how changes in the observed scores of the to-be-calibrated measure map onto the probability of performing a certain behavior. As a concrete example, one could examine how depression scores map onto whether an individual has or has not made a suicide attempt in the last six months. As Blanton and Jaccard (2006b) mention, one response function that could be found in this case is an exponential function, fitted using logistic regression (see Figure 2). As is seen in the figure, the probability of a suicide attempt is small and constant for depression scores below "10", but then start increasing around "15" and may reach unacceptable and dangerous levels around "25" and "30", respectively.

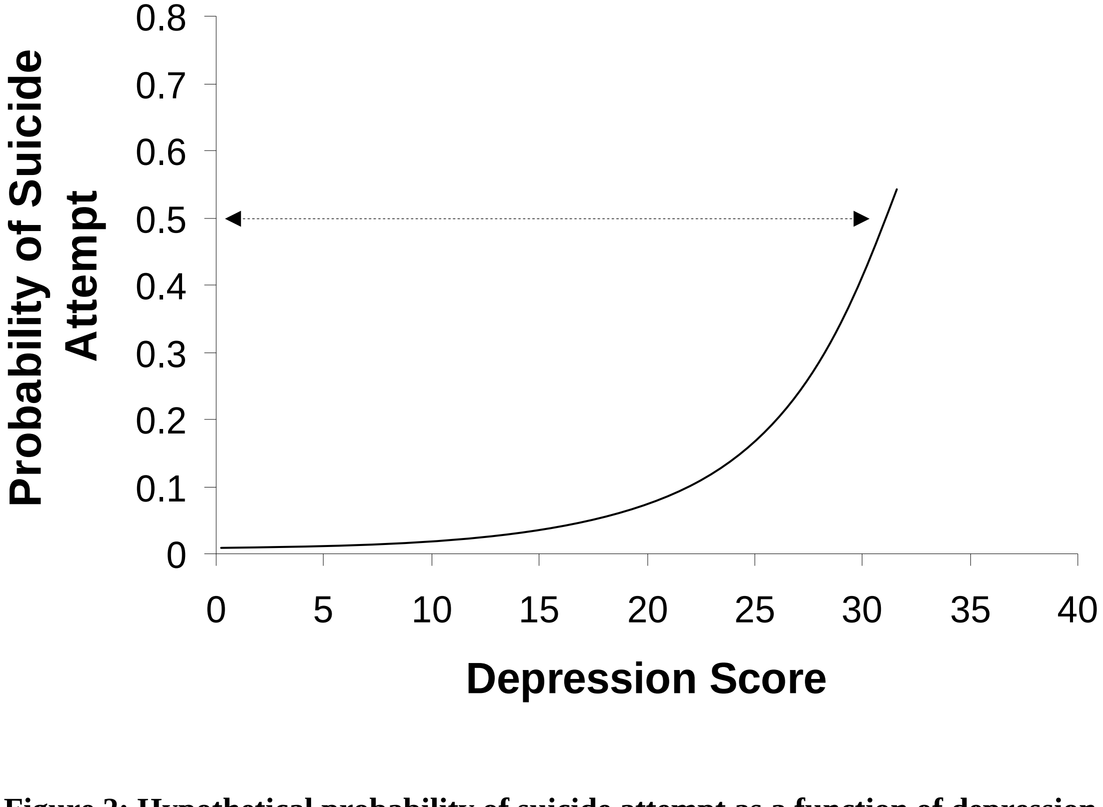


**Figure 2: Hypothetical probability of suicide attempt as a function of depression scores.**

It is important to note that, if this empirical mapping was found (and replicated), it would be clear that improving someone's depression score from a value of "30" to "25" would mean something quite different than improving someone's depression score from a value of "20" to "15". These sorts of meaningful inferences could not be made without this type of metric calibration information. Hence, as Blanton and Jaccard state more broadly, "as individual test scores are linked to meaningful external events [behaviors], the meaning and implications of a given test score become more apparent and the metric becomes less arbitrary" (p. 63).

Another example may help clarify the important concept involved in the notion of *noteworthy difference in behavior*. Consider a researcher interested in shedding light on the meaning of neuroticism scores by assessing individuals' emotional reactions to construct-relevant environmental stressors as a theoretically-relevant behavioral criterion to calibrate those scores. In particular, one would be interested in finding the mapping between neuroticism scores and qualitatively distinct emotional reactions to the relevant environmental stressor. In this context, the qualitatively distinct emotional reactions could be seen as the noteworthy difference in behavior. Trained independent judges could code the emotional reactions of participants to the environmental stressor as "calm, even tempered" or "irritated/angry." Assuming adequate inter-judge reliability, one could then find the approximate threshold neuroticism score that separates those who respond to the stressor with annoyance and irritation versus those who remain calm. Using a logistic regression, a graph could then be generated as depicted in Figure 3. As is evident in the non-linear mapping in Figure 3, a neuroticism score of approximately "18" could be viewed as representing a threshold value that distinguishes between individuals who respond to the environmental stressor with anger rather than calmness (given that a score of "18" maps onto a .5 probability of reacting with anger vs. calmness).

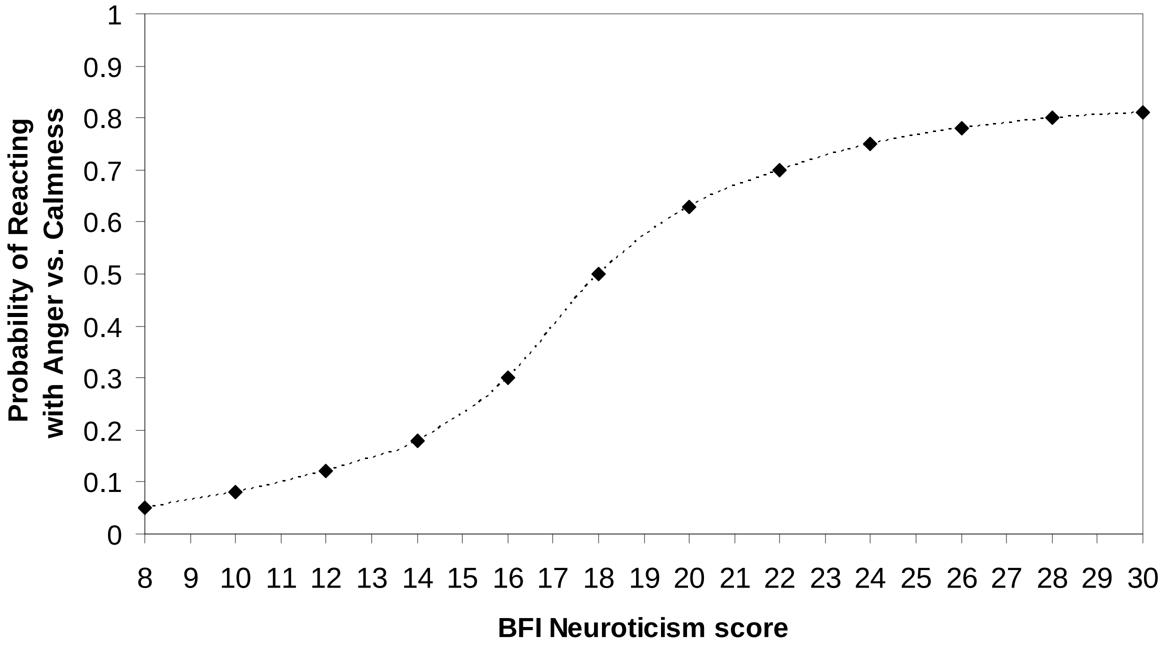


**Figure 3: Hypothetical mapping between neuroticism scores and probability of reacting to an environmental stressor with anger rather than calmness.**

That being said, it is also possible that a more nuanced three-category model of qualitatively distinct emotional reactions could be posited and found (using, for e.g., probit or logit regression). That is, strong construct theory and other considerations may predict three qualitatively distinct ways to respond to the environmental stressor (e.g., [1] calmness, [2] mild irritation, and [3] extreme anger). The example in this approach fits well with Blanton and Jaccard's (2006a) own recommendations, stating that meaningful metrics are "developed through the discovery of empirical thresholds that indicate noteworthy changes in the occurrence of observable events tied to the phenomenon in question" (p. 34).

Finally, it is also interesting to briefly note how this *noteworthy difference in behavior* approach parallels in a broad sense the general strategy of calibrating the metric of thermometers by using the qualitatively distinct changes in relevant external events as reference points (i.e., noteworthy change of states between non-boiling and boiling water or between non-frozen and frozen water). Although this parallel was not mentioned by Blanton and Jaccard (2006a, 2006b), I think it is informative to bear this in mind when considering the complex and abstract issues surrounding the metric calibration of psychological instruments. I will soon return to this issue in my section on inspirations from the history of the development of instruments in the natural sciences.

### 1.3.2 Strategy 2: Gradation in Behavior

The second general strategy to reduce metric arbitrariness involves finding empirical linkages between test scores and *gradation* of theoretically-relevant manifest behaviors. Using this strategy, particular scores on a measure can be mapped onto particular behavioral manifestations of the relevant construct, imbuing those particular scores with meaning. Gradation of such behaviors could take the form of an individual's performance on a behavioral task or a frequency count of the number of times a relevant behavior is performed. For example, one could examine the mapping between extraversion scores (e.g., using Eysenck's Introversion-Extraversion Scale [IES]; Eysenck & Eysenck, 1975) and number of hours spent socializing.

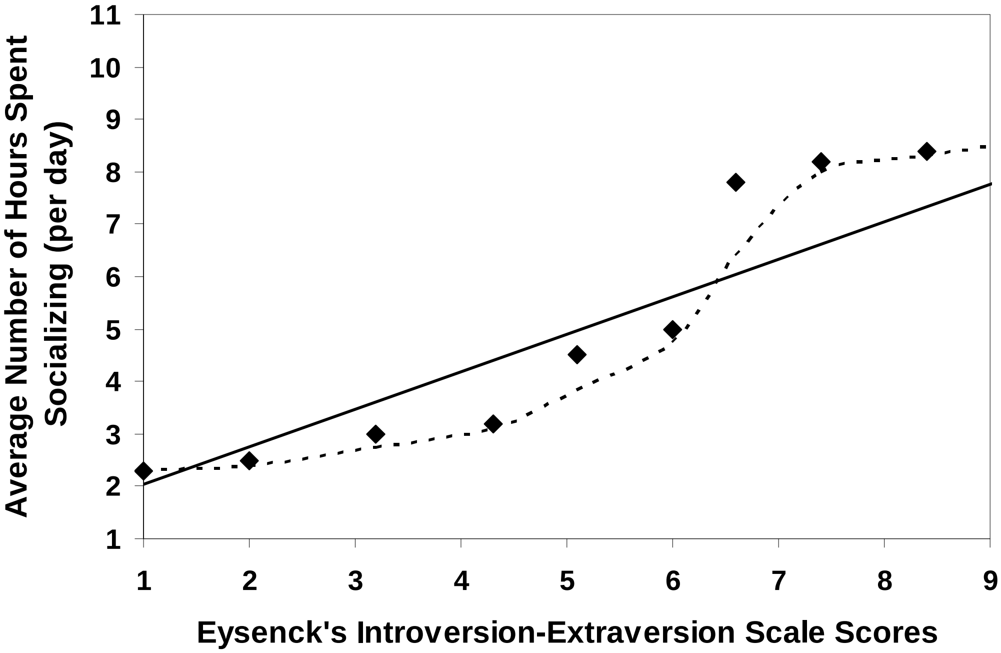


**Figure 4: Hypothetical linear (solid line) and non-linear (dotted line) mappings between extraversion scores and average hours spent socializing.**

As depicted in Figure 4, extraversion scores could gain meaning by way of their mapping to the behavioral manifestation of extraversion; that is, the average hours spent in the presence of others per day (as assessed, for e.g., using Mehl, Pennebaker, Crow, Dabbs, & Price's, 2001, electronically activated recorder [EAR]). As reflected by the linear mapping (solid line), an extraversion score of "9" would correspond to approximately 8 hours per day spent socializing whereas an extraversion score of "1" would correspond to approximately 2 hours of socializing per day. Hence, although in this approach no specific qualitatively distinct difference in behavior is available as a particular reference point, the scores nonetheless acquire meaning via the discovery of empirical mappings to relevant theoretically-relevant behaviors.

With systematic collaborative discussions among experts in the field, it is also possible that consensus eventually emerges as to what changes in values in the manifest behavior represent noteworthy differences. For instance, in the extraversion example, even though number of hours spent socializing per day is linear, it seems clear that spending an average of 2 hours socializing – versus spending an average of 9 hours – represents a qualitatively distinct state of affairs. Over time, as more is understood about extraversion, perhaps experts in the area could agree to even more nuanced "noteworthy" differentiations. Furthermore, even though average hours is a continuous variable, it is also entirely possible that the function relating extraversion scores to average hours spent socializing could be non-linear (e.g., cubic or exponential), in which case meaningful threshold values for the behavioral referent could be gleaned. Indeed, as reflected by a non-linear function (dotted line) in Figure 4, a closer inspection reveals some kind of discontinuity near the inflection point of the fitted non-linear curve, suggesting extraversion scores greater than approximately "6.5" may correspond to a qualitatively distinct manifestation of extraversion.

### 1.3.3 Strategy 3: Experimental Approach

A third strategy for calibrating the metric of psychological measures, which to my knowledge is a completely novel conceptual idea, involves adopting an experimental approach whereby the relevant construct is manipulated to extreme levels. I will propose two variants of the experimental approach to account for the fact that constructs in psychology are generally theorized to be either predominantly trait-like (e.g., personality constructs) or predominantly state-like. Hence, I will propose (1) an experimental strategy aimed to calibrate instruments that assess predominantly state-like constructs (i.e., a “strong” experimental strategy) and (2) an experimental approach aimed to calibrate instruments that assess predominantly trait-like constructs (i.e., a “weak” experimental strategy).

The strong experimental approach involves manipulating a certain construct to increasingly extreme levels and then simultaneously assessing the to-be-calibrated scores and manifest behavioral referents. Empirical mappings between scores and behaviors are then established by way of the manipulation levels (as depicted in Figure 5, panel A). This idea broadly derives from the history of the calibration of the thermometer and hygrometer (details covered below). In the physical sciences, it is common practice for scientists to calibrate their instruments by linking the instrument readings to reference points that involve extreme levels of the phenomenon. For example, in the case of the early hair hygrometer (i.e., a human hair used to index ambient humidity), scientists experimentally manipulated extreme degrees of humidity by creating conditions under which the ambient air was either extremely moist or extremely dry. Consequently, the arbitrary values of the early hair hygrometer (i.e., length of human hair expanding or contracting as the moisture in the air increased or decreased) gained meaning as they were linked to these extreme manifestations of humidity. Transporting this approach into the psychological arena implies that we can potentially increase the meaning of our metrics by experimentally manipulating a certain construct to levels as extreme as possible (both low and high) and then simultaneously assess changes in the to-be-calibrated measurement scores *and* the relevant behavioral manifestation of the construct. The empirical linkage between the to-be-calibrated measure scores and the behavioral manifestations would then be achieved by virtue of the experimentally manipulated levels. That is, the *mean* scores of the to-be-calibrated measure can be connected to the *mean* scores of the behavioral measure within each of the conditions (as depicted in Figure 5, panel A).

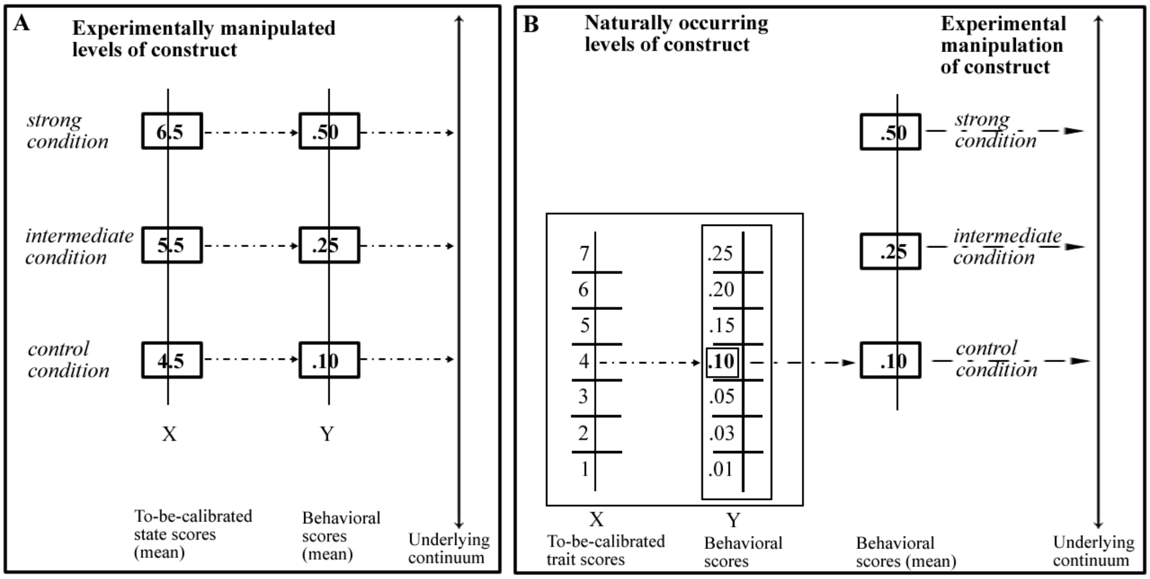


**Figure 5: Schematic diagrams of the key concepts of the strong (panel A) and weak (panel B) variants of the proposed experimental metric approach.**

For instance, strongly anxiety-provoking situations (vs. intermediate vs. control conditions) could be used to experimentally calibrate a state-measure of anxiety (Spielberger, 1983) to the probability of exhibiting a nervous tick, by examining the mapping between the mean state-anxiety scores and the mean behavioral reference point scores at each level of the anxiety manipulation (e.g., mean score of “6.5” on state-anxiety measure linked to a .5 probability of exhibiting a nervous tick). This strategy would roughly map onto the calibration of instruments in the physical sciences where the manipulation of the construct (e.g., increasing temperature of water via a flame) simultaneously impacts the to-be-calibrated measure (e.g., the glass-tube thermometer readings) and the reference point (e.g., presence or absence of boiling water).

The second “weak” variant of the experimental metric approach is proposed as a way to imbue further meaning into scores from measures posited to be predominantly trait-like, above and beyond the scores’ linkages to naturally occurring levels of theoretically-relevant behavioral referents covered in the first two strategies. Given that it may not be possible, or make theoretical sense, to attempt to manipulate trait-like measures to extreme levels (e.g., a self-report measure of extraversion), the strong form of the experimental approach is not appropriate. However, I propose that a “weaker” variant of the strong experimental strategy could nonetheless imbue additional meaning into the measurement scores of constructs theorized to be predominantly trait-like. This can be achieved by manipulating the behavioral expression of the construct to as extreme as possible levels and then assessing these behavioral manifestations (scores from self-report measure of the construct would be assessed before any manipulations given trait-like construct). Even though a construct may be posited to reflect a primarily trait-like component (e.g., extraversion), it may nonetheless be possible to increase the behavioral expression of the construct by manipulating theoretically-relevant situational factors. For instance, one could manipulate extraverted behavior with an alcohol manipulation (e.g., 0 vs. 2 vs. 4 units of alcohol per kg of body weight). Although personality psychologists may not see the expression of extraverted behavior due to alcohol as "true change" in the underlying construct, I contend that experimentally manipulating extraverted behavior to extreme levels can nonetheless provide valuable reference points that further increase the meaning of extraversion trait scores. This would be achieved by the fact that the experimentally manipulated extreme levels of the behavioral reference points would supplement the naturally occurring observed levels of the behavioral reference points.

The "weak" variant of the experimental approach is depicted in Figure 5 (panel B). Continuing with the extraversion example, we can see on the left-hand side of the diagram, the mapping between the naturally occurring levels of trait extraversion scores and a relevant extraverted behavioral reference point (e.g., probability of talking to a stranger). Hence, manipulating extraverted behaviors via alcohol could increase (on average) the probability of spontaneously talking to a stranger (values in boxes on the right-hand side). These manipulated behavioral reference points can then serve to add additional meaning to the naturally occurring levels of the behavioral reference point, which would lend further meaning to the interpretation of the trait scores. For example, the interpretation of the meaning of an extraversion trait score of "7" would be increased by virtue of the fact that the natural mapping to its behavioral reference point (i.e., .25 probability of talking to a stranger) can be interpreted with reference to the experimentally manipulated reference point (i.e., .50 probability).

In summary, following broadly the experimental logic of the calibration of instruments in the physical sciences, the experimental approach proposed (either in its strong or weak form) holds the potential to increase the meaning of psychological score metrics over and above the meaning gained from the first two non-experimental strategies.[2] The two primary strategies that will be used to reduce metric arbitrariness in the current research, however, will be the two non-experimental strategies first reviewed. Future studies should nevertheless explore the potential utility of my newly proposed experimental approach to metric calibration.

No matter what metric calibration strategy is used, it is important that the most appropriate statistical technique is used to model the response function that best connects the observed scores to the manifest behaviors. For instance, if a binary behavioral outcome is assessed, a logistic regression could be used to determine the logistic coefficient and intercept of the best fitting line; then predicted probability of performing the behavior can be determined for given test scores. If multiple category behavioral reference points are assessed, then a probit or logit regression could be used. For count-like data, for example the frequency of crying, poisson (or negative binomial) regression could be used. Regardless of the statistical strategy employed, the important issue is that a particular metric mapping is established and that the parameters of this function are then used to map the observed test scores to the criterion behaviors. It may also be informative to form prediction intervals (Neter, Kutner, Nachtsheim, & Wasserman, 1996) for each given test score, to gauge the amount of uncertainty inherent in the

[2] The known-groups approach sometimes used in construct validity research might come to mind to some readers in this section. Indeed, the known-groups approach could be seen as providing very preliminary information about the possible meaning of a measure's metric. For example, the finding that university professors score on average "6.5" on the need for cognition (NFC) scale (scale metric ranging from 1-7, for e.g.) whereas fashion designers score on average "3.5" could provide preliminary information to investigate what type of qualitatively distinct NFC-related behaviors distinguish university professors from fashion designers. Subsequent metric research could then systematically examine the empirical linkages between NFC scores and the relevant NFC-related behaviors identified in the known-groups stage.

empirical mapping for the *individual-level* scores. Prediction intervals estimate a range in which future observations will fall, given what has already been observed.[3]

## 1.4 Ideal Characteristics of Behavioral Reference Points

Regardless of the metric calibration strategy employed, it is important that behaviors chosen to act as reference points possess specific characteristics. In this section, I will elaborate on such particular characteristics, which behaviors serving as reference points should ideally possess. These characteristics are based on considerations that build upon past theorizing on the metric calibration process in the psychological arena. For instance, Sechrest et al. (1996), from an applied perspective, conceived criterion behaviors to be used in metric calibration as reflecting "external behavioral implications" (p. 1068) in relation to real-life events. More generally, Blanton and Jaccard (2006b) construed criterion behaviors as "meaningful and conceptually relevant behaviors or symptoms" (p. 63) or as "meaningful events that have gained consensus as being of relevance" (p. 68) with respect to certain locations on the underlying psychological dimension. But what particular characteristics render certain behaviors or psychological events "meaningful"? Going beyond considerations by Blanton and Jaccard and Sechrest et al., I contend that criterion behaviors to serve as behavioral reference points should ideally possess the following specific characteristics: *theoretically-relevant*, *objective*, *unambiguous construct-wise*, and *interpretationally clear*.

First, criterion behaviors should be theoretically-relevant in the sense that there is an expectation based on construct theory that a certain behavior reflects a relevant behavioral manifestation of the construct at hand. That is, the accepted working definition of the construct (itself ideally stemming from theoretical considerations surrounding the construct) should guide the decision of which particular behaviors one would theoretically expect to be connected to scores of the to-be-calibrated measure. Second,

[3] In a related vein, the role of random measurement error contaminating the scores of both the to-be-calibrated measure and criterion behavior should be considered and accounted for in the metric calibration process. Given that this issue has not yet been discussed in the metric calibration literature, future research should investigate how best to account for random measurement error in the metric calibration process.

behaviors to serve as reference points should be objective in the sense that independent observers can agree that a particular behavior was exhibited. Third, criterion behaviors should be chosen (and assessment situations configured) such that the observed behaviors are the most unambiguous as possible construct-wise. That is, the chosen behaviors should be assessed in such a manner whereby it can be argued that the observed behavior reflects primarily the construct of interest rather than also reflecting other constructs which are not of interest. Finally, criterion behaviors should be chosen and assessed such that they have a clear and intuitive interpretation, meaning that the scoring of the relevant behavior has a clear connection to the observed behavior in question (e.g., 1 = presence of a behavior and 0 = absence of a behavior; or number of times [or proportion of time] engaging in some behavior). For instance, if assessing time spent socializing with others (using e.g. Mehl et al.'s [2001] EAR), one would want to express the behavior in terms of hours spent socializing per day (for instance), rather than the number of seconds spent socializing per month. Also, another important consideration, as already mentioned, is that criterion behaviors to serve as behavioral reference points should be specifically chosen with the goal that the behaviors in question can be argued to reflect a particular location on the underlying dimension in an absolute sense. Taken together, I contend that it is the confluence of all of these characteristics that render certain behaviors strong candidates to be considered *meaningful* behavioral reference points.

Furthermore, it is also important to consider the features of the context in which criterion behaviors are assessed (the "interpretational context"). When searching for empirical mappings, it is critical that a researcher uses theory to guide his or her thinking about the particular contextual conditions that need to be in place to elicit the behavioral manifestation of the construct in question. Consequently, rather than modeling moderation, one must include the contextual moderators of a psychological phenomenon into the design of a metric research investigation. For example, in the abovementioned neuroticism example, it is crucial to configure the experimental situation to match the particular conditions under which neurotic individuals have been found to respond with negative emotions. Hence, any mapping found between neuroticism scores and manifest behavior can be viewed as being conditional with respect to the parameters of the experimental situation (i.e., the type and severity of the environmental stressor). Also, it is important to ensure that the measure being calibrated is tapping into the construct at the same level of *generality* and *temporality* as the manifest behaviors. Consistent with the specificity matching principle (Ajzen & Fishbein, 2005; Swann, Chang-Schneider, & McClarty, 2007), if the measure-to-be-calibrated taps into a relatively specific construct (e.g., attitudes toward potato chips), then the manifest behavior should be equally specific (e.g., how many grams of potato chips eaten in a year) whereas if the measure-to-be-calibrated taps into a relatively general construct, then the manifest behavior should be equally broad (e.g., number of social events attended per month). Similarly with temporality, if the measure-to-be-calibrated taps into behaviors or mental states over a long period of time, the manifest behaviors also need to be observed over an equally long temporal period whereas if the measure-to-be-calibrated taps into a transient mental state, then the manifest behavior used as reference point should reflect an equally transient manifest behavior (e.g., behavioral markers of transient anxiety). In addition, construct theory should be used to determine which particular facet of a construct is best suited to be calibrated to certain behavioral manifestations of the construct. For example, if number of mistakes in a detail-oriented task is to be used as a behavioral reference point to calibrate a conscientiousness measure, great care should be used to select the most appropriate lower-order facet of conscientiousness (e.g., Deliberation facet of the NEO-FFI; Costa & McRae, 1992).

## 1.5 Inspirations from the History of the Development of Instruments in the Natural Sciences

A brief glimpse into the history of the development of two important instruments in the natural sciences provides a useful context for discussing pertinent issues surrounding arbitrary metrics and the potential value of metric calibration in psychology. I will discuss, in turn, the history of the development of the thermometer and hygrometer.

The development of the thermometer was based on a basic principle – discovered in antiquity by Philo of Byzantium and Hero of Alexandria at about the end of the second century B.C. – that certain substances expand and contract under varying conditions (McGee, 1988). Many centuries later, the idea of developing an instrument to quantify this effect emerged. Although still contested, historians of science usually consider

Galileo Galilei, Santorio Santorio, Cornelius Drebbel and Robert Fludd as serious candidates for the honor of having "invented the thermometer" sometime in the 1500s (Middleton, 1966). These early thermometers or *thermoscopes* as they were properly called, used as their thermometric substance the expansion and contraction of air, which would displace water in an elongated tube. Of relevance to my dissertation, these early thermoscopes did not have a scale and hence lacked any systematic metric or numbering system (Middleton, 1966). Furthermore, when Francesco Sagredo and Santorio Santorio, around 1612, first put some kind of scale on their respective thermoscopes, these scales did not involve any meaningful metric. These primitive scales involved the gradation of lines drawn on the tube, sometimes with two moveable threads tied to the stem, presumably to detect a change in temperature (see Figure 6, left). The first systematic scale used with an air thermometer was developed by Jean Leurechon around 1625, which had a scale ranging from 1 to 9 "degrees" (see Figure 6, right) and an air thermometer reported by Telioux around 1613 which had a scale ranging from 1 to 8 (Middleton, 1969).

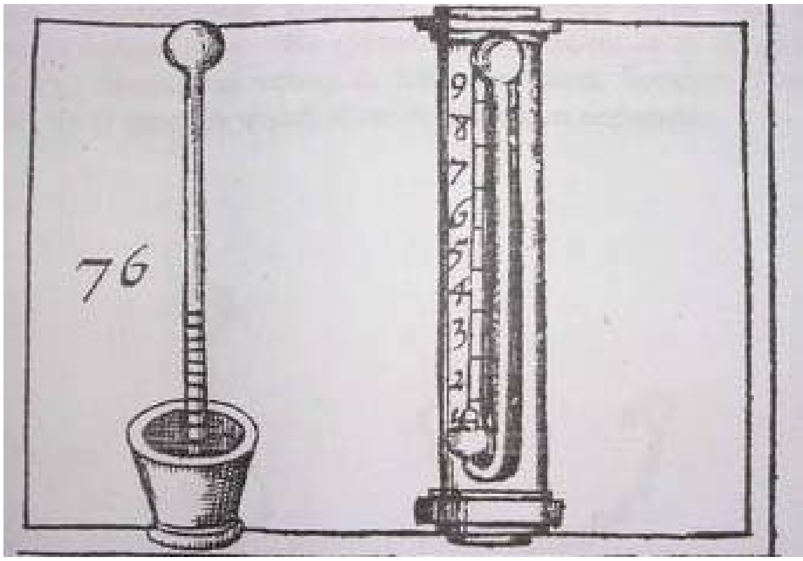


**Figure 6: Early thermometers having no metric (left) or arbitrary metric (right). Reprinted with permission of The John Hopkins University Press (© 1969) from Middleton (1969, p. 87, Figure 3.1).**

It is historically interesting to note that the scales of early air thermometers can be viewed as having had an arbitrary metric. That is, even though temperature measurements with these air thermometers may have been mostly valid (although see next paragraph) and reliable, without reference to other observable phenomena, the readings produced by these thermometers were devoid of much meaning. It is also very interesting to note that these early scales seemed strikingly similar to the Likert-type scales so pervasively used in modern day psychology.

Soon, however, a defect was discovered in the commonly used non-sealed air thermometers, such that they would respond to changes in air pressure as well as changes in temperature (i.e., early air thermometers were also barometers). To remedy this situation, Ferdinando II de Medici created a sealed liquid-in-glass thermometer in about 1654 that was immune to atmospheric air pressure. Subsequently, many different types of sealed thermometers were developed using different thermometric substances (e.g., water, wines and other alcoholic spirits, mercury) and using different scales. It soon became apparent, however, that it would be much more useful, both in terms of the interpretation of thermometer readings and for comparing thermometer readings across laboratories using differently constructed instruments, if thermometers could somehow be standardized. Hence, some time in the middle of the 1600s, scientists started proposing that thermometers should be standardized in their construction and in their calibration to certain fixed points (and hence the scale used). Robert Hooke was one of the first, around 1665, to propose that thermometers should be calibrated using one fixed point, namely the freezing point of distilled water; around the same time, Christiaan Huygens proposed to use as a reference point *either* the degree of cold at which water begins to freeze or the degree of heat of boiling water as a universal standard, so that degrees of heat and cold could be compared across laboratories without having to use the same instrument. A long debate, spanning almost a full century, thereafter ensued concerning which fixed points (and how many) should be used to calibrate thermometers. For example, fixed points proposed included (to name a few): constant temperature of deep cellars under the Paris Observatory (Mariotte, circa 1679), snow and boiling water (Bartolo, circa 1679), freezing point of water and melting point of butter (circa 1688), melting point of ice and salt and the temperature of very deep cellars (circa 1688), and melting point of ice and body temperature (Isaac Newton, circa 1701). Eventually, Daniel Fahrenheit (circa 1724), René-Antoine Réaumur (circa 1730), and Anders Celsius (circa 1742) proposed to use the freezing point of water and boiling point of water as universal reference points, although they each proposed their own scales (32 °F and 212 °F; 0 °R and 80 °R; 0 °C and 100 °C, respectively were proposed as values for the freezing and boiling points).

A similar story emerges from reading the history of the development of the early hygrometers (instruments to measure humidity). For instance, one of the first documented hygrometers, Santorio's string hygrometer (circa 1612) (see Figure 7), was a simple device composed of a stretched out cord attached on both ends to a wall, with a lead ball fixed in the middle with a scale drawn nearby (Middleton, 1966). The logic underlying this measurement instrument was that as the moisture in the air increased, the length of the cord expanded whereas as moisture decreased (or dryness increased) the length of the cord contracted, moving the lead ball up or down, which could be quantified by the scale drawn on the wall.

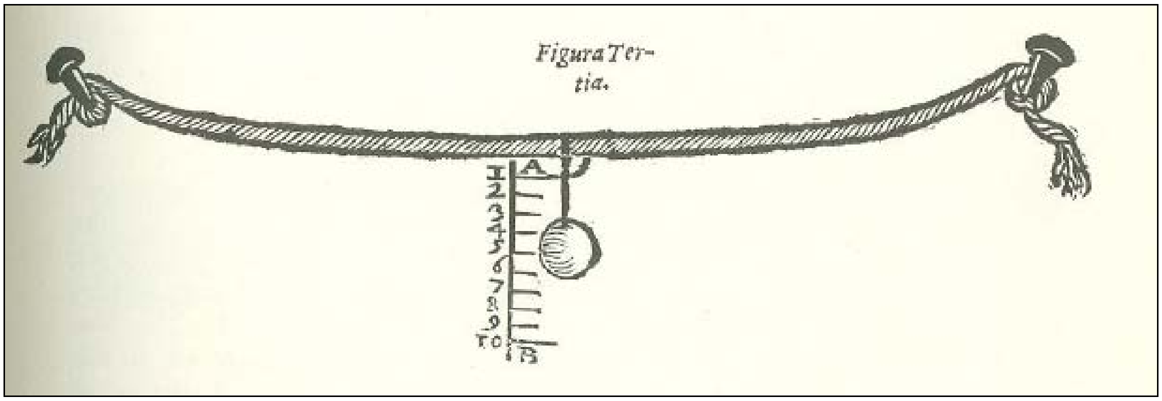


**Figure 7: Santorio's early string hygrometer having a scale with arbitrary metric. Reprinted with permission of The John Hopkins University Press (© 1966) from Middleton (1966, p. 21, Figure 1.9).**

As should be apparent, it is clear that the metric of this early string-hygrometer was arbitrary in nature, given that the scale values were not linked to any external reference points. Later hygrometers (e.g., de Saussure's hair-hygrometer and Deluc's whalebone-hygrometer), however, did include a meaningful metric by calibrating the devices to external reference points. For de Saussure's (circa 1778) hair hygrometer, for example, the reference points were an extreme condition of moisture (achieved by putting the hair apparatus under a bell-jar of which the sides and bottom were wet) and an extreme condition of dryness (achieved by enclosing a piece of sheet iron – previously made red-hot, cooled, and sprinkled with a mixture of powdered niter and cream of tartar – in a dry jar along with the hygrometer). Hence, the arbitrary values of de Saussure's early hair hygrometer (i.e., length of human hair) gained meaning as they were linked to these extreme manifestations of humidity.

Taken together, these historical sketches make clear that voluminous amounts of research was undertaken to calibrate thermometers and hygrometers to theoretically-relevant reference points (and many other instrument in the natural sciences). Such metric calibration research not only imbued temperature and humidity readings with more meaning, but also contributed in important ways to the cumulative knowledge base in these fields and correspondingly to theory development (e.g., theory of heat developing in step with the calibration of the thermometer, McCormmach, 2004). As it concerns my dissertation, the take-home message of these historical excerpts is: (a) that natural scientists agreed that the metric calibration of their measurement instruments was very important for the advancement of knowledge and (b) that this type of research involved the unique challenge of researchers reaching consensus as to the most theoretically-relevant and meaningful reference points to use in the calibration process. Consequently, a possible implication of these historical excerpts for my research is that perhaps it is time for psychological instruments to be improved in ways that are in a broad sense similar to the calibration of instruments in the natural sciences.[4] In particular, perhaps it is time for basic researchers in psychology to start considering the potential utility and feasibility of calibrating the metric of psychological instruments.

[4] It is important to keep in mind that I am not suggesting that psychologists follow strict parallelism to measurement and methodology used in the natural sciences. My goal is to use measurement examples from the natural sciences as metaphors (see Dooremalen & Borsboom, 2010) to help inspire novel ideas with respect to psychological measurement. Psychological measurement clearly involves unique methodological challenges that transcend measurement challenges in the physical sciences; hence, metric calibration in psychology needs to be tailored to these unique specific challenges.

## 1.6 Relevant Theorizing by Past Psychologists

*If I have seen further it is by standing on the shoulders of giants.* - Isaac Newton

Considering the scarcity of metric research in psychology, it may seem surprising that prominent psychology scholars have proposed theoretical ideas that are broadly consistent with my argument that reducing metric arbitrariness could potentially benefit basic psychological research. This observation seems even more remarkable given that specific research on metrics in psychology only emerged much later in the late 1990s. For instance, in a 1969 *American Psychologist* article, John Tukey propounded repeatedly that "amount, as well as direction, is vital" (p. 86). By this he meant that it is not just the direction of an experimental effect that is important, but by how much. In his own words:

> The physical sciences have learned much by storing up amounts, not just directions. If, for example, elasticity had been confined to "When you pull on it, it gets longer!" Hooke's law, the elastic limit, plasticity, and many other important topics could *not* have appeared (emphasis added) (p. 86).

It is important to keep in mind that Tukey is not simply arguing that we should be cognizant of the effect size of experimental findings. He is specifically making a plea that researchers should "store up amounts," which implies that simply reporting amounts is insufficient, and hence that researchers should actually keep track and become familiar with particular amounts (see also Tukey, 1991). Another quotation from the same article makes this point even more clearly:

> Measuring the right things on a *communicable scale* lets us stockpile information about amounts. Such information can be useful, whether or not the chosen scale is an interval scale. Before the second law of thermodynamics – and there were many decades of progress in physics and chemistry before it appeared – the scale of temperature was not, in any nontrivial sense, an interval scale. Yet these decades of progress would have been impossible had physicists and chemists refused either to record temperatures or to calculate with them (p. 87, emphasis added).

It seems clear from this passage that Tukey is espousing that we should keep track of the particular magnitude of an experimental effect in terms of scores that have a meaningfully interpretable metric. According to Tukey, it is valuable and important to keep track of the particular amount of an effect as expressed in the *units of measurement* of the instrument used. Of course, this "stockpiling" of information is only valuable if the units of measurement are indeed meaningful. If, on the other hand, the metrics of measures in a research domain are arbitrary and hence lacking in meaning, then it would not be surprising that researchers fail to stockpile this kind of information. In general, psychologists doing basic research fall into this category. Indeed, Tukey (1969) spoke on this matter and explicitly lamented that "being so disinterested in our variables that we do not care about their units can hardly be desirable" (p. 89). One way of explaining why psychology researchers have not heeded Tukey's ideas is indeed because virtually all psychological metrics used today are arbitrary (Blanton & Jaccard, 2006a). An important aspect of the current research, therefore, is to argue that given we now have a preliminary psychometric understanding of the steps required to make metrics of psychological instruments more meaningful (Blanton & Jaccard, 2006a, 2006b; Embretson, 2006), it is now time to tap into the great potential of Tukey's words of wisdom at a practical level.

Much inspired by Tukey's ideas, Jacob Cohen also had important things to say regarding arbitrary metrics in psychology. Resonating particularly well with Tukey's theorizing, Cohen (1994) emphasized that if all psychologists learn from a study is the direction of an effect, then we have not really learned much at all. In his own words: "But if all we, as psychologists, learn from a research is that A is larger than B ($p < .01$), we have not learned very much. And this is typically all we learn" (p. 1001). In a broad sense, this quote implies that we should be learning a lot more from a study than whether the groups differed in one direction or the other. That is, we should learn by how much the groups differed with respect to particular values of the dependent variable (DV) and also consider the departing value of the effect (e.g., manipulation increased DV scores by 2 units from a departing value of 4 units). In a strikingly similar style, Kirk (1996) made almost exactly the same remark while speaking about the severe limitations of the typical use of null hypothesis statistical significance testing (NHST), one of which is that it evaluates only ordinal relationships:

> …a rejection [of the null] means that the researcher is pretty sure of the direction of the difference. Is this any way to develop psychological theory? I think not. How far would physics have progressed if their researchers had focused on

> discovering only ordinal relationships [such as those tested by conventional NHST]? … knowing A is greater than B is not enough (p. 754).

Hence, in a broad sense, these poignant prods imply that we need to go beyond direction and start being cognizant of the units we work with. Indeed, Kirk specifically mentioned that the use of confidence intervals (CI) can help precisely because CIs use the same unit of measurement as the data, which “facilitates the interpretation of results and makes trivial effects harder to ignore” (p. 754). This kind of theorizing directly implies that researchers should be more acquainted with the metrics they work with. In fact, Cohen explicitly stated that psychologists need to “respect” the units they work with:

> To work constructively with “raw" regression coefficients and confidence intervals, psychologists have to start respecting the units they work with, or develop measurement units they can respect enough so that researchers in a given field or subfield can agree to use them. In this way, there can be hope that researchers' knowledge can be cumulative. There are few such in soft psychology. A beginning in this direction comes from meta-analysis, which, whatever else it may accomplish, has at least focused attention on effect sizes. But imagine how much more fruitful the typical meta-analysis would be if the research covered used the same measures for the constructs they studied. Researchers could get beyond using a mass of studies to demonstrate convincingly that "if you pull on it, it gets longer. (emphasis added, p. 1001)

“Respecting” the units one works with implies that one should become intimately acquainted with those units by first of all keeping track of them. Second, it also means that one should try to make sense of those units which requires that they actually be meaningfully interpretable. Indeed, the specific part of this last quote about developing measurement units that can be respected enough that different researchers can agree to use them specifically implies that researchers should develop non-arbitrary metrics that are respectable enough that different researchers can agree to use them (as occurred in the case of the development of thermometer and hygrometer scale metrics).

