A PSYCHOMETRIC INVESTIGATION OF A SELF-CONTROL SCALE: THE RELIABILITY AND VALIDITY OF GRASMICK ET AL.'S SCALE FOR A SAMPLE OF INCARCERATED MALE OFFENDERS

By

Christopher L. Gibson

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Criminal Justice

Under the Supervision of Dr. Ineke Haen Marshall

Omaha, Nebraska

August, 2005

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

DISSERTATION TITLE

A Psychometric Investigation of a Self-Control Scale: The Reliability and Validity of Grasmick et al.'s Scale for a Sample of Incarcerated Male Offenders

BY

Christopher L. Gibson

SUPERVISORY COMMITTEE:

APPROVED	DATE
Signature	6/20/2005
Ineke Haen Marshall, Ph.D. Typed Name	
Signature	6/20/2005
Miriam A. DeLone, Ph.D. Typed Name	
Signature	6/20/05
Dennis W. Roncek, Ph.D. Typed Name	
Signature	6/20/05
Russell L. Smith, Ph.D. Typed Name	
Signature	6/20/05
Jihong Zhao, Ph.D. Typed Name	
Signature	
Typed Name	

University of Nebraska Omaha

A PSYCHOMETRIC INVESTIGATION OF A SELF-CONTROL SCALE:

THE RELAIBILITY AND VALIDITY OF GRASMICK ET AL.'S SCALE FOR A

SAMPLE OF INCARCERATED MALE OFFENDERS

Christopher L. Gibson, Ph.D.

University of Nebraska, 2005

Advisor: Ineke Haen Marshall, Ph.D.

ABSTRACT

. This dissertation investigates the reliability and internal validity of one of the most commonly used measures of self-control, i.e., Grasmick et al.'s 24 item self-control scale. Using a sample of 651 male offenders residing in the Diagnostic and Evaluation center in Lincoln, Nebraska, this dissertation explores the psychometric properties of Gramsick et al.'s scale by answering the following questions. First, is Grasmick et al.'s scale reliable? Second, does it show observed differences across black and white offender groups? Third, is Grasmick et al.'s scale unidmensional? Fourth, is Grasmick et al.'s scale multidimensional? Fifth, can Grasmick et al.'s scale items discriminate among levels of self-control for a sample of incarcerated offenders? Sixth, do respondents' levels of self-control affect survey responses? Finally, are Grasmick et al.'s scale items invariant across black and white offender groups? These questioned are answered using several analytic methods including Cronbach's reliability coefficients, exploratory and confirmatory factor analyses, and a Rasch rating scale model.

Results from this study lead to several interesting conclusions. First, Grasmick et al.'s scale has high reliability for a sample of incarcerated male offenders. Second, racial differences were observed, but these differences were not what would be expected according to self-control theory. Second, confirmatory factor analysis and a Rasch model confirmed that Grasmick et al.'s scale was not measuring one construct, but was shown to measure six correlated dimensions. Third, a Rasch analysis showed that items were able to discriminate among offenders' levels of self-control. Fourth, a Rasch analysis revealed that level of self-control affects responses to survey items: low self-control offenders have unexpectedly lower scores. Finally, several of Gramsick et al.'s scale items exhibited differential item function or bias across black and white offenders. Directions for future research are discussed.

ACKNOWLEDGEMENTS

"One thing I have learned in a long life: that all our science, measured against reality, is primitive and childlike- and yet is the most precious thing we have."

- *Albert Einstein (1879-1955)*

Here I want to show my profound gratitude to those who have opened my eyes and led me to respect the power, beauty, parsimony, and importance of science; but yet have equally encouraged me to realize the limits, misuses, and abuses of this method. Many scholars, researchers, and laymen have directly and indirectly provided invaluable ideas, information, knowledge, and science that has challenged an affected not only my pathway to becoming a social scientist, but most importantly my worldview. To acknowledge every individual would be a daunting task. Therefore, I would like to generally thank all of those who have contributed to my evolution in becoming a social scientist and my growth as a person.

Particularly, I would like to give a special thanks to Dr. Ineke Haen Marshall, Dr. Dennis W. Roncek, Dr. Jihong Zhao, Dr. Miriam A. DeLone, and Dr. Russell Smith, for their insights, guidance and patience through this process. Furthermore, I would like to thank Dr. Julie Horney for providing access to her data for my dissertation. Finally, I would like to thank Leah E. Daigle for her love, support, and friendship.

Two more scholars and friends deserve acknowledgements. In pursuit of a MA degree in criminology, Dr. Steve Tibbetts and Dr. John Wright were most influential in my decision to choose an academic trajectory. How both of them shared their passion for knowledge and science with me will never be forgotten.

DEDICATION

I want to dedicate this research to my immediate family. These people have always given me their unconditional love, acceptance, encouragement, and support. They include my father, Harry, my mother, Nancy, my twin brother, Derrick, and my grandmother, Imogene.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

INTRODUCTION

Measurement is imperative, inevitable, and consequential across research in both physical and social science disciplines. Scientists must find ways to quantify particular phenomena of interest as accurately as possible before achieving their central research goals or before testing hypotheses. Each area of scientific exploration develops its own measurement procedures and devices. Physics, for example, uses established measures for mass, time, electric current, and luminous intensity. Neurologists and neuropsychologists use brain imaging techniques or instruments to assess the presence of brain abnormality, dysfunction, and tumors. For example, structural brain imaging instruments consist of computerized tomography (CT) and magnetic imaging (MRI), while functional brain imaging techniques consist of positron emission tomography and regional cerebral blood flow (RCBF) (Raine, 1993: 130-153). In contrast, measurement of psychological and social concepts in social science disciplines typically takes the form of a mark on a questionnaire, behavior documented in an observational study, answers obtained through an interview, or official records recorded by agencies and institutions (Carmines and Zeller, 1979; DeVellis, 1991). Although researchers in the physical sciences share a general sense of confidence in their measures, this has not been the case in the social sciences where psychological and psychosocial measurements are used (Bond and Fox, 2001: 2-3).