From a slightly different perspective, Paul Meehl indirectly argued for the importance of score metrics in the context of the nature of theory testing in psychology. Meehl (1978) mentions that for science in general, a theory is corroborated to the extent that it has been subjected to potentially risky tests. That is, “the more dangerous tests [a theory] has survived, the better corroborated it is” (p. 817). In other words, the higher the specificity of the predictions a theory confirms the more strength the theory acquires (see also Popper, 1968/1959). Hence, specific point predictions that involve estimating numerical point values are inherently more valuable for theory testing purposes than general directional predictions (see also Meehl, 1990a, 1990b). Indeed, Meehl specifically mentioned that "a theory that makes precise predictions and correctly picks out *narrow intervals* or *point values* out of the range of experimental possibilities is a pretty strong theory" (p. 818, emphasis in original). For psychology, this would translate into making specific theoretical predictions about how an experimental manipulation would pattern itself on specific locations of the metric of a DV. Meehl (1990a, 1990b) lambasted psychologists for invoking NHST in its weak form and laments that directional theory testing is highly sub-optimal because it subjects psychological theory to very weak or lenient tests.

Relatedly, Meehl (1967) has argued that the hurdle which physical sciences theory must surmount generally increases with improvement in experimental design and measurement whereas in the psychological sciences improvement in experimental design generally leads to easier hurdles for a theory to surmount (a situation Meehl calls a "methodological paradox"). This is the case because in the physical sciences, with increased knowledge, increasingly precise point-value predictions are made whereas in psychology only directional tests are ever made. Of course, Meehl mentions that most psychological theories (and the knowledge base from which theories are derived) may not be sufficiently quantitatively developed to be able to generate point-predictions (but see Granaas, 2002). Hence, he admits that although this state of affairs is surely unsatisfactory, it is "nobody's fault" given it is unclear how behavioral scientists would attempt to develop strong enough theory to be able to generate point-predictions that stand a larger risk of refutation (and hence would corroborate the theory in a much stronger way). That being said, Granaas (2002) questions whether psychologists' reluctance to make point predictions stems from (a) their theories not being sufficiently developed quantitatively or (b) from psychologists not being trained to think this way. From a metric perspective, however, I would argue that increased attention to the metric of our measures and research specifically aimed at calibrating our metrics to behavioral fixed points could represent the preliminary steps required to move our field into a direction that could make it eventually possible to adopt Meehl's recommendations of making more specific point-value estimate predictions (see also Mulaik, Raju, & Harshman, 1997). In fact, one could argue that metric research is a required and necessary first step to even consider the possibility of such kinds of point-prediction theory testing.

In summary, a fair bit of theoretical discourse exists that generally supports the idea that making metrics more meaningful could benefit the progress of basic psychological science. Ranging from theoretical ideas on more quantitatively-oriented theory testing, to pleas on developing units of psychological measures that researchers are willing to respect, to the stockpiling of information about particular amounts of experimental effects, taken together, these theoretical ideas provide the context from which my main thesis is derived. Ultimately, these broad-minded theorists had the foresight to discuss ideas that are broadly consistent with my argument that making metrics of psychological instruments more meaningful can move our field forward and benefit basic psychological research.

## 1.7 Past Research on Arbitrary Metrics

Past research that speaks more directly to the issue of metrics in the psychology arena is surprisingly limited. Based on my review of the literature, the only existing research on the issue of metrics is the relatively small amount of research done from an applied perspective in the area of clinical psychology (Kazdin, 1999; Kazdin, 2006; Sechrest et al., 1996), forensic psychology (Hanson, 2009; Pirelli, Gottdiener, & Zapf, 2011), sport psychology (Andersen, McCullagh, & Wilson, 2007), and individual-level diagnoses (Blanton & Jaccard, 2006a, 2006b; Blanton et al., 2009).[5] I discuss these in turn.

[5] Research approaches within the domain of industrial/organizational (I/O) psychology have also examined topics that could be viewed as related to metric calibration issues broadly construed (e.g., expectancy charts, utility analysis). However, given that this research does not directly tackle the issue of empirically developing meaningful units of measurement for one's measures, I will defer my discussion of these approaches until the General Discussion.

### 1.7.1 Clinical Psychology

Most of the research concerning arbitrary metrics from a clinical perspective involves papers that discuss the challenging task of evaluating the true effectiveness of clinical interventions on treating psychopathologies. Known under the rubric of *clinical significance*, this research aims at determining what constitutes proper evidence for showing *clinical significance* rather than merely showing *statistical* or *practical significance* (Chambless & Hollon, 1998; Kendall, 1999). For example, in a review of literature, Kazdin (1999) conceptualized the notion of clinical significance as the

> practical or applied value or importance of the effect of an intervention – that is, whether the intervention makes a real (e.g., genuine, palpable, practical, noticeable) difference in everyday life to the clients or to others with whom the clients interact (p. 332).

Kazdin conceptually analyzed the various ways that different researchers have operationalized *clinical significance* and concluded that all of these different meanings involve ambiguities that need to be clarified. Kazdin recommended that much more research effort should be focused on determining cut-off scores for outcome measures that can identify individuals who have changed in marked ways in everyday functioning by calibrating the metric of outcome measures commonly used in psychotherapeutic intervention studies.

Indeed, in his 2006 commentary on Blanton and Jaccard's (2006a) target article, Kazdin specifically delineated the numerous problems of using measures with arbitrary metrics for research on evidence-based psychotherapy. His main point was that the thorny issue of determining the clinical significance of intervention studies would be greatly improved if outcome measures used in intervention studies were calibrated against real-world referents as to reduce the metric arbitrariness of these outcomes measures (see also Kazdin, 1999, 2001). Kazdin argued that using outcome measures with non-arbitrary metrics would allow one to better gauge the actual impact of an intervention on a client's everyday functioning. For example, a depression inventory whose scores were linked to frequency of actual crying episodes would allow a better assessment of the actual impact of an intervention on client functioning.

Similarly, but said much more piercingly, Sechrest et al. (1996) delineated the interpretational problems in judging the effectiveness of clinical intervention studies when outcome measures have an arbitrary metric. For instance, Sechrest et al. reviewed a major national study of the treatment of depression, showing that BDI (Beck & Steer, 1987) scores decreased from about 26 points to about 17 points post-treatment (Watkins et al., 1993). What do these findings mean other than that depression decreased *to some extent* over the course of the study? Sechrest et al. stated bluntly: "nothing much we think, unless one has good understanding of just what is entailed by that *to some extent*" (p. 1065). They further stated that it is impossible to gauge the degree of effectiveness of most psychotherapy intervention studies (even psychotherapy meta-analyses, e.g., Lipsey & Wilson, 1993), because at best these studies express treatment effects in standard deviation units on whatever outcome measures were used. Sechrest et al. argued that for psychotherapy findings to be interpretable, findings must be expressed in terms of actual change in behavior or functioning rather than simply assuming change from a metric of uncertain meaning. Hence, they strongly recommended that the outcome measures used in intervention studies be calibrated against "external implications" reflected in actual behavior theoretically related to the construct at hand, as to imbue the metric of outcome measures with some inherent meaning and interpretability.

Although Sechrest et al. (1996) discussed issues pertaining to arbitrary metrics mostly from the applied context of clinical psychology, they also made statements about metrics directed to psychology more generally. For example, in their own words they stated that "science, [and] understanding of behavior, […] would be advanced by a better understanding of the measures by which the phenomena we concern ourselves are gauged" (p. 1068). In other words, it seems in the eyes of Sechrest et al. that understanding of human behavior more generally could be advanced by increasing our attention to the meaning of the metrics of the measures we use to assess psychological phenomena. This idea is stated more clearly and convincingly in their concluding paragraph, which will be included in its entirety due to its relevance and vigor:

> Our belief is that progress in psychology, [including the understanding of psychotherapy], like progress in all science, depends strongly on the quality of psychological measures. Psychologists cannot claim to have high-quality

> measures if they do not have full knowledge of their implications. Currently, that knowledge cannot be claimed for most measures used in psychology. We believe that knowledge, understanding, and progress in the science of psychology would be furthered greatly by concerted efforts to calibrate psychological measures in a variety of ways that are now available and that are sadly neglected. These methods include calibration of measures against each other so that it is possible to make accurate comparison across studies, but behavioral and other real-life implications should be accorded highest priority (p. 1071).

This concluding paragraph is most unambiguously consistent with the main thesis of my dissertation. Precisely how the calibration of measures could benefit basic psychological research, however, still remains unclear and unspecified. As mentioned, one of the goals of this dissertation is to specify precisely how metric research can potentially benefit basic psychology.

In a praiseworthy demonstration of Sechrest et al.'s (1996) general recommendations, Harman, Manning, Lurie and Liu (2001) published a large scale study that specifically linked mental health status measure scores (at time 1) to the probability of occurrence of subsequent major life events (at time 2). They framed the goal of their research as helping clinicians, researchers, and policy makers more easily interpret and gauge the actual significance of intervention outcomes. Harman et al. examined three mental health status scales including the Global Assessment Scale (GAS; individual life functioning), the Schedule for Affective Disorders and Schizophrenia (SADS; mood, anxiety, and delusions) subscale, and the Schizophrenia Subscale of the Brief Psychiatric Rating Scale (BPRS; emotional withdrawal, guilt, hostility, and disorientation). The major life events used in the study were psychiatric hospitalizations, victimizations, arrests, and suicide attempts, all assessed by patient self-report during face-to-face interviews.

Using a logistic robust regression, Harman et al. (2001) found, to list a few examples: that an 8 point increase on the GAS (metric range = 0 to 100) corresponded to a 24% decrease in probability of a suicide attempt; that a 5 point increase on the SADS depression subscale (metric range = 0 to 73) corresponded to a 19% increase in probability of psychiatric hospitalization and a 36% increase of a suicide attempt. It was argued that these kinds of linkages can help clinicians and policy makers interpret results of clinical interventions because rather than simply reporting that an intervention increased GAS scores by 8 points, one could report that the effect of the intervention was equivalent to a 24% reduction in probability of suicide attempt.

Another interesting implication of Harman et al.'s study is that it can be used (taking into account sampling differences) to interpret the clinical significance of past studies. For example, a study on the effect of risperidone versus haloperidol in treating refractory schizophrenia, showed that patients taking risperidone had post-treatment BPRS scores that were 2.3 points lower than patients taking haloperidol ($p < .05$, $d = .15$; Wirshing et al., 1999). According to Harman et al.'s calibration results, this effect translates into an approximate decrease of 5% in the probability of a psychiatric hospitalization, which is clearly more meaningfully interpretable than knowing the results of the study based solely on scores having an arbitrary metric, arbitrary effect sizes, and arbitrary *p*-values.

### 1.7.2 Forensic Psychology

Although still in its early stages, preliminary conceptual work has been done by Pirelli, Gottdiener, and Zapf (2011) with respect to the use of non-arbitrary metrics for competency to stand trial assessment instruments. In their review of the literature, it is concluded that each of the eleven competency to stand trial assessment instruments used in the forensic literature has an arbitrary metric and that this is problematic for both researchers and practitioners. Competency to stand trial instruments are especially important due to the costs associated with poor or flawed competency to stand trial evaluations should an incompetent defendant incorrectly be forced to stand trial or should a competent defendant be incorrectly committed to a forensic psychiatric facility. For example, the Competency Screening Test (CST; Lipsitt, Lelos & McGarry, 1971) is a self-administered measure containing 22 sentence completion items which are coded by independent judges as “0” (incompetent), “1” (marginally competent), and “2” (competent). Example items include “When I go to court the lawyer will…” and “When they say a man is innocent until proven guilty…” Composite score can range from 0 to 44, with a total score of 20 or below demarcating incompetent from competent defendants. Pirelli et al. argued that this measure (as all other competency to stand trial instruments) is seriously flawed given that the metric of the measure is arbitrary and thus it is quite unclear what the total scores really mean with respect to competency to stand trial. Without linking CST scores to specific competency to stand trial behaviors that are theoretically-relevant and meaningfully interpretable, CST scores remain ambiguous at best. As Pirelli et al. pointed out, it is axiomatic that the number “1” should always be located between “0” and “2” on any scale and that assuming a score of “1” on an instrument corresponds to a neutral point on the underlying construct of competency is completely unfounded.

Another important problem related to arbitrary metrics of instruments with composite scores like the CST is that, because the issue of arbitrary metrics applies both at the item and composite score level, two persons may arrive at the same total score via two completely different routes and this may mean quite different things in terms of actual behavior. For example, for the CST, a defendant could have a total score of 20 by receiving twenty “1”s and two “0”s or by receiving ten “2”s and twelve “0”s, which may translate into quite distinct competency to stand trial behaviors. Pirelli et al. concluded by recommending that it is imperative that researchers reduce the metric arbitrariness of competency to stand trial for the good of science and society, by empirically linking test scores to real-world competency to stand trial behavioral referents that are deemed by experts in that area to be theoretically-relevant and interpretable.

Paralleling Pirelli et al.’s (2011) general ideas, recent work by Hanson (2009) followed the same logic but was applied in the context of risk assessment measures used to predict crime and violence. Reviewing the literature on risk assessment tools for crime and violence (e.g., sexual deviancy, aggression measures), Hanson concluded that crime and violence risk assessment tools used to predict subsequent criminal and violent behaviors (sexual and violent recidivism, respectively) need stronger psychometric properties and would greatly benefit from having non-arbitrary metrics. Indeed, Hanson, Helmus, and Thornton (in press) reported research examining the empirical linkages between the scores from the most commonly used sexual recidivism risk tool in Canada and the U.S. (i.e., the Static-2002; McGrath, Cumming, & Burchard, 2003) and probability of re-committing a sexual offence. The Static-2002 tool combines objective (e.g., demographic, previous sexual offences) and self-report (e.g., deviant sexually interests) information for five content dimensions within a professional structured interview context (age, persistence of sex offending, deviant sexual interests, relationship to victims, and general criminality). For instance, an individual with 4 prior sexual convictions would receive a subscore of “3” for the “persistence of sex offending” dimension; an individual with any non-sexual convictions would receive a score of “1” for the “general criminality” dimension. These scores are then summed and recoded in a weighted fashion to produce the final total composite scores ranging from 0 to 14, with higher numbers representing higher levels of sexual crime risk.

In a large sample (*N* = 867), Hanson et al. (in press) found informative empirical mappings between Static-2002 scores and probability of sexual recidivism for both rapists and child molesters. For instance, Static-2002 scores of 0, 1, and 2 (for rapists) were associated with recidivism rates of roughly 10% or lower whereas Static-2002 scores of 9 and above corresponded to recidivism rates of 50% or greater. Although not framed as such in this particular report, evidence of this kind, which demonstrates empirical linkages between Static-2002 scores and meaningful external reference points, imbues the metric of Static-2002 scores with more meaning and hence increases score interpretability.

### 1.7.3 Sport Psychology

Researchers in sport psychology typically raise the issue of arbitrary metrics in the context of interpreting exercise intervention studies aimed at helping athletes improve the mental aspect of their sport with the ultimate goal of improving actual athletic performance. For example, Andersen et al. (2007) reviewed all articles in three of the top sport psychology journals published in 2005 and concluded that 86% of studies that used measures with arbitrary metrics did not discuss the results in terms of real-world sport behaviors and that this severely limits knowledge advancement in the field. In particular, Andersen et al. argued that if measures of mental subjective states are not calibrated against real-world sport behavior, then there is no way of knowing whether the effect (mean group difference between treatment and control group) of an exercise intervention study is meaningful or worth paying for. In their own words:

> Establishing that an intervention helps reduce competitive state anxiety by an average of 8 points on an inventory seems to be a diminished form of legitimization evidence. Is 8 points a big drop? Or better yet, is 8 points worth paying for? The answer to both those questions is that we really do not know. For a coach, which of the following would be more convincing: (a) with this relaxation and imagery program we can drop your runners' anxiety scores by 10 to 15 points, or (b) with this relaxation and imagery program we can reduce your runners' times by an average of 2.0%? We may be able to say the former, but the coach wants to hear the latter. And on the latter, in most cases, we must be silent, otherwise the aroma of snake oil will begin to waft across the sport and exercise psychology landscape (p. 666).

Hence, Andersen et al. make a strong case that exercise intervention studies that use outcome measures with arbitrary metrics are severely limited in terms of their interpretability. Strong recommendations are made for the calibration of sport psychology measures to real-world sport behaviors that are more meaningfully interpretable to coaches and practitioners.

### 1.7.4 Individual-Level Diagnoses

In a different vein altogether, Blanton and Jaccard (2006a) criticized the practice of giving individual-level diagnoses of "implicit racial preferences" to individuals based on their responses to an online instantiation of the race IAT (Greenwald et al., 1998).[6] The race IAT assesses "implicit preferences" by requiring individuals to classify certain types of stimuli (words and pictures) presented serially on a computer screen. In the case of the race IAT, the categories are Whites (pictures of Caucasian individuals) versus Blacks (pictures of African-American individuals) and pleasant versus unpleasant words (e.g., "sunshine" or "vomit", respectively). In a first task, participants' classify as quickly as possible (without making too many mistakes) whether the presented stimulus falls into the category of "White or pleasant" by pressing one key or whether the stimulus falls into the category of "Black or unpleasant" by pressing another key. The fundamental unit of analysis is the time taken to make these categorizations (i.e., response latency; RT). This task is generally referred to as the *compatible* task. In a second task, individuals classify as quickly as possible whether the stimulus falls into the category of "White or

[6] This website can be accessed via the following link: https://implicit.harvard.edu/implicit/ .

unpleasant" with one key or whether the stimulus falls into the category of "Black or pleasant" with another key. This task is generally referred to as the *incompatible* task. The "IAT effect" is calculated as the difference between the mean RTs from the incompatible task and the compatible task divided by the variability of the RTs (in addition to other transformations, see Greenwald, Nosek, & Banaji, 2003). Individuals who perform the compatible task faster than the incompatible task end up with positive race IAT scores and are characterized as having "automatic preferences for Whites over Blacks" whereas those performing the incompatible task faster than the compatible task end up with negative race IAT scores and are characterized as having "automatic preferences for Blacks over Whites."

As previously mentioned, Blanton and Jaccard (2006a, 2006b) provide strong arguments against the strategy of *meter reading* and *norming*, which the researchers at the Project Implicit website use to make absolute statements about individuals' standing on the underlying dimension of "implicit preferences." Blanton and Jaccard's main message is that it is both scientifically unfounded and ethically impermissible to make individual diagnoses of "implicit racial preferences" based on the scores of a measurement procedure that has a non-calibrated arbitrary metric. As previously elaborated upon, there are a multitude of factors that can conspire to shift the zero point of IAT scores away from the theoretical midpoint of no implicit preference (not to mention the logical reasons against meter reading; see Figure 1, panel B). In the particular case of the race IAT, Blanton and Jaccard mention that stimulus features can influence measurement scores, if for example, the pleasant words are more positive in character than the negative words are negative in character or if the pictures depicting African-Americans are more prototypical of Blacks than the pictures depicting Whites are prototypical of Whites (Bluemke & Friese, 2006). What's more, the various algorithmic transformations imposed on the raw RT data can also shift the zero point on the IAT away from the theoretical midpoint.

Indeed, in an analysis of race IAT data, Blanton and Jaccard (2006b) provided evidence against a *meter reading* strategy of IAT scores. Creatively, they aggregated scores from a standard race IAT using the original IAT scoring algorithm (Greenwald et al., 1998) and also using the new scoring algorithm (Greenwald et al., 2003) and then regressed the IAT scores from the new algorithm onto IAT scores aggregated using the original algorithm. Interestingly, a statistically significant non-zero intercept was found, demonstrating that a participant receiving a score of “0” on the old algorithm would now, on average, receive a positive IAT score with the new algorithm. Therefore, a participant diagnosed as lacking an implicit preference for Whites over Blacks in the year 2000, would now, based on exactly the same IAT responses, be diagnosed as having an implicit preference for Whites over Blacks. As this example demonstrates, the score on the IAT metric that maps onto the underlying dimension of no implicit preference must be empirically established rather than being embraced as a measurement assumption. To make individual diagnosis claims about the absolute standing of individuals on a psychological construct, which are defensible (both scientifically and ethically), it is imperative to have an empirically calibrated measure with a non-arbitrary metric. This is not the case with the IAT and no calibration research has been done with the IAT to achieve a non-arbitrary metric.

That being said, Blanton and Jaccard (2006a, 2006b) explicitly stated that metric arbitrariness is generally *not* an issue for theory testing purposes within the realm of basic psychological research. They mention that for most research purposes in psychology, the use of measures with arbitrary metrics is not problematic when, for instance, the focus of the research is on the study of basic processes which aims to test for the presence or absence of predicted linkages between theoretical variables. Blanton and Jaccard’s position is based on the fact that testing directional predictions derived from theory (which typically represents the bulk of psychological research) only requires a relative interpretation of measurement scores, which is permissible for measures with arbitrary metrics. These authors are mute, however, on whether using measures with non-arbitrary metrics could benefit basic psychology. Hence, my main thesis clearly goes beyond Blanton and Jaccard’s analysis, such that I make the specific claim that metric arbitrariness is also an important issue in basic psychological research and that calibrating the metrics of our instruments has the potential to benefit basic psychological research in several important respects.

# Chapter 2

## 2 Benefits of Metric Calibration for Basic Psychological Research

To re-iterate, the goal of my dissertation is to make the case that it is both useful and feasible to calibrate the metric of instruments commonly used in basic psychological research. To achieve this goal with regard to the utility of metric calibration, I will first present a conceptual analysis that elaborates on the potential benefits of non-arbitrary metrics for basic psychological research, given the premise that one is working with empirically established calibrated metrics. I will do so by delineating arguments for four distinct categories of benefits of non-arbitrary metrics for basic psychological research. Although distinct, the four benefits are hierarchically related with respect to how specific versus general the potential benefits are for basic psychological science. Hence, I will elaborate on these four distinct categories of benefits in an ascending order, from more specific benefits to more general benefits. I will argue that non-arbitrary metrics can benefit basic psychological research in the following four respects: (1) help in the interpretation of data, (2) facilitate construct validity research, (3) contribute to theoretical development, and (4) facilitate the general accumulation of knowledge (see Table 1 for a list of all proposed benefits).

**Table 1: List of proposed benefits of non-arbitrary metrics.**

| Proposed benefits |
|---|
| **1. Help in the interpretation of data** |
| a. Enhance the interpretability of statistical effects |
| b. Allow and facilitate the extraction of more information from data patterns |
| c. Help overcome important limitations of NHST |
| **2. Facilitate construct validity research** |
| a. Metric calibration can shed brighter light on psychological constructs |
| b. Metric approach can help with conceptual challenges that arise in construct validity research |
| c. Provide benchmark for detecting measurement problems and/or improving measures |
| **3. Contribute to theoretical development** |
| a. Aid and facilitate theoretical debates involving absolute claims |
| b. Allow for more precise theorizing via enhanced scientific language |
| c. Provide preliminary interpretive platform for quantitative testing of theories (Meehl, 1978) |
| **4. Facilitate general accumulation of knowledge** |
| a. Metric calibration findings are valuable information in their own right |
| b. Metric approach as guiding framework for cataloguing the magnitude of psychological effects |
| c. Facilitate phenomenon-based research (Rozin, 2001) |

To strengthen my case regarding these proposed benefits, in the General Discussion I will further demonstrate some of these benefits by applying some of my preliminary metric calibration findings to actual research findings in the literature.

## 2.1 Help in the Interpretation of Data

The first, and most specific, benefit involves facilitating the interpretation of data. To support my argument that non-arbitrary metrics could facilitate the process of interpreting data, I will elaborate on the following three ways that working with calibrated metrics could help the interpretation of data: (a) enhance the interpretability of basic statistical effects, (b) allow for the extraction of more information from data patterns, and (c) help overcome important limitations of NHST. I will elaborate on each of these aspects and support my reasoning with corresponding relevant examples.

### 2.1.1 Enhanced Interpretability of Statistical Effects

First, I put forth that working with calibrated metrics could enhance the interpretation of statistical effects for common statistical procedures. That is, if psychological variables were measured with instruments having non-arbitrary metrics, analyses using common statistical techniques (e.g., *t*-test for 2-group between-subjects design, moderated multiple regression) would be enhanced in the sense of being easier and more meaningful to interpret. I will unpack this point by focusing most of my attention on moderated multiple regression (MMR), which has become the preferred statistical procedure in basic research to analyze the interaction between continuous predictors or between a continuous and categorical predictor (rather than using the sub-optimal median split method and ANOVA; MacCallum, Zhang, Preacher, & Rucker, 2002). In the case where two continuous predictors (e.g., X and Z) are hypothesized to interact to predict an outcome variable (e.g., Y), predictors are typically mean-centered and a product term created (Aiken & West, 1991). Then, a statistically significant interaction term is typically followed up by plotting and statistically testing the simple slopes between X and Y at 1 standard deviation (SD) above the sample specific mean and 1 SD below the sample mean of Z (Aiken & West, 1991). Graphically, this would be depicted as in Figure 8 (panel A), which reflects the examination of the relation between X and Y at 1 SD above the mean on Z (i.e., the positive slope) and the relation between X and Y at 1 SD below the mean on Z (i.e., the negative slope). (Alternatively, one could examine the relation between Z and Y at 1 SD above the mean on X and the relation between Z and Y at 1 SD below the mean on X.)

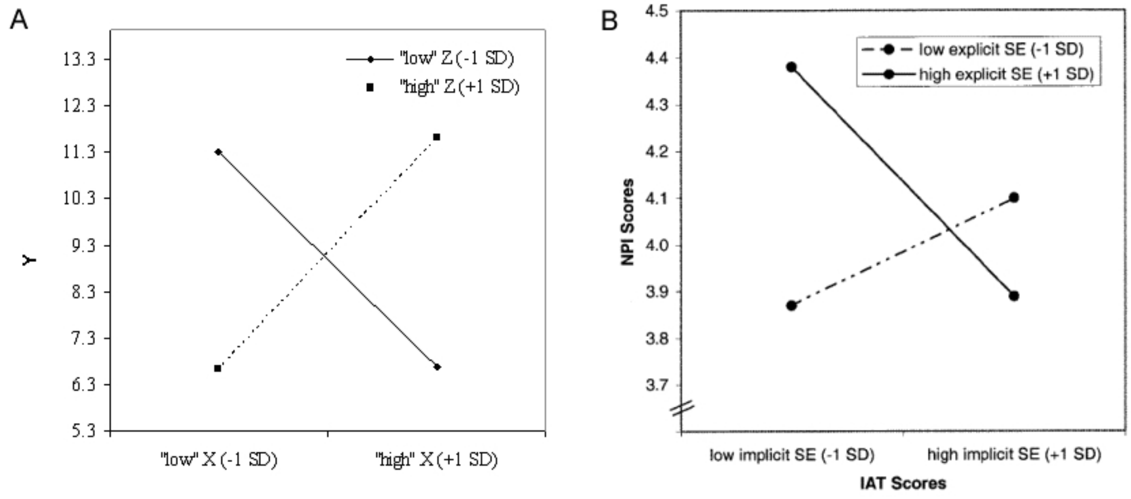


**Figure 8: A typical moderated multiple regression model when both predictors are continuous (panel A) with an actual example from the literature (panel B, Jordan et al., 2003; reproduced with permission).**

Using the +/-1 SD convention to examine the interaction between two continuous predictors has become common practice in the literature (Cohen, Cohen, West, & Aiken, 2003) and in general it does the job of explicating these types of interactions. From a metrics perspective, however, it becomes clear that the meaning of examining the relation between X and Y at the particular value of 1 SD above (or below) the mean of Z may be quite limited. This is the case because assuming the metric of the measured Z variable is arbitrary, it is unclear what a value of 1 SD above (or below) the mean of Z actually means with reference to the underlying dimension (other than a relative interpretation such that a value 1 SD above the mean implies greater levels of the underlying construct as compared to the mean, which in turn implies greater levels of the underlying construct as compared to 1 SD below the mean).

An actual case from the social psychological literature may more clearly demonstrate the limitations of using the arbitrary +/-1 SD convention when probing continuous predictor interactions. Jordan, Spencer, Zanna, Hoshino-Browne, and Correll (2003) examined the joint effect of explicit self-esteem (ESE) and so-called implicit self-esteem (ISE) on defensiveness, as reflected in narcissism, in-group bias, and cognitive dissonance reduction. As seen in Figure 8 (panel B), Jordan et al. found that the relation between ISE and self-reported narcissism became increasingly more negative as ESE scores increased. Probing the interaction more deeply by using the standard +/-1 SD convention, Jordan et al. found that "there was a significant negative relation between IAT scores and NPI scores for participants with high explicit SE (+1 SD)... [whereas]… among individuals with low explicit SE (-1 SD), the relation between implicit SE and narcissism was non-significantly positive" (p. 971). These results were argued to support their "nagging doubt" hypothesis, which states that individuals with "high explicit SE" differ markedly in their defensiveness depending on their levels of ISE. That is, "high explicit SE" individuals who have "low implicit SE" may experience "negative" implicit self-feelings as nagging doubts, which leads to defensive behavior.

It is important to note, however, that this interpretation hinges upon the assumption that ISE and ESE discrepancies at the statistical (or distributional) level translate into corresponding psychological discrepancies that are experienced subjectively as such. This is not necessarily the case when one appreciates the arbitrary nature of the metric of the self-esteem measures. Just because an individual has a high score on the ESE measure and a low score on the ISE measure, does not mean he or she will experience a subjectively felt discrepancy, as required by Jordan et al.'s position. This is so because the metrics of these self-esteem measures are arbitrary and hence it is completely unknown how particular scores on these measures map onto different locations on the underlying dimensions of the respective constructs. For instance, it is taken completely on faith that a score 1 SD below the mean on the Rosenberg (1965) self-esteem scale (RSES; typically a score of about "4" on a 1 to 7-point scale) actually reflects low self-esteem in absolute sense. The only way to know this is to empirically calibrate the metric of the measure to theoretically-relevant behaviors argued to be indicative of low self-esteem.

What's more, another typically unacknowledged limitation of the +/- 1 SD convention when working with arbitrary metrics is that the +1 and -1 SD values are relative to the sample specific range of scores. Thus, if the range of scores is different across two samples (e.g., scores ranging from 3 and 5 in one sample vs. 1 and 3 in another sample, on a 5-point scale), the +/- 1 SD values may refer to different levels of the underlying dimensions, further complicating the interpretation of these types of data. In other words, it is possible due to sampling error alone, to get different sample specific means (and sample specific standard deviations) for either predictors, which could lead to different theoretical interpretations of interaction patterns that would be spurious due to sampling error.

I argue, however, that if the instruments used to assess the predictors in these types of interaction analyses had non-arbitrary metrics, the interpretation of the data would be enhanced in at least three respects. First, rather than relying on the arbitrary +/-1 SD convention, calibrated values for the moderator variable could be used to statistically analyze these types of interactions. Hence, one could examine the relation of X and Y at particular calibrated values of Z, which correspond to external behaviors indicative of a high or low level of the underlying construct. This would enhance the interpretation because one would get a better sense of what the interaction means psychologically given the relevant slopes could be interpreted with respect to the behaviors corresponding to the calibrated values. Second, the data interpretation would be enhanced because the interaction analysis using calibrated values, which would be grounded in theoretically-relevant behavior, could yield different patterns of results that could have different and potentially important theoretical implications (see hypothetical example below). And finally, data interpretation would be improved because the use of (consensually agreed-upon) calibrated values would overcome the sampling error issue given that exactly the same calibrated values would be used across different samples rather than the fluctuating sample-specific values.[7] I will further discuss and demonstrate this potential benefit in the General Discussion.

Consider for the defensiveness research question from above, that the RSES measure had been calibrated to the probability of asking a clarifying question in a small group discussion (e.g., from a sociometer perspective), such that scores of “3.2”, “4.2”, “5.2”, and “6.2” on the RSES corresponded to probabilities of “0.2”, “0.6”, “0.7”, and “0.8”, respectively of asking a clarifying question. Given this calibration information, one could decide that the RSES value 1 SD below the mean (i.e., “4.2”) does not really reflect a condition of low self-esteem in a psychological sense, given that it is associated with a .6 probability of asking a clarifying question. Rather, one may decide that it would be more appropriate theoretically to examine the relation between ISE and narcissism at the RSES value of “3.2”, which corresponds to a qualitatively distinct behavioral manifestation of self-esteem which arguably is more diagnostic of the low end of the underlying self-esteem continuum (e.g., a .2 probability of asking a clarifying question).

### 2.1.2 Allow Extraction of More Information from Data Patterns

The second way non-arbitrary metrics could help data interpretation in psychological research is that non-arbitrary metrics would facilitate the process of extracting more information from data patterns. That is, using measures with calibrated metrics would allow researchers to glean more details from data patterns and hence facilitate more nuanced interpretations of data patterns. In a broad sense, this would be the case because of the more intuitive nature of calibrated metrics, which allow for a more natural focus on score interpretation. To unpack the reasoning behind this proposed benefit, I will again use specific relevant examples.

[7] This benefit could be demonstrated concretely by using a Monte Carlo simulation where an interaction with a certain known form (e.g., cross-over interaction) is defined in the population and then samples of size *n* (typical of sample sizes used in the literature) are repeatedly drawn from such population. Each sample could then be analyzed using both the conventional +/- 1 SD above the sample specific mean values and using calibrated values. Tabulated results would then show in concrete terms the superiority of the calibrated values approach in yielding more accurate conclusions with regard to the true interaction pattern.

First, in the case of a between-subjects design, I contend that if the DV is assessed using a measure with a meaningful metric, then between-group mean differences that occur at different locations on the measurement scale across studies would become more apparent and hence more easily noticed and interpreted accordingly. Also, and importantly, experimental effects that emerge at different locations on the scale could mean something quite different psychologically, hence DVs with non-arbitrary metrics may be quite useful in facilitating more nuanced interpretations of data. For example, consider a researcher studying the effect of self-construal on self-reported extraversion using Eysenck's IES scale (Eysenck & Eysenck, 1975). For a simple 2-group between-subjects design, imagine that a researcher finds that construing one's self in broader versus more concrete terms lead to higher levels of self-reported extraversion (mean of "7.1" vs. "6.0"; see Figure 9, Sample 1). Now imagine that the same researcher runs the same study again and finds the same general pattern except in the second sample the mean group difference is shifted down the scale (mean of "4.2" vs. "3.1" in the same direction; see Figure 9, Sample 2).

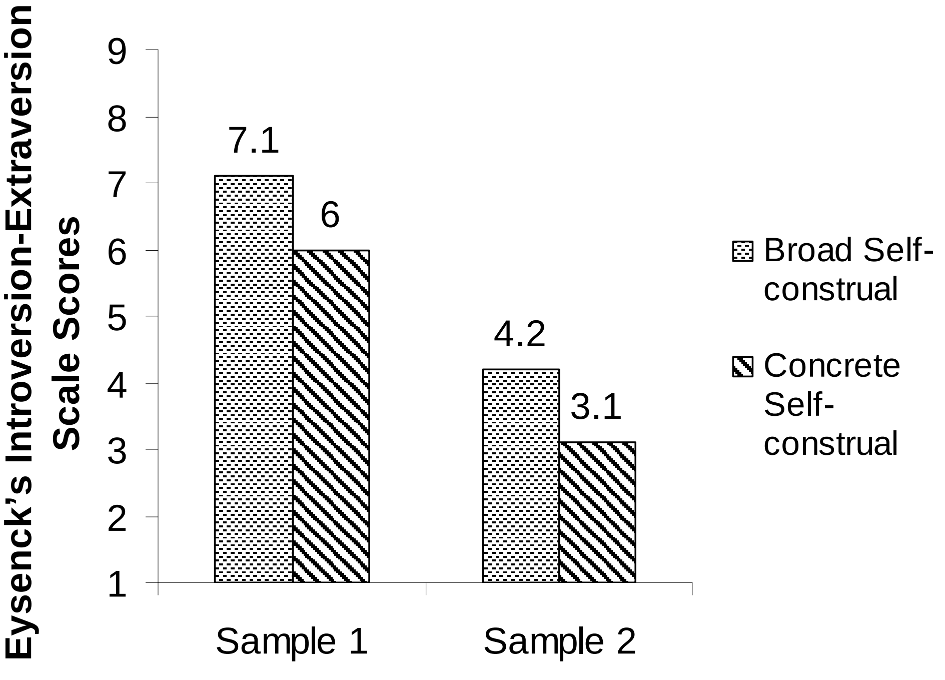


**Figure 9: Hypothetical experimental results across two samples at different locations on the DV scale.**

Although the effect was replicated in the predicted direction in Sample 2, the effect emerged at a different location on the scale of the DV measure. Standard research practice would typically ignore this fact and simply emphasize that the research finding was replicated across studies. If, however, the DV measure was calibrated to meaningful behavioral reference points (i.e., had a non-arbitrary metric), this difference across sample would become easier to notice. Most importantly, non-arbitrary metrics would allow a more nuanced interpretation of the data in these situations given that the experimental effects emerging at different locations on the DV scale can be interpreted with respect to the calibrated relevant behaviors. For instance, referring back to the hypothetical extraversion metric mapping in Figure 4 (dotted line), it can be seen that extraversion scores of "7.1" and "6.0" correspond to about 8 and 5 hours of socializing, respectively, whereas extraversion scores of "4.2" and "3.1" correspond to about 3 and 2.5 hours of socializing, respectively. It is clear that these experimental effects emerging at different locations on the DV scale would mean something quite different psychologically and hence should be interpreted as such. Hence, not only can non-arbitrary metrics make it more likely that these cross-sample differences are noticed in the first place, but the additional metric calibration information afforded by non-arbitrary metrics could also allow for more nuanced interpretations of data patterns that could also have theoretical value.

Furthermore, the exact logic above can also be applied to factorial designs and the interpretation of simple main effects that emerge at different locations on the DV scale. These subtle differences in simple main effects across data patterns would become much more salient with non-arbitrary metrics, and hence could also advance knowledge by facilitating more nuanced interpretations of data patterns for such factorial designs.

Finally, another illustration of the utility of non-arbitrary metrics in this context can be demonstrated using a slightly different scenario from the above example. Consider in that example, that the second sample revealed an effect opposite to the one in the first sample. That is, broad self-construal versus specific self-construal led to lower ("3.1" vs. "4.2"), rather than higher, levels of self-reported extraversion. With a calibrated metric, it would be easier to see that the second finding does not represent a failed replication of the first sample, but rather that the manipulation increases whatever personality characteristic is prevalent in the sample (i.e., extraversion vs. introversion). A broader point that is implied by this particular example is that non-arbitrary metrics may be especially useful for research investigations involving bipolar constructs. That is, it would be especially valuable to calibrate the scale midpoint on such bipolar constructs to diagnostic behavior that distinguishes individuals from the opposing poles of the construct. For instance, for Eysenck's IES scale (Eysenck & Eysenck, 1975), this could mean linking up the midpoint of the scale to a diagnostic behavior argued to distinguish an extravert from an introvert. The scale could then be converted into a more intuitive metric with the calibrated scale midpoint labeled as "0" and negative and positive values centered on this "0" value.

### 2.1.3 Help Overcome Limitations of NHST

The third and final way non-arbitrary metrics may help the interpretation of data is in the context of the limitations of NHST. Even though NHST is the dominant approach to hypothesis testing in psychology, it has, throughout its existence, been repeatedly attacked as a flawed or severely limited statistical practice (Berkson, 1942; Boring, 1919; Carver, 1978; Cohen, 1990, 1994; Cronbach, 1975; Cumming, 2008; Dracup, 1995; Eysenck, 1960; Falk & Greenbaum, 1995; Folger, 1989; Gigerenzer, 1998; Guttman, 1977, 1985; Hunter, 1997; Kirk, 1972, 1996; Lykken, 1968; McNemar, 1960; Meehl, 1967, 1978, 1990a, 1990b; Pedhazur & Schmelkin, 1991; Pollard, 1993; Rozeboom, 1960, 1997; Schmidt, 1996; Schmidt & Hunter, 1997; Shaver, 1993; Shrout, 1997; Signorelli, 1974; Thompson, 1993, 1996, 1997, 1998; for reviews see Harlow, Mulaik, & Steiger, 1997; Nickerson, 2000; Wagenmakers, 2007). Although criticism against NHST is multi-faceted and varied, the brewing controversy can be roughly summarized by five main criticisms: (a) NHST does not tell researchers what they want to know (i.e., it tells them the probability of obtaining a certain data point, *D*, given the null is true *rather* than telling them the probability that the null is true given the obtained *D*), (b) NHST is a trivial exercise given that the null hypothesis is *always* false with a large enough sample size, (c) rejecting the null in no way corroborates the substantive theory that implies the falsity of the null, (d) *p*-values yielded by NHST do not reflect result replicability, and (e) the dichotomous or all-or-none nature of NHST and the arbitrariness of the decision criterion α are problematic. Given these criticisms and limitations, methodologists have made different recommendations concerning NHST, ranging from an outright ban on the technique (Hunter, 1997; Schmidt, 1996), the use of the technique alongside the use of effect size and confidence intervals (Mulaik, Raju, & Harshman, 1997), the modified uses of the technique (Granaas, 2002; Kirk, 1996), or the use of alternative techniques such as Bayesian data analytic approaches (e.g., Kruschke, 2010; Lee & Wagenmakers, 2005).

Even though it can only be hoped that improved use of NHST will take place and/or that researchers will adopt superior alternative strategies, it seems clear that NHST is here to stay in some shape or form. Given this state of affairs, it seems fair to say even with improved usage and interpretation of NHST, that some limitations of NHST are inherently unavoidable. Given, for example, that the null hypothesis is always false with a large enough sample size and that the decision criterion α is an arbitrary value, NHST can be seen at best as insufficient or incomplete. One of the most frequently cited recommendations to overcome limitations of NHST is an increased focus on the estimation of effect sizes of experimental effects (Thompson, 2001, 2002; Wilkinson Task Force, 1999). Although I agree with this general recommendation, it is important to keep in mind that the quantification of effect sizes into *small* (e.g., $d = 0.2$), *medium* (e.g., $d = 0.5$), and *large* (e.g., $d = 0.8$) categories (Cohen, 1969, 1988) was proposed by Cohen only as a general guide to gauge the size of an effect (Thompson, 2002). As has been stated (Thompson, 2001, 2002, Kirk, 1996), if researchers sanctify these categories of effect sizes as much as they have sanctified *p*-value levels, we would "merely be being stupid in a different metric" (Thompson, 2001, p. 83). Hence, viewed from this perspective, effect size estimation can also be considered in a sense arbitrary and therefore insufficient in informing our research conclusions concerning empirical data sets.