Blalock (1968: 6) stated that social science theorists "often use concepts that are formulated at rather high levels of abstraction," and "the problem of bridging the

gap between theory and research is then seen as one of measurement error" (Blalock, 1968: 12). Blalock's statements proposed approximately thirty years ago suggest that operationalization and measurement in social sciences are challenging tasks that involve creating empirical indicators to represent elusive concepts, frequently leading to measurement error. Blalock's observations concerning measurement, theory, and research are still concerns for social scientists today.

Measurement validity and reliability are at the core of the research process, both having implications for theory and the interpretation of empirical findings. Inaccurate and unreliable measures may lead to many unintended problems. The use of "poor" measurement can impede the ability to make informed decisions that affect both theory and policy. For example, consequences of inadequate measurement might include inaccurate diagnosis of mental illness, misspecification of the empirical validity of a theory, or flawed police officer hiring decisions, to name only a few. Thus, theory specifying the relationship between concepts and empirical indicators is just as vital to social science research as the substantive components of theory that specify propositions or relationships between concepts (Carmines and Zeller, 1979: 11).

To prevent or minimize such problems, social scientists assess the validity and reliability of their measures using psychometric analysis. Psychometrics encompasses a wide range of methods and statistical techniques to examine measurement quality. Ranging from exploratory and confirmatory factor analysis to more modern techniques such as Rasch modeling, researchers use these tools to empirically derive the most accurate and reliable measures of theoretical concepts

(Andrich, 1988; Bollen, 1989; Bond and Fox, 2001; Kline, 1998). Psychometrics has a long history in disciplines such as psychology and education where researchers spend considerable effort attempting to quantify a range of intangible human traits. One excellent example of this can be found in the voluminous body of literature on the measurement of intelligence (See Gould, 1996).

Unlike some social science disciplines, criminology has not fully embraced the long-standing tradition of psychomterics. As such, rigorous psychometric assessments of measures representing elusive theoretical phenomena are often taken for granted in criminological research. Measures are sometimes employed haphazardly based on face validity, theoretical arguments, or minimal empirical examinations alone when testing relational hypotheses between criminological concepts.

Most criminological theories including contemporary strain theory (Agnew, 1992), social control theory (Hirschi, 1969), and self-control theory (Gottfredson and Hirschi, 1990) rely on concepts that are not directly observable (See Duncan, 1984). As such, criminologists, like other social scientists, have to rely on indirect indicators to represent ambiguous theoretical concepts. In the absence of well defined concepts, the result is often the inability of social scientists to reach a consensus on how to best pursue operationalization and measurement. This can often result in measures with questionable validity, unknown psychometric properties, and a general inability to compare findings across studies due to differential operationalizations of the same concept. Criminologists disproportionately tend to invest more time testing relational

propositions between constructs from these theories rather than focusing on the development of quality measures.

Early quantification and replication efforts should focus on a basic, yet fundamental, question: Are criminological measures accurately measuring what criminology theories imply? As will be shown, the concept and measure under investigation in this dissertation represent an excellent example in criminology for which a lack of conceptual clarity exists and, consequently, a lack of psychometric agreement emerges. This lack of conceptual clarity is thoroughly documented in later chapters.

A recent exception to the lack of psychometric investigations of constructs deemed to be important in the etiology of criminal behavior is the construct of self-control. This construct is entrenched in a hotly-debated theory known as self-control theory proposed approximately a decade ago in Gottfredson and Hirschi's (1990) book titled *A General Theory of Crime*. While the formulation of their theory is the subject of much criticism and empirical scrutiny, their theory more recently has sparked psychometric interest among researchers. Perhaps, this interest is due to the compelling explanatory power Gottfredson and Hirschi (1990) attribute to their construct of self-control.

According to the theory, low self-control is a disposition which forms early in life and consists of six elements—impulsivity, risk seeking, temperament, self-centeredness, preference for simple tasks, and preference for physical activities—that coalesce in similar individuals (Gottfredson and Hirschi, 1990). For example, those who are impulsive are more likely to also be risk seekers, self-centered and so on. To

date, there are several psychometric evaluations of the dimensionality of self-control (Arneklev, Grasmick, and Bursik, 1999; Grasmick, Tittle, Bursik, and Arneklev, 1993; Longshore and Turner, 1998; Longshore, Turner, and Stein, 1996; Piquero and Rosay, 1998; Piquero, MacIntosh, and Hickman, 2000). From these studies, an empirical debate centers on whether self-control is a unidimensional or multidimensional construct, i.e., whether self-control reflects one or several traits? Perhaps, this is both a conceptual and empirical question.

This dissertation will address empirically the dimensionality debate on self-control with data collected on a commonly used measure of the self-control construct, i.e., the Grasmick et al. (1993) scale. In addition, this dissertation will be the first extensive psychometric assessment of this self-control measure on a sample of incarcerated male offenders. Data for this dissertation are from a large random sample of incarcerated male offenders residing at the Diagnostic and Evaluation center in Lincoln, Nebraska during October 1997 and December 1998. Support for this research was provided by a grant from the National Institute of Justice awarded to Dr. Julie K. Horney, funded under Grant 96-IJ-CX-0015.

AN INTRODUCTION TO SELF-CONTROL AND THE DIMENSIONALITY DEBATE

In Gottfredson and Hirschi's (1990) "general theory," criminal behavior has six basic elements that underlie criminal conduct. They contend that offenders will resemble the nature of crime in that they "tend to be impulsive, insensitive, physical (as opposed to mental), risk-taking, short-sighted, and nonverbal" (Gottfredson and Hirschi, 1990: 90). They argue that this constellation of six elements develops in

early childhood due to a lack of adequate child rearing practices—deficient behavioral monitoring, inability to recognize deviant behavior when it occurs, and not appropriately punishing the behavior when it occurs—and that it will remain relatively stable throughout life. Gottfredson and Hirschi (1990) label this attribute low self-control.