The famous aspirin study (Belanger et al., 1988) provides a compelling case to support the claim that these standardized effect sizes can also be considered arbitrary. In this randomized controlled trial, physicians given low dosage of aspirin experienced fewer heart attacks (and hence deaths from heart attacks) than those in a placebo control condition. The trial had to be terminated early because the beneficial effect of the treatment was so clear that it was deemed unethical to continue to give the placebo to individuals in the control condition. What is most interesting, however, is that the magnitude of the experiment effect explained *less* than 1% of the variance in the DV. However, because the outcome variable had such a clear interpretation (255 heart attacks per 100,000 for the aspirin group compared to 440 heart attacks per 100,000 for the control group and hence fewer deaths), the importance of the experimental effect was easy to gauge. Hence, as this example clearly demonstrates, percentage of variability explained in a DV (i.e., effect size estimation) *alone* is insufficient to determine the importance or utility of an effect. Hence, if NHST and effect size estimation are both (in their most proper usage) insufficient, then where can we turn for additional guidance on how to interpret the importance or noteworthiness of a data pattern?

I contend the answer to this question relates to the meaning of measurement metrics. I argue that working with measures with calibrated metrics may facilitate the task of determining whether a particular experimental finding is worth paying attention to. That is, a researcher could use metric calibration information (e.g., a mean difference of 0.6 on the DV is equivalent to a certain difference in behavior) to help the decision process of determining whether the data pattern supports the research hypothesis and/or whether the pattern is non-trivial. Ultimately, the interpretation of empirical findings boils down to meaning. To answer the question of whether the results are “noteworthy” or “significant” (literally speaking), one must know what the results actually mean. But to know what results mean, one has to know the meaning of the measured variables, especially the meaning of the DV measure scores (in the context of experimental studies). My contention is that if the metric of the DV was more meaningful (i.e., non-arbitrary), this could help us gauge the meaning of an empirical finding. For instance, following an example used by Kirk (1996), consider a hypothetical situation where a researcher examining the effect of a drug for 12 Alzheimer patients on intelligence (vs. 12 patients in a control group), finds to her dismay that an increase of 13 IQ points in the treatment compared to the control group did not attain conventional levels of statistical significance (i.e., $p = .14$, $d = 0.90$). As Kirk mentions, the “non-significant” $p$-value does not necessarily mean that there is no IQ difference between groups, but rather that the effect size found in the sample is not large enough to yield a statistically favorable *p*-value given the sample size used (*N* = 24; 12 in each group).[8] However, because most researchers familiar with the IQ metric would likely agree that the 13 IQ points increase is a potentially noteworthy result, the study results can be seen as providing evidence for the hypothesis that the Alzheimer drug is effective. Is the effect real or rather simply due to sampling error? The only way to know is to attempt to replicate the observed effect.

My position that non-arbitrary metrics may help gauge the noteworthiness of experimental effects is broadly consistent with recent recommendations by Kashy, Donnellan, Ackerman, and Russell (2009) who emphasize that it is crucial to distinguish between findings that are theoretically important and those that are “significant” in a strictly statistical sense. My position is also in line with Kirk (1996) who argued that the decision of whether a certain data pattern support one’s research hypotheses should be a difficult one and that it is unrealistic to think that a completely objective statistical technique could ever be invented to do the job for us (see also Abelson, 1995; Thompson, 1996). Indeed, in Cohen’s (1994) own words: “…don’t look for a magic alternative to NHST, or some other objective mechanical ritual to replace it. It doesn’t exist” (p. 1001).

## 2.2 Facilitate Construct Validity Research

Metric calibration research may also benefit psychological research by facilitating the research process in the context of construct validity research. I will argue that metric

[8] Some may object to this conclusion, under the assumption that the greater than .05 *p*-value means that the results of the study could be attributable to mere chance. Although many researchers do interpret *p*-values in this fashion, some have argued that NHST cannot (unfortunately) separate “real” findings from those arising merely due to chance (Schmidt, 1996; Schmidt & Hunter, 1997). The argument goes like this. Meta-analytic reviews from many different research domains demonstrate that average statistical power (correctly concluding an effect exists) is in the range of .40 to .60 (Cohen, 1962, 1992; Schmidt, 1996; Schmidt, Hunter, & Urry, 1976). Hence, in these research domains, about 50% of all statistical conclusions that non-statistically significant effects are merely due to chance are actually erroneous. In some areas of research (e.g., job satisfaction), average statistical power is only .20 (Schmidt et al., 1976), which means using a coin flip would be a more accurate way of determining real from chance findings than using NHST. Furthermore, requiring sufficient statistical power as a solution to this problem does not work, because this would make it impossible to conduct the large proportion of studies that examine small effects, given that sufficiently large sample sizes to achieve power of .80 would be too costly or infeasible to run. This is a serious problem given that as knowledge increases in a certain research area, the effect sizes studied tend to become smaller (Schmidt, 1996).

calibration research could potentially facilitate construct validity investigations in at least three regards. That is, by (a) shedding more light on the construct itself, (b) aiding in conceptual challenges that arise in developing psychological instruments (e.g., construct definition), and (c) by providing a benchmark for detecting problems and/or improving psychological instruments.

### 2.2.1 Construct Illumination

First, the metric calibration approach in general, and metric calibration studies in particular, may help shed more light on a construct itself. The act of linking test scores to meaningful theoretically related behavioral referents can be seen as additional evidence supporting the validity of a construct. Indeed, Samuel Messick (1989), who wrote one of the most authoritative treaty on the topic of validity, actually mentions the idea of "criterion-referenced behaviors" as a strategy to help the process of interpreting scores: "…scores may be interpreted criterially in terms of performance standards or behavioral referents" (p. 44). Furthermore, in his discussion of the external component of construct validity, Messick states that the meaning of test scores is "substantiated externally by appraising the degree to which empirical relationships with other measures, …, are consistent with that meaning" (p. 45). It is important to keep in mind that Messick's conceptualization of construct validity differs from the conceptualization that has been entrenched in mainstream psychology, which views construct validity as the simpler question of whether a psychological instrument measures what it was intended to measure (but see Borsboom, Mellenbergh, & van Heerden, 2004). Rather, Messick sees validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores …" (p. 13, emphasis in original). This is similar to the conceptualization of the inventors of the concept of validity, who viewed validity as the complex question of whether test score interpretations are consistent with a nomological network involving theoretical and observational terms (Cronbach & Meehl, 1955). Hence, viewed from these two more nuanced conceptualizations of validity, metric calibration research can be seen as contributing to construct validity by providing additional evidence and support for the interpretation of test scores (Messick, 1995).

What’s more, the kind of metric research I propose can be seen as providing even more specific evidence for construct validity given that the focus is on the particular response functions or functional forms established between test scores and behavioral referents rather than an exclusive focus on zero-order correlations to other theoretically-related measures. For instance, successfully linking neuroticism test scores to the probability of reacting with anger to a mild stressor would provide further evidence for the construct validity of the neuroticism measure. Furthermore, a stronger case could be made that research providing empirical evidence for metric meaning should be seen as a requirement in the validation of any measure. This would be consistent with Messick’s (1989) idea that “test validation in essence is scientific inquiry into score meaning” (p. 56) and also consistent with Sechrest et al. (1996) who stated that one cannot claim to have a “high-quality measure” if one has no idea about the meaning of the metric (see also Messick, 1995).

### 2.2.2 Help with Conceptual Challenges

A second and related way that metric research may facilitate construct validity is by providing a general framework for clarifying difficult conceptual challenges that arise when developing or improving psychological instruments. In particular, metric calibration may help with conceptual issues involving construct definition and construct theory (Messick, 1989). That is, the process of designing and executing metric research aimed at empirically linking particular scores on one measure to external behavioral reference points, may help a researcher deal with the difficult questions surrounding what precisely a measure is measuring. For example, when attempting to find the most relevant behavioral reference points to calibrate test scores, one might realize that the construct has been defined too broadly or too narrowly, or that the construct suffers from other conceptual ambiguities (conceptual clarity issues, as discussed by Machado & Silva, 2007).

As an actual example, consider calibrating a measure of conscientiousness as assessed by the NEO-FFI (Costa & McCrae, 1992). Given the goal of metric research of linking scores from a measure to theoretically related behaviors that can serve as reference points, in the context of conscientiousness, researchers would need to ask themselves what relevant behaviors would best represent the most meaningful reference points to calibrate conscientiousness test scores. Clearly, this depends on the actual definition of the construct and the theoretical framework from which the construct was derived. In the case of conscientiousness, one quickly realizes that even though most researchers likely agree with the working definition of conscientiousness as the propensity of being painstaking and careful in acting according to the dictates of one's conscience (John & Srivastava, 1999), conscientiousness is actually posited to have many different facets including Self-Discipline, Carefulness, Thoroughness, Organization and Orderliness, Deliberation, Industriousness, Conventionality, Reliability, Virtue, and even Need for Achievement (Costa & McCrae, 1992; Goldberg, 1999; Roberts, Chernyshenko, Stark, & Goldberg, 2005). For instance, a researcher could decide to examine the lower-order facet of Orderliness, which has been conceptualized as the propensity to be organized and neat versus being messy (Jackson et al., 2009), by searching for empirical linkages between self-report scores of the Orderliness facet of the NEO-FFI (Costa & McCrae, 1992) and trained judges' ratings of the neatness of one's home or work office (Gosling, 2008). Hence, in the case of the conscientiousness construct, metric research may provide a useful framework for questioning the fundamental assumptions underlying conscientiousness, such as whether the construct is too broad in scope.

In fact, going through this process myself, I contend that a case could be made that conscientiousness (at least as it is conceptualized by most) might be too broad in scope and that it lacks tight construct theory given it posits the existence of so many heterogeneous lower-order facets under the rubric of conscientiousness. For example, is Industriousness (propensity to work hard), or Need for Achievement, or Virtue, really a reflection of conscientiousness? It would seem that more conceptual clarity could be achieved by relegating these lower-order facets to their proper distinct constructs; that is, Industriousness to the task persistence construct and Need for Achievement to the need for achievement construct (McClelland, 1951; McClelland, Atkinson, Clark, & Lowell, 1958). Some of these ambiguities likely lie in the fact that many personality inventories were derived using data reduction techniques that do not require strong construct theory (Borsboom, 2006). The strategy of linking distinct lower-order facets to distinct meaningful behavioral reference points, as required by metric research, could hence be seen as a superior method of determining the lower-order structure of constructs. At any rate, I argue that metric calibration research has the potential to facilitate the challenging (and often under-appreciated) task of working through the fundamental conceptual work underlying the measurement of constructs (see also Gawronski, Peters, & LeBel, 2008).

### 2.2.3 Measurement Benchmark

Finally, metric research may also facilitate construct validity research by providing a kind of benchmark for detecting problems with and/or improving psychological measures. That is, the empirical process of calibrating measures to relevant behavioral reference points may provide concrete information about psychometric problems plaguing a measure. In addition, the metric calibration approach may provide a concrete yardstick for improving measures by offering more diagnostic information than traditional validity investigations. This would be the case because finding empirical mappings between test scores and behavioral criteria has the potential to supply richer and more proximal information than standard convergent validity investigations. Metric research involves the discovery of a specific response function between test scores and behavioral manifestation of the construct, which stands in contrast to traditional criterion validity research which typically involves establishing zero-order correlations between test scores of the construct and test scores from other theoretically-related constructs. Indeed, in the General Discussion I will describe an actual example of this principle below, which I encountered when calibrating the scores of a task-persistence measure (Study 1), whereby the consideration of a metric calibration mapping revealed evidence suggestive of a construct validity issue.

As a further example of how metric research may aid in the detection of measurement issues, consider a researcher who has run a series of metric calibration studies and discovers that test scores are consistently *not* linking up with a theoretically-relevant behavioral referent. This may suggest that something is wrong with the measure and/or the construct theory that led to examining the particular behavioral referent. For example, imagine that no empirical linkage (linear, curvilinear, or otherwise) is found between extraversion facet scores and average time spent socializing per day (as assessed using Mehl et al.'s, 2001 EAR for e.g.). This may suggest that something is wrong with the extraversion measure or that the construct theory underlying extraversion needs revision (or both). Upon closer scrutiny, one may notice that a particular extraversion score of “36”, for example, corresponds to almost any value on the criterion measure (0.5, 1, 2, 3,4 and even 7 hours socializing per day). This could suggest that something is wrong with the measurement and/or conceptualization of extraversion. Hence, assuming solid measurement of the criterion behaviors and sound reasoning concerning important metric calibration principles (e.g., features of the context, level of measurement), metric research may provide a valuable benchmark for developing psychological measures. Furthermore, once an empirical mapping is established, all of the relevant calibration information can be used when improving a measure (or when attempting to improve the scoring algorithm of a measure) to ensure the integrity of the measure has not been compromised. In other words, the increased information provided by metric research may help researchers gauge their progress in improving a measure. Indeed, this form of construct validity research resonates well with the position of some theorists, who have argued that validity is more properly seen as a *continuous* research process that (a) aims to continually build a better evidence base to support score interpretation (Messick, 1989) and (b) strives to continually increase our understanding of the measurement error that contaminates test scores (and hence continually trying to reduce this error component; DeShon, 1998). In summary, metric calibration research could facilitate construct validity research by shedding more light on psychological constructs, aiding in challenging conceptual issues (e.g., construct definition), and by providing a clearer benchmark for developing and improving psychological instruments.

## 2.3 Contribution to Theoretical Development

In this section, I will elaborate on how non-arbitrary metrics could benefit psychological research by having the potential to advance theory development in basic psychological research. I will argue that using psychological measures with calibrated metrics could contribute to theory development by (a) shedding light on theoretical debates involving absolute claims, (b) allowing for more precise theorizing of psychological phenomena via enhanced scientific language, and (c) allowing researchers to test substantive theories more precisely (i.e., make specific point-value predictions).

### 2.3.1 Aid in Theoretical Debates Involving Absolute Claims

In a first sense, non-arbitrary metrics could contribute to theory development in potentially important ways by shedding light on theoretical claims based on assertions made about the absolute level of a certain psychological phenomenon. Though the bulk of basic psychological research involves testing directional hypotheses involving relative score comparisons across experimental conditions, certain theoretical questions in psychological research involve making claims of an absolute nature. I contend that metric calibration could contribute to theoretical development by providing empirical machinery to more directly tackle such theoretical questions. One such example comes from the research literature on the cross-cultural universality of self-enhancement. In this literature, researchers often examine whether individuals rate themselves more favorably on a series of culturally relevant desirable traits as compared to a hypothetical average other person, by testing mean ratings against the scale midpoint (e.g., 1 = *much less than the average person*; 4 = *about the same as the average person*; 7 = *much more than the average person*). If the mean ratings are statistically significantly greater than the scale midpoint, then it is inferred that self-enhancement is present in that culture (e.g., Gaertner, Sedikides, & Chang, 2008). However, as already stated, it is unfounded to assume that the scale midpoint coincides with the theoretical midpoint on the underlying dimension of self-enhancement. As mentioned, many factors may shift the numerical midpoint away from the theoretical midpoint on the underlying dimension (not to mention the aforementioned logical reasons against meter reading).

If trait ratings of self-enhancement, however, were linked to a behavioral index of self-enhancement (e.g., Paulhus, Harms, Bruce, & Lysy's, 2003 over-claiming technique), then the scores on the self-report self-enhancement measure would gain meaning that would potentially shed light on the theoretical claim regarding the universality of self-enhancement (Sedikides, Gaertner, & Toguchi, 2003). For instance, it would be critical to examine how scores typically interpreted as self-enhancement (e.g., a "5" on a 1 to 7 point scale with scale midpoint of "4") map onto the behavioral indices of self-enhancement, to get an actual sense of what kind of self-enhancement behaviors correspond to particular trait rating scores. If trait rating scores typically interpreted as self-enhancement do not map onto behavioral indices of over-claiming, this could cast doubt on the universality of self-enhancement claim based on the absolute interpretation of trait rating scores. This example demonstrates the potential that metric calibration research could hold in advancing or revising theoretical claims involving assertions of an absolute nature, which could contribute to theory development more broadly.

### 2.3.2 Allow More Precise Theorizing via Enhanced Scientific Language

A second way non-arbitrary metrics could contribute to theoretical development is by helping researchers more accurately and precisely theorize about psychological phenomena. This would be the case because non-arbitrary calibrated metrics would enhance our scientific language by empirically substantiating claims about the standing of individuals on underlying psychological dimensions. It is easy to find examples in the literature of theorizing that contain reference to “high-X individuals” or “low-X individuals” doing certain things under certain conditions (where the X can be any psychological construct). For example, “…it was found that individuals with a high need for closure were more likely to report having voted for conservative parties” (Chirumbolo & Leone, 2008, p. 1286); “our results provide support for Sedikides et al’s (2002) contention that people high in narcissism show a lack of contextual sensitivity…” (Collins & Stukas, 2008, p. 1629); “in such situations, high self-esteem individuals might be more resistant to persuasion than low self-esteem individuals…” (Briñol & Petty, 2005, p. 591). These kinds of claims, which are rampant in the literature, are unsubstantiated and potentially misleading given that claims about the standing of individuals on an underlying psychological dimension requires systematic empirical linkages to meaningful external referents (Blanton & Jaccard, 2006a, 2006b). Hence, theorizing that emerge from these unfounded claims can impede accurate theorizing about psychological phenomena and hence interfere with theory development.

Some readers may feel that it is unfair to characterize claims of the sort described above (“high-X individuals”) as unfounded. An astute reader could point out that researchers may indeed have some empirical knowledge to substantiate their claims about individuals being “high” or “low” on a certain psychological construct. For example, it could be brought up that the finding from the self-esteem literature showing that median-split "low" self-esteem individuals trust their relationship partners less than "high" self-esteem individuals (Murray, Holmes, & Griffin, 2000) actually provides support for using the labels "high" versus "low" self-esteem individuals. On this point, I would partially agree. Viewed in a broad sense, studies showing how a median-split continuous predictor variable patterns itself on a certain DV can actually be seen as a very coarse version of metric calibration research. I would argue, however, that such evidence is insufficient for making claims regarding individuals' standing on a construct for at least two reasons. First, the very rough (and arbitrary) nature of doing a median-split on the scores can mask/hide important information about the mappings between particular scores and the relevant behavioral referents. Furthermore, nowadays researchers typically avoid median-splits (as they should, see MacCallum et al., 2002), which means that these types of relations would be simply reported as the correlation between the variable in question and the DV. A correlation, of course, in and of itself, does not provide information about metric meaning (Blanton & Jaccard, 2006b). Second, more systematic thinking is required to generate the relevant behavioral referents to be used as reference points in the calibration process. Notwithstanding these limitations, I think it is potentially informative to review the literature for a certain construct (e.g., narcissism) and examine what kinds of DVs have been linked to it in these median-split or correlational studies. This could provide a starting point for determining what kinds of phenomena or manifest behaviors are related to the construct whose measure one is interested in calibrating. Actually, data from these studies could be re-analyzed (if recoverable from the authors) to examine the specific mapping between the non-median-split (i.e., full-range) scores of the predictor and the relevant behavioral DV.

### 2.3.3 Quantitative Testing of Psychological Theories

In a third and final way, non-arbitrary metrics have the potential to contribute to theory development by providing a guiding framework that might eventually allow basic psychologists to more precisely test psychological theories. That is, using more meaningful calibrated metrics may eventually allow researchers to make theoretical predictions about particular point-values of psychological phenomena. An important caveat to note, however, is that a pre-requisite for this potential benefit is that substantive theories need to be developed enough to actually be able to generate point-predictions (Meehl, 1967; but see Granaas, 2002). Although perhaps many would agree that in most areas of psychology theories may not be developed enough to make specific point predictions (Cook & Shadish, 1994), as previously mentioned, Granaas (2002) wonders whether it is the methodological training of psychological researchers that prevents psychologists from designing studies that make specific point-value predictions *rather* than the fact that most psychological theories are too weak to make such point-value predictions. Although the relatively young nature of psychological science undoubtedly plays a part, I argue that the use of arbitrary metrics also likely contributes to this problem. That is, psychologists are not trained to think about metric score meaning because virtually all metrics in psychology are arbitrary; and without paying attention to what particular scores mean, it seems unlikely one could develop a psychological theory that makes specific predictions about magnitude and particular values. Hence, I argue that the general metric calibration research approach proposed in this dissertation might eventually increase the possibility of testing psychological theory in a more quantitative fashion, whereby particular point-value predictions are made and then empirically tested.

In physics, specific point-value predictions involve comparing a theoretically predicted value $x_o$ (based on the particular experimental or natural factors embedded in a situation) with the empirically observed mean $\bar{x}_o$, and asking whether the predicted value falls within the band of probable error (due to random error of measurement) of the empirically observed mean (Meehl, 1967). For example, Mulaik et al. (1997) recounts the scenario, early in the 20th century, where Newton's theory of gravity predicted that gravitation would deflect light from a star passing near the edge of the sun by one-half the amount predicted by Einstein's theory of relativity (0".87 $r_0/r$ vs. 1".75 $r_0/r$, where $r_0$ = the radius of the sun and $r$ = closest distance of the star's light to center of the sun). Observed data from two independent observation sites during a total eclipse of the sun confirmed that Einstein's predicted value fell within the band of probable error of the observed value for both sites whereas Newton's predicted value fell outside the band, hence supporting Einstein's theory over Newton's theory.

Although it might be difficult to imagine that psychological research will ever reach this level of exactitude, I contend that we should perhaps nonetheless strive toward this general direction. This would be in line with recommendations by Harlow (1997) who suggested that more emphasis should be placed on creating very specific “defeatable” hypotheses, rather than the common practice of having a null hypothesis of no effect and a non-specific alternative hypothesis (see also Granaas, 2002, who recommends training psychologists to use theoretically meaningful null hypothesis values). Furthermore, perhaps we need to start thinking about how we can combine theories and/or design our studies such that we can derive possible ranges of values that we theoretically expect from placing individuals in a particular experimental situation (e.g., $1.2 < B < 1.6$; Meehl, 1967). Roughly paralleling the physics example from above, this could correspond to building a model (e.g., set of equations) based on a substantive theory that integrates how the different factors (impinging on the participant in the experimental situation), combine to influence the participant’s behavior. After a range of possible values is generated by the constructed model, one would empirically observe the behavior in the experimental situation and determine if the observed value fell within the model’s predicted range of values. If the value falls outside the range, then one would try to figure out why the prediction was not borne out, for example, by improving the relevant model and importantly ruling out other methodological and measurement issues (as is done in physics). If after repeated experimental tests, the observed value still lies outside the predicted range, one would be forced to revise the theory and/or auxiliary assumptions used to generate the substantive model.

## 2.4 Facilitation of General Accumulation of Knowledge

The last, and most general benefit of the metric calibration approach, reflects the proposition that metric calibration could potentially benefit basic psychological research by facilitating the accumulation of knowledge more broadly. I will expand on the three ways in which both the metric calibration approach and the resulting non-arbitrary metrics may facilitate the general accumulation of knowledge. That is, non-arbitrary metrics (a) may provide valuable general information in its own right, (b) may be a guiding framework for keeping track and cataloguing the magnitude of psychological effects, and (c) may facilitate phenomenon-based research.

### 2.4.1 Valuable Information in its Own Right

Metric research that seeks to find empirical linkages between a measure's scores and theoretically-relevant noteworthy behaviors can be seen as very useful information in its own right. That is, knowing that specific scores on a particular measure correspond to specific theoretically-relevant behavioral reference points can be viewed as providing valuable knowledge about human psychology, in the same way that scientists in the physical sciences seemed to think that calibration research provided valuable information in the early days of instrument development (e.g., thermoscope, hygrometer). Indeed, Sechrest et al. (1996) mention it is surprising that no psychological measure known to them has been systematically calibrated against relevant behavior. That is, Sechrest et al. state that it is strange that psychologists do not know, for example, what reduction in probability of being seen smiling is associated with each point increase on the BDI, or how many points on the Eysenck Introversion-Extraversion Scale (Eysenck & Eysenck, 1975) are associated with each additional hour spent alone per day or how many points on Scale 4 (Psychopathic Deviate) of the Minnesota Multiphasic Personality Inventory (Dahlstrom & Welsh, 1960) are associated with each arrest by age 25.

### 2.4.2 Guiding Framework for Cataloguing the Magnitude of Psychological Effects

Metric calibration research could also facilitate the accumulation of knowledge by providing a guiding framework for keeping track and cataloguing the magnitude or "quantity" of psychological effects, above and beyond direction, as Jacob Cohen has advocated (Cohen, 1994). Without metrics that have any inherent meaning, the "stockpiling" of information on quantity may not be very productive. That is, the utility of storing up information about the magnitude of experimental effects based on arbitrary effect size indices on scores from measures with arbitrary metrics might be quite limited. With metrics that do have some inherent meaning, however, the situation could be very different. Researchers would then have a guiding framework for systematically cataloguing information about the amount of an experimental effect, above and beyond its direction expressed in terms of arbitrary effect size. For example, one could catalogue that an intelligence-based self-concept threat decreased state self-esteem scores (Heatherton & Polivy, 1991) by 1.5 points (*d* = 0.5), which is behaviorally equivalent to a 50% increase in time spent on a self-affirmation task whereas a social-exclusion-based self-concept threat decreased state self-esteem scores by 2.0 points (*d* = 0.6), behaviorally equivalent to a 75% increase in time spent on the self-affirmation task. If systematically catalogued in the context of other related studies employing similar and different manipulations, this information – valuable in its own right – could become even more valuable in developing a database of "amounts" by which certain experimental manipulation impact different kinds of human behavior. This could facilitate the accumulation of knowledge by providing an organized system for structuring a research area's knowledge base in a much more information-rich manner.

### 2.4.3 Facilitate Phenomenon-Based Research

Finally, a third perspective on the way the metric calibration approach may facilitate the accretion of knowledge can be put forward from the perspective of phenomenon-based research (Asch, 1952/1987; Rozin, 2001; see also Funder, 2009; Rozin, 2009). From this perspective, it is critical to identify and describe phenomena and invariances (i.e., to describe *what is*), before engaging in modeling and hypothesis testing of complex research questions requiring sophisticated methodological designs and statistical techniques. A Soloman Asch quote (as cited in Rozin, 2001) reflects this idea poignantly: "Before we inquire into origins and functional relations, it is necessary to know the thing we are trying to explain" (Asch, 1952/1987, p. 65). Rozin reviewed objective data comparing research practices in the natural sciences versus psychology and demonstrates that natural scientists (a) much more often engage in descriptive research aimed at becoming familiar with the phenomenon at hand, (b) less often engage in specific model or hypothesis driven research aimed at testing specific hypotheses, and (c) less often use experimental designs to make statistical inferences. In addition, Rozin provided an interesting conceptual argument that the discovery of the molecular basis of genetic transmission, which Rozin claimed was the most important advance in the life sciences in the 20th century, occurred because scientists in this domain engaged heavily in descriptive, phenomenon-based research. Rozin states that the scientists' motive for early studies on x-ray diffraction and nucleotide was basically something along the lines of: "It looks like DNA is really important and a likely vehicle for genetic transmission. Let's find out more about it. What is its shape and what is it made of?" (p. 7). A potential implication of Rozin's arguments is that researchers in psychology have perhaps been too hasty or skipped altogether the valuable descriptive, phenomenon-based stage of research, and that this has interfered with the development of a cumulative knowledge base in psychology. Viewed from this perspective, I argue that metric calibration research can provide a useful framework to engage in this type of descriptive, phenomenon-based research. Indeed, metric research aimed at discovering the relevant behavioral manifestations of a construct, and how these manifest behaviors pattern themselves onto a corresponding measure's scores, could be viewed as accomplishing the goal of knowing in more depth, richer information about a certain phenomenon. Viewed in this light, metric research could facilitate the process of investing more energy in the fundamental early stages of science, argued to be critically needed for psychological science to reach its full potential (Asch, 1952/1987).

# Chapter 3

# 3 Empirical Demonstrations

I now turn to the empirical demonstration of the metric calibration approach applied to psychological instruments of constructs commonly examined in basic psychological research. These preliminary empirical demonstrations were meant to showcase in more concrete ways the steps required, both at the conceptual and procedural level, to calibrate the metrics of psychological measures typically used in psychological research. It is important to keep in mind, however, that these empirical demonstrations were executed primarily for illustrative purposes only, given that (a) collective agreement on the appropriateness of the behavioral criteria is a prerequisite and (b) much larger targeted samples are required in practice to ensure that the calibration values found are precise enough estimates of the population values (i.e., the particular mappings between test scores and behaviors are sufficiently stable). Study 1 focused on illustrating the metric calibration approach for need for cognition, task persistence, and conscientiousness instruments; Study 2 focused on the calibration of a self-enhancement measure in the context of the pan-cultural debate of self-enhancement (Sedikides et al., 2003); and Study 3 focused on the metric calibration of behavioral instruments of risk-taking. Finally, re-analyses from two shared datasets further illustrated the metric calibration approach for instruments assessing self-control, extraversion, and once more conscientiousness.

## 3.1 Study 1

The primary goal of the first study was to provide preliminary empirical demonstrations of the metric approach applied to three constructs commonly used in basic psychological research: need for cognition, task persistence, and conscientiousness. In a broad sense, these constructs fall under the broader umbrella concept of cognitive effort, which plays an important role in dual-process models that have become increasingly popular in many areas of psychology (e.g., Chaiken & Trope, 1999; Devine, 1989; Epstein, 1990; Fazio, 1990; Fiske & Neuberg, 1990; Gawronski & Bodenhausen, 2006; Gilbert, 1989; Petty & Cacioppo, 1986; Sloman, 1996; Smith & DeCoster, 2000; Strack & Deutsch, 2004;

Trope, 1986). In these dual-process models, cognitive effort (i.e., cognitive elaboration) is posited to impact how different pieces of information influence social or self judgments and/or behaviors. For instance, in the Elaboration Likelihood Model (ELM; Petty & Caciooppo, 1986), the extent to which individuals process and analyze information (i.e., the "elaboration continuum") contained in a persuasive message is posited to influence the impact of central versus peripheral cues on attitude change. For example, under conditions of high elaboration (e.g., individuals high in need for cognition), central cues, such as argument strength, are posited to be the main determinants of attitude change whereas under conditions of low elaboration, peripheral cues, such as source credibility, are posited to be the primary determinants of attitude change. Given the emphasis on cognitive elaboration in such models, it becomes apparent why shedding light on the metric meaning of measures of constructs in such category would be important. These constructs were also selected for study because our current level of understanding of the constructs and related phenomena seemed sufficiently developed to be good candidates for engaging in metric calibration research.

To illustrate the feasibility of increasing metric meaning in the context of basic psychological research, respective measures for each of these three constructs were calibrated to each of their own theoretically-relevant behavioral referents (details of the particular behavioral referents below). The calibration process involved examining the particular response function that connected the measurement scores (e.g., need for cognition scores) with the relevant behavioral reference points (e.g., probability of choosing cognitively challenging vs. simpler task). The goal in this first step was to illustrate empirically the practical feasibility of this type of metric calibration approach for constructs studied in this area of psychology.

### 3.1.1 Theoretical Derivation of Relevant Behavioral Referents

In this section, I will elaborate on my theoretical reasoning for examining the particular external behavioral referents chosen to calibrate the scores of the respective measures. When going through the derivation of behavioral reference points for the three constructs, it is important to keep in mind the broader context of dual-process models in which the concept of cognitive elaboration plays an important role.

First, let’s consider need for cognition (NFC). NFC is conceptualized as the tendency for an individual to engage in cognitively effortful activities and enjoy thinking in its own right (Cacioppo & Petty, 1982; Cacioppo, Petty, & Kao, 1984). The NFC construct originated based on earlier research examining individuals’ behavioral tendencies in how they organize, understand, and evaluate information in their environments (Cohen, Stotland, & Wolfe, 1955; Cohen, 1957). NFC is typically measured using the revised and shortened 18-item scale (Cacioppo et al., 1984), which is based on the original 34-item scale (Cacioppo & Petty, 1982). Based on the abovementioned conceptual definition (and the conceptual framework from which the construct arose), one potential external behavior to examine as a possible behavioral reference point to imbue NFC scores with meaning, is the probability of choosing to complete a cognitively challenging versus simple task. Individuals high in the underlying dimension of NFC should be more likely to engage in a cognitive task that is described as being more cognitively challenging compared to one that is described as cognitively simpler, given that these individuals find cognitively effortful activities intrinsically enjoyable. However, the type of task and specific features of the task would need to be specifically configured, so that it is able to capture a behavioral manifestation of NFC. For instance, it is critical to choose a type of cognitive task from which two versions of the task (i.e., a cognitively challenging and cognitively simpler version) could be constructed. The cognitively challenging version would need to actually appear cognitively more challenging, but not so much that most individuals would think it was too difficult; conversely, the cognitively simpler task version would need to appear cognitively simpler, but not so much that most individuals would consider it too boring. In addition, the presentation of the two versions of the task should be done in a way that minimizes the possibility of having the vast majority of individuals choosing one task over the other.

The task chosen in the present study was a modification of the Remotes Association Test (RAT; Mednick & Mednick, 1967), originally used to measure word-based creativity. In this task, individuals are presented with three distinct words and are asked to generate a fourth word that relates in some way to each of the three stem words. For example, if presented with the words “turkey”, “freezing”, and “war”, a possible answer would be “cold”. It was expected that this task would be relevant in capturing variations in the underlying dimension of NFC because the type of thinking required to solve these puzzles resonates well with the fundamental conceptualization of the NFC construct (i.e., engaging in cognitively effortful activities involving the organization, processing, and understanding of information). More importantly, with this task it was possible to develop two versions (varying in cognitive difficulty) in a way that could potentially discriminate between those high and low on the underlying dimension of NFC. These two versions were presented to participants via actual example items for each task and specific explanations were given as to how the two tasks differed. This was done in a way that specifically matched the conceptualization of NFC, such that the more cognitively challenging task was framed as requiring more intricate thinking than the simpler task. That is, it was explained that in the simple task the fourth word would generally relate to the three words in the same way (e.g., semantically related) whereas in the more cognitively challenging task, the fourth word could relate to each of the three words in a different way (e.g., semantic, conceptual, visual). Hence, it was predicted, based on construct theory, that individuals high in NFC would see the more cognitive challenging task as more cognitively effortful than the simpler task and hence be more likely to seize the opportunity to engage in and enjoy the more effortful task (see below for more details on the specific parameters of the tasks).

With regard to conscientiousness, although the construct originated from personality psychology (Costa & McCrae, 1992; Goldberg, 1999; John & Srivastava, 1999; Tellegen & Waller, 1994), it is now examined in the context of many other research areas including social (Kelly & Conley, 1987), health (Booth-Kewley & Vickers, 1994), and personnel psychology (Hogan, Rybicki, Motowidlo, & Borman, 1998). According to dictionary definitions, conscientiousness refers to the “trait of being painstaking and careful” or the “quality of being in accord with the dictates of conscience” (Merriam-Webster; Princeton WordNet Dictionary). Psychological research, however, seems to have progressed to a more differentiated conceptualization that views conscientiousness as reflecting the tendency to follow socially prescribed norms and rules, to be goal-directed, planful, able to delay gratification, and to control impulses (John & Srivastava, 1999) or, worded slightly differently, as the degree of organization, persistence, and motivation in goal-directed behavior (Costa & McCrae, 1992). This heterogeneity in conceptualization makes it somewhat difficult to theorize about which behavioral referents to use to calibrate the scores of the conscientiousness measure. As elaborated in the introduction, however, during the calibration process, one needs to carefully consider the (possible) multi-faceted nature of a construct, and if applicable, to find separate score-behavior linkages for each facet. Given the multi-faceted nature of conscientiousness, I decided as a starting point to examine facets relating to detail-orientedness and tried to link conscientiousness scores from relevant facets to performance on a difficult proof-reading essay task that requires high levels of detail-orientedness. After careful scrutiny of the many different facets of conscientiousness available via different assessment instruments, the most theoretically-relevant facets were deemed to be the Deliberation and Self-Discipline facets of the NEO-FFI (Costa & McCrae, 1992), the Self-Control facet from the MPQ (Tellegen & Waller, 1994), and the Impulse Control facet from Goldberg's Abridged Big Five Dimensional Circumplex (AB5C; Goldberg, 1999). For example, the Deliberation facet of the NEO-FFI contains items such as "I avoid mistakes" and "I choose my words with care" and the self-control facet of the MPQ contains items such as "I am exacting in my work" and "I pay attention to details." Hence, it was hypothesized that an empirical mapping would emerge between the conscientiousness scores from those particular facets and the number of mistakes found in a difficult 4-page essay proof-reading task.

The essay task, adopted from Glass, Singer, and Friedman (1969), was configured to capture the behavioral manifestation of detail-orientedness. For example, special care was taken to systematically introduce unambiguous mistakes that do not require grammatical knowledge. That is, only clear typographical errors (e.g., "aspetcs", "hows to") and unambiguous punctuation errors ("The Style; Template", "stage of, publication") were introduced into the text. In addition, the task length (i.e., 4-pages with approximately 200 words per page) and task time (i.e., 8 minutes) were chosen to create optimal conditions for capturing conscientiousness, such that the task was difficult enough so that non-conscientious individuals did not have enough time to find most of the mistakes (see below for more details of the essay task). Also, and critically, to attempt to normalize levels of motivation for performance in the task, instructions stressed that it was important that participants tried their best in finding as many mistakes as possible in the allotted time.

Finally, concerning task persistence, this construct is of broad relevance to different areas of psychology, having been investigated in the context of addictive behaviors and distress tolerance (Rodman, Daughters, & Lejuez, 2009; Quinn, Brandon, & Copeland, 1996; Steinberg et al., 2007), human motivation and goals (e.g., achievement motivation; Feather, 1961), health psychology (e.g., mindfulness; Evans, Baer, & Segerstrom, 2009), and social psychology (e.g., narcissism; Wallace, Ready, & Weitenhagen, 2009). Task persistence is typically conceptualized as the tendency to persist in an effortful behavior or frustration-inducing activity (Steinberg et al., 2007) and has been measured both behaviorally (anagram persistence task: Brandon et al., 2003; mirror tracing persistence task: Quinn et al., 1996) and via self-report (Steinberg et al., 2007; Pomerleau, Pomerleau, Flessland, & Basson, 1992). As a potential demonstration of metric calibration research, I decided to calibrate a self-report task persistence measure with a theoretically-relevant manifest behavior. I decided to use Steinberg et al.'s (2007) 2-item self-report measure, developed from Cloninger's (1987) Tridimensional Personality Questionnaire (TPQ), which were specifically derived from the theoretical framework of Learned Industriousness Theory (Eisenberger, 1992). This brief self-report measure of persistence is desirable because of its clear practical advantages over behavioral persistence measures in terms of its ease of use, lower cost, and portability (Ditre & Brandon, 2008). The items are as follows: "I will keep trying the same thing over again even when I have not had success the first time" and "I will often continue to work on something, even after other people have given up". Steinberg et al. successfully used this measure in the context of teenager smoking, showing that self-reported task persistence was greater among adolescent non-smokers as compared to current adolescent smokers (see also Ditre & Brandon, 2008, who also successfully used this measure).

As a behavioral manifestation of task persistence, I used a commonly used anagram persistence task (Brandon et al., 2003; Evans et al., 2009; Nes, Segerstrom, & Sephton, 2005; Quinn et al., 1996), which involves unscrambling near-impossible ("X L Y I K" = "K Y L I X") and easy ("B E A H C" = "B E A C H") anagram puzzles, with the average time persisting on the near-impossible items used as an index of task persistence. Critically, participants are instructed that (a) they have 3 minutes to solve each anagram, (b) as many attempts as desired can be made, and that (c) they can give up and skip to the next item before the maximum time has elapsed. The assumption was that the self-report measure would tap into some aspect of task persistence that would share some overlap with the behavioral index of task persistence. Hence, it was predicted that a mapping would emerge between scores of the self-report measure and actual behavioral persistence exhibited in the anagram persistence task.

### 3.1.2 Method

#### 3.1.2.1 Participants and Design

Ninety four (94) University of Western Ontario introductory psychology undergraduates participated for partial course credit (69 females, 25 males; mean age = 18.46, *SD* = 2.18, range = 17 to 30). No restrictions were imposed on participant sex, age, or ethnicity. No experimental conditions were examined, hence all participants completed the same measures and tasks in the same order (see below for details).

#### 3.1.2.2 Procedure and Materials

Participants were run in groups of two to five in a large testing room where each participant was seated in a separate cubicle in front of a PC computer. The experimenter (myself) individually gave brief verbal instructions before participants started, stating:

> All of the tasks you will complete today, except one, will be completed on the computer. It is very important to thoroughly read all of the instructions before starting any of the tasks. If you have any questions about any of the tasks, feel free to ask me for clarification. Also, before starting, please turn off your cell phone (or any other electronic devices).

Participants then followed on-screen instructions and completed each task in a serial fashion in the following order: the measures of conscientiousness, the need for cognition scale, the self-report measure of task persistence, the word association decision task, the essay proofreading task, the anagram persistence task, and then demographics and debriefing questions.

### 3.1.2.2.1 Conscientiousness

To assess conscientiousness, I measured the Deliberation and Self-Discipline facets of the NEO-FFI (Costa & McCrae, 1992), the Self-Control facet from the MPQ (Tellegen & Waller, 1994), and the Impulse Control facet from Goldberg's Abridged Big Five Dimensional Circumplex (AB5C; Goldberg, 1999). I used the International Personality Item Pool (IPIP; Goldberg et al., 2006) version of the relevant MPQ and NEO-FFI facets (see Appendix A for actual instructions and items). For the NEO-FFI and MPQ items, participants rated the extent to which each statement was an accurate description of themselves, using a 5-point Likert scale, with the response categories 1 = "Very Inaccurate",  2 = "Moderately Inaccurate", 3 = "Neither Accurate Nor Inaccurate", 4 = "Moderately Accurate", and 5 = "Very Accurate". Instructions emphasized that participants should describe themselves as they generally are, to answer honestly, and to answer in relation to other people the same sex and roughly the same age. For the AB5C (impulse control), participants rated the extent to which 12 trait adjectives (e.g., "Careful", "Cautious") described them, using a 5-point Likert scale, with the response categories 1 = "Strongly Disagree", 2 = "Somewhat Disagree", 3 = "Neither", 4 = "Somewhat Agree", and 5 = "Strongly Agree" (see Appendix B for instructions and items). Instructions emphasized that participants should describe themselves at the present time and to describe themselves as they are generally or typically. About half of the items on the four facets were negatively worded and hence were reverse-scored. Reliability of the scores from the four facets were generally acceptable (deliberation, 10 items, α = .80; self-discipline, 10-items, α = .87; self-control, 10-items α = .76; impulse-control, 12-items, α = .76). For ease of interpretation, the ratings for each facet were averaged, therefore creating a metric ranging from 1 to 5, with decimal numbers in between.