In theory, low self-control, in the presence of opportunity, should account for disparities in offending rates and analogous behaviors such as promiscuous sex, gambling, smoking, involvement in accidents, and academic dishonesty. According to Gottfredson and Hirschi (1990: 96), the presence of low self-control will hinder "the achievement of long-term individual goals." Furthermore, Gottfredson and Hirschi (1990: 96) argue that the elements of low self-control, "impede educational and occupational achievement, destroy interpersonal relations, and undermine physical health and economic well being." From these claims, Gottfredson and Hirschi (1990) reverse the causal order of many theories of crime and delinquency by endorsing a population heterogeneity perspective; that is, by claiming that the relationships between social failure, prior criminal behavior, and antisocial behaviors are not causal but spurious due to an underlying latent trait.

To date, empirical studies testing self-control theory largely attend to two core propositions. The most prevalent is testing relational propositions linking low self-control to a host of outcomes (Pratt and Cullen, 2000). As such, these studies reveal moderate support for the theory and conclude that a disposition of low self-control has associations with criminal and imprudent behaviors and many negative social outcomes (Arneklev, Grasmick, Tittle, and Bursik, 1993; Evans, Cullen, Burton,

Dunaway, and Benson, 1997; Gibson and Wright, 2001; Grasmick, Tittle, Bursik, and Arneklev, 1993; Hay, 2001; Keane, Maxim, and Teevan, 1993; LaGrange and Silverman, 1999; Piquero and Tibbetts, 1996; Sellers, 1999; Wood, Pfefferbaum, and Arneklev, 1993). While these studies are important and implicate the role of low self-control, many of them presume self-control scales to be both valid and reliable because conventional analyses e.g., internal consistency and principal components analyses, reveal that these measures are unidimensional and reliable.

Another group of studies focus on the measurement and dimensionality propositions of self-control. In the first attempt to measure self-control, Grasmick and colleagues (1993) develop a twenty-four item, self-report scale specifically designed to reflect Gottfredson and Hirschi's (1990) concept. Grasmick and colleagues (1993) use exploratory analyses to conclude that their scale appears to measure one underlying factor that is consistent with self-control theory.

Since the publication of Grasmick et al.'s scale, a number of studies testing its psychometric properties have been published. Overall, these studies have led to inconsistent conclusions (Arneklev, Grasmick, and Bursik, 1999; Grasmick, Tittle, Bursik, and Arneklev, 1993; Longshore and Turner, 1998; Longshore, Turner, and Stein, 1996; Piquero and Rosay, 1998; Piquero, MacIntosh, and Hickman, 2000). The following conclusions have been drawn. First, evidence has emerged of good scale reliability and the presence of one underlying factor or a unidimensional structure (Nagin and Paternoster, 1993; Piquero and Tibbetts, 1996; Piquero and Rosay, 1998). Second, evidence has supported the presence of multiple dimensions, rejecting the notion of unidimensionality (Longshore, Turner, and Stein, 1996;

Vazsonyi, Pickering, Junger, and Hessing, 2001). Third, Arneklev, Grasmick , and Bursik's (1999) confirmatory factor analyses have shown that six unique dimensions exist that can be explained by a common, second-order factor.

While conflicting results exist regarding the scale's dimensionality, these findings are largely from classical factor analytic methods that are test-based. Test-based, classical approaches have only been able to assess how scale items relate to each other, thus, not considering the interaction between persons and items. In other words, the statistical methods that produce what is known about Grasmick et al.'s scale do not separate person-ability from item difficulty. Conventional factor analytic approaches can not provide information on how people respond to items. Therefore, respondent characteristics cannot be separated from test characteristics. Each can only be interpreted in the context of the other (Bond and Fox, 2001; Hambleton, Swaminathan, and Rogers, 1991; Wright and Masters, 1982). A respondent's ability, e.g., level of self-control, is defined only in terms of a particular test, e.g., Grasmick et al.'s scale. When the test is difficult or has items that are difficult to endorse, i.e., with which to agree, then the respondent may appear to have a low score, in this case high self-control. When the test is easy or items are easy to endorse, the respondent may appear to have a high score, in this case low self-control (Hambleton, Swaminathan, and Rogers, 1991). In both cases, it is important for researchers to use methods to detect whether or not items are measuring the full range of abilities within a sample of respondents. This problem can be addressed with more modern psychometric approaches that use methods such as Rasch modeling. As such, the

Rasch model considers the ability of persons and difficulty of items, whereas, factor analysis only assesses correlations between items.

One of the most current empirical statements regarding Grasmick et al.'s scale takes the above concerns into account by using a Rasch model. Piquero and colleagues (2000) show that Grasmick et al.'s scale does not conform to a mathematically defined unidimensional model after jointly considering item difficulty and person ability. Furthermore, responses to self-control items are contingent on an individual's underlying level of ability. For example, Piquero and colleagues (2000) find that scale items function differently for low and high self-control groups. This finding could only emerge from using a Rasch model.

In sum, two consequences emerge from studies testing the dimensionality of self-control using Grasmick et al.'s scale. First, these studies increase the exposure to and importance of psychometric evaluation in criminology. Second, they enhance a debate on the conceptual and empirical dimensionality of self-control and, consequently, the scale's validity. From a construct validation perspective (Cronbach and Meehl, 1955; Loevinger, 1958), however, much work remains to be done. The purpose of this study is to accomplish some of this work by replicating and also extending past psychometric studies of Grasmick et al.'s self-control scale.

PROJECT SIGNIFICANCE AND CONTRIBUTION

This dissertation provides a reliability assessment and construct validation test of Grasmick et al.'s (1993) self-control scale using multiple methods that encompass all techniques used in past studies on this self-control scale. In doing so, this dissertation will test different conceptualizations of the self-control construct. This

dissertation will also journey beyond past studies to make several original contributions to the empirical debate surrounding this scale. This dissertation attempts to understand the validity and reliability of Grasmick et al.'s scale for a sample of incarcerated male offenders and across racial groups of these offenders.

Data for this dissertation are from a large random sample of incarcerated offenders collected from 1997 through 1998 during their initial stay at the Diagnostic and Evaluation Center in Lincoln, Nebraska while waiting for permanent assignments to state institutions (Horney, 2001). The selection of this particular sample is important because much research on Grasmick et al.'s scale uses student and community samples. The use of such samples affects the known usefulness and generalizability to other populations that have higher levels of involvement in crime and delinquency. The data in this dissertation allow the examination of Grasmick et al.'s scale for individuals who have engaged in severe and chronic criminal behavior.

It is acknowledged that generalizability of the findings from this dissertation is limited. The sample of offenders used in this study may share unique characteristics that distinguish them from other offender populations that are not incarcerated or receiving alternative forms of punishment, presenting a selection bias. Nevertheless, using such a sample can be viewed as a unique opportunity to assess the measurement quality of a scale originally validated on a community sample. For example, Grasmick et al.'s scale items may not be well suited to measure self-control for individuals that are likely to already have very low levels of self-control. As such, this sample enhances the importance of this dissertation.

Exploring the psychometric properties of Grasmick et al.'s (1993) scale across racial categories for a sample of offenders is an important contribution to the dimensionality debate for several reasons. First, Gottfredson and Hirschi (1990) make predictions about how self-control should vary across racial categories. For example, Gottfredson and Hirschi (1990) state that "...differences in self-control probably far outweigh differences in supervision in accounting for racial or ethnic variations [in crime]," implying that minority groups will have substantially lower levels of self-control. Therefore, the composite self-control scale and its dimensions should reveal differences across racial categories among a group of serious offenders. Second, for a scale to be objective its psychometric properties should be similar across different subgroups.

Several analytic techniques in this dissertation have only rarely been used in criminology. For example, this dissertation adds to the small but growing body of knowledge on how respondent ability can influence survey item responses. This will be accomplished by employing a Rasch statistical model that has been recently coined as the most objective measurement validation method (Fox and Bond, 2001). Such models avoid problems associated with classical test theory of measurement by considering person ability and item difficulty parameters or the interaction between persons and scale items, which traditional exploratory and confirmatory factor analytic methods can not accomplish.

A dissertation testing the dimensionality of Grasmick et al.'s scale is much more than a statistical exercise that has no bearing on the merit and substance of self-control theory. Including the auhor of this dissertation, some argue (Piquero et al.,

2000) that such a study is important for several reasons. First, it is imperative from a construct validity framework that replication of previous studies be conducted using different samples so that increased confidence can be gained on how items of a scale represent the construct under investigation (Loevinger, 1958; Nunally and Bernstein, 1994). Second, the concept of self-control is at the core of Gottfredson and Hirschi's (1990) self-control theory. As such, they propose specific predictions pertaining to the internal structure of their main construct, i.e., self-control. Hence, the psychometric importance of its indicators is of crucial importance. Third, such a study has implications for the predictive validity of the self-control construct. Hence, the psychometric importance of its indicators is of crucial importance. For example, if several dimensions of self-control exist, some dimensions of self-control could be more important than others in predicting outcomes such as crime and deviance. Finally, this dissertation will broaden criminologists' understanding of psychometric theory as it relates to criminological scales of measurement.

The subsequent chapters include the following discussions. Chapter 2 provides an overview of psychometric theory, a discussion on measurement validity and reliability, and the use of psychometrics in criminology. Chapter 3 discusses the construct of self-control, its conceptualization, and documents the reliability and validity of Grasmick et al.'s scale. The study's research design and analytic strategy will be discussed in Chapter 4. This chapter will include a discussion of the data, sample, and measures being used. Additionally, it describes the statistical techniques that will be used and compares the benefits and appropriateness of each. Most notably, Chapter 4 discusses the Rasch model and how it is better equipped to meet

standards of measurement in psychometric theory. Chapter 4 also discusses why and how the Rasch model is able to answer questions about Grasmick et al.'s scale that other techniques such as classical confirmatory factor analysis cannot. Finally, Chapters 5 and 6 discuss results and conclusions from the analyses, respectively.

CHAPTER 2:

PSYCHOMETRICS AND ITS HISTORY: RELIABILITY AND VALIDITY IN SOCIAL SCIENCE MEASUREMENT

The achievement of precise and accurate measurement in the social sciences is a complex problem facing both applied and pure researchers pursuing explanations of social and psychosocial phenomena (Blalock, 1982; Duncan, 1984). Some scientific disciplines are less fraught with having to question the validity and reliability of their measurement practices. A physical scientist can often measure the length of an object by using a standardized metric tool via inches, centimeters, or millimeters to derive a meaningful numeric score, compare it to the length of another object, take the difference and make various inferences with relative simplicity (Piquero et al., 2002). For example, theoretical physics is a science most known for its coherent systems of measures that have powerful dimensional properties (Duncan, 1984: ix). Social scientists usually do not have such measures at their disposal.