For the essay task behavioral measure, participants were told that they would receive a 4-page essay on actual paper and that their task would be to circle as many mistakes as possible within a span of eight minutes (approximately 10 mistakes per page were systematically introduced in the text, with a total of 42 mistakes). Participants were clearly informed to look only for mistakes such as misspellings, incorrect punctuations, and typographical errors (but not formatting issues, e.g., spacing). They were also informed that it was important that they do their best in finding as many mistakes as possible in the allotted time. After reading these instructions, the experimenter brought the 4-page essay and pen to the participant and started the 8-minute timer set in the MediaLab software. In terms of scoring, I counted (using an answer overlay for accuracy and expediency) the number of mistakes correctly circled by the participant. Scores were expressed as a percentage of the total mistakes found by the participant. Scores ranged from 17 to 88%.

### 3.1.2.2.2 Need for Cognition

As mentioned, Cacioppo et al.'s (1984) revised and shortened 18-item NFC scale was used. The questionnaire was introduced as a tool assessing people's individual thinking styles and it was mentioned that there were no right or wrong answers. Participants rated the extent to which each item was characteristic of them, using the response categories 1 = "Extremely Uncharacteristic", 2 = "Somewhat Uncharacteristic", 3 = "Uncertain", 4 = "Somewhat Characteristic", and 5 = "Extremely Characteristic" (see Appendix C for items). Negatively worded items were reverse scored and the mean of the 18 items was computed, yielding a metric ranging from 1.0 to 5.0 ($\alpha$ = .88).

For the behavioral reference point, as mentioned, a choice was given to participants to complete one of two word association tasks that were described as varying in terms of their level of cognitive challenge. Introduced as a task used to measure individuals' conceptual ability to solve problems involving the connections between words, participants were explained the basic logic of the task (i.e., three words given, must find a related fourth word) and told that they would complete 10 questions (having 20 seconds for each). Then, an example problem for Task 1 (cognitively simpler task: "FRIES" "KISS" "TOAST") and Task 2 (cognitively more challenging task: "BOARD" "MAGIC" "DEATH") were presented without the answer, emphasizing that these example questions reflected the level of cognitive challenge that would be found in the items of the respective tasks. A minimum of 5 seconds was imposed before participants were allowed to proceed to the next screen to see the answers to the Task 1 and 2 example problems. This was done to ensure that participants would accurately perceive the level of cognitive challenge of the example questions (i.e., prevent people from thinking even the challenging problem was obvious if they were presented with the problem *and* the answer simultaneously). On the screen with the answer to the Task 1 and 2 examples, it was clearly explained how the two examples differed in level of cognitive challenge by explaining that in the Task 1 example, the answer (i.e., "FRENCH") related to the three words in the *same* way whereas in the Task 2 example, the answer (i.e., "BLACK") related to the three words in a *different* way (i.e., "BLACK" relates to "BOARD" in a semantic way whereas "BLACK" relates "DEATH" in a visual way, etc.). Participants then clicked on the respective button to make their choice and proceeded to complete the 10 questions. Overall, 58 individuals (62%) chose Task 1 and 36 individuals (38%) chose Task 2.

#### 3.1.2.2.3 Task Persistence

As described earlier, Steinberg et al.'s (2007) 2-item self-report measure of task persistence was administered ("I will keep trying the same thing over again even when I have not had success the first time" and "I will often continue to work on something, even after other people have given up"). Participants rated their degree of persistence using a 4-point scale, using the response categories 1 = *Very untrue, not at all like me*, 2 = *Somewhat untrue or not like me*, 3 = *Somewhat true or like me*, and 4 = *Very true, very much like me*. The reliability estimate of the measurement scores was α = .67, which is somewhat lower than the α = .73 reported by Steinberg et al. For ease of interpretation, ratings from the two items were averaged, creating a metric ranging from 1 to 4, with half steps in between.

For the behavioral measure, an anagram persistence task (Brandon et al., 2003; Nes et al., 2005; Quinn et al., 1996) was used. Participants were told that they would complete anagram puzzles and that although some of these would be very difficult, that they were all solvable. Participants were given an example (i.e., re-arrange the letters "O B A T" into "B O A T") and then told that the task would contain 11 anagrams presented serially. Participants were told that they would be allowed as many attempts as they wished to figure out the correct answer, but that they would have a maximum of 3 minutes to work on each question (at which time the program would automatically proceed to the next question). Importantly, participants were informed that if they wished to give up on an anagram question before the maximum time, they could skip to the next question by clicking the “SKIP” button. Unbeknownst to the participants, the Visual Basic software program recorded the amount of time spent on each question and also how many attempts were tried for each question (and the content of each attempt). The task contained 6 near-impossible anagrams (e.g., “Q Y U I A” = “Y A Q U I”) and 5 relatively easy anagrams (e.g., “B E C H A” = “B E A C H”; see Appendix D for all items). As done in past research, average time spent persisting on the 6 near-impossible questions was used as the index of task persistence. However, upon closer scrutiny of participants’ actual answers on the near-impossible items, I noticed that many participants actually correctly guessed the answer often within the first minute (surprisingly, 20.3% of the near-impossible anagrams were correctly guessed within the allotted time; however, all individuals left at least 2 anagrams unsolved, meaning no one had to be excluded due to correctly guessing all near-impossible anagrams).[9] To control for this contamination, only unsolved anagrams were used to compute behavioral indices of task persistence.

### 3.1.3 Results

#### 3.1.3.1 Preliminary Data Treatment

Data from all measures were first screened informally for any evidence of non-compliance. That is, I examined the time taken (as recorded by MediaLab) on instruction screens and actual questionnaire items for any evidence of consistently short latencies suggestive of non-compliance. No unambiguous cases were identified and hence all participants were retained.

[9] It is interesting to note that I did not find mention of this issue in any of the past research using this persistence task. Indeed, Brandon et al. (2003) specifically mentioned that “the dependent measure was the mean time spent on the six difficult anagrams, which were never solved by the participants” (p. 450). I suspect this discrepancy may have arisen due to differences in the administration of the task, which in past research has typically been done with an experimenter and a stop-watch, whose presence may drastically reduce the amount of trial-and-error guessing and hence considerably reduce the number of correctly guessed answers.

### 3.1.3.2 Main Analyses

I now present the main metric mapping results. For rhetorical reasons, I will present results for the constructs in the following order: need for cognition, task persistence, and conscientiousness. Table 2 presents descriptive statistics and zero-order correlations for variables in Study 1.

**Table 2: Descriptive statistics and correlations for variables in Study 1 (*N* = 94).**

| Variable | Minimum | Maximum | Mean | *SD* | NFC | Cautious. | Self-disc. | Self-cont. | Impulse | Errors | TP | APT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NFC (metric = 1-5) | 1.83 | 4.78 | 3.37 | 0.65 | | | | | | | | |
| Conscientiousness | | | | | | | | | | | | |
| Deliberation (NEO-FFI; 1-5) | 1.70 | 4.70 | 3.39 | 0.62 | .31* | | | | | | | |
| Self-Discipline (NEO-FFI; 1-5) | 1.80 | 4.80 | 3.16 | 0.74 | .34* | .53* | | | | | | |
| Self-Control (MPQ; 1-5) | 2.10 | 4.80 | 3.55 | 0.57 | .23* | .84* | .57* | | | | | |
| Impulse Control (AB5C; 1-5) | 2.50 | 4.58 | 3.71 | 0.49 | .28* | .73* | .49* | .68* | | | | |
| Essay errors found (%; 0-100) | 17.0 | 88.0 | 45.0 | 14.0 | .12 | -.02 | -.02 | .04 | -.03 | | | |
| Task Persistence (TP) (self-report; 1-4) | 1.00 | 4.00 | 3.07 | 0.64 | .41* | .31* | .51* | .25* | .29* | -.14 | | |
| Anagram Persistence Task (avg mins / question; 0-3) | 0.16 | 3.00 | 1.51 | 0.78 | .26* | .22* | .02 | .12 | .26* | -.01 | .15 | |
| Cognitive task choice | 0.0 | 1.0 | - | - | .34* | .13 | .03 | .15 | .06 | .06 | .02 | .29* |

#### 3.1.3.2.1 Need for Cognition

The empirical linkages between NFC scores and probability of choosing the cognitively challenging task over the simpler task was modeled using a logistic regression, with NFC scores as the predictor and behavioral choice as the dichotomous outcome. Results from the logistic regression revealed that indeed NFC scores were able to successfully discriminate (72.3% classification accuracy compared to both the 50.0% baseline and 62.0% largest-group baseline) between those who chose the more cognitively challenging task (Task 2) and those who chose the cognitively simpler task (Task 1), Wald's $\chi^2$ = 9.71, *B* = 1.20, *odds ratio (OR)* = 3.33, *p* = .002. This indicates that for every unit increase in NFC scores, the odds of choosing the challenging over simpler task more than triples (i.e., OR = 3.33). This can be seen visually, by plotting the predicted probability of choosing Task 2 over Task 1 (calculated using the coefficient and intercept values of the best fitting exponential regression line) for each particular NFC score obtained in the sample (see Figure 10).[10] At least three important things can be gleaned from Figure 10, which are critical to imbuing meaning to the metric of NFC scores. The first observation that can be made is to garner general information about how the NFC scores along the metric range map onto the relevant behavioral referent. In this case, it can be seen that NFC scores of "2.0", "2.5", "3.0", "3.5", "4.0", and "4.5" map onto about a 10%, 17%, 27%, 40%, 55%, and 70% chance of choosing the more cognitively challenging over simpler task (i.e., Task 2 over Task 1).

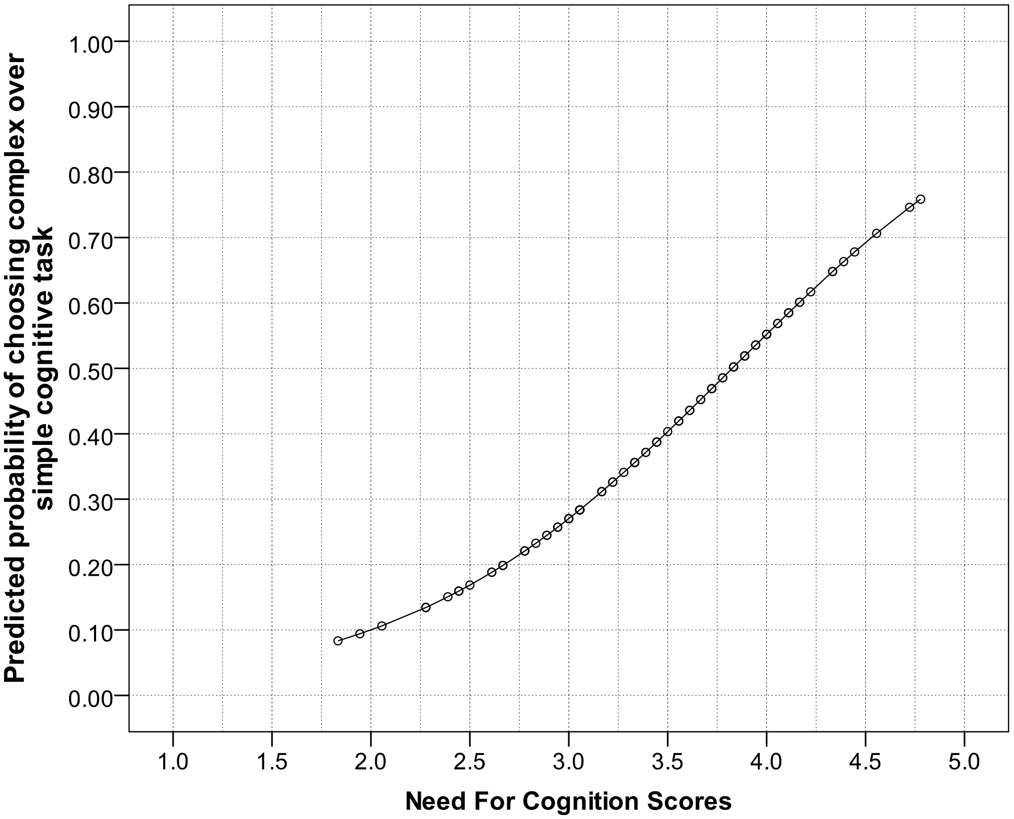


**Figure 10: Predicted probability of choosing a cognitively challenging task (Task 2) over a cognitively simpler task (Task 1) given need for cognition (NFC) scores.**

[10] Interpretation of the odds ratio (in this case OR = 3.33) can be visualized in this figure. For example, an NFC score of "2.0" yielded a predicted probability of choosing Task 2 over Task 1 of approximately 0.10, equivalent to an odds of about 0.11 (0.10/0.90) whereas an NFC score of "3.0" yielded a predicted probability of choosing Task 2 over Task 1 of approximately 0.27, equivalent to an odds of 0.37 (0.27/.73). Hence, the odds of choosing Task 2 over Task 1 more than triples (0.37/0.11 = 3.33) for every 1-unit increase in NFC scores.

The second important observation that can be made from Figure 10, which follows the general logic implied by Blanton and Jaccard's (2006b) analysis, is to use the 50% probability of choosing Task 2 over Task 1 as a qualitatively distinct behavioral reference point to imbue meaning into NFC scores. That is, to the extent that one interprets 50% as a meaningful behavioral reference point, the mapping between an NFC score of approximately "3.8" and 50% can be used to imbue NFC scores with meaning, by interpreting other NFC scores relative to the value of "3.8". In other words, the approximate location of the NFC scores on the underlying dimension of NFC can be inferred relative to the "3.8" threshold value. Overall, this metric meaning analysis implies that the metric range of NFC scores seems to be capturing the lower end of the underlying continuum of NFC. This is based on the logic that the NFC numerical midpoint (i.e., "3.0") is associated with a value on the behavioral reference point that is considerably *lower* than the behavioral threshold value (27% lower than 50%). If the opposite would have emerged (i.e., the NFC numerical midpoint of "3.0" was associated with a 75% chance of choosing the more cognitively challenging task), then this would have suggested that the NFC measure would be capturing the higher end of the underlying continuum of NFC.

A final observation that can be made from Figure 10 is that the scale midpoint (i.e., "3.0") maps onto a 27% chance of choosing Task 2 over Task 1. This observation can be seen as providing preliminary empirical evidence that the scale midpoint should not be interpreted as reflecting a neutral position on the underlying dimension of NFC, assuming that a 50% chance of choosing Task 2 over Task 1 is interpreted as a meaningful NFC behavioral reference point. In other words, this result implies that characterizations of individuals above the scale midpoint as "high" in NFC and those below the midpoint as "low" in NFC would be misleading, given that the scale midpoint of "3.0" was associated with an approximate 27% chance of choosing the more cognitive challenging task. In addition, given that the median of NFC scores in the sample was "3.4", the empirical mapping found also implies that characterization of individuals above the median as "high" or those as below the median as "low" on NFC could also be seen as misleading, given that a score of "3.4" was associated with an approximate 32% chance of choosing the more cognitive challenging task. Granted, these implications only hold to the extent that experts in this research area agree that this behavioral choice reflects a theoretically meaningful reference point that actually reflects the underlying dimension (and is also provisional on replicating the finding). Hence, as NFC scores are linked to other meaningful and/or more extreme need for cognition behavioral reference points, NFC scores could gain even more meaning and interpretability.

### 3.1.3.2.2 Task Persistence

Plotting mean time spent on the unsolved near-impossible anagram puzzles against self-reported task persistence (TP) scores revealed a linear trend (see Figure 11). A linear regression confirmed the presence of a positive relation (B = 0.18, $\beta = r = .15$), though the effect did not attain conventional levels of statistical significance ($p = .15$). Interestingly, however, this effect is consistent with results from Ditre and Brandon (2008) who found a small positive relation between the same self-report TP measure and a mirror-tracing persistence task ($r = .17$, $p = .056$) and a breath-holding persistence task ($r = .16$, $p = .07$). Given that Brandon et al. (2003) found a positive relation between this mirror-tracing persistence task and the exact anagram persistence task used in my study ($r = .27$, $p = .001$), this suggests that the correlation found in my study is a meaningful effect given that it is consistent with past research. From Figure 11 (solid line), it can be seen that a self-report TP score of “2.0” corresponded to approximately 1 minute and 19 seconds of mean time persisting on the near-impossible anagrams whereas a self-report TP score of “3.0” corresponded to about 1 minute and 30 seconds of persistence. This linear relationship can be gleaned more precisely by directly interpreting the unstandardized regression coefficient (B = 0.18), which indicates that a 1-unit increase in self-report TP scores is associated with an increase of 11 seconds (11 = 0.18 × 60) in persistence.[11] Hence, although in this case the behavioral referent does not have a clear

[11] This particular finding demonstrates the conceptual idea of how metric research can help us overcome limitations of NHST. In this case, a researcher strictly relying on *p*-values or effect size would have likely dismissed the result given the greater than .05 *p*-value and the “small” effect size (2.2% of the variance explained). Focusing our attention on the meaning of measurement scores, however, reveals a potentially more useful perspective on assessing the actual noteworthiness of this empirical finding. In this particular case, a researcher must ask themselves the (potentially difficult) question of whether an 11-second increase in actual task persistence for every 1-unit increase in the self-report measure scores is noteworthy.

qualitatively distinct threshold cut-off value, the mapping nonetheless imbues the self-report TP scores with some meaning.

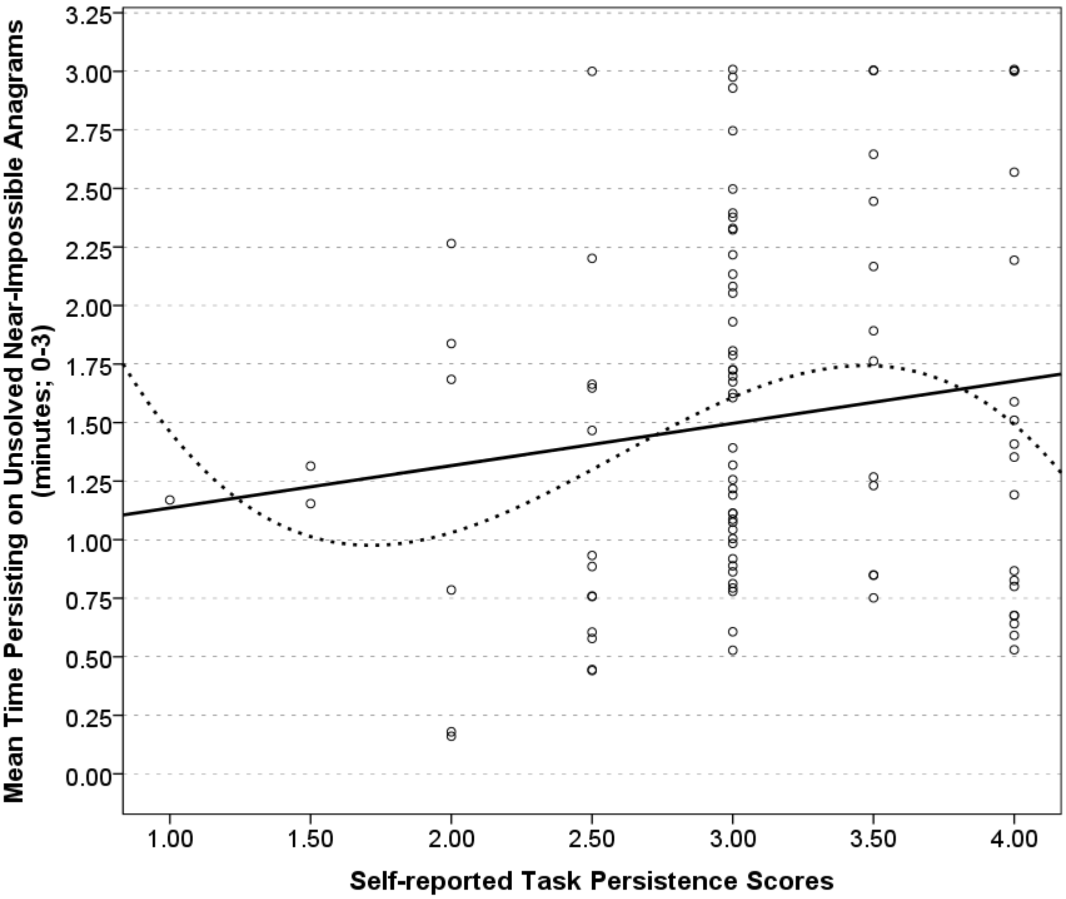


**Figure 11: Mean time spent on unsolved near-impossible anagrams (in minutes) in the anagram persistence task (APT) plotted against self-reported task persistence scores using a linear (solid line) or cubic function (dotted line).**

For example, knowing that a self-report TP score of “1.0” is associated with approximately 1 minute and 8 seconds of actual persistence whereas a score of “4.0” is associated with approximately 1 minute and 40 seconds of persistence helps one get a sense of what these scores might mean in relation to the underlying dimension of task persistence.

An interesting second perspective on the mapping between self-reported and actual persistence can be gained when one takes a closer look at the scatterplot in Figure 11.

Although overall it appears that as self-reported TP scores increase time persisting also increases, for values greater than “3.5” on the self-report measure, time spent persisting actually seems to *decrease* rather than *increase*. This could be the case if individuals who endorsed the highest possible response option for both self-report items (“4”s on both items) exhibited an over-reporting bias when indicating their typical persistence, implying that this select group of individuals (those scoring “4”s on both items) exhibited less persistence behaviorally than their “3.5” counterparts. Indeed, a cubic function applied to the data seemed to fit the data reasonably well, explaining approximately three times more variance than the linear function (6.2% versus 2.2%) (see Figure 11, dotted line). Hence, using this response function, a score of “3.0” corresponded to about 1 minute and 36 seconds of persistence; a score of “3.5” corresponded to about 1 minute and 44 seconds; whereas a score of “4.0” corresponded to only about 1 minute and 29 seconds. This could be seen as preliminary evidence suggesting that responses on the self-report measure of task persistence may suffer from sub-optimalities that further construct validity research should clarify. This curvilinear effect could also potentially explain the consistently small positive linear relations observed between the self-report measure and the three behavioral persistence measures (as found by Ditre and Brandon, 2008, and in my study).

Indeed, by re-analyzing Ditre and Brandon’s (2008) data (kindly provided by the authors upon request), I was able to replicate this non-linear cubic response function in Ditre and Brandon’s data. That is, a cubic function modeling the relation between the same self-reported task persistence scores and mean time persisting on unsolved mirror-tracing trials (which has shown a sizable correlation to the exact anagram persistence task used in my study) explained approximately twice the amount of variance than a linear function (5.3% vs. 2.8%). The nature of the cubic function followed the same “dipping pattern” as in my data, whereby a score of “3.0” corresponded to about 2 minutes and 52 seconds of persistence; a score of “3.5” corresponded to about 3 minutes and 1 seconds; whereas a score of “4.0” corresponded to only about 2 minutes and 48 seconds (all values numerically higher because participants were allowed up to 5 minutes per trial compared to a maximum of 3 minutes in my task). Hence, this finding provides further evidence supporting the hypothesis that certain individuals (i.e., those scoring “4”s on both items) may be exhibiting a reporting bias. Furthermore, this finding implies that further construct validity research could examine whether a strong accuracy instruction eliminates the alleged over-reporting bias, which could be reflected in the data if a linear function would explain more variance than the cubic function.

### 3.1.3.2.3 Conscientiousness

Contrary to expectations, scatterplots of percentage of errors found in the essay task plotted against the scores from the relevant facets of conscientiousness did not reveal any clear mappings for any of the four facets. That is, higher conscientiousness scores did not necessarily correspond to a higher percentage of mistakes found in the essay task for none of the facets (*r* = -.02, *r* = .04, *r* = -.02, *r* = -.03, for the Impulse Control, Self-Control, Self-Discipline, and Deliberation facets, respectively).

Many factors could underlie why no meaningful mappings were found between the conscientiousness facets and performance in the specifically designed detail-oriented task. I will elaborate on a few possibilities. First, although carefully constructed, the essay task may not have been the best task to capture the intended “detail-orientedness” manifestation of conscientiousness. The primarily language-based component of the task may not have captured detail-orientedness in those less linguistically inclined. Indeed, consistent with this idea, a post-experiment debriefing question tapping the extent to which individuals read books, revealed a small positive correlation between reading and percentage of errors found (*r* = .18, *p* = .09, overall and *r* = .22, *p* = .03 for errors found on the first page, which is likely more diagnostic given most people did not get to the last page of the essay task). Second, perhaps the essay task was mostly driven by (i.e., confounded with) motivation rather than conscientiousness. Third, my choice of conscientiousness facets may have contributed to the null mappings. Although I took great care in selecting facets that seemed to be most theoretically-relevant to detail-orientedness, admittedly I was not satisfied with my final choices of measures. For example, closer inspection of the items comprising the facets revealed odd items that do not seem to reflect the facet I intended to measure (e.g., “I do crazy things,” which was part of the Deliberation facet).[12] Hence, perhaps the chosen facets did not accurately reflect the construct I intended to capture. Similarly, the fact that I used the (briefer) IPIP instantiations of the MPQ and NEO-FFI might have played a part. Finally, perhaps partial violation of the matching principle with respect to temporality may have contributed to the null mappings. Conscientiousness self-reports are typically assumed to capture individuals' recall of how they characteristically behave with respect to the conscientiousness facets whereas the criterion behavior assessed in the study captured individual's transient momentary detail-oriented inclination at the present moment in time. One possibility to overcome this issue would be to take the average performance on the essay task across three independent occasions and then examine the mappings between conscientiousness facet scores and behavioral performance.

Although the conscientiousness facet scores did not reveal meaningful mappings to essay performance, the Impulse Control facet of conscientiousness did reveal some interpretable linkages to the behavioral persistence in the APT (see Figure 12). As can be seen in Figure 12, every unit increase in Impulse Control self-report scores corresponded to roughly a 25 second increase in persistence on the APT (B = .41, $t$ = 2.56, $p$ = .01; .41 × 60 = 25 seconds). For example, individuals self-reporting conscientiousness around the scale midpoint (i.e., "3") persisted for an average of about 1 minute and 12 seconds on the near-impossible unsolved anagrams whereas those self-reporting around "4" persisted for about 1 minute and 37 seconds.

[12] Although psychometrically hazardous, I also examined empirical mappings between specific items deemed most theoretically relevant in predicting detail-oriented behavior in the essay task (I *a priori* picked four questions from the MPQ and NEO-FFI (i.e., "I am exacting in my work", "I pay attention to details", "I avoid mistakes", and "I choose my words with care") and four items from the adjective ratings (i.e., "careful", "cautious", "conscientious", and "systematic") that seemed the most theoretically relevant). Interestingly, I did find two meaningful mappings: one between the adjective ratings for the "cautious" item and overall errors found ($r$ = .30, $p$ = .004) and another between the adjective ratings for the "conscientious" item and errors found on the first page ($r$ = .25, $p$ = .02). These mappings, however, must be interpreted with caution given that they involve a 1-item self-report measure. Nonetheless, this could be seen as preliminary evidence suggesting, in line with the matching principle, that more specific self-report measures would be more useful in predicting the specific expression of conscientiousness presumably required for the essay task.

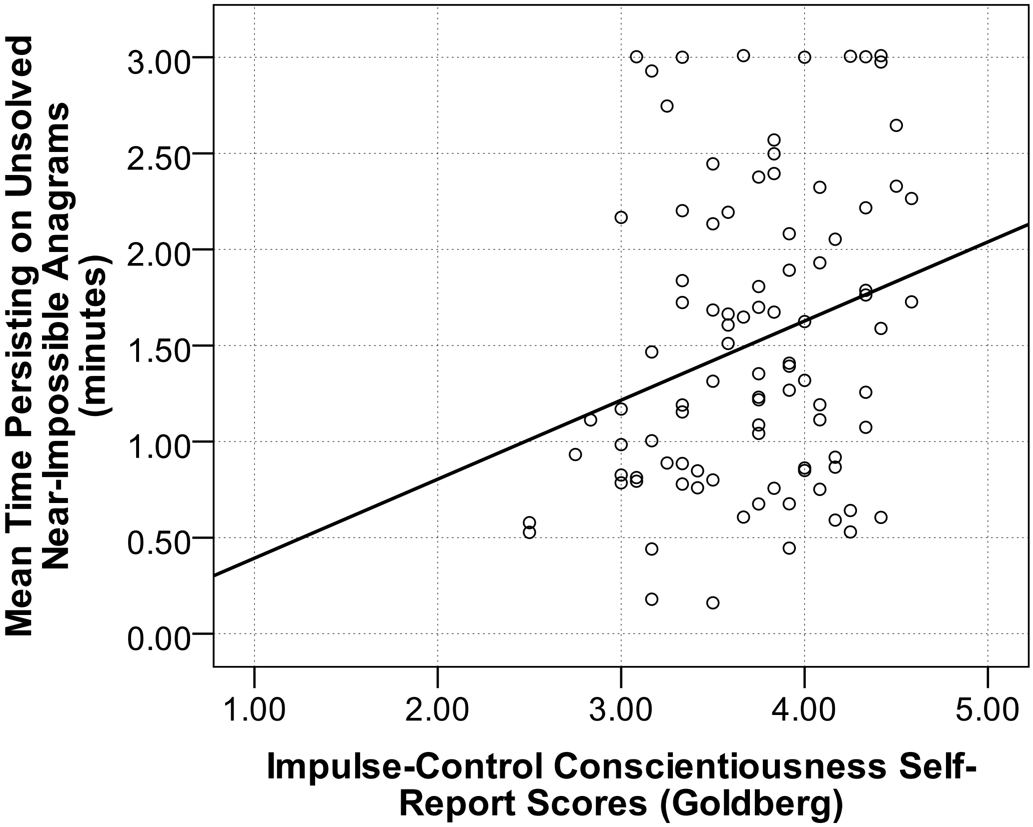


**Figure 12: Mean time spent on unsolved near-impossible anagrams in the anagram persistence task plotted against self-reported Impulse-Control scores (Goldberg).**

This metric mapping can actually be seen as consistent with the broad conceptualization of conscientiousness which includes as a facet the degree of organization, persistence, and motivation in goal-directed behavior (Costa & McCrae, 1992). Assuming that some individuals had it as a goal to persist in the task, the conscientiousness-persistence metric mapping could make sense theoretically to the extent that the particular conscientiousness facet captured this tendency to persist in goal-directed behavior.

### 3.1.4 Discussion

Overall, Study 1 demonstrated the feasibility of calibrating the metric of measures commonly used in basic psychological research by employing two metric calibration strategies inspired by past research. In summary, I found a meaningful and illuminating empirical mapping between NFC scores and probability of choosing a cognitively challenging versus simpler task. I also found an informative linkage between a pragmatically useful self-report task persistence measure and actual behavioral persistence in a commonly used persistence task. Finally, although no clear connection was found between the conscientiousness facets and errors found in the essay task, I did find some theoretically interpretable linkages between scores from the Impulse Control facet (Goldberg) and behavioral persistence in the anagram task. Taken together, these findings suggest that the metric calibration approach is feasible in achieving the goal of calibrating the metric of measures commonly used in psychological research. Hence, these promising results suggest that the benefits proposed in my conceptual analysis could one day potentially bear fruit.

Supposing replication and consensus from the field as to the meaningfulness of the behavioral reference points, these metric findings could speak to theoretical issues in the research literatures involving the constructs of need for cognition, conscientiousness, and task persistence. For instance, one could delve into the attitudes and persuasion literature and find published studies that involved NFC as a moderating variable of attitude change. One could then attempt to re-analyze moderated multiple regression analyses in these studies using calibrated NFC metric values, which could shed additional light on the research questions tested (e.g., NFC as a moderator of the impact of central vs. peripheral cues on attitude change). In fact, in the implications section of the General Discussion, I will report results of this kind of re-analysis using precisely such an approach.

## 3.2 Study 2

The primary goal of Study 2 was to provide a preliminary demonstration of the feasibility and utility of the metric approach with regard to contributing to theoretical development. More specifically, the goal was to illustrate how the metric calibration approach could be used to shed light on theoretical debates that involve claims of an absolute nature. Another goal of Study 2 was to further demonstrate the metric approach for a distinct construct commonly studied in basic psychological research. The study centered on the topic of self-enhancement in the context of theoretical controversies regarding the question of pan-cultural self-enhancement. An important aspect of this theoretical debate is that it involves making claims regarding the presence versus absence of self-enhancement within and across cultures, which involves making theoretical claims about absolute levels of self-enhancement based on scores from trait rating measures which have arbitrary metrics. Consequently, this topic was chosen precisely to attempt to illustrate how the metric approach can potentially shed new light on the debate by providing specific information regarding the metric meaning of trait rating scores for those measures. Toward this end, two self-enhancement measures that play a focal role in the debate were calibrated to a theoretically-relevant behavioral reference point.

### 3.2.1 Pan-cultural Self-Enhancement Debate

Although research on cultural differences in the self-enhancement motive has a long history (Heine, Lehman, Markus, & Kitayama, 1999; Markus, Kitayama, & Heiman, 1996; Pepitone & Triandis, 1987), the pan-cultural self-enhancement debate intensified when Sedikides et al. (2003) asserted as misguided the idea that self-enhancement is pervasive in individualistic cultures (Westerners) but absent collectivistic cultures (Easterners). Rather, Sedikides et al. proposed that Westerners and Easterners use different tactics to achieve the same goal of self-enhancement, such that Westerners self-enhance on individualistic attributes whereas Easterners self-enhance on collectivistic attributes. This tactical self-enhancement hypothesis was supported empirically in a set of two studies showing that Americans self-enhanced on individualistic attributes whereas Japanese self-enhanced on collectivistic attributes. This led Sedikides et al. to conclude that self-enhancement is a universal human motive. Heine (2005) challenged this claim on empirical and methodological grounds. Specifically, Heine argued that Sedikides et al. ignored numerous conflicting past findings that contradict their main conclusion and that they used inappropriate cross-cultural samples. More relevant to the present dissertation, Heine also called into question Sedikides et al.'s pan-cultural claim based on an important methodological issue. Heine argued that the better-than-average paradigm used by Sedikides et al. to index self-enhancement is flawed given that it is confounded with a general cognitive bias. That is, past research has shown that individuals view not only themselves as better than average, but they also view any randomly chosen individual as better than average (the "everyone is better-than-average-effect" [EBTA]; Klar & Giladi, 1999; Sears, 1983). It is clear that rating a random other as better than average has nothing to do with self-enhancement and hence this cognitive bias contaminates the self-enhancement index scores in an upward fashion and seriously calls into question Sedikides et al.'s main theoretical claim.

Although Sedikides, Gaertner, and Vevea (2005) responded to Heine's (2005) challenge with a meta-analytic investigation showing the same overall patterns of results across 27 combined samples, Heine, Kitayama, and Hamamura (2007a) further challenged Sedikides et al.'s claim by showing their meta-analytic conclusions do not hold with the inclusion of missing cross-cultural studies. Heine et al. also once more challenged Sedikides et al. on methodological grounds (see also Heine, Kitayama, & Hamamura, 2007b; Sedikides, Gaertner, & Vevea, 2007a, 2007b). Importantly, when considering all cross-cultural studies available in the literature, Heine et al. (2007a) noticed a striking pattern: the studies most consistently yielding evidence supportive of the pan-cultural hypothesis used the better than average (BAE) method (either combining self-vs.-other judgments or rendering the judgments separately). And, given that this method over-estimates self-enhancement because of the abovementioned methodological artifact conflated with the method (i.e., the EBTA), this calls into question the conclusion that self-enhancement is universal. Taken together, one important aspect of the debate involves the fact that much of the evidence in support of the pan-cultural self-enhancement hypothesis comes from studies using BAE measures, which suffer from a methodological artifact that inflates estimates of self-enhancement.

From a metric perspective, I would argue that calibrating measures of self-enhancement (e.g., BAE paradigm: self-vs.-other judgments simultaneously or self minus other separately) could shed light on the pan-cultural theoretical controversy, by attempting to provide evidence about what types of self-enhancement behaviors correspond to different scores on these trait rating measures of self-enhancement. In the process, metric research may also reveal more directly how much construct-irrelevant cognitive biases are contaminating these self-enhancement trait measures. Critically, one would examine what type of self-enhancement behaviors actually correspond to particular scores on the trait measures, including critically scores that are typically interpreted as self-enhancement (e.g., scores statistically greater than the scale midpoint for the self-vs.-other judgments made simultaneously).

The main goal of Study 2 was to calibrate two of the trait self-enhancement measures (Alicke et al., 1995; Sedikides et al., 2003) that feature prominently in this debate, namely a trait rating measure involving self versus other judgments made simultaneously and a trait rating measure where other judgments are subtracted from self judgments. This was achieved by calibrating the metric of these two measures to behaviors reflective of self-enhancement, which could be used as a reference point. I used Paulhus et al.'s (2003) over-claiming technique (OCT) as the behavioral index of self-enhancement. The calibration of these trait rating self-enhancement measures could potentially shed light on the universality of self-enhancement debate because part of this debate involves trait rating measures of self-enhancement that are often interpreted in an absolute fashion (e.g., Gaertner et al., 2008). As reviewed in the introduction, this strategy of meter reading is unfounded given that the metrics of these trait rating measures are arbitrary and hence it is unknown what region of the underlying dimension of self-enhancement is captured by these trait rating measures. Applying a metric approach to this topic, however, could potentially shed new light on this controversy by examining what kinds of behavioral manifestations of self-enhancement correspond to scores typically interpreted as self-enhancement. If, for instance, scores above the midpoint traditionally interpreted as self-enhancement (e.g., mean trait rating of "4.5" tested against a scale midpoint of "3.5"; see Gaertner et al., 2008) are associated with negligible behavioral manifestations of self-enhancement, then this would cast doubt on claims of self-enhancement based on greater-than-the-midpoint analyses of trait rating measures. Granted, this would represent only the first step toward speaking to this controversy, given that consensus would need to be reached with respect to the behavioral criterion of self-enhancement. An auxiliary goal of Study 2 was also to calibrate the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984), another commonly used self-enhancement measure, to behavioral markers of self-enhancement manifested in the OCT.

## 3.2.2 Method

### 3.2.2.1 Design and Participants

One hundred ninety-four (194) University of Western Ontario undergraduates participated for partial course credit (97 males, 97 females; mean age = 18.9, *SD* = 1.2, range = 17 to 25). No restrictions were imposed on participant sex, age, or ethnicity. The study included two experimental conditions, having altered instruction sets for the OCT as a first exploration into the proposed experimental approach to metric calibration. However, given that this instruction manipulation had virtually no effect on OCT scores, I do not discuss these data further.[13] Hence, the final sample was composed of 97 undergraduates that completed the OCT with standard instructions (47 males, 50 females; mean age = 18.9, *SD* = 1.3, range = 17 to 25). Thus, all participants completed the same measures and tasks in the same order (see below for details).

### 3.2.2.2 Procedure and Materials

Participants were run in groups of two to five in a lab testing room where each participant was seated in a separate cubicle in front of a PC computer. Before starting, the experimenter verbally instructed each participant to carefully read all instructions and to ask the experimenter for clarification if any questions arose. The experimenter also instructed each participant to turn off any and all electronic devices before starting. Participants then followed on-screen instructions and completed each measure in a serial fashion in the following order: combined trait rating measure of self-enhancement (self-vs.-other judgments), separate trait rating measure (self and other judgments made separately), a filler task (Remotes Association Test), the OCT, the BIDR, and then demographics and debriefing questions.

#### 3.2.2.2.1 Combined Self vs. Other Judgments

The first trait measure of self-enhancement involved self-versus-other judgments following the logic and structure of the trait rating measures used in the literature on the

[13] Specifically, the mean OCT scores across conditions were $M_{warning}$ = .09 (*SD* = .08), $M_{control}$ = .11 (*SD* = .12), and $M_{\text{confidence}}$ = .09 (*SD* = .10).

pan-cultural debate (Sedikides et al., 2003), except with culture-independent traits given the present non-cross-cultural sample (the standard traits as used by Alicke et al., 1995). The measure originated from the better-than-average effect literature and was subsequently adopted by researchers in the pan-cultural debate. The better-than-average effect has been argued to represent a fundamental type of self-enhancement reflected in the tendency to view one's behaviors, opinions, and characteristics more favorably than those of others (Alicke et al., 1995). Assessment of these views typically involve making ratings of oneself relative to an "average other" on a series of traits on a bipolar scale anchored at the extremes by self and other (e.g., "To what extent does Trait A describe you relative to the average other?": 1 = *much worse than the average other,* 4 = *as well as the average other,* 7 = *much better than the average other*). However, given research showing that better-than-average effects can be inflated when the comparison target is ambiguous (Alicke et al., 1995), researchers have often tried to make the comparison target more concrete to permit more stringent analyses. Given that some of the past pan-cultural studies (but not all) have adopted these more stringent assessment conditions, I decided to also adopt these more stringent measure configurations. Hence, following Gaertner et al. (2008), participants rated the extent to which each listed trait described themselves relative to the average Western university student of their own age and gender on an 9-point scale (1 = *much worse than the average university student of my age and gender,* 4 = *as well as the average university student of my age and gender,* 7 = *much better than the average university student of my age and gender*). Participants rated the following 10 positive and 10 negative traits (adopted from Alicke et al., 1995): *dependable, intelligent, considerate, observant, polite, respectful, cooperative, reliable, friendly, creative* and *gullible, disobedient, snobbish, lazy, disrespectful, mean, unforgiving, vain, uncivil, unpleasant* (α = .75).

#### 3.2.2.2.2 Separate Self vs. Other Judgments

For the second measure of self-enhancement, participants rated the extent to which a series of statements were true of themselves and subsequently of others, assessed separately (e.g., Hornsey & Jetten, 2005; Heine & Lehman, 1999). Mean ratings of others were then subtracted from ratings of the self to form an index of self-enhancement.

Participants rated the extent to which the following 10 traits were true of themselves (5 positive, 5 negative) and subsequently the extent to which these same traits were true of others: *clear-headed, resourceful, reliable, perceptive, trustful* and *insecure, spiteful, unstudious, maladjusted, complaining* (again from Alicke et al., 1995). Because past research has failed to find order effects for whether self versus other ratings are completed first, participants completed the ratings in one order only (self ratings first) (Brown & Kobayashi, 2002; Hornsey & Jetten, 2005). For self-ratings, participants were informed to indicate the extent to which the traits were true of themselves whereas for ratings of others, participants were informed to indicate the extent to which the traits were true of the average Western student of their age and gender (using the scale anchors: 1 = *not at all true* and 7 = *completely true*) (α = .62).

#### 3.2.2.2.3 Over-claiming Technique

After completing a brief filler task alleged to assess their creativity (i.e., the RAT; about 5 minutes), participants completed the OCT as a behavioral reference point of self-enhancement. The OCT involves the presentation of a series of words allegedly describing people, places, and objects, some of which, unbeknownst to the participants, refer to non-existent items (i.e., some words are foils). In the standard OCT task, participants are instructed to indicate the extent to which they are familiar with each word (0 = "never heard of it" to 6 = "completely familiar"). Independent indices are then typically computed, using signal detection theory, to estimate actual knowledge (i.e., accuracy: hit rate – false alarm rate) and self-enhancement (i.e., over-claiming bias: [hit rate + false alarm rate] / 2). Although the signal detection theory framework provides a mathematically rigorous estimate of knowledge exaggeration, unfortunately these estimates are interpretationally ambiguous given they are composed of the mean of both the hit rate and false alarm rate. Given that the goal of metric research is to link test scores to maximally meaningful (i.e., unambiguous) behavioral scores, I chose to use the raw false-alarm rate to index the self-enhancement bias, given it provides the clearest and most meaningfully interpretable face-valid operationalization of self-enhancement (i.e., the proportion of non-existent items claimed as familiar). Hence, this implied a slight change in the response format of the task, whereby participants in this study indicated whether the item was familiar or unfamiliar using a binary choice format rather than the usual 7-point polytomous choice format.

Participants completed a variant of the 150-item OCT (Paulhus et al., 2003), which is broken down into 10 categories of 15 items. Three out of every 15 items per category were foils, that is, they do not actually exist (see Appendix E). Standard instructions were used, whereby participants were instructed to indicate whether each item was familiar to them or unfamiliar to them. The number of non-existent words indicated as familiar served as the main behavioral index of self-enhancement (metric range = 0 to 30).

#### 3.2.2.2.4 Balanced Inventory of Desirable Responding

Participants completed the 40-item BIDR (Paulhus, 1984; 1991), a measure of social desirable responding that has previously been used in the context of the over-claiming technique (Paulhus et al., 2003). The BIDR involves the subscales of self-deception (honestly held exaggeration of one's positive attributes) and impression management (positive self-presentation targeted at a public audience). Items were answered using a 7-point Likert scale (1 = *Not true,* 4 = *Somewhat true,* and 7 = *Very true*) (BIDR Version 6, Form 40; Paulhus, 1991). By convention, BIDR scoring involves adding one point for each "6" or "7" item response indicated by the participant (Paulhus, 1984; 1991). Hence, the metric for the total scores can range from 0 to 40. Scores from this measure were then calibrated to the behavioral indices of self-enhancement in the OCT. After the BIDR, participants completed a few debriefing and demographics questions (α = .76).

## 3.2.3 Results

### 3.2.3.1 Preliminary Data Treatment

Data from all measures were first screened informally for any evidence of non-compliance. I examined the time taken (as recorded by MediaLab) on instruction screens and actual questionnaire items for any evidence of consistently short latencies suggestive of non-compliance. No unambiguous cases were identified and hence all participants were retained.

### 3.2.3.2 Main Analyses

The metric mapping results will be presented in the order the measures were introduced in the Methods section. Table 3 presents descriptive statistics and zero-order correlations for all Study 2 variables.

**Table 3: Descriptive statistics and correlations for variables in Study 2 (*N* = 97).**

| Variable | Minimum | Maximum | Mean | *SD* | Combined | Separate | Self | Other | BIDR |
|---|---|---|---|---|---|---|---|---|---|
| Combined trait rating scores (1-7) | 3.85 | 6.05 | 4.93 | .56 | | | | | |
| Separate trait rating scores (self – other; -6 to +6) | -1.00 | 2.9 | .77 | .82 | .12 | | | | |
| Self (1-7) | 3.4 | 6.5 | 5.04 | .68 | .34* | .67* | | | |
| Other (1-7) | 2.3 | 5.9 | 4.27 | .63 | .21* | -.59* | .21* | | |
| BIDR (0-40) | 1.00 | 23.0 | 8.54 | 4.80 | .23* | .36* | .51* | .08 | |
| OCT (# of non-existent items claimed as familiar; 0-30) | 0.0 | 16.7 | 3.23 | 3.67 | .29* | .16 | .25* | .06 | .27* |

#### 3.2.3.2.1 Combined Trait Rating Measure

A regression analysis revealed that trait rating scores on the combined self-vs.-other judgments were positively predictive of behavioral over-claiming scores (B = 1.88, $\beta = r = .29$, $p = .004$). This metric mapping can be unpacked by visualizing Figure 13 (solid line), whereby a 1-unit increase on the trait rating scale corresponds to over-claiming knowledge of about 2 more non-existent words. Specific metric mappings for particular trait rating scores revealed that a trait rating score of “4”, “5”, and “6” corresponded to claiming familiarity of about 1.5, 3, and 5 non-existent words, respectively (derived from using the intercept and regression coefficient). These particular metric mappings start to give us a rough sense of the kinds of self-enhancement behaviors associated with particular trait rating scores, hence imbuing the metric of the trait rating measure with meaning. Once replicated on larger and culturally-appropriate samples, these metric mappings could then potentially shed new light on the pan-cultural debate of self-enhancement by examining the kinds of self-enhancement behaviors associated with trait rating scores typically interpreted as self-enhancement (e.g., trait rating score of “5” [on a 7-point scale] associated with over-claiming about 3 non-existing words).

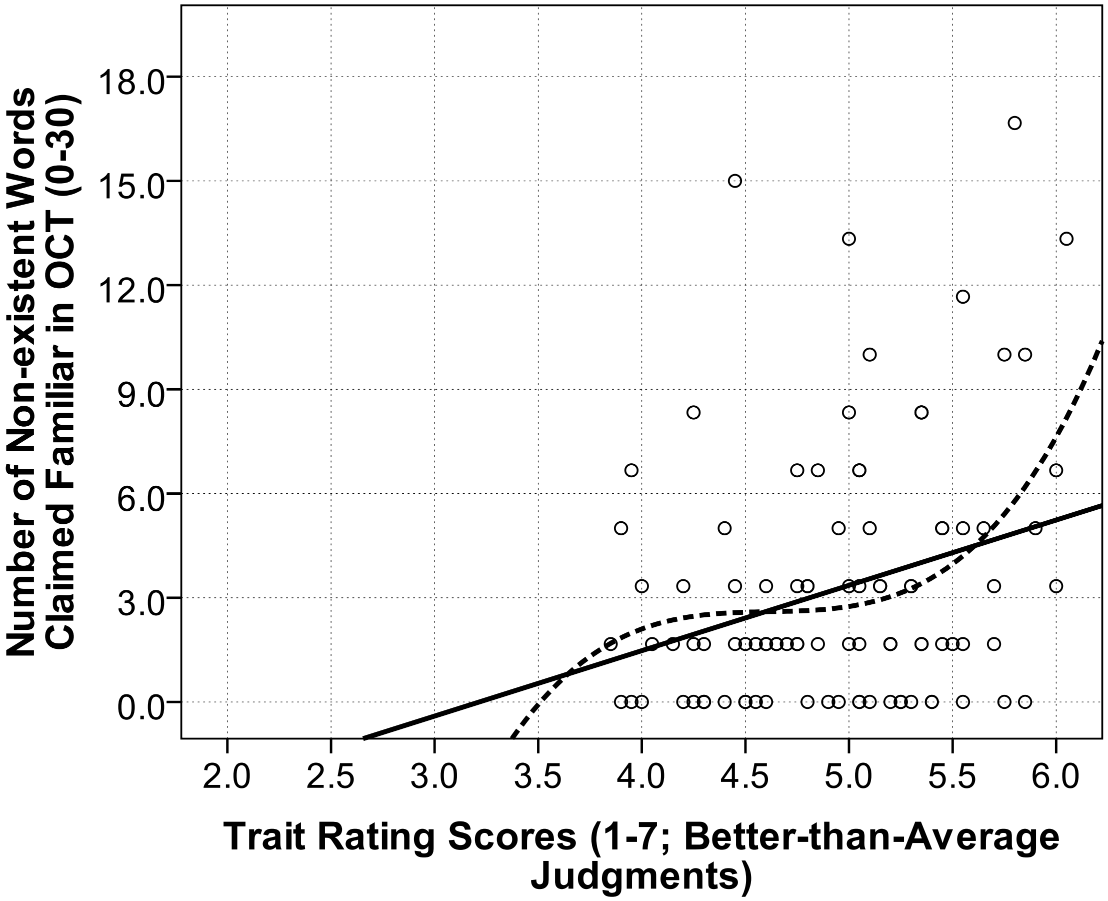


**Figure 13: Number of non-existent words claimed familiar in OCT plotted against trait rating scores using a linear (solid line) or cubic function (dotted line).**

Given the odd shape of the scatterplot of trait rating scores and over-claiming behavior in OCT, I also decided to examine the metric mapping using a non-linear cubic function. This analysis revealed that the cubic function accounted for about 50% more variance than the linear function ($R^2$ = .12 vs. $R^2$ = .083). As depicted in Figure 13 (dotted line), a cubic functional form shows that trait rating scores between “4” and “5” corresponded to over-claiming of approximately 3 words whereas trait rating scores greater than “5.5” corresponded to a sharp increase in over-claiming (“5.5” was linked to about 4.5 words whereas “6” was linked to about 8 words). Consistent with the linear metric mapping, the non-linear metric mapping suggests that very little self-enhancement behaviors corresponded to trait rating scores typically interpreted as self-enhancement.

### 3.2.3.2.2 Separate Trait Rating Measure

Trait rating scores from the self versus other judgments made separately revealed a small positive trend with OCT scores, however this was not statistically significant, B = .71, *β* = *r* = .16, *p* = .12 (see Figure 14). A possible explanation for this less robust relation is that trait rating scores in this measure involved difference scores. Such aggregate scores are known to suffer in reliability in direct proportion to the correlation between the individual components scores (Cronbach, 1958; Edwards, 2002).

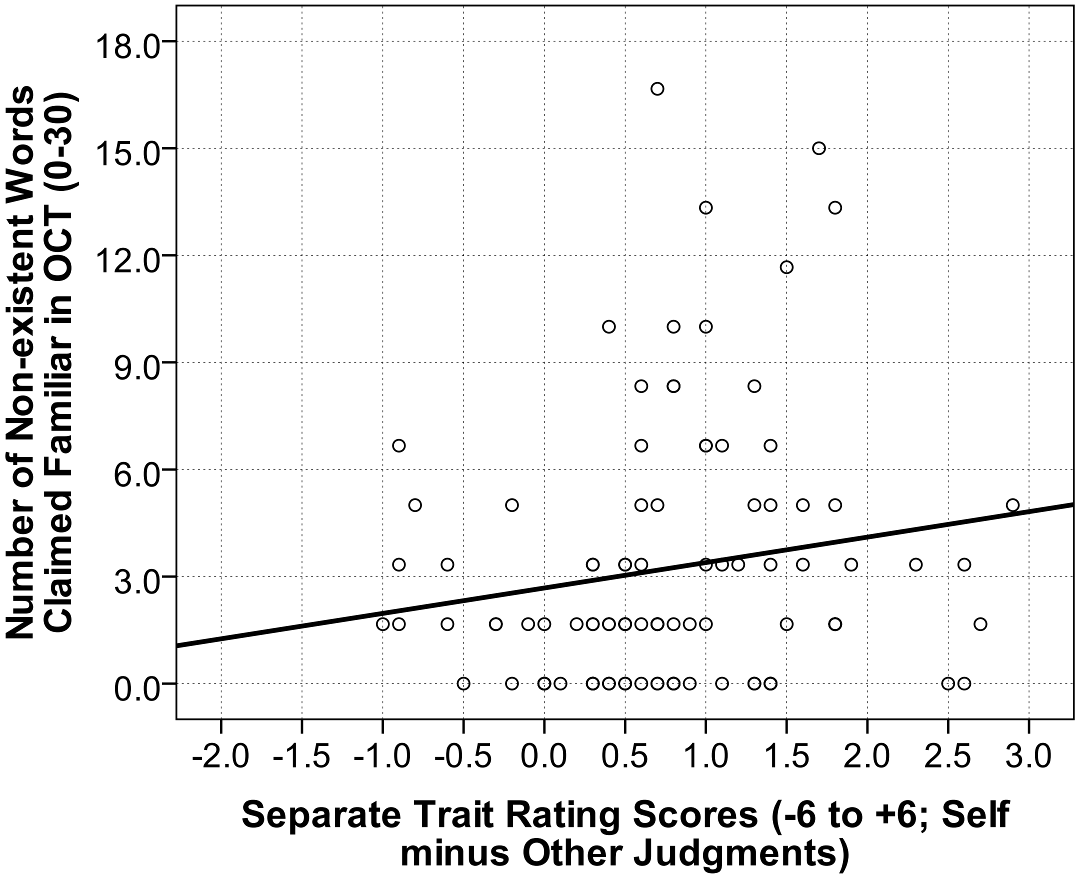


**Figure 14: Number of non-existent words claimed familiar in OCT plotted against trait rating scores made separately using a linear function.**

### 3.2.3.2.3 Balanced Inventory of Desirable Responding

Finally, a regression analysis revealed a linear relation between BIDR scores and OCT scores (B = .21, *β* = *r* = .27, *p* = .008). That is, a 5-unit increase in BIDR scores corresponded to over-claiming of about one more non-existent word (5 × .21 = 1.05; given the much wider metric range of the BIDR). For instance, a BIDR score of "1" corresponded to about two non-existent words claimed as familiar, a mid-range BIDR score of "21" corresponded to about six non-existent words claimed as familiar, whereas a maximal BIDR score of "40" would have corresponded to about 10 non-existent words claimed as familiar (see Figure 15). An examination of BIDR's two subscales revealed that this relation was primarily driven by the self-deception rather than the impression management facet.

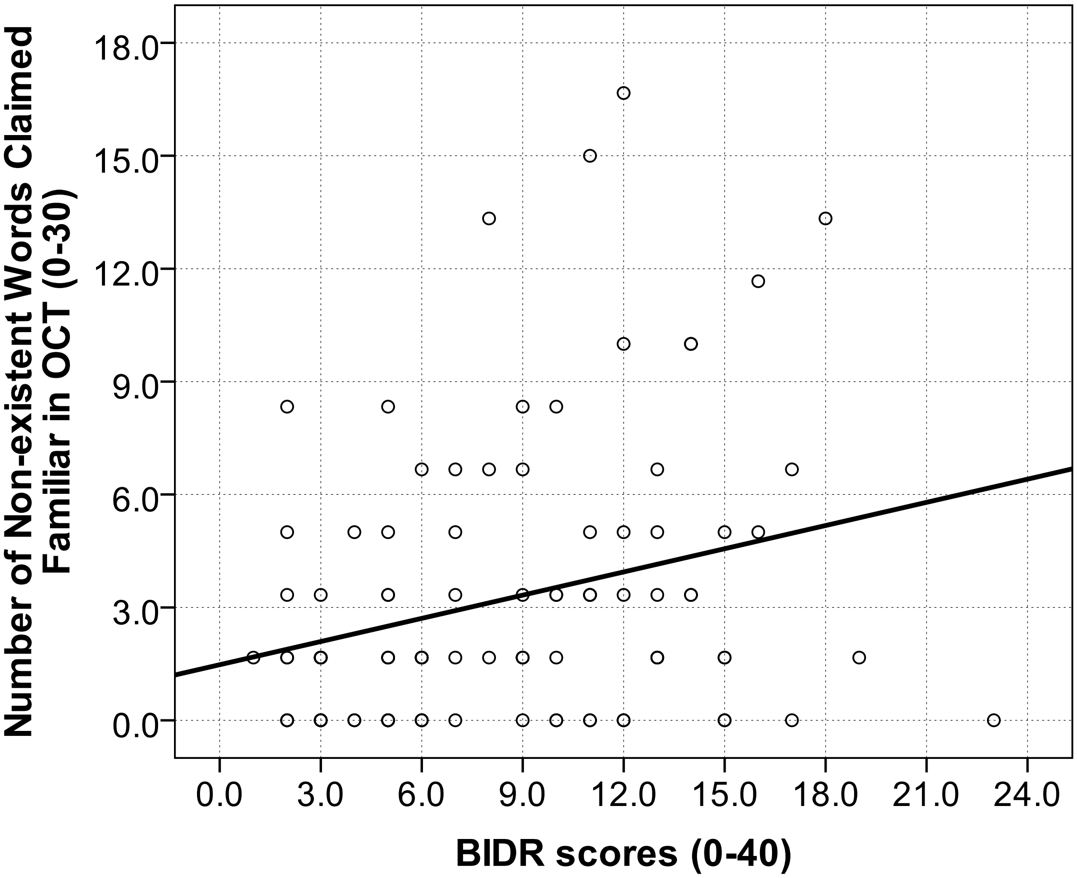


**Figure 15: Number of non-existent words claimed familiar in OCT as a function of BIDR scores using a linear function.**

That is, a regression analysis revealed a more positive relation between BIDR self-deception scores and OCT scores (B = .35, $\beta = r = .29$, $p = .004$) than between BIDR impression management scores and OCT scores (B = .17, $\beta = r = .14$, $p = .18$).

### 3.2.4 Discussion

Study 2 successfully applied the metric approach to the construct of self-enhancement. More importantly, however, the current study illustrated how the metric approach could potentially be valuable in shedding light on theoretical debates involving absolute claims, by focusing on the pan-cultural debate of self-enhancement (Sedikides et al., 2003; Heine, 2005). The metric mapping results for a trait rating measure of self-enhancement commonly used in this debate, showed that very little behavioral evidence of self-enhancement corresponded to trait scores typically interpreted as self-enhancement. That is, a trait rating score of “5” sometimes interpreted as self-enhancement (e.g., Gaertner et al., 2008), corresponded to the over-claiming of only about 3 non-existent words. These results are consistent with Heine’s (2005) concern that such better-than-average trait rating judgments provide inflated estimates of self-enhancement. I will elaborate more on these details and the broader implications of these findings in the General Discussion.

## 3.3 Study 3

The goal of Study 3 was twofold. First, Study 3 sought to demonstrate the utility and feasibility of calibrating the scores of measures capturing predominantly *state*-like constructs. In all of the empirical demonstrations thus far, predominantly *trait*-like constructs have been examined. As previously mentioned, however, the potential benefits of non-arbitrary metrics apply to the measurement of any construct, *state*-like or *trait*-like, or anywhere in between. That being said, some of the benefits proposed in my conceptual analysis are best demonstrated using predominantly *state*-like constructs, which are commonly assessed in the context of experimental studies. Consequently, the primary goal of Study 3 was to empirically reveal the calibration process for a predominantly *state*-like construct, in order to better demonstrate the proposed benefits relevant in experimental contexts (e.g., extracting more information from data patterns). Second, Study 3 was designed to illustrate the calibration approach for behavioral measures. Up to now, all calibrated measures have happened to be self-report measures; it is important to emphasize, however, that issues involving arbitrary metrics are pertinent to any measure, whether self-report, behavioral, or unobtrusive (Blanton & Jaccard, 2006a, 2006b).

To achieve these two primary goals, I examined the construct of risk-taking. In the literature, risk-taking is typically defined as the purposive enacting of a behavior that involves the possibility of some positive consequences or gains (e.g., personal thrill, monetary gain), but with some potential negative consequences (e.g., danger, harm, financial loss; Ben-zur & Zeidner, 2009; Lejuez et al., 2002). Empirical investigations of risk-taking have been executed in different areas of basic psychological research including developmental (Boyer, 2006; Steinberg, 2010), cognitive (Pleskac, 2008; Pleskac, Wallsten, Wang, & Lejuez, 2008), and social psychology (Hamilton, 1974; Leith & Baumeister, 1996). Furthermore, risk-taking has often been studied in an experimental context, supporting the idea that risk-taking involves a substantial *state*-like component amenable to change by situational manipulations (e.g., Goudriaan et al., 2010; Maner, Gailliot, Butz, & Peruche, 2007), although this does not preclude the possibility for temporal stability of the construct (e.g., White, Lejuez, & de Wit, 2008). Taken together, these considerations rendered the construct of risk-taking as an ideal candidate for a metric calibration study with the aforementioned goals.

### 3.3.1 Risk-taking Measures to be Calibrated

As mentioned, risk-taking involves behaviors that involve potential gains at the cost of potential negative consequences. To capture this defining feature of risk-taking, the primary behavioral measure calibrated in Study 3 was the Balloon Analogue Risk Task (BART; Lejuez et al., 2002), which is currently the most widely used and tested sequential risk-taking instrument in the literature (Pleskac et al., 2008). In this task, participants inflate a series of simulated balloons presented on the computer screen. For each balloon, participants can choose the risky option of pumping up the balloon, which inflates the balloon and rewards the participant with a constant amount of money (typically 5 cents), placed in a temporary bank. Naturally, pumping up the balloon sometimes causes it to explode, causing the loss of the accumulated money and the end of the trial. The safe option is to stop inflating the balloon at some point and collect the earned money (which is placed in a permanent bank) and begin the next balloon trial. The task adopts a sequential risk-taking paradigm whereby for each trial, participants must sequentially choose between a risky play option and a safe stop option. Within each trial, risk therefore increases over time in a dynamic way such that choices within a trial become incrementally risky. This is to be contrasted with many other risk-taking tasks involving the choice between gambles involving static non-changing levels of risk (e.g., Brand et al., 2005). Importantly, the BART's dynamic nature models real-world situations in which excessive risk often results in diminishing returns. The BART then was chosen as the primary measure of the study due to these valuable attributes, in addition to its prominence in the literature and the fact that it is the most widely tested and understood sequential risk task in the literature (Pleskac et al., 2008).

In addition to the BART, I also sought to calibrate scores from the hot version of the Columbia Card Task (CCT; Figner, Mackinlay, Wilkening, & Weber, 2009), a recently developed behavioral risk-taking measure that has shown promising results in understanding age-related changes in risk-taking and the informational use processes underlying risk-taking. Similar to the BART, the CCT involves a sequential and dynamic paradigm whereby risk-taking is assessed via participants' voluntary stopping point behavior in a series of incrementally risky choices. However, the CCT goes beyond the BART by (a) assessing the complexity of the decision maker's information use and (b) providing a more optimal probabilistic environment to observe risk-taking behavior (see below for more details). The CCT involves a series of trials in which participants turn over 32 cards presented face down on the computer screen (arranged in four rows of eight cards). The object of the game is to turn over as many cards as possible to accumulate points (each card turned is worth a trial-specific amount). Participants are told they can continue to turn cards over as long as gain cards (smiling face) are encountered. The moment a loss card (frowning face) is encountered, the trial is over and the accumulated points are deducted from the permanent bank. Similar to the BART, participants must decide when to stop and collect their earned points.

### 3.3.2 Theoretical Derivation of Behavioral Reference Points

The search for the most theoretically-relevant and meaningful behavioral reference point to calibrate the BART and CCT was guided primarily by (a) conceptual analysis, (b) past empirical research, and (c) theorizing regarding the cognitive processes underlying risk-taking. First, and consistent with the theoretical derivation of behavioral referents for

Study 1 and 2, the starting point involved the careful consideration of the working-definition of the construct of risk-taking. To re-iterate, risk-taking is most typically defined as behavior entailing the possibility of positive consequences, while at the same time involving the possibility of negative outcomes (Lejuez et al., 2002). Hence, a relevant behavioral reference point to ground the metric of the BART and CCT, first and foremost, must fit within such conceptualization of risk-taking. Second, and following from points raised in the introduction, the reference point must be an objective behavior that has a clear interpretation. *Objective*, meaning that two independent observers could agree that the behavior in question occurred and *clear* meaning that the observed scoring of the behavioral reference point is directly interpretable (e.g., 1 = presence of a behavior and 0 = absence of a behavior; or number of times engaging in some behavior). Following from these considerations, I combed the literature in search of a behavioral measure of risk-taking that satisfied these requirements.

After an extensive search, I decided on a task involving lottery risky choices (adapted from Hsee & Weber, 1999, based on the classic lottery tasks from Tversky & Kahneman, 1981). These lottery choices typically involve a series of choices between two choice options and participants must choose which option they would prefer to receive. For instance, one could be faced with a choice between option A ($4 for certain) or option B (a 50% chance of winning $10 or $0).[14] The behavior of choosing the risky option (rather than the safe option) can then be used as a clear behavioral reference point to calibrate the metric of the relevant target measures. In addition to satisfying all of the aforementioned requirements, this behavioral measure was chosen because of empirical precedence demonstrating that these lottery risky choices were successfully used as a criterion measure to validate a risk propensity scale (Weber, Blais, & Betz, 2002). Theorizing regarding the cognitive processes underlying risk-taking behavior in the BART also supports the theoretical adequacy of the lottery choices as a behavioral reference point (Bishara et al., 2009; Wallsten, Pleskac, & Lejuez, 2005). Wallsten et al. developed and

[14] Following Tversky and Kahneman (1981), to enhance the realism of these behavioral choices, participants were informed that two of the 100 participants would actually receive the monetary sum of one of their realized choices.

successfully substantiated a mathematical model of the multiple cognitive processes underlying behavior in the BART. One of the four parameters in this model involves the extent to which participants are sensitive to changes in the payoffs associated with pumping a balloon on a particular trial (i.e., payoff sensitivity, $\gamma$). Participants with larger values of $\gamma$ are assumed to be more sensitive to the changing payoffs involving gains and losses. Hence, to the extent that lottery choices are at least partially governed by attending to the payoffs of the choice options, one would expect empirical linkages between BART scores and choice behaviors in the lottery task.

For the sake of completeness, two self-report trait measures of risk propensity were also included in the current study and calibrated to the same lottery choices. The measures were the Risk Propensity Scale (RPS; Meertens & Lion, 2008) and the Domain-Specific Risk-Taking Scale (DOSPERT; Blais & Weber, 2006) (see below for details of these measures). This provided the opportunity to further demonstrate the calibration of self-report measures to relevant behavioral fixed points for the distinct construct of risk-taking.

## 3.3.3 Method

### 3.3.3.1 Participants and Design

Ninety nine (99) individuals from the University of Western Ontario campus participated in the current study (58 males, 39 females, 2 non-specified; mean age = 24.46, *SD* = 5.48, range = 17 to 46). Participants were compensated $5 (CDN) in addition to the money earned in the balloon task (mean BART earnings = $7.37, *SD* = 2.67). Two a priori randomly selected participants also received the money associated with one of their lottery risk choices (both lucky participants chose the gamble; one of them won the $10 whereas the other lost). No restrictions were imposed on participant sex, age, or ethnicity. No experimental conditions were examined, hence all participants completed all measures and tasks in the same order (see below for details).

#### 3.3.3.2 Procedure and Materials

Participants were run in groups of two to four in a lab testing room where each participant was seated in a separate cubicle in front of a PC computer. Before starting, the experimenter verbally instructed each participant to carefully read all instructions and to ask the experimenter for clarification if any questions arose. The experimenter also instructed each participant to turn off any and all electronic devices before starting and to put on the headphones for the first two tasks. Participants then followed on-screen instructions and completed each measure in a serial fashion in the following order: BART, game of dice (GDT) task, CCT, lottery risk choices, affect misattribution procedure (AMP), RPS, DOSPERT, a volunteering questionnaire, and then demographics and debriefing questions.[15]

##### 3.3.3.2.1 Balloon Analogue Risk Task

The balloon task was designed specifically to provide a diagnostic context to observe actual risky behavior (Lejuez et al., 2002). The task involved 30 consecutive trials of inflating balloons by clicking a button labeled “Pump up the balloon”. Each pump earned participants exactly 1¢, which accrued in a temporary bank. Participants decided how many pumps to inflate each balloon before collecting their accumulated earnings for that trial by clicking a button labeled “Collect $$$”. If the balloon was inflated past its explosion point, all earnings in the temporary bank were lost and the next trial started. The balloon number, the current number of pumps, total winnings, and potential earning for that trial were all displayed on the right-hand side of the screen. The task was implemented and run using Inquisit 3.0 and featured a real-life picture of a red balloon that inflated slightly with each pump (about 0.125 in. [0.3 cm] in all directions). To add realism to the task, the computer generated the sound of a real balloon inflating for each pump and also produced a balloon popping sound in the event of an explosion. Participants were instructed that the explosion point for each balloon would be different.

[15] The AMP and volunteering questionnaire were assessed for two unrelated investigations. The GDT was assessed for the current study but was not correlated with any of the behavioral referents and hence will not be discussed further.

Also, following recent instruction improvements of the BART (Pleskac et al., 2008), participants were explicitly informed that each balloon explosion point could range anywhere from the first pump to a maximum of 128 pumps. The actual explosion point for each balloon was determined randomly by the computer, by choosing a random number between 1 and 128 without replacement from an array (the value 1 indicating an explosion). Hence, the probability that a balloon would explode on the first trial was 1/128, second trial 1/127, and so on up until the 128th pump at which the probability of an explosion was 1/1 (i.e., 100%). According to these parameters, the average explosion point in the long run would be 64 pumps (i.e., if one were to pump a large number of balloons 64 times each, one would expect about half to explode in the long run). As previously mentioned, the sequential nature of this task models real-world situations whereby excessive risk often results in diminishing returns (e.g., pumping the balloon on the 3rd trial would only risk losing 2¢ and would possibly increase the total earnings by 50%; after the 60th pump, however, a subsequent pump would risk 60 cents but possibly increase total earnings by only 1.6%). After reading two instruction screens, participants pressed a button to begin the task. At the conclusion of the task, participants were paid the amount earned in the task (rounded up to the nearest 25¢).

#### 3.3.3.2.2 Columbia Card Task

The goal of the "hot" version of the Columbia Card Task is to accumulate as many points as possible by sequentially turning over as many cards as possible (presented face down in a 4 × 8 array on the computer screen). Participants are informed that the cards can be either gain (smiling face) or loss (frowning face) cards. If a gain card is turned over, participants earn a specified gain amount for that trial and are able to continue the trial. If a loss card is turned over, participants lose a specified loss amount for that trial from their total points earned up to that point, and the trial ends. Trials vary in terms of the following parameters: number of loss cards (1, 2, or 3 loss cards out of the 32 cards total), gain amount (10, 20, or 30 point per gain card), and loss amount (-250, -500, or -750 points from total points earned up to that point). This information is presented at the top of the screen. These three game parameters are varied using a full factorial within-subject design, presenting each of the 27 parameter combinations twice resulting in 54 trials.

However, to maximize the assessment of voluntary stopping, the game is rigged such that loss cards are always the last possible card turned over. To conceal this, nine additional trials are randomly interspersed amount the 54 experimental trials whereby the probability of turning a loss card anywhere in the array is very high. Hence, participants completed a total of 63 trials. To turn over cards, participants were instructed to simply click on the card. To stop turning cards and end the trial, participants were instructed to click a button on the bottom of the screen labeled “STOP”. The CCT was also implemented and run using Inquisit 3.0.

### 3.3.3.2.3 Lottery Risk Choices

Participants subsequently completed five lottery risk choices involving the choice between two options (adapted from Hsee & Weber, 1999, based on the classic lottery tasks from Tversky & Kahneman, 1981). Instructions informed participants that they would have to indicate which of the two (option A or B) lottery options they preferred. Following Tversky and Kahneman (1981), it was explicitly mentioned (and emphasized on two different screens) that two of the 100 participants would actually receive the money associated with their preferred option for one of the lottery questions, and hence, that they should make their choices as if they were actually playing these lotteries. As displayed in Table 4, the lottery choices were as follows:

**Table 4: Lottery options format in lottery risk task.**

| Lottery | Option A | Option B |
|---|---|---|
| 1 | $6 for certain | Flip a coin. Receive $10 if heads, receive $0 if tails. |
| 2 | $2 for certain | Flip a coin. Receive $10 if heads, receive $0 if tails. |
| 3 | $8 for certain | Flip a coin. Receive $10 if heads, receive $0 if tails. |
| 4 | $5 for certain | Flip a coin. Receive $10 if heads, receive $0 if tails. |
| 5 | $4 for certain | Flip a coin. Receive $10 if heads, receive $0 if tails. |

It was explained to participants that if option B was selected, the experimenter would actually flip a coin, and the participant would receive the dollar amount associated with the coin flip outcome. Choice behaviors in this task were interpreted such that gambles on questions involving larger sure bets reflected incrementally higher reference points for higher levels of risk-taking (i.e., choosing the 50% gamble to win $10 over $8 for certain represents a higher level of risk-taking then choosing the 50% gamble to win $10 over $6, and so on).

### 3.3.3.2.4 Risk Propensity Scale

Participants completed the 7-item risk propensity scale (Meertens & Lion, 2008), which is a self-report measure that attempts to capture general risk-taking tendencies. The items are: “Safety first”, “I do not take risks with my health”, “I prefer to avoid risks”, “I take risks regularly”, “I really dislike not knowing what is going to happen”, “I usually view risks as a challenge”, and “I view myself as a …” Participants were instructed to indicate the extent to which they agreed or disagreed with the statements, and following standard procedure, were asked not to think too long before answering each question (scale anchors 1 = *totally disagree* and 9 = *totally agree* except for the last item where 1 = *risk avoider* and 9 = *risk seeker*). The first, second, third and fifth items were reverse-scored. Responses were averaged for each participant, with higher scores reflecting higher levels of risk-propensity (α = .78).

### 3.3.3.2.5 Domain-Specific Risk-Taking Scale

Participants completed the 30-item version (Blais & Weber, 2006) of the domain-specific risk-taking scale for adult populations (Weber et al., 2002). The measure attempts to capture risk-taking tendencies across five distinct and commonly encountered content domains including ethical, health/safety, social, recreation, and financial (further decomposed into gambling and investment domains). Example items include: “Going camping in the wilderness” (recreation), “Drinking heavily at a social function” (health/safety domain), and “Having an affair with a married man/woman” (ethical; see Appendix F for all items). Participants were informed to indicate the likelihood that they would engage in the described activity or behavior if they were to find themselves in that situation (using scale anchors 1 = *Extremely Unlikely,* 2 = *Moderately Unlikely,* 3 = *Somewhat Unlikely*, 4 = *Not Sure*, 5 = *Somewhat Likely,* 6 = *Moderately Likely*, and 7 = *Extremely Likely*). Responses were averaged, with higher scores reflecting higher levels of risk-taking (α = .82).

## 3.3.4 Results

### 3.3.4.1 Preliminary Data Treatment

Data from all measures were first screened informally for any evidence of non-compliance. I examined the time taken (as recorded by MediaLab) on instruction screens and actual questionnaire items for any evidence of consistently short latencies suggestive of non-compliance. One case was identified for the DOSPERT questionnaire (latencies < 350 milliseconds for the last 4 items) and so this participant was excluded in analyses involving this measure. In addition, in the debriefing, 10 participants indicated unambiguous suspicion regarding the rigged nature of the CCT and so were excluded in analyses involving that measure.

### 3.3.4.2 Main Analyses

Table 5 presents descriptive statistics and zero-order correlations for all of the main variables in Study 3. The metric mapping results will be presented in the following order: BART scores, CCT scores, RPS scores, and DOSPERT scores, all calibrated to the risky lottery choices (DOSPERT scores will also be linked to behavior in the BART).

**Table 5: Descriptive statistics and correlations for variables in Study 3 (*N* = 99).**

| Variable | Minimum | Maximum | Mean | *SD* | BART | CCT | RPS | DOSPERT | RLC $2 | RLC $4 | RLC $5 | RLC $6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BART (mean # of pumps on non-explosion trials) | 2.24 | 88.33 | 39.0 | 17.95 | | | | | | | | |
| CCT (mean # of cards turned over; 0-30) | 2.29 | 28.70 | 22.86 | 5.24 | .17 | | | | | | | |
| RPS (1-9) | 1.71 | 8.29 | 4.41 | 1.36 | .16 | -.01 | | | | | | |
| DOSPERT (1-7) | 2.27 | 5.80 | 3.65 | .73 | .21* | -.01 | .65* | | | | | |
| Risky lottery choices (RLC) | | | | | | | | | | | | |
| $10 gamble vs. $2 for certain | 0 | 1 | - | - | .15 | .13 | -.08 | .23* | | | | |
| $10 gamble vs. $4 for certain | 0 | 1 | - | - | .23* | .20† | .22* | .36* | .31* | | | |
| $10 gamble vs. $5 for certain | 0 | 1 | - | - | -.07 | .18 | .02 | .04 | .26* | .33* | | |
| $10 gamble vs. $6 for certain | 0 | 1 | - | - | .11 | .29* | .24* | .20* | .10 | .03 | .25* | |
| $10 gamble vs. $8 for certain | 0 | 1 | - | - | -.24* | .10 | .02 | -.01 | -.11 | -.04 | .01 | -.03 |

To determine which choice behaviors in the lottery task to use as behavioral reference points, in a first step, I examined the point bi-serial correlations between the different measures and the five choice behaviors (see Table 5 for all correlations). Overall, results revealed generally positive correlations (small to moderate in size) between the scores of the different measures and the five binary choices (with some correlations, however, negative in sign). Consideration of the frequencies of the particular choices for the different lotteries revealed, however, that responses for the $2 and $8 lotteries were highly polarized compared to the other lotteries, hence potentially explaining the unexpected negative correlations (e.g., for the $8 lottery, 94% of participants selected the safe bet). Given that correlations were the most consistent for the $4 lottery, I decided to use this behavior as the main behavioral reference point for the current study and sometimes used choices on the $6 lottery as a secondary reference point.

### 3.3.4.2.1 Balloon Analogue Risk Task

The empirical linkages between BART scores and probability of choosing the risky gamble in the $4 lottery was modeled using a logistic regression, with BART scores as the predictor and behavioral choice as the dichotomous outcome. Results from the logistic regression revealed a statistically significant positive predictive relation between BART scores and behavioral choice in the $4 lottery (Wald's $\chi^2$ = 4.85, *B* = .03, *odds ratio (OR)* = 1.03, *p* = .03). This indicates that for every unit increase in BART score, the odds of choosing the risky gamble over the safe bet of $4 increases by 3% (i.e., OR = 1.03; in other words, an increase of 10-units in BART scores is associated with a 30% odds increase of choosing the gamble, OR = 1.29). This can be seen visually in Figure 16, which plots the predicted probability of choosing the $10 gamble over the $4 sure bet (calculated using the coefficient and intercept values of the best fitting exponential regression line) for each particular BART score obtained in the sample. That is, BART scores of “10”, “30”, “50”, “70”, and “90” approximately corresponded to a .37, .50, .64, .76, and .84 probabilities, respectively, of choosing the gamble over the $4 safe bet. These empirical mappings imbue the metric of BART scores with meaning in a general sense. More specific meaning can be gleaned by focusing on the particular BART score that corresponds to a .50 probability of choosing the risky gamble over the safe bet (i.e., a

BART score of approximately “29”). To the extent that this behavior is interpreted as a qualitatively distinct risky behavior (i.e., consensus among experts), then BART scores gain meaning with respect to this threshold value of “29” (Blanton & Jaccard, 2006b).

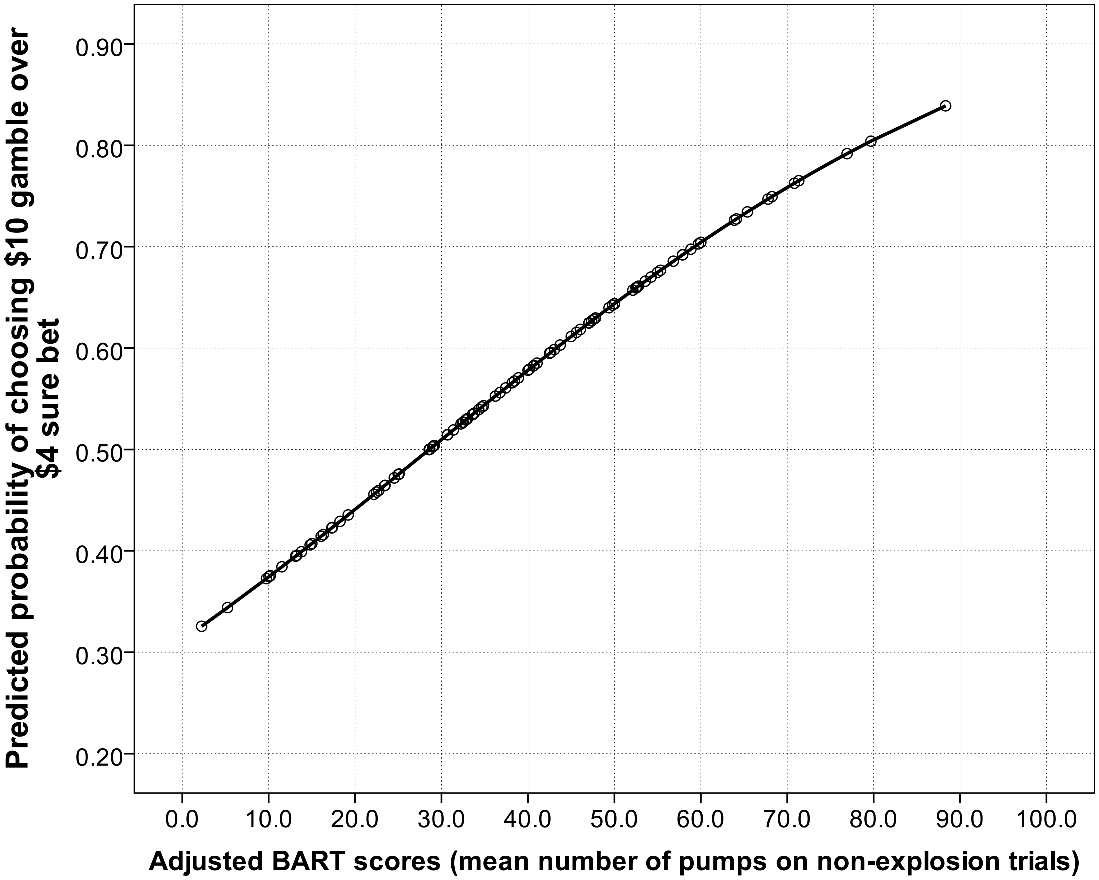


**Figure 16: Predicted probabilities of choosing $10 gamble over $4 safe bet plotted against adjusted BART scores.**

### 3.3.4.2.2 Columbia Card Task

The empirical linkages between CCT scores and probability of choosing the risky gamble in the $4 lottery were similarly modeled using a logistic regression. Results of this analysis revealed a positive predictive relation between CCT scores and probability of choosing the risky gamble, though the *p*-value only reached marginal statistical significance (Wald’s $\chi^2 = 3.24$, $B = .08$, *odds ratio (OR)* = 1.08, $p = .07$). The particular functional form of this mapping can be visualized in Figure 17 (solid line). Of note, an approximate CCT score of “20” was associated with a 50/50 chance of choosing the $10 gamble over the safe $4 option.