Social science measures often do not possess measurement properties that are as desirable and straightforward as those used in physical sciences. A criminologist can't journey to the closest criminological hardware outlet and invest in a delinquency measuring tape to use when measuring his or her subjects' behaviors (Hickman, Piquero, and Piquero, 2004). As such, a host of problems challenge social scientists when attempting to measure indicators of their concepts. One particular quandary in social science disciplines, although not foreign to other disciplines, is that what is measured is not typically a physical attribute, but rather an abstract entity that cannot be directly observed. For example, in assessing whether a test of "self-

control" is actually measuring "self-control" a social scientist cannot compare and contrast a respondent's score on the test directly with his or her actual level of self-control. The social scientist is restricted to observing how test scores differentiate between low and high self-control individuals according to some other ideas of how low self-control people should behave or what attitudes they might harbor. Thus, valid and reliable measurement of constructs pose challenges to the most seasoned social scientists that devote their careers to creating and empirically testing measures of abstract constructs.

As in other scientific disciplines, measurement in social science should entail both a conceptual and quantitative process (See Blalock, 1982; Rust and Golombok, 1999). Blalock (1982) made two very important points concerning conceptualization and measurement in social science. First, conceptualization, as defined by Blalock (1982: 11), "involves a series of processes by which theoretical constructs, ideas, and concepts are clarified, distinguished, and given definitions that make it possible to reach a reasonable degree of consensus and understanding of the theoretical ideas we are trying to express." In other words, it is necessary that theoretical constructs are defined as clearly and concisely as possible to avoid misinterpretation and poor operationalization. Second, measurement, as defined by Blalock (1982: 11), "refers to the general process through which numbers are assigned to objects in such a fashion that it is also understood just what kinds of mathematical operations can be legitimately used..." In sum, conceptualization implies a theoretical process which guides research operations, and measurement entails the linkage between physical research operations and mathematics. The complete process should involve a linkage

between three things: theoretical constructs, physical measurement operations, and statistical operations. It is necessary that careful attention be given to all three elements.

The development of measures in social science requires a meticulous, multi-stage process that can require much research. Once developed, social scientists must subject their measures to a series of reliability and validity assessments across many samples before establishing measurement quality. Thorough investigations of the quality of measuring instruments are essential and must not be overlooked. While rigorous standards for measurement quality appear to be the rule in social science disciplines such as education and psychology (See Bond and Fox, 2002), they, unfortunately, tend to be the exception in criminology.

One major advantage of quality measurement is that it takes the guess work out of scientific observation (Blalock, 1982). Scientific statements must be independently confirmable by other scientists. The principle is violated if scientists can not agree on a measuring device. For example, it would be nearly impossible, and definitely premature, to come to any foregone empirical conclusion about the effect of self-control on delinquency and crime until self-control can be measured in a way that is agreed upon by those testing relational propositions of Gottfredson and Hirschi's (1990) general theory of crime (See Pratt and Cullen, 2000).

A theory can be tested only to the extent that its concepts can be adequately measured. With this said, it could be argued that measurement is *the* most serious issue in social science. According to Nunnally and Bernstein (1994: 7), "scientific results inevitably involve functional relations among measured variables, and the

science of psychology can progress no faster than the measurement of its key variables." Much the same can be said for the discipline of criminology.

This chapter will discuss several aspects of psychometrics. First, it will define psychometrics. Second, it will provide a brief history of the evolution of psychometrics as a discipline and a discussion of its common uses in academic and applied settings today. Third, the chapter will discuss several important concepts, theories, and methods that are embodied by psychometrics. The discussion will lay the foundations of the current dissertation. In particular, a discussion of measurement reliability and validity of measurement will be provided. Finally, the use of psychometrics in criminology will be discussed.

Defining Psychometrics

Historically, psychometrics has been referred to as an area of study that sets forth standards and tools for assessing measurement quality in psychology to aid in achieving precise, accurate, and objective measures. Several definitions for the word psychometrics exists. As stated on the cover of Rust and Golombok's (1999) book, psychometrics is "the science of psychological assessment." As defined by the Chambers Twentieth Century Dictionary, psychometrics is the " branch of psychology dealing with measurable factors', but also as the occult power of defining the properties of things by mere contact" (as cited in Rust and Golombok, 1999: 4) Others have indicated that psychometrics is merely"... the measurement of human characteristics..." (http://www.fordham.edu/aps/whatpsy.html).

Psychometrics is a complex discipline with an extensive history of improvement and refinement over the past century that is not made apparent from the

above definitions. The study of psychometrics consists of a general collection of techniques for evaluating the development and use of psychological traits, axioms, and theories. Also known as test theory, psychometrics consists largely of both applied mathematics and statistics used together to solve a vast array of measurement problems. Psychometrics has been the origin of intelligence testing, personality testing, and vocational testing, and has contributed to advances in psychological measurement ever since its inception (http://www.fordham.edu/aps/whatpsy.html).

Psychometricians often seek to answer questions about the quality of multiple-item measures. The quality of a 24 item measure (Grasmick, 1993) of Gottfredson and Hirschi's (1990) construct of self-control is the main focus of this dissertation. Hypothetically, several questions could be proposed by psychometricians regarding the Grasmick et al. scale. First, do different items measure different dimensions of the trait? how should test scores be created from item responses? Second, do item responses add up to make a general self-control score? Third, how accurately do the items measure the concept? Fourth, how many items are needed to measure the trait? Finally, is the measurement instrument equally valid and reliable across different groups of people? In confronting such issues, psychometricians use quantitative techniques that have been created to refine, formalize, and clarify questions such as those above into more precise statements to provide empirical answers. The most important of these approaches will be discussed and used in this dissertation.

Regardless of the semantics used in defining psychometrics, psychometricians share a common set of fundamental ideas about the qualities that a measurement instrument should possess: the measure should be reliable, valid, and free from bias.