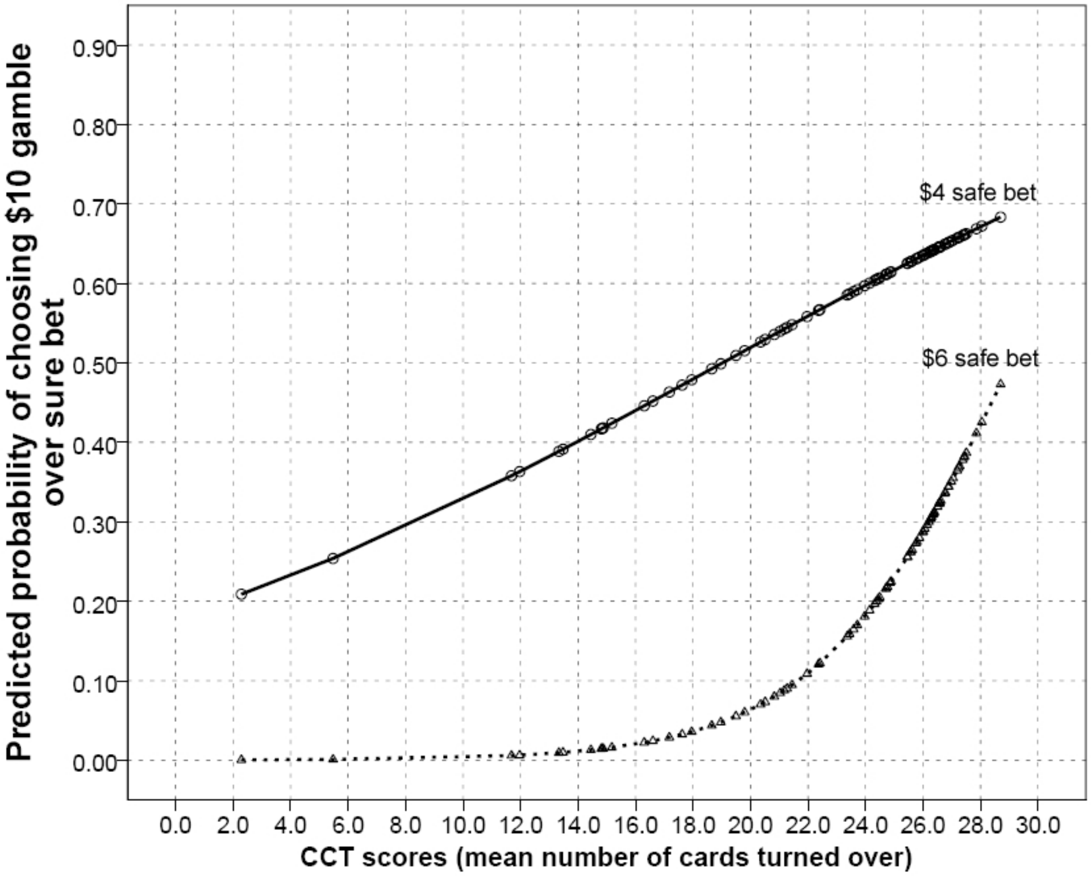


**Figure 17: Predicted probabilities of choosing $10 gamble over $4 safe bet (solid line) or over $6 safe bet (dotted line) given CCT scores.**

Hence, those who turned an average of 20 more cards in the task were statistically more likely to choose the gamble than the safe bet whereas those who turned less than an average of 20 cards were statistically more likely to choose the safe bet rather than the gamble. Further insights into the meaning of the metric of CCT scores can be achieved by examining how CCT scores map onto the act of choosing the $10 gamble over the $6 safe bet. Choosing the gamble on this lottery clearly involves more risk than the previous lottery (i.e., one can lose $6 vs. $4); hence, theoretically, one would expect a higher threshold value for CCT scores that maps onto the probability of choosing this gamble. Indeed, a logistic regression analysis revealed a positive predictive relation between CCT scores and probability of choosing the \$10 gamble over the \$6 safe bet (Wald's $\chi^2$ = 5.78, *B* = .30, *odds ratio (OR)* = 1.35, *p* = .02), such that the 50/50 gamble point mapped onto an approximate CCT score of "29" (see Figure 17, dotted line). In other words, scoring virtually the highest score in this card task ("30" is the maximal behavioral score on this instrument) is associated with "only" a 50/50 chance of choosing the \$10 gamble over the \$6 safe bet. Hence, this metric mapping result vividly illustrates the power of the metric approach to shed light on the possible meaning of metrics which were previously void in meaning in a non-relative sense. What's more, this metric mapping provides information that could have important implications for the interpretation of experimental studies using CCT scores as the DV. That is, a mean difference (of a certain effect size) at the upper range of the CCT metric could have very different interpretations than a mean difference at the mid range (more on this in the Discussion section).

#### 3.3.4.2.3 Risk Propensity Scale

Turning to the self-report measures, a logistic regression was used to determine the mapping between scores from the risk-propensity measure and act of choosing the \$10 gamble over the \$4 safe option. The analysis revealed a meaningful positive relation between RPS scores and probability of choosing the risky \$10 gamble over the \$4 sure shot (Wald's $\chi^2$ = 4.57, *B* = .35, *OR* = 1.42, *p* = .03). A similar analysis also revealed a meaningful positive relation between RPS scores and the \$6 lottery choice (Wald's $\chi^2$ = 5.02, *B* = .45, *OR* = 1.56, *p* = .03). These mapping results provide preliminary empirical evidence about the meaning of the metric of the RPS scale. For instance, with regard to the \$6 lottery, the metric mapping helps us gauge the meaning of RPS scores in a non-relative sense by showing that an almost maximal score on the RPS (i.e., "8.5" out of a maximum of "9") corresponded to no more than a slightly higher than 50/50 chance of choosing a \$10 gamble over a \$6 safe bet.

#### 3.3.4.2.4 Domain-Specific Risk-Taking Scale

Given the multi-faceted nature of the Domain-Specific Risk-Taking Scale, a metric analysis was performed on the most theoretically specific and relevant subscale of the DOSPERT, that is, risk-taking in the financial domain. A logistic regression analysis revealed a meaningful positive relation between DOSPERT financial scores and probability of choosing the risky \$10 gamble over the safe \$4 option (Wald's $\chi^2$ = 9.42, *B* = .69, *OR* = 1.99, *p* = .002) and also a meaningful positive relation between the financial facet scores and probability of choosing the risky gamble over the \$6 safe bet (Wald's $\chi^2$ = 4.93, *B* = .47, *OR* = 1.60, *p* = .03).[16] This analysis revealed, for example, that a maximal score on the financial DOSPERT (i.e., "7" out of "7") corresponded to a slightly higher than 50/50 chance (predicted probability = .58) of choosing a \$10 gamble over a \$6 safe bet.

For further illustrative purposes, I executed a final metric analysis calibrating DOSPERT scores onto mean number of pumps (on non-exploding trials) in the BART. A meaningful metric mapping emerged between total DOSPERT scores and BART scores (*r* = .21, *p* = .05), with a slightly stronger mapping between DOSPERT scores in the recreational domain and BART scores (*r* = .27, *p* = .01). A regression analysis specified the particular functional form of this relation, whereby a 1-unit increase on the DOSPERT recreational scale (metric range = 1 to 7) corresponded to an increase of about 3 balloon pumps averaged across trials in the BART (unstandardized regression coefficient: *B* = 3.29, *p* = .01). Furthermore, a maximal DOSPERT recreational score of "7" corresponded to inflating the balloons in the BART an average of 48 times.

### 3.3.5 Discussion

The primary goal of Study 3 was to demonstrate the calibration process for the metric of two behavioral measures of risk-taking presumed to involve a state-like component. Overall, the study generally achieved this goal, showing promising results in illustrating the feasibility of calibrating behavioral measures of risk-taking to meaningful behavioral fixed points, as to reduce the metric arbitrariness of the behavioral measures. In summary, I found meaningful metric mappings between BART and CCT scores to the

[16] Surprisingly, very similar patterns of results were also found for general DOSPERT scores, computed across all domains (i.e., point bi-serial correlations between total DOSPERT scores and the \$4 and \$6 lottery choices of r = .36, p = .001 and r = .20, p = .05, respectively, compared to r = .33, p = .001 and r = .24, p = .02, respectively for the financial DOSPERT scores).

probability of choosing risky gambles in the lottery choices. For instance, the observed mapping between an almost maximal score on the card task and a 50/50 chance of choosing the $10 gamble over the $6 safe choice, demonstrated the ability of the metric calibration approach to shed light on metric meaning. In addition, Study 3 successfully demonstrated the calibration of the metric of two self-report measures of risk-taking. The RPS, a general risk-propensity scale and the DOSPERT, a domain-specific risk-taking measure, both yielded meaningful calibration results to the same lottery choice behavioral reference points.

Study 3 findings have at least two important implications regarding the potential utility of systematically calibrating the metrics of measures used in experimental studies. First, the metric findings in Study 3 demonstrate the potential utility of non-arbitrary metrics to allow for the extraction of more information from data patterns. In particular, the current findings illustrate how using measures with calibrated metrics can enhance the interpretation of experimental mean differences that emerge at different locations on the scale of the DV measure. Second, the findings from Study 3 speak to the issue that non-arbitrary metrics may help us overcome some of the limitations of NHST. The basic idea is that measures with a calibrated non-arbitrary metric could potentially help us gauge the theoretical "noteworthiness" of an experimental effect, above and beyond the statistical significance and effect size indices, by interpreting the effect with respect to relevant behavioral reference points. In the Implications section of the General Discussion, I will elaborate on concrete examples of these two benefits by applying my metric calibration findings to actual experimental effects from the literature.

A final point worth mentioning with regard to implications of Study 3 findings is that valuable information can also be gleaned about the metric meaning of BART and CCT scores by using the same $4 lottery gamble choice as a common reference point for both measures. For instance, using a .50 probability of choosing the risky gamble as a common reference point, one can infer that a CCT score of about "20" is approximately equivalent to a BART score of about "30" with respect to the underlying dimension of risk-taking (see Figure 16 and Figure 17 [solid line]). This would also further suggest that a CCT score of about "20" likely reflects a higher level of risk-taking than a BART score of about “20.” Future research should investigate the psychometric and scientific value of this type of approach of calibrating the metric of different measures of the same construct to a common reference point.[17]

## 3.4 Other Analyses

To bolster the empirical substance of the current dissertation, I sought other empirical datasets that could further demonstrate the feasibility of applying the proposed metric calibration approach to constructs commonly examined in basic psychology. Toward this end, I searched the literature for published datasets that contained the necessary components to allow for a metric calibration analysis (i.e., independent assessment of test scores and relevant behavior) and e-mailed authors requesting their data. I was able to acquire one such dataset from Tangney, Baumeister, and Boone (2004) and another from Hong and Paunonen (2009). I present the results of my re-analyses of these datasets in turn.

### 3.4.1 Trait Self-control

Tangney et al. (2004) investigated the benefits of self-control by examining the psychological correlates of a new individual difference measure of the trait. In a sample of 157 undergraduates, they found that higher scores on the new self-control measure correlated with better adjustment, less binge eating and alcohol abuse, better relationship and interpersonal skills, and higher grade point average (GPA). From a metrics perspective, this study is a good candidate because GPA can be seen as a behavioral reference point which has a fairly intuitive meaning to most psychologists. Hence, scores from the self-control measure can be calibrated to GPA as a way to increase the meaning

[17] This approach can be contrasted to traditional approaches to test equating where, for instance, scores from two measures of the same construct are equated using a simple linear regression prediction equation. Given the different logic of the two approaches, as expected, applying this approach to the scores of my two behavioral measures of risk-taking resulted in different equivalence mappings between the two measures (e.g., BART score of “50” equivalent to a CCT score of about “24” using a traditional test equation approach whereas the same BART score of “50” was equivalent to a CCT score of about “30” using my common reference point approach). Importantly, such kind of traditional test equating cannot be used to develop non-arbitrary metrics given scores are not linked to reference points external to the tests.

of the metric of the self-control measure.[18] In their study, Tangney et al. (Study 1) had undergraduate students complete the self-control measure in addition to a host of other theoretically-relevant measures. Example items from their self-control scale were "I am good at resisting temptation", "I often interrupt people"(R), and "I sometimes drink or use drugs to excess" (R), using a 5-point Likert scale (1 = *Not at all* and 5 = *Very much*). They found a correlation of $r = .39$, $p = .001$ between self-control scores and GPA.[19]

To probe the metric mapping between self-control scores and GPA, a regression analysis was performed and revealed an unstandardized regression coefficient of B = .40 ($\beta = r = .39$, $p = .001$). As can be seen in Figure 18, a 1-unit increase in self-control scores corresponded to almost a .50 increase in GPA. For instance, a self-control score of "3" (scale midpoint) corresponded to a GPA of about 2.9, a "4" to a GPA of about 3.3, and a "5" to a GPA of about 3.7. What is potentially most illuminating about this metric mapping is that according to the regression equation, a 4.0 GPA would correspond to a hypothetical self-control score of "6." Linking the self-control scores to external criteria such as GPA, which have real-world meaning, therefore helps make apparent the meaning and implications of a given self-control score and hence the metric becomes less arbitrary. Through such linkages between self-control scores and GPA, we start getting a rough sense of the approximate location of the scores of the measure on the underlying dimension of self-control (Blanton & Jaccard, 2006b).

[18] Although this strategy deviates from the main metric calibration strategies elaborated in the introduction, such that scores from the self-control measure are linked to behavioral expressions of a different rather than the same construct, this different approach is nonetheless consistent with theorizing by Blanton and Jaccard (2006b), who stated that metric calibration can also be achieved by linking scores to theoretically-relevant behaviors "so that one can better appreciate the real-world implications of obtaining one test score versus another" (p. 63; see also Sechrest et al., 1996).

[19] Unfortunately, GPA was self-reported in this study which is not ideal for a metric study. However, as mentioned by Tangney et al., the self-control GPA relation was virtually unchanged when controlling for social desirability using the Marlowe-Crowne Social Desirability Scale (Crowne &Marlowe, 1960), $r_{\text{partial}} = .32$.

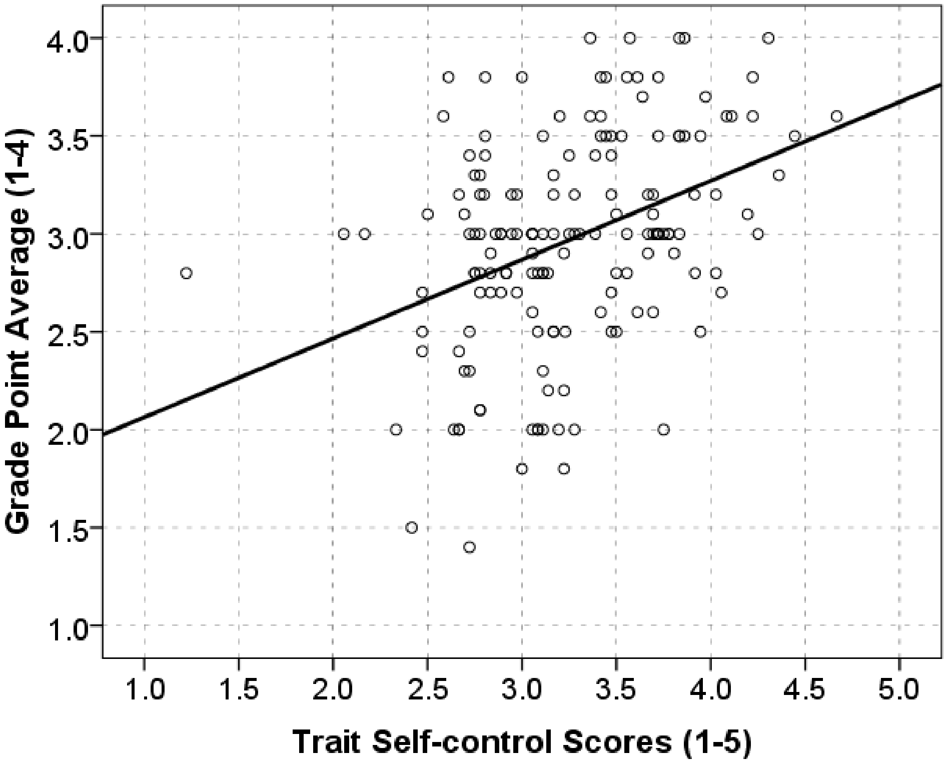


**Figure 18: Grade point average plotted against Tangney et al.'s (2004) trait self-control scores.**

### 3.4.2 Extraversion and Conscientiousness

Hong and Paunonen (2009) investigated and found reliable associations between various personality facets and health-risk behaviors, which provides the necessary components to further illustrate metric calibration for the constructs of extraversion and conscientiousness. In a sample of 124 undergraduate students, Hong and Paunonen had participants complete the NEO-PI-R (Costa & McCrae, 1992) and a behavior report form (BRF; Paunonen, 2003) that included various behaviors theoretically-relevant to different personality facets. For my purposes, I focused on behaviors that seemed the most interpretationally meaningful with respect to the available personality facets. Consequently, I examined the behavior of attending social parties as a behavioral reference point for the extraversion facet of Gregariousness and I also examined speeding behavior as a reference point for the conscientiousness facet of Dutifulness. The personality items were answered using a 5-point *strongly disagree-strongly agree* scale

(Costa & McCrae, 1992) and each facet score was a sum score based on 8 items. The behavior of attending social parties was assessed by the question "Estimate the average number of *parties per month* that you attend." For participant who had a driver's license, speeding behavior was assessed by the question "What is the fastest you have driven?" measured in kilometers per hour.

Regression analyses revealed illuminating metric calibration patterns for both behaviors. For the extraversion facet, the regression analysis revealed a robust positive relation between Gregariousness facet scores and number of social parties attended per month, (B = 2.01, *β* = *r* = .39, *p* = .0001). As is visually depicted in Figure 19 (panel A), a 1-unit increase in Gregariousness scores corresponded to attending 2 more parties per month. For instance, a Gregariousness score of "3" (scale midpoint) corresponded to attending approximately 3 social parties whereas a Gregariousness maximal score of "5" corresponded to attending about 7 social parties per month.

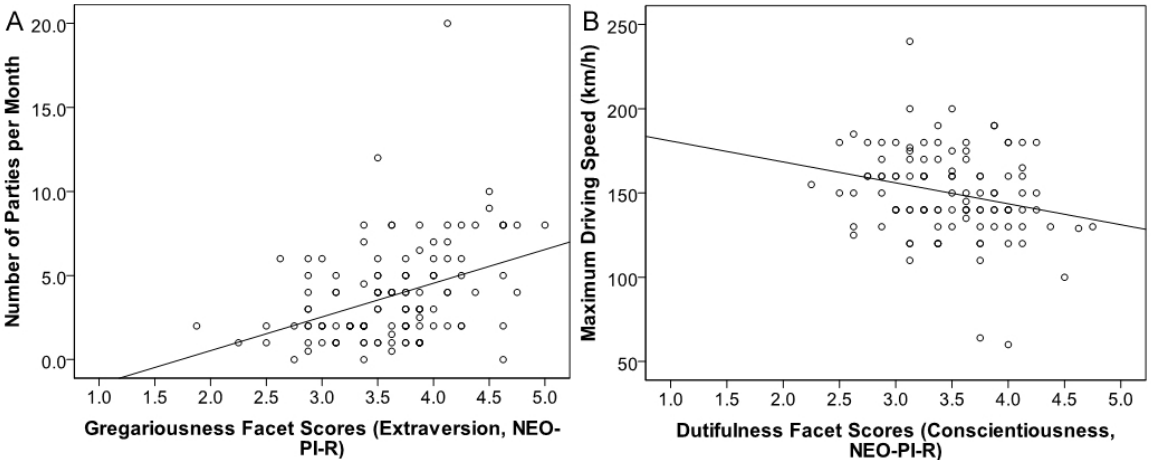


**Figure 19: Number of parties per month given gregariousness facet scores (panel A) and maximum driving speed given dutifulness facet scores (panel B).**

For the conscientiousness facet, the regression analysis revealed a reliable negative relation between Dutifulness facet scores and speeding behavior (B = -12.4, *β* = *r* = -.24, *p* = .01). As shown in Figure 19 (panel B), a 1-unit increase in Dutifulness scores corresponded to a maximum driving speed that was about 12 km/h slower. For example, a Dutifulness score of "3" (scale midpoint) corresponded to a maximum driving speed of about 156 km/h whereas a Dutifulness maximal score of “5” corresponded to a maximum driving speed of about 131 km/h. These metric linkages between lower-order personality facet scores and external theoretically-relevant behaviors hence provide preliminary information about the meaning of the metric of these personality inventories. To the extent that personality theorists can agree on where certain behaviors locate an individual on a relevant underlying personality dimension, then linking personality scores to those behaviors can provide information about the meaning and implication of particular personality scores.

Taken together, the metric mappings presented in these additional analyses provide more empirical substance to strengthen my contention that the metric calibration approach is feasible when applied to constructs commonly studied in basic psychological research. Furthermore, these additional analyses will also provide more empirical examples to use as illustrations to further demonstrate some of the proposed benefits and utility of the metric calibration approach, which I will discuss further in the General Discussion.

# Chapter 4

# 4 General Discussion

In a world where the metrics of our instruments are meaningful, psychological research could be done in importantly different ways, ranging from the way data are analyzed and interpreted, to how psychological theories are tested, to how psychological findings are catalogued. The overarching goal of the current dissertation is to make the case that it is both *useful* and *feasible* to calibrate the metric of psychological instruments commonly used in basic research, as to render their metrics non-arbitrary. In this section, I will first review and summarize the findings from my empirical demonstrations that speak to the feasibility of the metric calibration approach and then elaborate on the broader implications of the metric approach with respect to the usefulness of non-arbitrary metrics by reviewing several potential benefits they may one day provide.

## 4.1 Feasibility

Across seven distinct constructs assessed in five different samples (including two samples graciously provided by other researchers), I demonstrated that it is empirically possible to reduce the metric arbitrariness of instruments commonly used in basic research. In these metric calibration studies, I illustrated how to apply a metric calibration approach to a variety of psychological instruments, whether self-report or behavioral, whether for predominantly trait-like (e.g., conscientiousness) or state-like (e.g., risk-taking) constructs, and whether the construct is commonly studied in social psychology (e.g., self-enhancement), cognitive psychology (e.g., risk-taking), or personality psychology (e.g., extraversion). In summary, Study 1 showed a meaningful metric calibration result for the metric of the instrument most commonly used to assess need for cognition, an important individual difference variable in the research area of attitudes and persuasion (Cacioppo, Petty, Feinstein, & Jarvis, 1996). Scores from Cacioppo et al.'s (1984) need for cognition measure were calibrated to the probability of choosing to complete a cognitively effortful over a cognitively simpler task. Study 1 also found an interesting metric mapping between scores of a practically useful self-report measure of task persistence (Steinberg et al., 2007) and actual persistence in an anagram persistence task, whereby a non-linear cubic function explained three times more variance than a linear metric mapping (which incidentally was replicated in another relevant dataset shared by Ditre and Brandon, 2008). The “dipping shape” in the metric calibration relation (see Figure 11), suggested that individuals indicating the maximal score on the self-report measure may be over-reporting their typical task persistence (how the metric approach can help in detecting measurement problems such as these will be elaborated below).

Set in the context of the pan-cultural debate of self-enhancement (Sedikides et al., 2003), Study 2 found theoretically interesting metric linkages (linear and cubic, see Figure 13) between a trait rating measure of self-enhancement (which figures prominently in the debate) and a specifically configured behavioral measure of self-enhancement. More specifically, the metric mappings showed that trait rating scores above the scale midpoint (typically interpreted as self-enhancement) corresponded to very little evidence of actual self-enhancement behavior as assessed by over-claiming of knowledge in the OCT (see Figure 13). This finding suggests that researchers should not interpret trait scores above the scale midpoint as evidence for self-enhancement. Rather, metric calibration research is required for making these kinds of claims whereby trait self-enhancement scores are empirically connected to consensually agreed upon behaviors argued to reflect self-enhancement.

Study 3 extended the metric calibration approach to commonly used behavioral measures of risk-taking and found meaningful metric mappings to risky gambles in binary lottery choices involving the possibility of winning real money. For instance, the observed mapping between an almost maximal score on the Columbia card task measure and a 50/50 chance of choosing the $10 gamble over the $6 safe choice was illuminating in demonstrating how the metric approach can imbue meaning into scores and hence reduce metric arbitrariness (see Figure 17). In addition, Study 3 successfully demonstrated the calibration of the metric of two self-report measures of risk-taking (Risk Propensity Scale and DOSPERT scale) to the same risky gamble choices.

Finally, the feasibility of the metric approach was further demonstrated by re-analyzing relevant datasets from samples shared by other researchers. Meaningful metric mappings were found for instruments assessing extraversion, conscientiousness, and self-control. Gregariousness facet scores were linked to number of social parties attended per month, Dutifulness facet scores (conscientiousness) were connected to maximum driving speed, and trait self-control scores were calibrated to GPA.

Taken together, these empirical demonstrations across several constructs and samples provide concrete evidence that it is possible to apply the metric calibration approach (and hence increase metric meaning) to constructs commonly studied in basic psychological research. Though many challenges exist in the calibration process of psychological instruments, the empirical illustrations reported herein should nonetheless reveal to researchers that the metric calibration approach espoused in this dissertation is possible. The next big question, then, is whether metric calibration is worth it? That is, what concrete benefits do we gain from metric calibration and non-arbitrary metrics? I turn to this next.

## 4.2 Implications

In this section, I review the implications of the metric calibration approach with respect to utility following the structure and order used in the Introduction (see Table 1 for a summary of the proposed benefits). Consequently, I will briefly elaborate on the potential usefulness for each of the proposed benefits and further support my claims by drawing on some of my empirical demonstrations or by providing additional re-analyses of yet other shared datasets from the literature.

### 4.2.1 Help in the Interpretation of Data

In my first category of proposed benefits, I argue that non-arbitrary metrics would facilitate the process of interpreting data in three main respects.

#### 4.2.1.1 Enhance Interpretability of Statistical Effects

First, I contend that calibrated metrics would help in the interpretation of data by enhancing the interpretability of statistical effects for statistical procedures commonly used in basic research. I will focus my attention on moderated multiple regression (MMR), which has become the statistical procedure of choice in basic research to probe interactions involving continuous predictors (rather than the sub-optimal approach of using median splits; MacCallum et al., 2002). To demonstrate my point, I will re-analyze a finding from the psychological literature involving need for cognition as a moderator using calibrated NFC values found in Study 1, rather than the values of +/- 1 SD above the sample specific mean used by convention. The psychological finding that I examined is from a study by O'Hara, Walter, and Christopher (2009), who, in the context of understanding the personality underpinnings of political behavior, found that NFC moderated the relation between conscientiousness and political behavior.[20] Figure 20 (panel A) shows the main conscientiousness × NFC interaction from their study plotted at +/- 1 SD (SD = .72) above the sample specific mean of "3.4" (predictors mean-centered and product term created; Aiken & West, 1991).

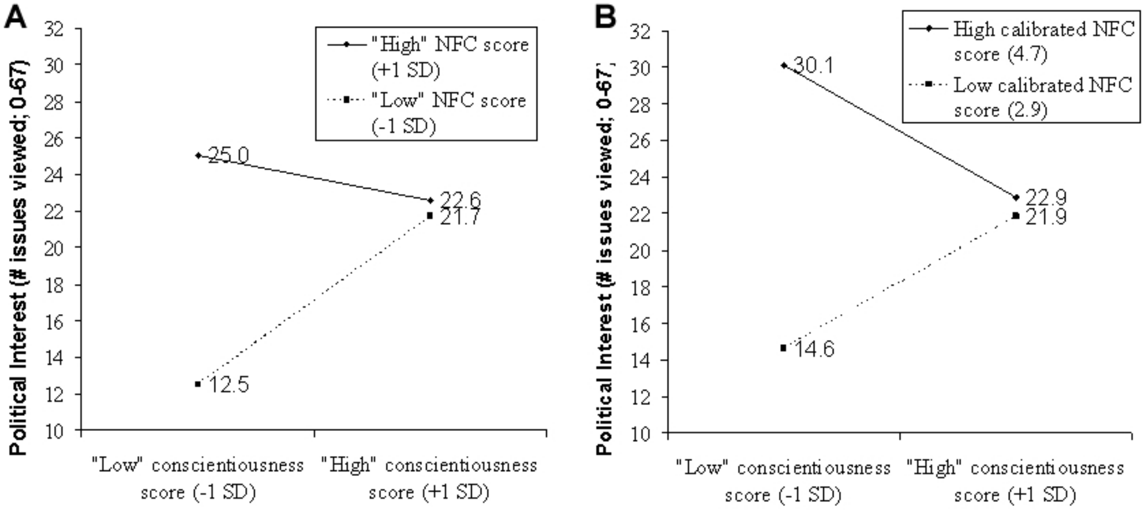


**Figure 20: Moderated multiple regression re-analysis of O'Hara et al. (2009) using conventional +/-1 SD (panel A) or calibrated values (panel B).**

[20] To find this dataset, I first combed the literature for relatively recent articles reporting data patterns involving NFC as a moderator of some psychological effect. I found about 20 such articles and e-mailed the corresponding authors requesting their datasets. Only two of such 20 requests resulted in the acquisition of the relevant datasets (typically no reply or datasets unavailable). I report re-analyses from the O'Hara et al. (2009) paper given that it best demonstrated the principles at hand.

As can be seen, for those scoring 1 SD below the sample specific NFC mean, individuals with higher conscientiousness scores exhibited higher levels of political interest, whereas those scoring 1 SD above the sample specific NFC mean exhibited relatively high levels of political interest regardless of their conscientiousness scores. With the calibrated NFC metric in hand from Study 1 (for illustrative purposes), however, the interpretation of this data pattern can be enhanced considerably. This can be achieved by analyzing the conscientiousness × NFC interaction at calibrated NFC values, which have gained meaning via empirical linkages to corresponding NFC behavior (i.e., completing cognitively effortful task). For illustrative purposes, I re-analyzed O'Hara et al.'s interaction pattern by centering the NFC scores around the NFC score associated with a 50/50 chance of choosing to complete the cognitively challenging task (mean = 3.8; see Figure 10). Then, to plot and test simple slopes, I analyzed the relation between conscientiousness and political interest at the calibrated NFC values associated with a 25% and 75% chance of completing the cognitively challenging task (i.e., NFC values of "2.9" and "4.7", respectively). The data pattern from this re-analysis is displayed in Figure 20 (panel B). This illustration renders three major things apparent. First, it should be evident that the interpretation of the MMR data pattern is enhanced given that the interaction is analyzed using NFC values that have been grounded to actual NFC behavior rather than arbitrary NFC values which have no meaning other than a relative interpretation. That is, one gets a better sense of what the interaction might mean psychologically because the relevant slopes can be interpreted with respect to the probabilities of exhibiting a relevant behavior. In this particular case, that is, the relation between conscientiousness and political interest is positive for the calibrated NFC value of "2.9", which corresponds to a 25% chance of exhibiting an NFC behavior, whereas the relation is negative for the calibrated NFC value of "4.7", which corresponds to a 75% of exhibiting the NFC behavior.

The interpretation of the MMR data pattern is also enhanced because the interaction analysis involving calibrated values, which are grounded to theoretically-relevant behavior, may reveal different patterns of results which could have different, but potentially important, theoretical implications. As is evident in Figure 20 (panel B), the slope between conscientiousness and political interest is markedly more negative at the calibrated high NFC value (i.e., “4.7”) compared to the +1 SD NFC value (i.e., “4.1”). Even though in this particular case, the different interaction pattern may not imply a drastically different theoretical implication, it is quite possible that it could in a more theoretically driven research situation. In research on implicit versus explicit attitudes, for instance, a negative slope (rather than a flat slope) between implicit and explicit attitudes is sometimes interpreted as over-correction of the implicit attitude on the explicit attitude measure (Fazio & Olson, 2003). Hence, if the use of calibrated NFC values in such MMR analyses consistently yielded negative implicit-explicit attitudes slopes at the calibrated high NFC value (whereas the +1 SD above the mean value did not yield such negative slopes), then this could have important theoretical implications regarding over-correction processes underlying attitude judgments.

A final way calibrated values could enhance data interpretation in MMR analyses is by overcoming sampling error issues inherent in the conventional MMR approach. The issue involves the fact that the +/-1 SD approach is based on sample-specific values of the mean and standard deviation. Thus, it is possible, due solely to sampling error, that an interaction analysis yields a different pattern of results from previous research, which a researcher incorrectly interprets in a theoretically substantive way (hence obfuscating the accumulation of knowledge). For instance, returning to O’Hara et al.’s (2009) interaction pattern (Figure 20, panel A), consider a follow-up extension study on the same topic, but this time the mean of the NFC scores is “4.0” rather than “3.4”. In this situation, a negative conscientiousness-political interest slope may be found and interpreted theoretically even though the result would have been due solely to the NFC-aberrant sample of individuals. If, on the other hand, it would be standard convention to use consensually-agreed upon calibrated NFC values to analyze these types of MMR analyses, this sampling error issue would be overcome, and hence the interpretation of data would be enhanced.[21]

[21] Strictly speaking, even a consensually-agreed upon convention of always using particular non-calibrated scale scores when executing such MMR analyses could also overcome the sampling error issue (though researchers would also need to agree to always use the same number of scale points). From a metric calibration perspective, however, it is clear that striving toward consensually agreed upon calibrated metric values is the most useful approach.

#### 4.2.1.2 Allow Extraction of More Information from Data Patterns

Another way non-arbitrary metrics could facilitate the interpretation of data is by allowing the extraction of more information from data patterns. As alluded to in the Discussion section for Study 3, the use of calibrated metrics in experimental studies could allow more fine-grained interpretations of experimental effects that emerge at different locations on the scale of the DV. That is, with non-arbitrary metrics, it becomes apparent that an experimental effect in different ranges of the DV metric implies something different psychologically and hence should be interpreted as such theoretically. For example, referring back to Figure 17 (dotted line), it can be seen that an experimental effect found at the upper range of the CCT metric (e.g., *M* = 29.0 vs. *M* = 26.0, *d* = .5) would mean something quite different psychologically than an experimental effect of the same size found at the mid range (e.g., *M* = 15 vs. *M* = 12, *d* = .5). This would be so because a mean difference at the upper range of the CCT metric is associated with a much larger difference in lottery choice behavior than the same mean difference in the middle range (same logic as in Blanton & Jaccard, 2006b).

To make the proposed benefit more concrete, the calibrated CCT metric can be applied to an actual (quasi-) experimental study of risk-taking in the literature, whereby Figner et al. (2009, Experiment 3) found that teenagers turned over statistically significantly more cards in the hot CCT than adults (*M* ~= 25 vs. *M* ~= 20, *d* = .65). If this study were replicated, however, and one found an experimental effect of the same magnitude, but in the middle or lower range of the CCT metric, it would be much more apparent (given the calibrated CCT metric) that this experimental effect implies something different psychologically and hence should be interpreted as such. Furthermore, even in the case where the metric mapping for a certain metric is linear (e.g., calibrated trait rating metric to OCT behavior, Study 2, Figure 13), it can be argued that using a calibrated metric makes it much more apparent that an experimental effect has occurred at a different location on the DV because one will naturally pay more attention to metric values if they have some kind of meaning (rather than if the metric is arbitrary, which tends to encourage just focusing on *p*-values and effect sizes). Hence, using calibrated non-arbitrary metrics could allow the extraction of more information from experimental data patterns and thus facilitate how data are interpreted and catalogued.

### 4.2.1.3 Help Overcome Limitations of NHST

A final way calibrated metrics could aid with data interpretation is by potentially overcoming some of the limitations of NHST. The basic idea is that measures with a calibrated metric could help us gauge the theoretical "noteworthiness" of an experimental effect, above and beyond the statistical significance and effect size indices, because the experimental effect could be interpreted with respect to the relevant calibrated behavior. To illustrate this concretely, I will apply my metric findings for the BART to an actual experimental effect from the literature involving the BART as DV. For instance, Benjamin and Robbins (2007) investigated the impact of framing effects on risk-taking in the BART and found that a loss frame led to higher scores in the BART compared to a gain frame ($M_{loss}$ = 48.8 vs. $M_{gain}$ = 42.3, $p < .05$, $d = .57$). The metric calibration results of BART scores in the present Study 3 help to add meaning to this experimental effect via its linkages to behavior in the risky lottery task (i.e., predicted probability of choosing the $10 gamble over the $4 safe bet). That is, Benjamin and Robbins' experimental effect in the particular range of the BART metric (i.e., BART score of "49" and "42") corresponds to probabilities of .64 and .60 of choosing the risky $10 gamble (over the $4 safe bet), respectively (see Figure 16). Hence, one can interpret the increase in risk-taking in the BART due to Benjamin and Robbins' particular framing manipulation as roughly equivalent to a 7% increase in the probability of choosing that particular risky gamble. As illustrated in this example, non-arbitrary metrics information can be seen as providing additional information to consider (over and above *p*-values, sample size, and effect sizes) when faced with the difficult task of deciding on the "noteworthiness" of an experimental result (Kirk, 1996). Hence, in this sense, using calibrated metrics could be seen as helping us with data interpretation in the context of the limitations of NHST.

### 4.2.2 Facilitate Construct Validity Research

In my second category of proposed benefits, I argue that the metric calibration approach could help with construct validity research in three main regards.

#### 4.2.2.1 Construct Illumination

Metric calibration could help construct validity research by shedding brighter (i.e., more illuminating) light on the construct at hand. Consistent with more nuanced conceptualizations of construct validity by past theorists (i.e., Cronbach & Meehl, 1955; Messick, 1989), the process of linking test scores to theoretically-relevant and meaningful behaviors can be seen as a more compelling form of evidence supporting the construct validity of a psychological instrument (Messick, 1995). This is so because the connection between test scores and theoretically-relevant behavior that results from metric calibration reveals more illuminating construct validity evidence given that test scores are linked directly to specifically-configured interpretable behaviors rather than just another theoretically-related measure. Furthermore, and importantly, construct validity evidence adduced by the metric calibration approach is more fine-grained because it involves modeling particular response functions (linear or non-linear) between test scores and behavior expressed in meaningful unstandardized regression coefficients (or odds ratio in the case of a binary behavioral outcome) rather than the conventional (and arguably impoverished) zero-order "validity" correlations. For example, the Impulse-Control (conscientiousness) task persistence metric mapping found in Study 1 provides a useful demonstration of this point. As can be seen in Figure 12, every unit increase in Impulse Control self-report scores corresponded to roughly a 25 second increase in persistence on the near-impossible anagrams. This metric mapping translates to about 1 minute and 12 seconds of persistence for individuals reporting Impulse Control scores at the scale midpoint of "3" and about 1 minute and 37 seconds of persistence for individuals reporting Impulse Control scores of "4". Therefore, these kinds of mappings between test scores and specific behaviors provide more illuminating and hence stronger construct validity evidence for the psychological instrument at hand.

#### 4.2.2.2 Help with Conceptual Challenges

Metric calibration could also help construct validity research by aiding with challenging conceptual issues that often arise in the development of psychological instruments. In particular, the process of metric calibration may help with conceptual challenges related to construct definition and basic theorizing of the construct (Messick, 1989). One conceptual challenge that often arises, for example, is the issue of how broad a construct should be defined. That is, finding the most optimal construct definition that is neither too broad nor too narrow in scope (Gawronski et al., 2008). Some of these issues became readily apparent when going through the metric calibration process for the construct of conscientiousness in Study 1. For example, even though most researchers seem to accept the construct definition of conscientiousness as the propensity of being painstaking and careful in acting according to the dictates of one’s conscience (John & Srivastava, 1999), conscientiousness in the literature is actually posited to have many different facets including Self-Discipline, Carefulness, Thoroughness, Organization and Orderliness, Deliberation, Industriousness, Conventionality, Reliability, Virtue, Dutifulness, and Need for Achievement (Costa & McCrae, 1992; Goldberg, 1999; Roberts et al., 2005). Given that metric calibration requires researchers to focus on only a few diagnostic behavioral manifestations of the construct, it has the potential to help with conceptual issues such as whether a construct is too broad in scope.

In fact, I contend that a case can be made that conscientiousness is too broad in scope and that conceptual clarity could be achieved by relegating most of those lower-order facets of conscientiousness to their proper distinct constructs (e.g., relegate “need for achievement” conscientiousness facet to actual “need for achievement” construct; McClelland, 1951). Indeed, these types of conceptual challenges relate very closely to what Jack Block termed the “jingle-jangle fallacy” (1995, 2000) whereby the same term is used by different researchers to refer to different psychological entities and where different terms are used by different researchers to refer to the same psychological entity. In this respect, I contend that the metric calibration approach has the potential to help researchers work through these difficult conceptual challenges.

#### 4.2.2.3 Measurement Benchmark

A final way metric calibration studies could also help with construct validity research is by providing a sort of measurement benchmark for detecting measurement problems and/or to further improve psychological instruments. The logic underlying this idea stems from the recently mentioned fact that the empirical metric linkages between test scores and theoretically meaningful behaviors provide richer and arguably more diagnostic information than traditional convergent or criterion validity approaches. The non-linear metric mapping between self-reported task persistence and persistence in the anagram task provides a good example of this proposed benefit. As depicted in Figure 11 (dotted line), a non-linear cubic function was found which explained almost three times more variance than a linear function. This "dipping" pattern (which was subsequently replicated in Ditre & Brandon's, 2008 data) suggested that many individuals endorsing the highest possible response on the self-report measure ("4" out of 4) exhibited some kind of over-reporting bias given that these individuals showed less persistence in the anagram task than those with a lower self-report persistence score (i.e., "3.5"). This metric calibration finding demonstrates how the metric calibration approach may facilitate the process of detecting measurement issues and also help in improving measurement instruments. Indeed, a straightforward implication of the task persistence self-report measure problem would be to examine whether a strong accuracy or honesty instruction would eliminate the alleged over-reporting bias, as would be reflected if a linear function would explain more or just as much variance as a cubic function.[22]

### 4.2.3 Contribute to Theoretical Development

In my third category of proposed benefits, I argue that non-arbitrary metrics could contribute to theoretical development more broadly, by aiding in theoretical debates involving absolute claims, allowing for more precise theorizing in our scientific language, and providing a platform for more quantitative testing of theories.