Ultimately, the goal is to achieve objective measurement by subjecting a measure to a number of logical and quantitative analyses. The Program Committee of the Institute for Objective Measurement (IOM) defined objective measurement as follows:

> Objective measurement is the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured. An objective measurement estimate of amount stays constant and unchanging (within the allowable error) across the persons measured, across different brands of instruments, and across instrument users. The goal of objective measurement is to produce a reference standard common currency for the exchange of quantitative value, so that all research and practice relevant to a particular variable can be conducted in uniform terms. Objective measurement research tests the extent to which a given number can be interpreted as indicating the same amount of the thing measured, across persons measured, and brands of the instrument (http://www.rasch.org/define.htm).

Research on objective measurement is directed at testable hypotheses concerning the quantitative status of psychosocial variables; the ultimate goal being the creation of standardized measures free from bias. In attempting to achieve such goals, research investigations of this type often start out with a measurement instrument, data, and a theory. These three components are used together to make improvements on one another when needed. The process of deriving an objective measure may take years if not decades to achieve, extensive financial support to fund, and many replication efforts to produce (http://www.rasch.org/define.htm).

Modern day measurement and psychometric theory encompass a wide array of purposes that are academic as well as pragmatic, with the ultimate goal of achieving objective measures. For example, students are tested to monitor their performance and achievement, those of legal driving age are subjected to written and practical tests to obtain drivers licenses, job entrance and promotions are often determined through

performance on skills tests and personality instruments, prospective graduate students are often expected to score within a particular numerical range on the Graduate Record Examination for entrance into graduate school, and risk assessment instruments are administered to convicted criminals for placement purposes. Despite the broad assortment of psychometric applications, all measures when subjected to psychometric analysis should share a general set of characteristics: reliability, validity, and be unbiased.

Tracing the Evolution of Psychometrics

Social and psychological measurement is pervasive in modern society, but it has historical roots in many practical purposes in ancient cultures dating back to early China. The Chan dynasty during 1000 B.C. utilized assessment instruments to screen officials of the emperor every three years requiring the demonstration of proficiency in a wide array of skills and abilities including: arithmetic, horsemanship, music, and writing. Similar to modern times, procedures, although rudimentary, were invoked to enhance confidentiality, validation, and standardization. This general framework has remained stable for over 3,000 years and was in extensive use on other continents before the industrial revolution (Rust and Golombok, 1999: 4).

Duncan (1984) traced the invention of social measurement back to mid 400 B.C where measurement techniques were used by the Greeks and Romans to solve practical problems. For example, Duncan (1984: 1-3) identified that the Greeks used voting as a practical way to measure and replace earlier methods for ascertaining the collective preference for election and military decisions. Albeit a crude measure, judges were used to record volume of applause for electing senate members to serve

on councils. Duncan (1984) also noted that this example was one of the earliest, traceable forms of social measurement using a psychophysical method because the quantity directly measured (perceived loudness of applause) is not the one of concern. Although it is unknown how the judges recorded applause, it is known that they were using the method to measure votes for prospective senators.

While some of the earliest documented forms of social measurement were used for practical purposes such as voting, it wasn't until an eruption of mathematical social science in France during the late 18th century that social measurement was to be systematically pursued. Some of the fundamental inventions that identified basic ideas of social science measurement that precede modern science were: counting to measure groups size as an aid to taxation or size of military forces; defining or labeling social rank as documented by the Greeks and Romans; and appraising people according to games, contests, achievement, and grades, as in the ancient Chinese employment examinations and the Greek sporting events (Duncan, 1984: chapter 3). However, while some of the basic ideas of modern social measurement can be found in earlier civilizations, many of these ideas are the product of focused scientific endeavors to create index numbers, scaling techniques, measures of statistical distributions, and measures of properties of social networks (Duncan, 1984).

It wasn't until the mid 19th to early 20th century that sociologists and psychologists started thinking about methods of observation and measurement more seriously. One line of thought in sociology came from Durkheim who used the term "social facts." "Social facts" are social phenomena that have an existence of their own and should be separated from subjective reality. According to Durkheim (1934),

"social facts" are not ideas, but rather things because they have an objective quality to them; they are general throughout society, external to individuals, and observable. Examples of social facts could be religious doctrines, laws, educational institutions, morality, and collective conciousness. In his book, *The Rules of the Sociological Method*, Durkheim (1938) discussed the process of conceptualizing a social fact as an objective reality through avoiding subjectivity, defining clearly what will be investigated, and that the phenomena being studied is defined by characteristics external to a person. His work consisted of deductions that have been a central focus of sociology since its publication, but lacked mathematical precision to ignite quantification. However, it could be argued that Durkheim's work on suicide was an attempt at quantifying "social facts," but he did not necessarily focus on the psychometric properties of suicide rates as a measure of a "social fact." Nonetheless, his theoretical contributions and thoughts have been essential for advancing ideas about social phenomena and how to study them objectively.

Psychologists, however, have made some of the most important contributions to psychometrics through understanding the measurement-conceptualization process of human traits and the construction of mathematical models for assessing the quality of measures (Blalock, 1982: 8). Psychometrics was not institutionally developed and embraced as a way of thinking about measurement until influential works appeared in the mid 19th to early 20th century by psychologists and statisticians. Many key developments during this time period form the foundation of psychometric theory. Particularly, these contributions were largely made through intelligence testing and the development of statistical techniques to assess measurement quality of these tests.

In combination, these advances represent the theoretical and mathematical basis of measurement assessment today (e.g., Binet and Simon, 1911; Gilbert, 1894; Spearman, 1904a, 1904b).

Although psychometrics has led to wonderful achievements in measurement, it has been criticized due to the substantive focus of early works—much of which was linked with the eugenics movement. The following paragraphs will outline an abridged historical account of major researchers, discoveries, and events contributing to the development of psychometric theory; thus, documenting the somewhat chaotic development of this discipline over the past 150 years.

Darwin's (1871) seminal work on evolution in the Galapagos Islands became interpreted in a popular positivistic view in European political culture that evolutionary theory could be generalized to man, ultimately defining a social hierarchy based on genetic superiority. White, English, middle-class citizens were at the top of the evolutionary pyramid while other races, including those of Irish decent and those of English working class strata, were given a label of being genetically inferior. These ideas were pursued empirically by the eugenics movement founded by Galton in 1893 (http://www.eugenicsarchive.org/), i.e., a so-called scientific movement that dealt with the improvement of hereditary qualities of a race or breed.

Many eugenicists promoted an ideology of Social Darwinism. Social Darwinism, defined by Gould (1996:368) is:

> a general term for any evolutionary argument about the biological basis of human differences, but the initial meaning referred to a specific theory of class stratification within industrial societies, particularly to the idea that a permanently poor underclass consisting of genetically inferior people had precipitated down into their inevitable fate.

Darwin, of course, never endorsed these ideas. Rather, his science was interpreted wrongly and endorsed by the prevailing opinion in late 19th century England to justify the construction of social classifications that resulted in some groups of people labeled as biologically and genetically inferior. Eugenics landed on American shores in the early 20th century. A large majority of American eugenicists pursued the segregation of people they deemed as being not fit for breeding by using sterilization and racial segregation methods. Such classifications were justified using mental or intelligence testing (http://www.eugenicsarchive.org/).

Intelligence emerged as an important concept that was of great interest to scientists, eugenicists, philosophers, and politicians alike; many endorsing an ideology that ascribed the attribute to a hereditable process under the cloak of Social Darwinism. The objective of intelligence testing was to express numerically differences among persons in their ability to perform a variety of mental operations.

The ever-mystical concept of intelligence was of particular interest to Sir Francis Galton, a cousin of Charles Darwin, who coined the term eugenics and who is generally credited with the title "founder of psychometrics" (Rust and Golombok, 1999: 5). Galton's accomplishments and contributions to psychometrics were believed to be the motivating, inspiring, and igniting spark for the curiosity of other scientists during this period. Among his accomplishments, Galton established an anthropometric laboratory at South Keningston in 1880. Data produced from his laboratory were used to develop crude intelligence tests. Furthermore, his efforts helped initiate the construction of analytic tools in psychometrics. For example, in collaboration with Karl Pearson, Galton created the Pearson Product-Moment

Correlation Coefficient now used across scientific disciplines and essential to psychometric analysis today (Rust and Golombok, 1999: 5-6).

The evolution of psychometric theory can not be divorced from the history of intelligence testing. In the beginning, many scholars other than Galton contributed to psychometrics by doing research on intelligence testing. In the late 19th century, Cattell and Gilbert continued the study and measurement of intelligence. Both of them carried out studies building on Galton's work using correlation analysis on large samples of university students to assess the validity of intelligence measures (Rust and Golombok, 1999: 5).

In the early 20th century, Alfred Binet began to play a vital role in early intelligence testing. Unlike others before him, Binet took intelligence testing out of the laboratory and applied it to the persisting problem of retardation of children in the Paris Schools with the goal of identifying those with learning difficulties so that good intentioned interventions could be implemented. Binet created one of the first well known intelligence tests and before his death he published three versions of the instrument (Gould, 1996: 178-184). In America, Binet's test was translated and popularized by H.H Goddard of the New Jersey training school for the feebleminded at Vineland. Unlike the well intentioned Binet, Goddard used this instrument to label people as "morons," claimed that it measured a single entity, and largely attributed variability of this entity across people to heredity factors (Gould, 1996: 188-194). While Goddard introduced the Binet scale in America, Lewis Terman, a Stanford University professor, probably played the largest role in its wide spread popularity in the United States. During the early 20th century, Terman (1906, 1916) published

what became known as the most widely used revision of Binet's test-the Stanford Revision and extension of the Binet-Simon Scale. Similar to Goddard, Terman was a hereditarian who used intelligence testing to justify the eradication of people with low intelligence (Gould, 1996: 204-212).

As Binet and others were developing items for intelligence tests, other psychologists were struggling with a conceptually and mathematically tougher problem of quantitatively defining the structure of intelligence. These scientists were exploring the internal structure of intelligence tests or whether such tests measured one or several entities. For example, Spearman (1904a) used factor analytic models, which he created, to conclude that an underlying unidimensional cognitive process, that he titled "g", was driving the correlations between test items. While Spearman's substantive findings have marginal importance now, the methodological tools, namely factor analysis and correlation analysis, that he and others created were important contributions to psychometrics (Gould, 1996:265-350).

The creation of analytic tools driving psychometric theory was well underway during the early years of the twentieth century. Pearson continued to build on the mathematical development of his correlation coefficient, while also deriving proofs for the chi-square test and partial and multiple correlation coefficients. Charles Spearman (1904a) was refining mathematical formulas for more complex analyses of correlation matrices known as factor analysis that were advanced by others (Thurstone, 1924, 1947), while also publishing papers that gave rise to both common factor theory and classical true score theory for measurement reliability. By the first decade of the 20th century, foundations for psychometric theory and analysis were

essentially established by psychologists focusing on the measurement and dimensionality of intelligence (Rust and Golombok, 1999: 6). These analytic tools and measurement assumptions are still vital to psychometric analysis today.

Unlike Binet, many scholars in the United States supported some of the early European notions that intelligence was a fixed, inborn, real entity resulting in its reification. Group intelligence testing in America gained wide spread use following World War I, and the use of psychometrics in the United States started to gain momentum. The credibility of intelligence testing and psychometrics, however, would eventually be seriously questioned due to horrific uses of testing to justify policies that caused great physical and emotional damage to humanity.