#### 4.2.3.1 Aid in Theoretical Debates Involving Absolute Claims

First, I argue that metric calibration could contribute to theoretical development by helping in theoretical debates that involve making absolute claims about psychological phenomena. That is, the metric approach demonstrated in the current dissertation can contribute in new ways to theoretical development in basic psychology by providing methodological machinery for researchers to more directly tackle theoretical questions

[22] In fact, Ditre and Brandon (personal communication, February 10, 2010) showed great interest in my finding and in wanting to execute such a follow-up construct validity study.

that involve making claims of an absolute nature. Many interesting theoretical questions are absolute in nature, for instance: “Is (implicit) self-esteem universally positive?” (Yamaguchi et al., 2007), “Are young people narcissistic?” (Twenge, Konrath, Foster, Campbell, & Bushman, 2008), and “Are most people unconscious racists?” (Blanton & Jaccard, 2006a, 2006b). With arbitrary metrics, however, we cannot tackle these important research questions directly and so researchers either tippy-toe around the question in less effective indirect ways or avoid the questions altogether. The systematic use of a metric calibration approach, however, could potentially open the door to more directly tackling these important questions about human psychology.

Study 2 illustrated how the metric calibration approach could potentially be valuable in contributing to theoretical debates involving absolute claims, by focusing on the pan-cultural debate of self-enhancement (Heine, 2005; Sedikides et al., 2003). The metric mapping results for a trait rating measure of self-enhancement commonly used in the debate showed that very little behavioral evidence of self-enhancement corresponded to trait scores typically interpreted as self-enhancement. That is, a trait rating score of “5” (on a 1 to 7-point scale) typically interpreted as self-enhancement (e.g., Gaertner et al., 2008), corresponded to the over-claiming of only about 3 non-existent words.[23] Of course, this mapping should be interpreted with some caution given the small sample size and lack of consensus on the OCT behavioral reference point. For the sake of illustration, however, if one would put stock into this metric mapping, then a theoretical implication for the pan-cultural debate could be that researchers should not use trait rating scores tested against the scale midpoint to examine self-enhancement within or across cultures.

[23] This conclusion becomes even more pronounced if one uses the minimum value that is statistically greater than the scale midpoint. In my sample, a trait rating score of “4.2” would be statistically significantly greater than the scale midpoint of “4” ($p < .05$), but would only be associated with over-claiming of about one non-existent word. This was manifested in Gaertner et al. (2008), for example, where a trait-rating mean of “3.7”, statistically greater than the scale midpoint of “3.5” ($p < .05$), was taken as evidence of self-enhancement in a Taiwanese sample.

#### 4.2.3.2 Allow More Precise Theorizing via Enhanced Scientific Language

Metric calibration could contribute to theoretical development in a second sense by making our scientific language more precise, which in turn will increase the precision in our theorizing. That is, calibrating the metric of psychological instruments empirically substantiates claims about the standing of individuals on the underlying psychological dimensions captured by those instruments. At present, theorizing containing references to "high-X individuals" or "low-X individuals" doing certain things under certain conditions are rampant in the literature (where X can be any construct). For example, "…high-SE individuals possess self-doubts and insecurities…" (Jordan et al., 2003, p. 975). These kinds of meter-reading claims are strictly unsubstantiated given that they are based on scores with arbitrary metrics (Blanton & Jaccard, 2006a, 2006b), and hence impede accurate theorizing and potentially interferes with theory development. With arbitrary metrics, all that one can say is that individuals who "scored high (or low)" on a certain instrument acted in certain ways. Only when scores are empirically calibrated to behaviors consensually agreed-upon as reflecting high (or low) levels of the construct does one's theorizing involving expressions such as "high-X individuals" become empirically substantiated.

For example, consider the preliminary metric calibration finding from Study 3 showing that an almost maximal CCT score of "29" (out of "30") corresponded to only a 50/50 chance of choosing a risky $10 gamble over $6 for certain (see Figure 17, dotted line). As this example nicely demonstrates, it would be potentially misleading to assume that a high score on this measure reflects a high level of risk-taking in an absolute sense. This problem becomes even more apparent when one considers adopting this kind of meter-reading strategy to different instruments of the same construct, for instance, Study 3's BART. It should be clear that a high score on the BART does not necessarily reflect the same level of risk-taking than a high score on the CCT. This important issue, however, is obfuscated when unsubstantiated expressions such as "high-X individuals" are made based on measures with arbitrary metrics. The metric calibration approach could overcome these problems and lead to more precise scientific language in describing psychological phenomena. Ultimately, this could facilitate more accurate theorizing about human psychology and hence contribute to theoretical development more broadly.

#### 4.2.3.3 Quantitative Testing of Psychological Theories

A final way metric calibration could potentially contribute to theoretical development is by providing a platform for testing psychological theories in a more quantitative manner rather than the strictly directional hypothesis testing approach typically used in psychology (Meehl, 1978). In particular, using more meaningful calibrated metrics in day-to-day research activities may eventually get researchers into the habit of paying much more attention to the meaning of particular scores and metric meaning more broadly. And in conjunction with the more nuanced interpretations of data patterns calibrated metrics could allow, this different mentality may eventually lead to more fine-grained integrated theoretical accounts of a research area which could facilitate the process of testing psychological theories more quantitatively by developing hypotheses involving particular point-value predictions. In physics, specific point-value predictions involve comparing a theoretically predicted value $x_0$ (based on theoretical considerations of the particular experimental or natural factors embedded in a situation) with the observed mean $\bar{x}_0$, and determining whether the predicted value falls within the band of probable error (due to random measurement error) of the empirically observed mean (Meehl, 1967). Although it might be difficult to imagine that psychological theory will ever be developed enough to be able to generate these types of point-value predictions, I contend that the metric calibration approach may provide a developmental platform in striving toward this general direction.

### 4.2.4 Facilitate the General Accumulation of Knowledge

The metric calibration approach may facilitate the general accumulation of knowledge more broadly in three main respects. These are described below.

#### 4.2.4.1 Valuable Information in its Own Right

The empirical findings that result from the metric calibration approach can be seen as valuable information in its own right. That is, knowing what kinds of particular theoretically-relevant behaviors correspond to certain scores on a particular measure provides, in itself, valuable knowledge about human psychology. For example, my re-analysis of Hong and Paunonen (2009) revealed a meaningful metric mapping between gregariousness facet scores and number of social parties attended per month, such that a maximal score of “5” corresponded to attending about 7 social parties per month. By building a network of these kinds of metric mappings for various instruments assessing diverse psychological constructs, valuable information could be systematically amassed regarding our general knowledge of the psychological landscape.

#### 4.2.4.2 Guiding Framework for Cataloguing the Magnitude of Psychological Effects

The metric calibration approach could also help the general accumulation of knowledge by providing a framework to help systematically catalogue the magnitude of psychological effects, as strongly advocated by Jacob Cohen (1994). With arbitrary metrics and effect size indices of limited meaning, however, storing up information about the magnitude of experimental effects would likely be quite unproductive. With calibrated metrics, on the other hand, this storing up of information could potentially be much more useful because one could express the magnitude of a psychological effect with respect to the calibrated (and consensually agreed-upon) behavioral reference points. That is, one could express the magnitude of a particular manipulation on a set of DV scores in terms of meaningfully interpretable behaviors. For example, pulling from my previous re-interpretation of Benjamin and Robbins’ (2007) study involving a framing manipulation on BART scores, one could catalogue the magnitude of the experimental effect in terms of the increased probability of choosing the risky gamble. In this way, it is hoped that we could perhaps finally heed to John Tukey’s (1969) plea for psychologists to store up “amount[s], not just direction” (p. 86).

#### 4.2.4.3 Facilitate Phenomenon-Based Research

A final way the metric calibration approach could contribute to the general accumulation of knowledge is by facilitating phenomenon-based research (Asch, 1952/1987; Rozin, 2001). Adherents of the phenomenon-based research perspective argue that it is critical to identify and describe phenomena and invariances before engaging in more sophisticated types of modeling and hypothesis testing. Viewed from this perspective, I contend that the metric calibration approach could provide a useful general framework to engage in this type of descriptive, phenomenon-driven research. In fact, basic metric calibration studies, such as those executed in this dissertation, can be viewed as providing descriptively-rich information about psychological phenomena, given that the goal of the metric approach is to discover how the scores of a certain measure map onto meaningful behaviors argued to reflect different levels of the underlying construct. View in this light, metric calibration can be seen as having the potential to contribute more broadly to the accumulation of psychological knowledge.

## 4.3 Relatedness to Other Past Measurement Approaches

The metric calibration approach espoused in this dissertation is broadly consistent with, and can be seen as extending, other past measurement approaches, which are worth mentioning for the sake of knowledge continuity. Importantly, however, notable differences exist between these methods and the metric approach and ultimately only the proposed metric calibration approach can render the units of measurement of psychological instruments non-arbitrary.

For instance, Guttman's (1950) scalogram approach involves finding a series of behaviors such that all individuals exhibiting a set of cumulatively-ordered behaviors belong to the same "level" of the underlying construct, whereas individuals exhibiting those same behaviors and at least one additional behavior belong in the next higher "level" of the construct. To achieve such Guttman scaling, the set of behaviors must be ordered cumulatively such that exhibiting a set of such behaviors is assumed to reflect a lower level of the construct than someone exhibiting that same set of behaviors in addition to one other cumulatively-ordered behavior. Consider as an example the Guttman scale developed to measure fear of battle in World War II soldiers (Stouffer, 1950). For this scale, individuals who did not experience "violent pounding of the heart" during battle formed the lowest level of the construct, while those who did were part of the next higher level of the construct. If a soldier also reported a "sinking feeling in the stomach" during battle as well as violent pounding of the heart, the solider belonged to the next higher level of the construct, and so on.

The Guttman (1950) scalogram approach can be seen as similar to the metric calibration approach in the sense that it focuses on the meaning of particular behaviors that are assumed to reflect different levels of the underlying construct. An important difference, however, between such approach and the metric calibration approach is that the cumulatively-ordered behaviors are typically self-reported in the Guttman approach whereas metric calibration emphasizes the objective manifestation of meaningfully interpretable behaviors. A more important difference, however, lies in the fact that no metrics or scores are calibrated in the Guttman approach because the behaviors are used in and of themselves to reflect the different levels of the underlying construct whereas the goal in metric calibration is to empirically connect particular scores to behaviors argued to reflect different levels of the underlying construct.

Thurstone's (1927) method of equal-appearing intervals is another measurement approach that metric calibration is broadly consistent with. In such an approach, individuals indicate their agreement or disagreement with attitudinal statements that have been empirically judged to vary with respect to favorability toward the attitude object. Measurement scores are then calculated by examining the items that individuals agreed with and computing the average item favorability score for each of those endorsed items. The Thurstone approach can be seen as similar to the metric calibration approach in the broad sense that effort is put into creating endorsable statements that have been empirically judged to reflect different levels of the underlying construct. Nonetheless, and similar to the Guttman (1950) approach, the Thurstone approach is clearly different from metric calibration because the endorsed statements (or more accurately the average item favorability of the endorsed statements) are used in and of themselves to reflect different levels of the underlying construct. Hence, it is unknown how resulting scores map onto the underlying dimension because scores are based on subjective favorability judgments from independent judges.

The metric calibration approach can also be seen as an extension of “concept mapping,” which is sometimes used in the item generation stage when developing self-report instruments in the social sciences (Trochim, 1989). In this approach, concept mapping refers to a “type of structured conceptualization” (p. 1) that facilitates the process of conceptualizing the domain of a construct by using concept maps. From this perspective, a group of experts of a certain construct would generate statements that describe behaviors a person high in the construct would exhibit, distinct from behaviors a person low in the construct would exhibit. Self-report items are then constructed based on those statements. As can be seen, the concept mapping approach involves asking very similar questions that a researcher from the metric calibration approach would ask, with regard to behaviors reflective of high or low levels of the construct. The major difference, of course, is that in the metric approach the relevant behaviors are used to calibrate test scores rather than simply being used to generate self-report items.

Finally, the metric calibration approach espoused in this dissertation can also be seen as an extension of two approaches used in the area of industrial/organizational psychology. For instance, metric calibration can be viewed as an extension of the expectancy chart approach sometimes used in the context of personnel selection (Lawshe & Bolda, 1958; Lawshe, Bolda, Brune, & Auclair, 1958). In this approach, charts are constructed such that the expected likelihoods of successful job performance of an individual are tabulated for different ranges of predictor scores on some kind of assessment tool (e.g., a personality measure). For instance, obtaining a score in the 50th percentile could correspond to an expected probability of .2 of successful job performance whereas a score in the 90th percentile could correspond to an expected probability of .6 of successful job performance. Such approach is similar to the metric approach in the sense that an empirical mapping is sought between particular test scores and the probability of exhibiting a particular relevant behavior. The approaches are also similar in the sense that similar types of statistical analyses are used to empirically connect test scores to behavior (e.g., logistic regression). The two approaches differ in important ways, however, in that only the metric calibration approach focuses on developing meaningful units of measurement for different instruments of a construct. In this sense, the metric calibration approach is significantly different from the expectancy chart approach given that only the metric calibration approach requires specifically choosing and configuring theoretically-relevant behaviors to serve as reference points that can be argued to reflect particular locations on the underlying dimension of interest. Also, only in the metric calibration approach is it relevant to select several distinct behaviors to reflect ordered reference points. Finally, the metric calibration approach uniquely focuses on calibrating different instruments to the same behavioral reference points as to discover what ranges of the underlying dimension the different instruments are capturing.

Utility analysis is another approach in industrial/organizational psychology that could be seen as a past research approach extended by the metric calibration approach. Originally introduced by Brogden and Taylor (1950) and further developed by Cronbach and Gleser (1965), utility analysis refers to a quantitative method that estimates the benefits in dollar figures that would be gained by an organization if an intervention or selection procedure designed to increase worker productivity was used. In a selection context, for instance, a utility analysis approach would allow one to estimate how much money an organization would gain if a person selected for a job had a particular score on a selection test compared to another score (e.g., hiring a person with a score of “45” on a selection measure could benefit the company $50,000 versus hiring a person with a score of “40”). The utility analysis approach is similar to metric calibration in the sense that scores from the selection measures are linked to external criteria that describe the implications of receiving one score versus another. As with the expectancy chart approach, the two approaches are also alike in that similar types of statistical analyses are used (e.g., linear regression). The two approaches differ in important ways, however, in that the goal in utility analysis is to specifically link selection scores to dollar figures to help managers evaluate the financial impact of their decisions whereas the more general goal in the metric calibration approach is to link the scores of different instruments to a common set of behavioral reference points as to render the units of measurement of the different instruments meaningful and comparable. Hence, only in metric calibration is the focus on developing meaningful units of measurements that researchers can collectively use.[24]

In summary, the metric calibration approach can be seen as broadly consistent with all of these past measurement approaches. More specifically, the metric calibration approach can be viewed as an extension and refinement of these past measurement approaches rather than being viewed as a completely novel approach.

## 4.4 Limitations and Caveats

At the empirical level, the two most important limitations of the current dissertation are related to the sample size of the present metric calibration studies and the need for consensus in choosing appropriate behavioral reference points. In metric calibration studies proper, large samples are required to ensure that the parameter estimates of the metric mapping functions are accurate (i.e., stable) estimates of the relevant parameters of the targeted population. This is critical because in metric calibration the ultimate goal is to find meaningful empirical linkages between test scores and the probability or frequency of theoretically-relevant behaviors. Hence, if the sample size is small and parameter estimates of the metric function are contaminated with large amounts of sampling error, then a metric mapping found in a particular sample may not be very meaningful and hence useful. For instance, in the case of a binary behavioral reference point, the log-odds coefficient and intercept upon which the metric mapping is calculated may be too imprecise of a population estimate to put much stock in. Though what constitutes a "large" sample may be difficult to pinpoint exactly, I would say sample sizes in the range of 300 or more should be considered as a lower bound.

[24] Another approach in the I/O psychology area that relates to metric calibration broadly construed is the Productivity Measurement and Enhancement System (ProMES; Pritchard, Jones, Roth, Stuebing, & Ekeberg, 1989), which seeks to develop integrative sets of utility functions for different aspects of successful job performance (e.g., % of circuit boards completed) using a common organizational effectiveness metric as outcome of the utility functions. This approach differs in important ways from metric calibration, however, in that metric mappings are not established by empirically connecting test scores to independently measured behavioral reference points but rather the utility functions between the different aspects of job performance and organizational effectiveness are decided by discussion and consensus (though the different aspects of job performance are measured).

The second most important limitation at the empirical level is the need for some kind of collective consensus in agreeing on behaviors that are most theoretically appropriate to serve as behavioral reference points. Consequently, it is important to realize that my empirical demonstrations are limited by the extent to which relevant experts agree with my choice of behavioral reference points. I tried my best possible to choose behavioral reference points that were theoretically derived and conceptually consistent with the most commonly accepted working definition of each construct, including sometimes contacting relevant experts and soliciting their opinions (e.g., in the case of choosing lottery risk choices as behavioral reference points for risk-taking; T. Pleskac, personal communication, June 15, 2010). Nonetheless, strictly speaking without some kind of consensus on the appropriateness of the behavioral reference points, at best the empirical demonstrations should be seen as simply that: illustrative empirical examples of the metric calibration process assuming some kind of consensus exists.

At a more conceptual level, two limitations are worth briefly discussing here. First, the metric approach may be limited in utility for broad personality constructs often studied in personality psychology. That is, given the sometimes explicit goal in personality research to assess and understand broad behavioral trends (typically assessed via self-report; Paunonen, 2009), rather than more circumscribed and particularly meaningful behaviors, it could seem that the metric calibration approach, which requires the selection of only a few relevant behaviors to act as reference points, is of limited utility for the calibration of instruments in the personality area. Though there may be a grain of salt in that position, I contend that a possible alternative view on this issue is that the metric approach may suggest that such modal measurement approach in personality is in itself limited. This perspective would be consistent with pleas by certain theorists who have recently called for much more direct behavioral observations in personality research (Back & Egloff, 2009; Furr, 2009; Mehl, 2009). Ultimately, these theorists argue that focusing more of our attention on direct behavioral observations would bring us closer to the key mission of psychology: “understanding the determinants and consequences of what people actually do” (Back & Egloff, 2009, p. 405).

The metric calibration approach could also be seen as limited in utility for instruments assessing highly phenomenological constructs that tap into psychological states not directly reflected in any observable behavior. For instance, instruments used to assess phenomenological or "experiential" constructs such as consciousness or sensory color perceptions may not be amenable to the metric calibration approach. That being said, I would put forward that many constructs that appear at first glance to be too experiential or subjective for metric calibration may, upon further consideration, actually be amenable to metric calibration. For example, constructs such as personal values, inner motivations, and transient feelings all could, upon deeper consideration, be argued to nonetheless have correspondent behavioral manifestations that could be used as reference points to calibrate the scores of such subjective and experiential measures (e.g., attending a pro-life rally as a behavioral manifestation of holding anti-abortion values).

Another caveat worth mentioning is that in certain research situations, the metric issue may be less relevant if researchers are more simply interested in the description of behavior rather than using behavior as a proxy for an underlying latent construct. For instance, in the psychological literature of addiction research, researchers may be interested in assessing the number of daily cigarettes smoked after an intervention as a completely descriptive measure of that specific behavior. In this very specific research situation, metric calibration is not relevant because the metric (i.e., number of daily cigarettes smoked) is meaningful given the strict descriptive nature of the assessment (Blanton & Jaccard, 2006a). That being said, if the number of daily cigarettes is used as a proxy to assess self-regulation, then the metric becomes arbitrary and metric calibration becomes relevant if one wants to get a sense of where on the underlying dimension of self-control the number of cigarettes metric falls. It seems safe to say that the vast majority of psychological research falls in this latter category whereby behaviors are used as a proxy to an underlying latent construct (Borsboom, 2005; Embretson, 2006).

## 4.5 Future Directions

In this section, I want to briefly elaborate on a few different future directions that I believe constitute potentially fruitful avenues to explore to increase the feasibility and ultimate utility of the metric calibration approach. First, future metric calibration research should seriously consider using more sophisticated methodology to assess richer behaviors to serve as reference points. This could include, for example, using eye-tracking technology to assess particular eye gaze behaviors that could serve as diagnostic behavioral reference points to calibrate test scores of an instrument assessing a relevant construct. For instance, in the context of calibrating test scores for a measure of goal activation, one could use the percentage of time individuals' gaze focus on goal-relevant features of serially presented pictures. Eye-tracking methodology may turn out to be a powerful general tool for metric calibration research because eye gaze behaviors may be more diagnostic reflections of the construct at hand, given that early saccades have been argued to be relatively unfiltered "up-stream" components of behavior (Guitton & Volle, 1987). Using an observational approach whereby independent judges code behavior observed in carefully constructed laboratory situations could also be another fruitful avenue to explore to provide richer behavioral reference points.

In addition, future research should also consider utilizing more sophisticated methodology to assess ecologically valid behaviors that emerge in naturalistic settings to serve as behavioral reference points. Though at first glance this future direction seems methodologically prohibitive, recent technological developments have made possible the assessment of human behaviors *in vivo* as they naturally occur in the lives of individuals tracked over time. For instance, Mehl and colleagues (2001) have developed the electronically activated recorder (EAR) as a naturalistic observation sampling method that allows researchers to unobtrusively "observe" actual behavior as it unfolds in natural environments. This is achieved by individuals wearing a pocket-sized audio-recorder that captures snippets of ambient sounds in individuals' momentary environments at random intervals throughout the day, which can then be coded by independent judges. As an example, in one study Mehl and colleagues coded the percentage of people's waking hours spent socializing, in the context of a cross-cultural study on whether Mexicans are more or less sociable than Americans (Ramirez-Esparza, Mehl, Alvarez Bermudez, & Pennebaker, 2009). From a metric calibration perspective, these ecologically valid behavioral observations provide arguably the most meaningful and compelling behavioral reference points for which to calibrate test scores.

Another avenue to explore is the utility of my proposed experimental approach to metric calibration inspired by the calibration of instruments in the physical sciences such as the thermometer and hygrometer. The basic logic here is to experimentally manipulate the construct at hand to extreme levels and look to identify any qualitatively distinct behavioral manifestations of the construct that can serve as additional or better reference points. This would potentially supplement the standard metric calibration approach in important ways because it is possible that more diagnostic behavioral reference points exist outside the range of naturally-occurring levels of the construct (akin to how the calibration of thermometers to naturally-occurring levels of temperature could be seen as limited, because the fixed points of boiling and freezing water do not necessarily arise within naturally-varying temperature levels).

Finally, it is worth considering the application of more advanced psychometric procedures as future avenues to supplant the extant metric calibration approach. One angle to take in this vein is to explore the utility of a within-subjects approach to metric calibration by employing psychometrically-inspired state-space models (e.g., Commandeur & Koopman, 2007). From this perspective, the construct at hand is assessed using a repeated-measure design using both the to-be-calibrated measure and relevant behavioral assessments. Then, individual-specific slopes and intercepts can be estimated which can then be used to construct person-specific metric mappings to some common behavioral reference point. This approach could be very powerful given that it would allow for the consideration and comparison of metric calibration patterns at both the intra- and inter-individual levels (which could turn out to be critically important).

A final psychometric future direction to consider is the application of item response theory (IRT; Embretson & Reise, 2000; Lord, 1980) to provide a more sophisticated modeling of relevant behavioral reference points. An important conceptual obstacle in the metric calibration approach is that sometimes several different behaviors may be seen as theoretically meaningful in serving as reference points and so the choice of behaviors could turn out to be difficult for certain constructs. It may be possible, however, to use an IRT approach to model a set of hierarchically-ordered behaviors (treated as "items" varying in "item difficulty"), which could yield "behavior" characteristic curves, that would reveal the predicted probability of exhibiting each behavior as a function of a person's level on the underlying construct.[25] This could be seen, in a sense, as an IRT approach applied to a Guttman-like behavioral scale. Ultimately, this approach could provide a valuable tool allowing for a more fine-grained use of multiple behaviors to act as distinct and ordered reference points to calibrate scores of psychological instruments.

## 4.6 Coda

In closing, given the advent of new technological developments in both methodological assessment tools and psychometric advances, the future is bright for the metric calibration approach to contribute in important ways to the betterment and advancement of basic psychological research. I leave you with the hope that, in the spirit of John Tukey, the metric calibration approach may one day finally allow psychological researchers to care about their units of measurement. Ultimately, this would solidify the cornerstone of measurement that so critically underlies empirical psychological research.

[25] I would like to thank Patrick Shrout for discussions that directly inspired this future direction research idea (P. Shrout, SPSP 2011, January 28, 2011).

# References

Abelson, R. P. (1995). *Statistics as principled argument.* Mahwah, NJ: Lawrence Erlbaum Associates.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* London: Sage.

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracı´n, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change: Basic principles* (pp. 173–222). Mahwah, NJ: Erlbaum.

Alice, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average-effect. *Journal of Personality and Social Psychology, 68*, 804-825.

Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers really tell us?: Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport & Exercise Psychology, 29*, 664-672.

Asch, S. E. (1952/1987). *Social psychology.* New York: Oxford University Press. (Original work published 1952).

Back, M. D., & Egloff, B. (2009). Yes we can! A plea for direct behavioral observation in personality research. *European Journal of Personality, 23,* 403-405.

Bech, P., Rasmussen, N.-A., Olsen, L. R., Noerholm, V., & Abildgaard, W. (2001). The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. *Journal of Affective Disorders, 66*, 159-164.

Beck, A. T., & Steer, R. A. (1987). *Beck Depression Inventory manual*. San Antonio, TX: Harcourt Brace.

Belanger, C., Buring, J. E., Eberlein, K., Goldhaber, S.Z., Gordon, D., Hennekens, C. H., Mayrent, S. L., Peto, R., Rosner, B., Stampfer, M., Stubblefield, F., & Willett, W. (1988). Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 318*, 262-264.

Benjamin, A. M., & Robbins, S. J. (2007). The role of framing effects in performance on the Balloon Analogue Risk Task (BART). *Personality and Individual Differences, 43,* 221-230.

Ben-zur, H., & Zeidner, M. (2009). Threat to life and risk-taking behaviors: A review of empirical findings and explanatory models. *Personality and Social Psychology Review, 13,* 109-128.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66*, 5–20.

Bishara, A. J., Pleskac, T. J., Fridberg, D. J., Yechiam, E., Lucas, J., Busemeyer, J. R., Finn, P. R., & Stout, J. C. (2009). Similar processes despite divergent behavior in two commonly used measures of risky decision-making. *Journal of Behavioral Decision Making, 22*, 435-454.

Blais, A-R., & Weber, E. U. (2006). A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgment and Decision Making, 1,* 33-47.

Blanton, H., & Jaccard, J. (2006a). Arbitrary metrics in psychology. *American Psychologist, 61*, 27-41.

Blanton, H., & Jaccard, J. (2006b). Arbitrary metrics redux. *American Psychologist, 61*, 62-71.

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology, 94*, 567–582.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117,* 187-215.

Block, J. (2000). Three tasks for personality psychology. In L. R. Bergman, R. B. Cairns, L-G. Nilsson, & L. Nystedt (Eds.), *Developmental science and the holistic approach* (pp. 155-164). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology, 42*, 163-176.

Booth-Kewley, S., & Vickers, R. R. (1994). Associations between major domains of personality and health behavior. *Journal of Personality, 62*, 281-298.

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin, 16*, 335-338.

Borsboom, D. (2005). Measuring the mind. Conceptual issues in contemporary psychometrics. Cambridge: Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425–440.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Boyer, T. W. (2006). The development of risk-taking: A multi-perspective review. *Developmental Review, 26,* 291-345.

Brand, M., Fujiwara, E., Borsutzky, S., Kalbe, E., Kessler, J., & Markowitsch, H. J. (2005). Decision-making deficits of Korsakoff patients in a new gambling task with explicit rules: Associations with executive functions. *Neuropsychology, 19,* 267-277.

Brandon, T. H., Herzog, T. A., Juliano, L. M., Irvin, J. E., Lazev, A. B., & Simmons, V. (2003). Pretreatment task persistence predicts smoking cessation outcome. *Journal of Abnormal Psychology, 112*, 448−456.

Briñol, P., & Petty, R.E. (2005) Individual differences in persuasion. In D. Albarracín, B. T. Johnson, & M. P. Zanna (eds), *The Handbook of Attitudes and Attitude Change* (pp. 575—616). Hillsdale, NJ: Erlbaum.

Brogden, H. E., & Taylor, E. K. (1950). The dollar criterion: Applying cost accounting concepts to criterion selection. *Personnel Psychology, 3*, 133-154.

Brown, J. D. & Kobayashi, C. (2002). Self-enhancement in Japan and America. *Asian Journal of Social Psychology, 5*, 145–167.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116-131.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306-307.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, B. W. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119,* 197-253.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7–18.

Chirumbolo, A., & Leone, L. (2008). Individual differences in need for closure and voting behaviour. *Personality and Individual Differences, 44,* 1279-1288.

Cloninger, C. R. (1987). A systematic method for clinical description and classification of personality variants: A proposal. *Archives of General Psychiatry, 44*, 573−588.

Cohen, A. R. (1957). Need for cognition and order of communication as determinants of opinion change. In C. I. Hovland (Ed.), *The order of presentation in persuasion* (pp. 79-97). New Haven, CT: Yale University Press.

Cohen, A. R., Stotland, E., & Wolfe, D. M. (1955). An experimental investigation of need for cognition. *Journal of Abnormal and Social Psychology, 51,* 291-294.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1,* 98-101.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for Behavioral Sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Collins, D. R., & Stukas, A. A. (2008). Narcissism and self-presentation: The moderating effects of accountability and contingencies of self-worth. *Journal of Research in Personality, 42,* 1629-1634.

Commandeur, J. J. F., & Koopman, S. J. (2007). *An introduction to state space time series analysis.* Oxford: Oxford University Press.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past 15 years. *Annual Review of Psychology, 45*, 545-580.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1958). Proposals leading to analytic treatment of social perception scores. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 353-379). Stanford, CA: Stanford University Press.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116-127.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Cronbach, L. J., & Gleser, G.C. (1965). *Psychological tests and personnel decisions.* (2nd ed.). Urbana: University of Illinois.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting and Clinical Psychology, 24,* 349-354.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3,* 286-300.

Dahlstrom, W. G., & Welsh, G. S. (1960). *An MMPI handbook: A guide to use in clinical practice and research*. Minneapolis: University of Minnesota Press.

De Houwer, J. (2006). What are implicit measures and why are we using them? In R.W. Wiers & A.W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA: Sage.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3,* 412-423.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18.

Ditre, J. W., & Brandon, T. H. (2008). Does self-reported task persistence predict performance on behavioral measures of task persistence and distress tolerance? Paper presented at the meeting of the Society for Research on Nicotine and Tobacco, Portland, OR, USA.

Dooremalen, A. J. P. W., & Borsboom, D. (2010). Metaphors in psychological conceptualization and explanation. In A. Toomela & J. Valsiner (Eds.), *Methodological thinking in psychology: 60 years gone astray?* (pp. 121-144). Charlotte: Information Age Publishers.

Dracup, C. (1995). Hypothesis testing—What it really is. *The Psychologist, 8,* 359-362.

Edwards, J. R. (2002). Ten difference score myths. *Organizational Research Methods, 4*, 265-287.

Eisenberger, R. (1992). Learned industriousness. *Psychological Review, 99*, 248−267.

Embretson, S. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61*, 50–55.

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

Epstein, S. (1990). Cognitive-experiential Self-theory. In L. Pervin (Ed.), *Handbook of personality theory and research: Theory and research* (pp. 165-192). NY: Guilford Publications, Inc.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and under-confidence: The role of error in judgment processes. *Psychological Review, 101*, 519–527.

Evans, D. R., Baer, R. A., & Segerstrom, S. C. (2009). The effects of mindfulness and self-consciousness on persistence. *Personality and Individual Differences, 47,* 379-382.

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review, 67*, 269-271.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. San Diego, CA: Hodder & Staughton/DIGITS.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*, 75-98.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75-109). San Diego: Academic Press.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327.

Feather N. T (1961). The relationship of persistence at a task to expectation of success and achievement related motives. *Journal of Abnormal and Social Psychology, 63*, 552-561.

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and Deliberative Processes in Risky Choice: Age Differences in Risk Taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 709-730.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1-74). New York: Academic Press.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin, 106*, 155-160.

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8*, 49–57.

Funder, D. C. (2009). Naïve and obvious questions. *Perspectives on Psychological Science, 4,* 340-345.

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality, 23,* 369-401.

Gaertner, L., Sedikides, C., & Chang, K. (2008). On pancultural self-enhancement: Well adjusted Taiwanese self-enhance on personally valued traits. *Journal of Cross-Cultural Psychology, 39*, 463–477.

Gannon, K. M., & Ostrom, T. M. (1996). How meaning is given to rating scales: The effects of response language on category activation. *Journal of Experimental Social Psychology, 32*, 337-360.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692-731.

Gawronski, B., Peters, K. R., & LeBel, E. P. (2008). What makes mental associations personal or extra-personal? Conceptual issues in the methodological debate about implicit attitude measures. *Social and Personality Psychology Compass, 2*, 1002-1023.

Gigerenzer, G. (1998). Surrogates for theories. *Theory and Psychology, 8*, 195-204.

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189–211). New York: Guilford Press.

Glass, D. C, Singer, J. E., & Friedman, L. N. (1969). Psychic cost of adaptation to an environmental stressor. *Journal of Personality and Social Psychology, 12*, 200-210.

Goffin, R. D., Gellatly, I. R., Paunonen, S. V., Jackson, D. N., & Meyer, J. P. (1996). Criterion validation of two approaches to performance appraisal: The behavioral observation scale and the relative percentile method. *Journal of Business and Psychology, 11*, 23–33.

Goffin, R. D., & Olson, J. M. (2011). Is It All Relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science, 6,* 48-60.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84-96.

Gosling, S. (2008). *Snoop: What your stuff says about you*. New York: Basic Books.

Goudriaan, A. E., Lapauw, B., Ruige, J., Feyen, E., Kaufman, J. M., Brand, M., Vingerhoets, G. (2010). The influence of high-normal testosterone levels on risk-taking in healthy males in a 1-week letrozole administration study. *Psychoneuroendocrinology, 35*, 1416-1421.

Granaas, M. (2002). Hypothesis testing in psychology: Throwing the baby out with the bathwater? Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197-216.

Guitton, D., & Voile, M. (1987). Gaze control in humans: Eye-head coordination during orienting movements to targets within and beyond the oculo-motor range. *Journal of Neurophysiology, 58,* 427-459.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26*, 81-107.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3-10.

Guttman, L. L. (1950). The basis for Scalogram analysis. In S. A. Stouffer, L. L. Guttman, E. A. Suchman, P. W. Lazarsfeld, S. A. Star, & J. A. Clausen, *Studies in social psychology – World War II* (Vol. 4). Princeton, NJ: Princeton University Press.

Hamilton, J. O. (1974). Motivation and risk taking behavior: A test of Atkinson’s theory. *Journal of Personality and Social Psychology, 29,* 856-864.

Hanson, R. K. (2009). The psychological assessment of risk for crime and violence. *Canadian Psychology, 50,* 172-182.

Hanson, R. K., Helmus, L., & Thornton, D. (in press). Predicting recidivism among sexual offenders: A multi-site study of Static-2002. *Law and Human Behavior*.

Harlow, L. L. (1997). Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Erlbaum.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests*? Hillsdale, NJ: Erlbaum.

Harman, J. S., Manning, W. G., Lurie, N., & Liu, C.-F. (2001). Interpreting results in mental health research. *Mental Health Services Research, 3,* 91-97.

Heatherton, T. E, & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology, 60*, 895-910.

Heine, S. J. (2005). Where is the evidence for pancultural self-enhancement?: A reply to Sedikides, Gaertner, and Toguchi (2003). *Journal of Personality and Social Psychology, 89*, 531–538.

Heine, S. J., & Lehman, D. R. (1999). Culture, self-discrepancies, and self-satisfaction. *Personality and Social Psychology Bulletin, 25*, 915–925.

Heine, S. J., Kitayama, S. & Hamamura, T. (2007a). Inclusion of additional studies yields different conclusions: Comment on Sedikides, Gaertner & Vevea (2005), Journal of Personality and Social Psychology. *Asian Journal of Social Psychology, 10*, 49–58.

Heine, S. J., Kitayama, S. & Hamamura, T. (2007b). Which studies test whether self-enhancement is pancultural? A reply to Sedikides, Gaertner, and Vevea, 2007. *Asian Journal of Social Psychology, 10,* 198-200.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review, 106*, 766–794.

Hogan, J., Rybicki, S. L., Motowidlo, S. J., & Borman, W. C. (1998). Relations between contextual performance, personality, and occupational advancement. *Human Performance, 11*, 189-207.

Hong, R. Y., & Paunonen, S. V. (2009). Personality traits and health-risk behaviors in university students. *European Journal of Personality, 23,* 675-696.

Hornsey, M. J. & Jetten, J. (2005). Loyalty without conformity: Tailoring self-perception as a means of balancing belonging and differentiation. *Self and Identity, 4*, 81–95.

Hsee, C. K., & Weber, E. U. (1999). Cross-national differences in risk preferences and lay predictions. *Journal of Behavioral Decision Making, 12,* 165-179.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3-7.

Jackson, J. J., Bogg, T., Walton, K. E., Lodi-Smith, J., Wood, D., Harms, P. D., et al. (2009). Not all conscientiousness scales change alike: A multi-method, multi-sample examination of age differences in conscientiousness. *Journal of Personality and Social Psychology, 96*, 446–459.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York: Guilford Press.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology, 85,* 969-978.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin, 35,* 1131-1142.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332–339.

Kazdin, A. E. (2001). Almost clinically significant (p < .10): Current measures may only approach clinical significance. *Clinical Psychology: Science and Practice, 8*, 455–462.

Kazdin, A. E. (2006). Arbitrary metrics: Implications for identifying evidence-based treatments. *American Psychologist, 61*, 42–49.

Kelly, E. L., & Conley, J. J. (1987). Personality and compatibility: A prospective analysis of marital stability and marital satisfaction. *Journal of Personality and Social Psychology, 52,* 27-40.

Kendall, P. C. (Ed.). (1999). Clinical significance [Special section]. *Journal of Consulting and Clinical Psychology, 67*, 283–339.

Kirk, R. E. (1972). *Statistical issues*. Monterey, CA: Brooks/Cole.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Klar, Y. & Giladi, E. E. (1999). Are most people happier than their peers, or are they just happy? *Personality and Social Psychology Bulletin, 25*, 585–594.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science, 1,* 658-676.

Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19,* 317-322.

Lawshe, C. H., & Bolda, R. A. (1958). Expectancy charts I: Their use and empirical development. *Personnel Psychology, 11*, 353-365.

Lawshe, C. H., Bolda, R. A., Brune, R. L., & Auclair, G. (1958). Expectancy charts II: Their theoretical development. *Personnel Psychology, 11,* 545-559.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112*, 662-668.

Leith, K, P., & Baumeister, R. F. (1996). Why do bad moods increase self-defeating behavior? Emotion, risk taking, and self-regulation. *Journal of Personality and Social Psychology, 71*, 1250-1267.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*, 75–84.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209.

Lipsitt, P. D., Lelos, D., & McGarry, A. L. (1971). Competency for trial: A screening instrument. *American Journal of Psychiatry, 128*, 105-109.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Mahwah, NJ: Erlbaum.

Luce, R. D. (1997). Quantification and symmetry: Commentary on Michell, quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 395-398.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151-159.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19–40.

Machado, A., & Silva, F.J. (2007). Toward a richer view of the scientific method: The role of conceptual analysis. *American Psychologist, 62*, 671–681.

Maner, J. K., Gailliot, M. T., Butz, D. A., & Peruche, B. M. (2007). Power, risk, and the status quo: Does power promote riskier or more conservative decision making? *Personality and Social Psychology Bulletin, 33*, 451–463.

Markus, H. R., Kitayama, S., & Heiman, R. J. (1996). Culture and basic psychological principles. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 857–914). New York: Guilford Press.

McClelland, D. C. (1951). *Personality*. New York: Dryden Press.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1958). A scoring manual for the achievement motive (Chapter 12) in J. W. Atkinson (Ed.), *Motives in Fantasy, Action and Society*. New York: Van Nostrand.

McCormmach, R. (2004). *Speculative Truth*. New York: McGraw-Hill.

McGee, T. D. (1988). *Principles and methods of temperature measurement.* Hoboken, NJ: Wiley.

McGrath, R. J., Cumming, G. F., & Burchard, B. L. (2003). *Current practices and trends in sexual abuser management: The Safer Society 2002 Nationwide Survey*. Brandon, VT: Safer Society Foundation, Inc.

McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist, 15*, 295-300.

Mednick, S.A., & Mednick, M.T. (1967). *Examiner's manual: Remote Associates Test.* Boston: Houghton Mifflin.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834.

Meehl, P. E. (1990a). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Meehl, P. E. (1990b). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.

Meertens, R. M., & Lion, R. (2008). Measuring an individual's tendency to take risks: The risk propensity scale. *Journal of Applied Social Psychology, 38*, 1506-1520.

Mehl, M. R. (2009). Naturalistic observation of daily behavior in personality psychology. *European Journal of Personality, 23*, 414-416.

Mehl, M. R., Pennebaker, J. W., Crow, M. D., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, and Computers, 33*, 517-523.

Merenda, P. F. (2007). Update on the decline in the education and training in psychological measurement and assessment. *Psychological Reports, 101,* 153-155.

Messick, S. (1989). *Validity.* In R. L. Linn (Ed.), Educational measurement (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Middleton, W. E. K. (1966). *A history of the thermometer and its use in meteorology.* Baltimore, MD: The Johns Hopkins Press.

Middleton, W. E. K. (1969). *Invention of the meteorological instruments.* Baltimore, MD: The Johns Hopkins Press.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Erlbaum.

Murphy, K. R., & DeShon, R. P. (2000). Progress in psychometrics: Can personnel psychology catch up? *Personnel Psychology, 53,* 913-924.

Murray, S. L., Holmes, J. G., & Griffin, D. W. (2000). Self-esteem and the quest for felt security: How perceived regard regulates attachment processes. *Journal of Personality and Social Psychology, 78*, 478-498.

Nes, L., Segerstrom, S., & Sephton, S. (2005). Engagement and arousal: Optimism's effects during a brief stressor. *Personality and Social Psychology Bulletin, 31*, 111–120.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

O'Hara, R. E., Walter, M. I., & Christopher, A. N. (2009). Need for cognition and conscientiousness as predictors of political interest and voting strategy. *Journal of Applied Social Psychology, 39,* 1397-1416.

Olson, J. M., Goffin, R. D., & Haynes, G. A. (2007). Relative versus absolute measures of explicit attitudes: Implications for predicting diverse attitude-relevant criteria. *Journal of Personality and Social Psychology, 93,* 907-926.

Parducci, A., & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 496-516.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego,CA: Academic Press.

Paulhus, D.L., Harms, P.D., Bruce, M.N., & Lysy, D.C. (2003). The over-claiming technique: Measuring bias independent of accuracy. *Journal of Personality and Social Psychology, 84*, 681-693.

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology, 84*, 411–424.

Paunonen, S. V. (2009). Behaviours, non-behaviours and self-reports. *European Journal of Personality, 23,* 419-421.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Pepitone, A., & Triandis, H. C. (1987). On the universality of social psychological theories. *Journal of Cross-Cultural Psychology, 18*, 471–499.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change.* New York: Springer-Verlag.

Pirelli, G., Gottdiener, W. H., & Zapf, P. A. (2011). The use of arbitrary metrics in competence to stand trial assessment instruments. Manuscript submitted for publication.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34*, 167–185.

Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. W. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and Clinical Psychopharmacology, 16*, 555-564.

Pollard, P. (1993). How significant is "significance"? In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Erlbaum.

Pomerleau, C. S., Pomerleau, O. F., Flessland, K. A., & Basson, S. M. (1992). Relationship of Tridimensional Personality Questionnaire scores and smoking variables in female and male smokers. *Journal of Substance Abuse, 4*, 143−154.

Popper, K.R. (1968/1959). *The logic of scientific discovery*. New York: Harper & Rowe.

Pritchard, R. D., Jones, S. D., Roth, P. L., Stuebing, K. K., & Ekeberg, S. E. (1989). The evaluation of an integrated approach to measuring organizational productivity. *Personnel Psychology, 42*, 69-115.

Quinn, E. P., Brandon, T. H., & Copeland, A. L. (1996). Is task persistence related to smoking and substance abuse? The application of learned industriousness theory to addictive behavior. *Experimental and Clinical Psychopharmacology, 4*, 186-190.

Ramirez-Esparza, N., Mehl, M. R., Alvarez Bermudez, J., & Pennebaker, J. W. (2009). Are Mexicans more or less sociable than Americans? Insights from a naturalistic observation study. *Journal of Research in Personality, 43*, 1-7.

Roberts, B. W., Chernyshenko, O., Stark, S., & Goldberg, L. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58*, 103–139.

Rodman, S. A., Daughters, S. B., Lejuez, C. W. (2009). Distress tolerance and rational-emotive behavior therapy: A new role for behavioral analogue tasks. *Journal of Rational-Emotive & Cognitive Behavior Therapy, 27*, 97-120.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rotter, G. S., & Tinkleman, V. (1970). Anchor effects in the development of behavior rating scales. *Educational and Psychological Measurement, 30*, 311–318.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-391). Hillsdale, NJ: Erlbaum.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*, 2–14.

Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward? A different perspective. *Perspectives on Psychological Science, 4,* 435-439.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*? (pp. 37-64). Hillsdale, NJ: Erlbaum.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validity studies. *Journal of Applied Psychology, 61*, 473-485.

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49*, 388-395.

Sears, D. O. (1983). The person-positivity bias. *Journal of Personality and Social Psychology, 44*, 233–250.

Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist, 51*, 1065-1071.

Sedikides C., Gaertner, L., & Vevea, J. L. (2007b). Evaluating the evidence for pancultural self-enhancement. *Asian Journal of Social Psychology, 10*, 201-203.

Sedikides, C., Gaertner, L. & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology, 84*, 60–70.

Sedikides, C., Gaertner, L. & Vevea, J. L. (2005). Pancultural self-enhancement reloaded: A meta-analytic reply to Heine (2005). *Journal of Personality and Social Psychology, 89*, 539–551.

Sedikides, C., Gaertner, L. & Vevea, J. L. (2007a). Inclusion of theory-relevant moderators yield the same conclusions as Sedikides, Gaertner, and Vevea (2005): A meta-analytical reply to Heine, Kitayama, and Hamamura (2007). *Asian Journal of Social Psychology, 10*, 59–67.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61*, 293-316.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8*, 1-2.

Signorelli, A. (1974). Statistics: Tool or master of the psychologist? *American Psychologist, 29*, 774-777.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108–131.

Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory (STAI).* Palo Alto, CA: Consulting Psychologists Press.

Steinberg, L. (2010). A dual systems model of adolescent risk-taking. *Developmental Psychobiology. Special Issue: Psychobiological models of adolescent risk, 52*, 216-224.

Steinberg, M. L., Krejci, J. A., Collett, K., Brandon, T. H., Ziedonis, D.M., & Chen, K. (2007). Relationship between self-reported task persistence and history of quitting smoking, plans for quitting smoking, and current smoking status in adolescents. *Addictive Behaviors, 32*, 1451−1460.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Stouffer, S. A. (1950). An overview of the contributions to scaling and scale theory. In S. A. Stouffer, L. L. Guttman, E. A. Suchman, P. W. Lazarsfeld, S. A. Star, & J. A. Clausen, *Studies in social psychology – World War II* (Vol. 4). Princeton, NJ: Princeton University Press.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220–247.

Swann, W. B., Chang-Schneider, C., & McClarty, K. L. (2007). Do people's self-views matter? *American Psychologist, 62*, 84–94.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72,* 271-324.

Tellegen, A., & Waller, N. (1994). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs & J. M. Cheek (Eds.), *Personality measures: Development and evaluation* (Vol. 1, pp. 133–161). Greenwich, CT: JAI Press.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*, 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*, 29-32.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Straw arguments move the field. *Journal of Experimental Education, 70*, 80–93.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology, 38,* 368-389.

Trochim, W. M. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning, 12*, 1-16.

Trope, Y. (1986). Self-assessment and self-enhancement in achievement motivation. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol 1, pp. 350-378). New York: Guilford Press.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83-91.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453-458.

Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W. K., & Bushman, B. J. (2008). Egos inflating over time: A cross-temporal meta-analysis of the Narcissistic Personality Inventory. *Journal of Personality, 76*, 875-901.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779-804.

Wallace, H. M., Ready, C. B., & Weitenhagen E. (2009). Narcissism and task persistence. *Self and Identity, 8,* 78-93.

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review, 112*, 862–880.

Watkins, J. T., Leber, W. R., Imber, S. D., Collins, J. F., Elkin, I., Pilkonis, P. A., Stotsky, S. M., Shea, M. T., & Glass, D. R. (1993). Temporal course of change in depression. *Journal of Consulting and Clinical Psychology, 61,* 858-864.

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making, 15*, 263–290.

Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55*, 341–356.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*, 806–820.

White, T. L., Lejuez, C. W., & de Wit, H. (2008). Test-retest characteristics of the Balloon Analogue Risk Task (BART). *Experimental and Clinical Psychopharmacology, 16,* 565-570.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594-604.

Wirshing, D. A., Marshall, B.D., Jr., Green, M. F., Mintz, J., Marder, S. R., & Wirshing, W. C. (1999). Risperidone in treatment-refractory schizophrenia. *American Journal of Psychiatry, 156*, 1374–1379.

Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F., Chen, D., Shiomura, K., Kobayashi, C., Cai, H., & Krendl, A. (2007). Apparent universality of positive implicit self-esteem. *Psychological Science, 18,* 498-500.

Zung, W.W. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12*, 63–70.

# Appendices

## Appendix A: Conscientiousness items (MPQ and NEO-FFI) used in Study 1.

Conscientiousness facets (MPQ; Tellegen & Waller, 1994; NEO-FFI; Costa & McCrae, 1992) (IPIP version of the MPQ and NEO-FFI facets; Goldberg et al., 2006)

HOW ACCURATELY CAN YOU DESCRIBE YOURSELF?

The next part of the experiment involves describing yourself as you generally are now, not as you wish to be in the future. Describe yourself as you HONESTLY see yourself, in relation to other people you know of the same sex as you are, and roughly the same age.

So that you can describe yourself in an honest manner, your responses will be kept anonymous in absolute confidence. You will be presented with a series of statements of behavioral descriptions. For each statement, indicate (using the scale options) whether the statement is:

| | | | | |
|---|---|---|---|---|
| 1<br>Very Inaccurate | 2<br>Moderately Inaccurate | 3<br>Neither Accurate Nor Inaccurate | 4<br>Moderately Accurate | 5<br>Very Accurate |

as a description of you.

MPQ1. I like to plan ahead.
MPQ2. I make a mess of things.*
MPQ3. I am exacting in my work.
MPQ4. I pay attention to details.
MPQ5. I often make last-minute plans.*
MPQ6. I jump into things without thinking.*
MPQ7. I make plans and stick to them.
MPQ8. I like to act on a whim.*
MPQ9. I do things by the book.
MPQ10. I make rash decisions.*

NEO1. I get chores done right away.
NEO2. I find it difficult to get down to work.*
NEO3. I am always prepared.
NEO4. I waste my time.*
NEO5. I start tasks right away.
NEO6. I postpone decisions.*
NEO7. I get to work at once.
NEO8. I need a push to get started.*
NEO9. I carry out my plans.
NEO10. I avoid mistakes.
NEO11. I rush into things.*
NEO12. I choose my words with care.
NEO13. I do crazy things.*
NEO14. I stick to my chosen path.
NEO15. I act without thinking.*
NEO16. I have difficulty starting tasks.*
*Note.* Asterisks (*) denotes reverse-scored items.

Scoring:
*MPQ Self-Control facet*: MPQ1, **MPQ2r,** MPQ3, MPQ4, **MPQ5r, MPQ6r**, MPQ7, **MPQ8r**, MPQ9, **MPQ10r**
*NEO-FFI Self-Discipline facet*: NEO1, **NEO2r**, NEO3, **NEO4r**, NEO5, **NEO6r,** NEO7, **NEO8r,** NEO9, **NEO16r**

*NEO-FFI Deliberation (IPIP cautiousness) facet:* NEO10, **NEO11r**, NEO12, **NEO13r**, NEO14, **NEO15r, MPQ5r, MPQ6r, MPQ8r**, **MPQ10r**

# Appendix B: Conscientiousness items (AB5C) used in Study 1.

Conscientiousness Impulse-Control facet (Goldberg's Abridged Big Five Dimensional Circumplex [AB5C]; Goldberg, 1999)

HOW DO SEE YOURSELF IN GENERAL?

In the next task, you will see a series of common human traits. Please use these traits to describe yourself as accurately as possible. Describe yourself as you see yourself at the present time, not as you wish to be in the future. Describe yourself as you are GENERALLY or TYPICALLY.

For each trait that you will see, please indicate whether that trait describes you using the following rating scale:

| | | | | |
|---|---|---|---|---|
| 1<br>Strongly Disagree | 2<br>Somewhat Disagree | 3<br>Neither | 4<br>Somewhat Agree | 5<br>Strongly Agree |

Gold1. Careful
Gold2. Careless *
Gold3. Cautious
Gold4. Conscientious
Gold5. Erratic *
Gold6. Impulsive *
Gold7. Particular
Gold8. Rash *
Gold9. Reckless *
Gold10. Ritualistic
Gold11. Systematic
Gold12. Uncautious *

*Note.* Asterisks (*) denotes reverse-scored items.

Scoring:
*Goldberg Impulse-Control facet*: Gold1, **Gold2r**, Gold3, Gold4, **Gold5r**, **Gold6r**, Gold7, **Gold8r**, **Gold9r**, Gold10, Gold11, **Gold12r**

## Appendix C: NFC items used in Study 1.

Need for Cognition revised scale (NFC; Cacioppo, Petty, & Kao, 1984)

The next task involves answering questions that are designed to assess your thinking style. There are no right or wrong answers.

For each of the statements below, please indicate to what extent the statement is characteristic of you, using the following scale options:

| | | | | |
|---|---|---|---|---|
| 1<br>Extremely<br>Uncharacteristic | 2<br>Somewhat<br>Uncharacteristic | 3<br>Uncertain | 4<br>Somewhat<br>Characteristic | 5<br>Extremely<br>Characteristic |

NFC1. I prefer complex to simple problems.
NFC2. I like to have the responsibility of handling a situation that requires a lot of thinking.
NFC3. Thinking is not my idea of fun.*
NFC4. I would rather do something that requires little thought than something that is sure to challenge my abilities.*
NFC5. I try to anticipate and avoid situations where there is a likely chance I will have to think in depth about something.*
NFC6. I find satisfaction in deliberating hard for long hours.
NFC7. I only think as hard as I have to.*
NFC8. I prefer to think about small daily projects rather than long-term ones.*
NFC9. I like tasks that require little thought once I've learned them.*
NFC10. The idea of relying on thought to make my way to the top appeals to me.
NFC11. I really enjoy a task that involves coming up with new solutions to problems.
NFC12. Learning new ways to think doesn't excite me much.*
NFC13. I prefer my life to be filled with problems that I must solve.
NFC14. The notion of thinking abstractly is appealing to me.
NFC15. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.
NFC16. I feel relief rather than satisfaction after completing a task that requires a lot of mental effort.*
NFC17. It's enough for me that something gets the job done; I don't care how or why it works.*
NFC18. I usually end up deliberating about issues even when they do not affect me personally.

*Note.* Asterisks (*) denotes reverse-scored items.

# Appendix D: Anagram persistence task (APT) materials used in Study 1.

Anagram Persistence Task Materials (APT; Brandon et al., 2003)

| Anagram | | Solution |
|---|---|---|
| 1. | BEAHC | BEACH |
| 2. | KLYXI * | KYLIX |
| 3. | LMAAE * | MALAE |
| 4. | QYUIA * | YAQUI |
| 5. | NTRAI | TRAIN |
| 6. | CINAI * | INIAC |
| 7. | LBFUE * | FULBE |
| 8. | DPSUA * | PADUS |
| 9. | EOCVI | VOICE |
| 10. | AEWTR | WATER |
| 11. | IFNLG | FLING |

*Note.* Items with an asterisk (*) indicate the 6 critical near-impossible items used to compute the behavioral index of task persistence.

## Appendix E: Words used in over-claiming technique (OCT) of Study 2.

**Over-claiming Technique 150 (variant of Version 2005.1)**

Paulhus, D.L., Harms, P. D., Bruce, M.N., & Lysy, D.C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology, 84*, 890-904.

**PLEASE INDICATE FOR EACH ITEM WHETHER YOU ARE FAMILIAR WITH THE ITEM OR NOT, BY CIRCLING THE APPROPRIATE NUMBER.**

| | |
|---|---|
| **0** | **1** |
| **Never heard** | **Familiar** |
| **of it** | **with it** |

EXAMPLES:

1. If you're asked about POLITICIANS and the item said "Bill Clinton", you would probably circle **'1'** to indicate that you are familiar with him.

2. If the category was FAMOUS ATHLETES and the item said "Fred Gruneberg", you would probably circle '**0**' if you have never heard of him.

| **Historical Names and Events** | **Fine Arts** |
|---|---|
| 1. Napoleon | 16. Mozart |
| 2. Robespierre | 17. a cappella |
| 3. El Puente* | 18. Pullman paintings* |
| 4. My Lai | 19. art deco |
| 5. The Lusitania | 20. Paul Gauguin |
| 6. Ronald Reagan | 21. Mona Lisa |
| 7. Prince Lorenzo* | 22. La Neige Jaune* |
| 8. The Luddites | 23. Mario Lanza |
| 9. Neville Chamberlain | 24. Verdi |
| 10. Vichy Government | 25. Vermeer |
| 11. Queen Shattuck* | 26. Jackson Howell* |
| 12. Bay of Pigs | 27. Grand Pooh Bah |
| 13. Torquemada | 28. Botticelli |
| 14. Wounded Knee | 29. harpsichord |
| 15. Clara Barton | 30. dramatis personae |

| **Language.** | **Books and Poems** |
|---|---|
| 31. subjunctive | 46. Antigone |
| 32. hyperbole | 47. Murphy's Last Ride* |
| 33. alliteration | 48. Catcher in the Rye |
| 34. sentence stigma* | 49. The Bible |
| 35. euphemism | 50. Hiawatha |
| 36. double entendre | 51. Trapnell Meets Katz* |
| 37. blank verse | 52. Mein Kampf |
| 38. pseudo-verb* | 53. The Aeneid |
| 39. ampersand | 54. Faustus |
| 40. myth | 55. The Boy Who Cried Wolf |
| 41. aphorism | 56. Pygmalion |
| 42. shunt-word* | 57. Hickory Dickory Dock |
| 43. simile | 58. The Divine Comedy |
| 44. acronym | 59. Windermere Wild* |
| 45. synonym | 60. The Raven |

| Authors and Characters | Social Science and Law |
|---|---|
| 61. Adonis | 76. yellow journalism |
| 62. Mephistopheles | 77. angst |
| 63. Shylock | 78. nationalism |
| 64. Ancient Mariner | 79. megaphrenia* |
| 65. Doctor Fehr* | 80. acrophobia |
| 66. Venus | 81. pulse tax* |
| 67. Romeo and Juliet | 82. pork-barreling |
| 68. Bulldog Graziano* | 83. prejudice |
| 69. Norman Mailer | 84. Christian Science |
| 70. Horatio Alger | 85. ombudsman |
| 71. Charlotte Bronte | 86. consumer apparatus* |
| 72. Artemis | 87. superego |
| 73. Lewis Carroll | 88. trust-busting |
| 74. Admiral Broughton* | 89. behaviorism |
| 75. Mrs. Malaprop | 90. Oedipus complex |

| Physical Sciences | Life Sciences |
|---|---|
| 91. Manhattan Project | 106.mammal |
| 92. planets | 107.adrenal gland |
| 93. nuclear fusion | 108.sciatica |
| 94. cholarine* | 109.insulin |
| 95. atomic number | 110.meta-toxins* |
| 96. hydroponics | 111.intestine |
| 97. alloy | 112.bio-sexual* |
| 98. plate tectonics | 113.meiosis |
| 99. photon | 114.ribonucleic acid |
| 100.ultra-lipid* | 115.electrocardiograph |
| 101.centripetal force | 116.amniotic sac |
| 102.plates of parallax* | 117.hemoglobin |
| 103.nebula | 118.retroplex* |
| 104.particle accelerator | 119.antigen |
| 105.satellite | 120.recessive trait |

| Century Culture Names | Philosophy |
|---|---|
| 121.Gail Brennan* | 136.logistic heresy* |
| 122.Jackie Robinson | 137.creationism |
| 123.Houdini | 138.Goedel's theorem |
| 124.Ginger Rogers | 139.social constructionism |
| 125.Greta Garbo | 140.Platonic sense*<br>141.hermeneutics |
| 126.Dale Carnegie | |
| 127.Scott Joplin | 142.esoteric deduction* |
| 128.Rube Goldberg | 143.ghost in the machine |
| 129.George Gershwin | 144.Hegel |
| 130.Mae West | 145.Socrates |
| 131.Jesse Owens | 146.categorical imperative |
| 132.Oliver Marjorie* | 147.free will |
| 133.Louis Lapointe* | 148.Ayn Rand |
| 134.King Kong | 149.situational ethics |
| 135.P.T. Barnum | 150.Principia Mathematica |

*Note. * Indicates items that are foils (i.e., non-existent items:* 3, 7, 11, 18, 22, 26, 34, 38, 42, 47, 51, 59, 65, 68, 74, 79, 81, 86, 94, 100, 102, 110, 112, 118, 121, 132, 133, 136, 140, 142).

## Appendix F: Domain-specific risk-taking scale (DOSPERT) items used in Study 3.

Domain-specific risk-taking scale (DOSPERT; 30-item version, Blais & Weber, 2006)

For each of the following statements, please indicate the likelihood that you would engage in the described activity or behavior if you were to find yourself in that situation.

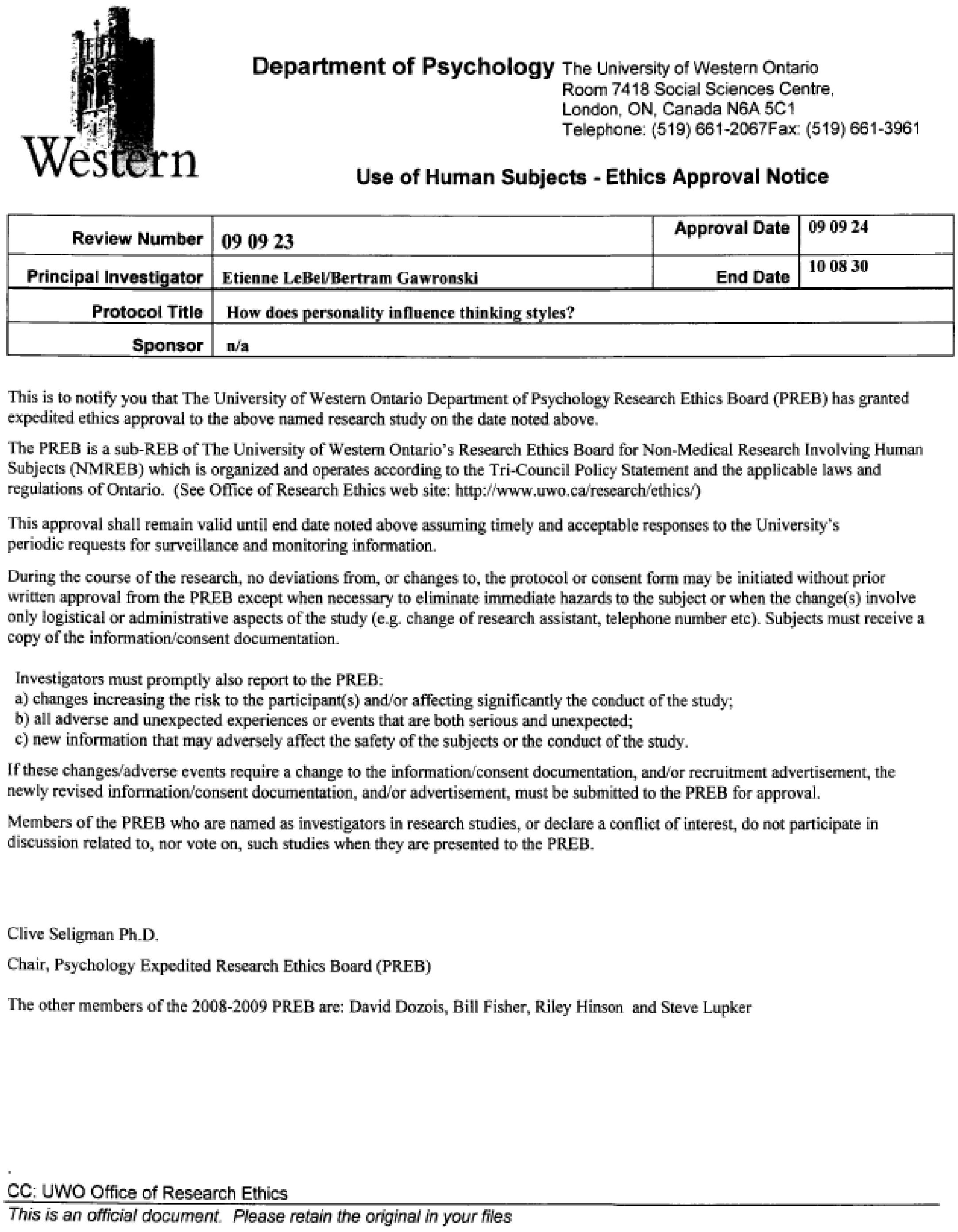
Provide a rating from using the following scale:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| *Extremely Unlikely* | *Moderately Unlikely* | *Somewhat Unlikely* | *Not Sure* | *Somewhat Likely* | *Moderately Likely* | *Extremely Likely* |

1. Admitting that your tastes are different from those of a friend. (S)
2. Going camping in the wilderness. (R)
3. Betting a day's income at the horse races. (F)
4. Investing 10% of your annual income in a moderate growth mutual fund. (F)
5. Drinking heavily at a social function. (H/S)
6. Taking some questionable deductions on your income tax return. (E)
7. Disagreeing with an authority figure on a major issue. (S)
8. Betting a day's income at a high-stake poker game. (F)
9. Having an affair with a married man/woman. (E)
10. Passing off somebody else's work as your own. (E)
11. Going down a ski run that is beyond your ability. (R)
12. Investing 5% of your annual income in a very speculative stock. (F)
13. Going whitewater rafting at high water in the spring. (R)
14. Betting a day's income on the outcome of a sporting event (F)
15. Engaging in unprotected sex. (H/S)
16. Revealing a friend's secret to someone else. (E)
17. Driving a car without wearing a seat belt. (H/S)
18. Investing 10% of your annual income in a new business venture. (F)
19. Taking a skydiving class. (R)
20. Riding a motorcycle without a helmet. (H/S)
21. Choosing a career that you truly enjoy over a more secure one.11 (S)
22. Speaking your mind about an unpopular issue in a meeting at work. (S)
23. Sunbathing without sunscreen. (H/S)
24. Bungee jumping off a tall bridge. (R)
25. Piloting a small plane. (R)
26. Walking home alone at night in an unsafe area of town. (H/S)
27. Moving to a city far away from your extended family. (S)
28. Starting a new career in your mid-thirties. (S)
29. Leaving your young children alone at home while running an errand. (E)
30. Not returning a wallet you found that contains $200. (E)

*Note.* E = Ethical, F = Financial, H/S = Health/Safety, R = Recreational, and S = Social.

# Appendix G: Ethics approval for Study1.



Western

**Department of Psychology** The University of Western Ontario
Room 7418 Social Sciences Centre,
London, ON, Canada N6A 5C1
Telephone: (519) 661-2067Fax: (519) 661-3961

**Use of Human Subjects - Ethics Approval Notice**

| | | | |
|---|---|---|---|
| **Review Number** | 09 09 23 | **Approval Date** | 09 09 24 |
| **Principal Investigator** | Etienne LeBel/Bertram Gawronski | **End Date** | 10 08 30 |
| **Protocol Title** | How does personality influence thinking styles? | | |
| **Sponsor** | n/a | | |

This is to notify you that The University of Western Ontario Department of Psychology Research Ethics Board (PREB) has granted expedited ethics approval to the above named research study on the date noted above.

The PREB is a sub-REB of The University of Western Ontario's Research Ethics Board for Non-Medical Research Involving Human Subjects (NMREB) which is organized and operates according to the Tri-Council Policy Statement and the applicable laws and regulations of Ontario. (See Office of Research Ethics web site: http://www.uwo.ca/research/ethics/)

This approval shall remain valid until end date noted above assuming timely and acceptable responses to the University's periodic requests for surveillance and monitoring information.

During the course of the research, no deviations from, or changes to, the protocol or consent form may be initiated without prior written approval from the PREB except when necessary to eliminate immediate hazards to the subject or when the change(s) involve only logistical or administrative aspects of the study (e.g. change of research assistant, telephone number etc). Subjects must receive a copy of the information/consent documentation.

Investigators must promptly also report to the PREB:
a) changes increasing the risk to the participant(s) and/or affecting significantly the conduct of the study;
b) all adverse and unexpected experiences or events that are both serious and unexpected;
c) new information that may adversely affect the safety of the subjects or the conduct of the study.

If these changes/adverse events require a change to the information/consent documentation, and/or recruitment advertisement, the newly revised information/consent documentation, and/or advertisement, must be submitted to the PREB for approval.

Members of the PREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the PREB.

Clive Seligman Ph.D.

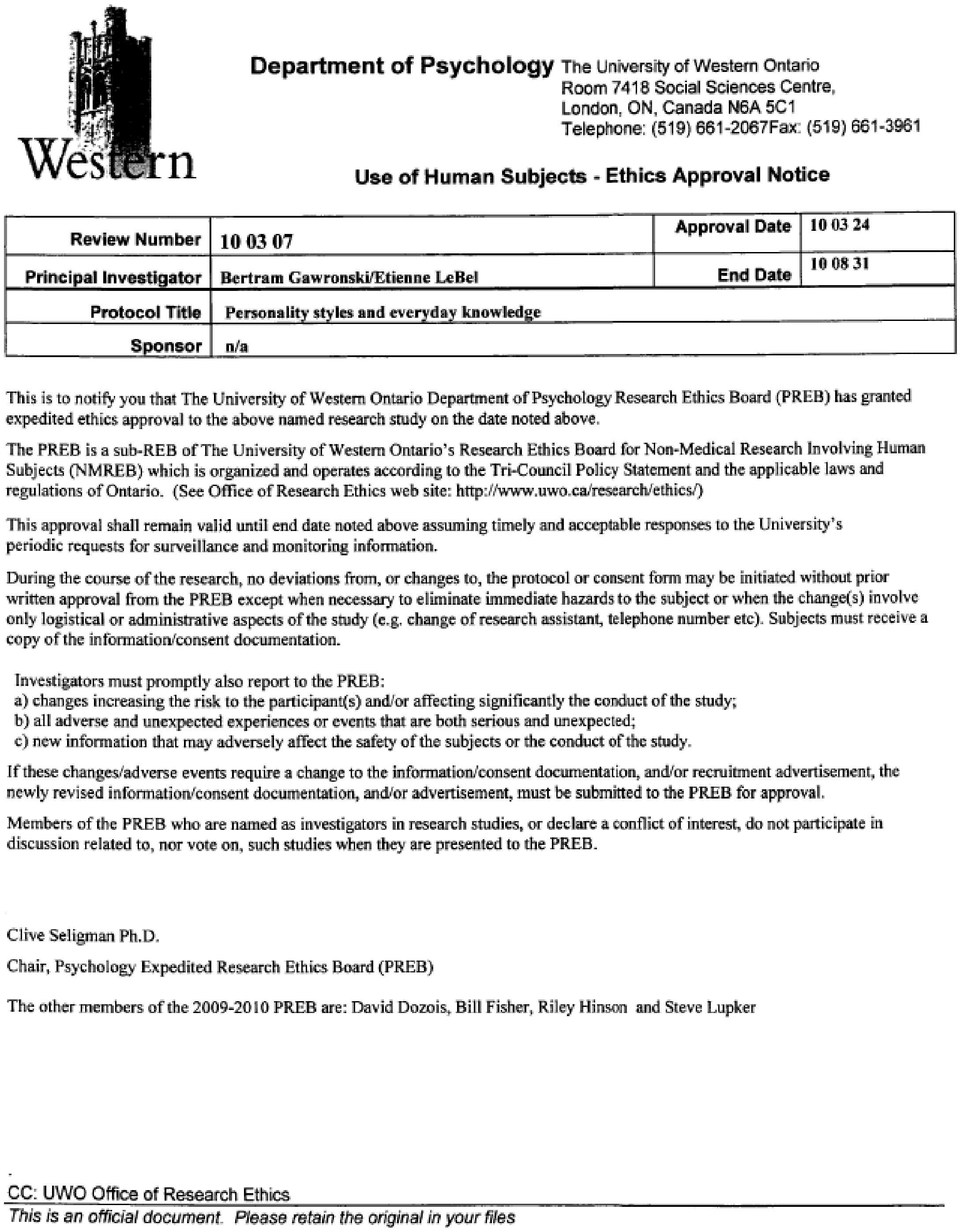
Chair, Psychology Expedited Research Ethics Board (PREB)

The other members of the 2008-2009 PREB are: David Dozois, Bill Fisher, Riley Hinson and Steve Lupker

CC: UWO Office of Research Ethics

*This is an official document. Please retain the original in your files*

## Appendix H: Ethics approval for Study 2.



Western

**Department of Psychology** The University of Western Ontario
Room 7418 Social Sciences Centre,
London, ON, Canada N6A 5C1
Telephone: (519) 661-2067Fax: (519) 661-3961

**Use of Human Subjects - Ethics Approval Notice**

| | | | |
|---|---|---|---|
| **Review Number** | 10 03 07 | **Approval Date** | 10 03 24 |
| **Principal Investigator** | Bertram Gawronski/Etienne LeBel | **End Date** | 10 08 31 |
| **Protocol Title** | Personality styles and everyday knowledge | | |
| **Sponsor** | n/a | | |

This is to notify you that The University of Western Ontario Department of Psychology Research Ethics Board (PREB) has granted expedited ethics approval to the above named research study on the date noted above.

The PREB is a sub-REB of The University of Western Ontario's Research Ethics Board for Non-Medical Research Involving Human Subjects (NMREB) which is organized and operates according to the Tri-Council Policy Statement and the applicable laws and regulations of Ontario. (See Office of Research Ethics web site: http://www.uwo.ca/research/ethics/)

This approval shall remain valid until end date noted above assuming timely and acceptable responses to the University's periodic requests for surveillance and monitoring information.

During the course of the research, no deviations from, or changes to, the protocol or consent form may be initiated without prior written approval from the PREB except when necessary to eliminate immediate hazards to the subject or when the change(s) involve only logistical or administrative aspects of the study (e.g. change of research assistant, telephone number etc). Subjects must receive a copy of the information/consent documentation.

Investigators must promptly also report to the PREB:
a) changes increasing the risk to the participant(s) and/or affecting significantly the conduct of the study;
b) all adverse and unexpected experiences or events that are both serious and unexpected;
c) new information that may adversely affect the safety of the subjects or the conduct of the study.

If these changes/adverse events require a change to the information/consent documentation, and/or recruitment advertisement, the newly revised information/consent documentation, and/or advertisement, must be submitted to the PREB for approval.

Members of the PREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the PREB.

Clive Seligman Ph.D.

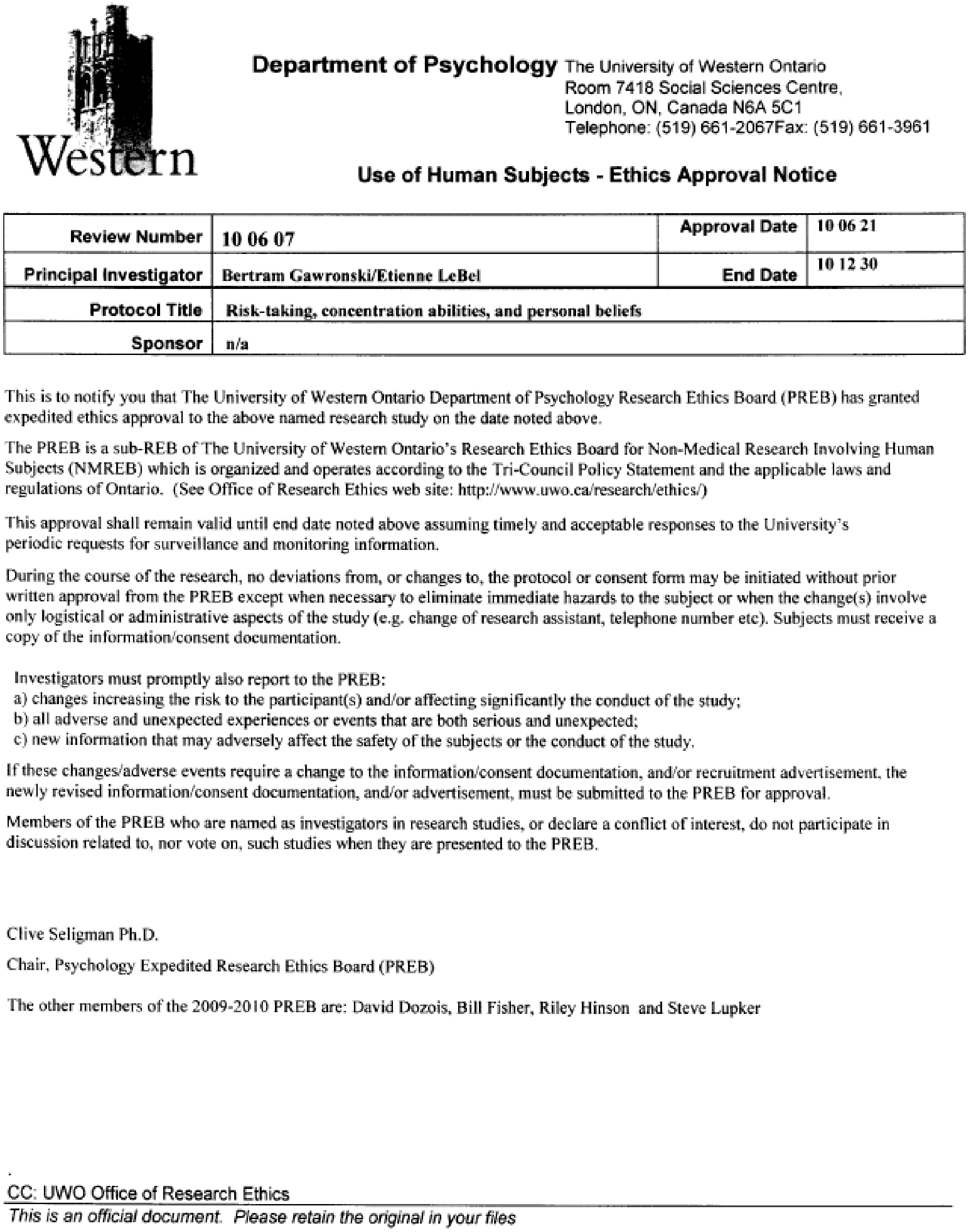
Chair, Psychology Expedited Research Ethics Board (PREB)

The other members of the 2009-2010 PREB are: David Dozois, Bill Fisher, Riley Hinson and Steve Lupker

CC: UWO Office of Research Ethics

*This is an official document. Please retain the original in your files*

# Appendix I: Ethics approval for Study 3.



Western

**Department of Psychology** The University of Western Ontario
Room 7418 Social Sciences Centre,
London, ON, Canada N6A 5C1
Telephone: (519) 661-2067Fax: (519) 661-3961

**Use of Human Subjects - Ethics Approval Notice**

| | | | |
|---|---|---|---|
| **Review Number** | **10 06 07** | **Approval Date** | **10 06 21** |
| **Principal Investigator** | **Bertram Gawronski/Etienne LeBel** | **End Date** | **10 12 30** |
| **Protocol Title** | **Risk-taking, concentration abilities, and personal beliefs** | | |
| **Sponsor** | **n/a** | | |

This is to notify you that The University of Western Ontario Department of Psychology Research Ethics Board (PREB) has granted expedited ethics approval to the above named research study on the date noted above.

The PREB is a sub-REB of The University of Western Ontario's Research Ethics Board for Non-Medical Research Involving Human Subjects (NMREB) which is organized and operates according to the Tri-Council Policy Statement and the applicable laws and regulations of Ontario. (See Office of Research Ethics web site: http://www.uwo.ca/research/ethics/)

This approval shall remain valid until end date noted above assuming timely and acceptable responses to the University's periodic requests for surveillance and monitoring information.

During the course of the research, no deviations from, or changes to, the protocol or consent form may be initiated without prior written approval from the PREB except when necessary to eliminate immediate hazards to the subject or when the change(s) involve only logistical or administrative aspects of the study (e.g. change of research assistant, telephone number etc). Subjects must receive a copy of the information/consent documentation.

Investigators must promptly also report to the PREB:
a) changes increasing the risk to the participant(s) and/or affecting significantly the conduct of the study;
b) all adverse and unexpected experiences or events that are both serious and unexpected;
c) new information that may adversely affect the safety of the subjects or the conduct of the study.

If these changes/adverse events require a change to the information/consent documentation, and/or recruitment advertisement, the newly revised information/consent documentation, and/or advertisement, must be submitted to the PREB for approval.

Members of the PREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the PREB.

Clive Seligman Ph.D.

Chair, Psychology Expedited Research Ethics Board (PREB)

The other members of the 2009-2010 PREB are: David Dozois, Bill Fisher, Riley Hinson and Steve Lupker

CC: UWO Office of Research Ethics

*This is an official document. Please retain the original in your files*

# Curriculum Vitae

**NAME** Etienne P. LeBel

## EDUCATION

| | |
|---|---|
| 2011 | Doctor of Philosophy (Ph.D.), Social Psychology<br>The University of Western Ontario; Supervisor: Dr. Bertram Gawronski |
| 2007 | Master of Science (M.Sc.), Social Psychology<br>The University of Western Ontario; Supervisor: Dr. Bertram Gawronski |
| 2005 | Bachelor of Arts (B.A.), Honors Psychology Program<br>University of Waterloo; Supervisor: Dr. Steve Spencer |

## HONOURS & AWARDS (selection)

| | |
|---|---|
| 2011-2013 | SSHRC Post-Doctoral Fellowship Award (Value: $81,000)<br>Social Sciences and Humanities Research Council (SSHRC) |
| 2007-2010 | Canadian Graduate Scholarship Doctoral Scholarship (Value: $105,000)<br>Social Sciences and Humanities Research Council (SSHRC) |
| 2006-2007 | Canadian Graduate Scholarship Master's Scholarship (Value: $17,500)<br>Social Sciences and Humanities Research Council (SSHRC) |

## PUBLICATIONS (selection)

**LeBel, E. P**., & Paunonen, S. V. (2011). Sexy but often unreliable: Impact of unreliability on the replicability of experimental findings involving implicit measures. *Personality and Social Psychology Bulletin, 37,* 570-583.

**LeBel, E. P.** (2010). Attitude accessibility as a moderator of implicit and explicit self-esteem correspondence. *Self and Identity, 9*, 195-208.

**LeBel, E. P**., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. *Journal of Experimental Social Psychology, 45*, 1291-1294.

**LeBel, E. P**., & Gawronski, B. (2009). How to find what's in a name: Scrutinizing the optimality of five scoring algorithms for the name-letter task. *European Journal of Personality, 23*, 85-106.

Gawronski, B., & **LeBel, E. P.** (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology, 44*, 1355-1361.

Gawronski, B., **LeBel, E. P**., & Peters, K.P. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2*, 181-193.

## TEACHING EXPERIENCE (selection)

| | |
|---|---|
| 2011 | Course Instructor, The University of Western Ontario<br>Course: Psychology of Persuasion (Distance Education) |
| 2010-2011 | Graduate-level Teaching Assistant, The University of Western Ontario<br>Course: Research Design and Statistics (Graduate Level) |