As discussed earlier, many of the primary uses of intelligence testing, although viewed by some as valid and reliable methods to establish individual differences in mental functioning, were to sort people into groups with the goal of institutionalizing those with less than normal intelligence and prohibit them from reproducing. Historically, many atrocious ideologies and practices existed supporting the control of those perceived to have low intelligence. Most of these ideas and practices, at least in America, came after the emergence of intelligence testing and the inception of psychometrics as scientific justifications to support a political and social agenda of control. The following will document two of these horrid ideas and practices in America. In the example shown below, Goddard used the Binet scale to justify the prohibition of mating between what he called feeble-minded people or people with subnormal mental ages (as cited in Gould, 1996: 193):

> If both parents are feeble-minded all the children will be feeble-minded. It is obvious that such matings should not be allowed. It is perfectly clear that no

> feeble-minded person should ever be allowed to marry or to become a parent. It is obvious that if this rule is to be carried out the intelligent part of society must enforce it.

Goddard also proposed several solutions upon identification of feeble-minded people. Particularly, he recommended that something must be done to restrict marriage of feeble-minded people stating, "...to this end there are two proposals: the first is colonization, the second is sterilization (as cited in Gould, 1996: 194)." Goddard's use of Binet's tests had dire consequences for immigrants who were returned home and for people who were forced into mental institutions and sometimes sterilized.

One of the most documented historical cases is the notorious Buck vs. Bell ruling where the United States Supreme Court supported sterilization of humans-salpingectomy for women and vasectomy for men. In 1927, Oliver Wendell Holmes Jr. announced the Supreme Courts decision to uphold the Virginia sterilization law. As an occupant of the State Colony for Epileptics and Feeble Minded, Carrie Buck, who had a child diagnosed as having a feeble-mind, scored a nine on the Stanford-Binet; whereas, her mother, then fifty-two years of age, scored a seven. At the time, such scores indicated mental ages that represented subnormal mental incompetence. Carrie Buck was the first person to be sterilized under Virginia's Eugenical Sterilization Act of 1924. In one of the most alarming and significant statements of last century, Holmes wrote (as cited in Gould, 1996:365):

> We have seen more than once that the public welfare may call upon the best citizens for their lives. It would be strange if it could not call upon those who already sap the strength of the state for these lesser sacrifices. ... Three generations of imbeciles are enough.

As of February 1980, the *Washington Post* printed that over 7,500 people were sterilized in Virginia alone. Most of these procedures were conducted in mental health institutions. Sterilizations were performed on men and women. Specifically, both children and adults identified as feebleminded through testing and possessing behavioral tendencies to engage in petty crime and disciplinary problems were sterilized (See Gould, 1996: 365). The impact of the Buck v. Bell decision was felt throughout the United States. By the early 1930's, thirty states had adopted similar eugenics laws. Some estimates indicate that from 1907 onward approximately 60,000 people were sterilized involuntarily, with California and Virginia having the most sterilizations per state (http://www.healthsystem.virginia.edu).

While the above examples represent only two of the most documented atrocities stemming from intelligence testing and the use of psychometrics, similar incidents were quite prevalent across the world into the mid 20th century. Eventually, misuse of the science of psychometrics to justify intelligence testing as a way to select people for inhumane intervention resulted in a negative stigmatization of its study within the scientific and academic communities. Psychometrics became so unpopular that its teaching was deemphasized or even abandoned in psychology and education courses throughout the world (see Rush and Golombok, 1999: 6).

Fundamental Concepts in Psychometrics

Despite the adverse, socio-historical events linked to psychometrics and intelligence testing, several fundamental concepts, ideas, and theories on measurement emerged and became the staples of psychometric theory and analysis. Two of these concepts are reliability and validity. Although both are related and vital

to constructing good measures, they address different aspects of a measurement instrument using both empirical and theoretical strategies. As will be discussed, it is critical to develop an understanding of the relationship between reliability and validity as they relate to measurement, as this is the main thrust of the current dissertation.

Measurement Reliability

Reliability refers to the consistency and/or reproducibility of a construct's measure that is the extent to which a score is free from random error. According to some criminologists (Huizinga and Elliott, 1986), "the reliability of a measuring instrument is commonly defined as the level of precision of the instrument....the extent to which the measuring instrument would produce identical scores if it were used to make multiple measures of the same object or equivalently, the amount of measurement error" (295). This definition implies that a person's score on a measure should, if the measuring device is reliable, reproduce itself from time 1 measurement to time 2, with only random error causing minor fluctuations.

Although the above definition is correct, psychometricians prefer a more detailed definition of reliability that hinges on reduction of measurement error (American Psychological Association, 1985). For example, Nunnally and Bernstein (1994: 213) stated, "one definition of reliability is freedom from error, i.e., how repeatable observations are (1) when different persons make the measurement, (2) with alternative instruments to measure the same thing, and (3) when incidental variations exist in the conditions of measurement." Reliability is a classical issue in scientific inference that is achieved once similar results are produced even when

opportunity for variation has occurred. Regardless of the semantics, it is necessary that all measurement instruments should possess a certain degree of reliability as indicated by internal consistency of items or repeatability of the measure.

Reliability is necessary but not sufficient in achieving valid measurement. A highly reliable measure does not guarantee validity; however, a valid measure can not be unreliable. For example, imagine a scientist who intends to measure intelligence by having participants throw a football as far as possible. Multiple throwing observations for one person would most likely produce similar distances—24 yards, 23 yards, 22 yards, 24 yards, and 22 yards- resulting in highly repeatable observations. Although repeatability is observed and reliability is achieved, most people would know that football throwing does not measure intelligence. This is the same for all sciences, whether using multiple items to measure a construct or a single item, correlated indicators or repeatable scores can not be interpreted as an accurate reflection of what is intended to be measured. Unfortunately, high reliability may be so alluring that even scientists can mistakenly interpret such a measure as valid. Sardonically, Rozeboom (1966: 375) labeled reliability as "the poorman's validity coefficient [or] instant validity."

The underlying mechanics of the theory driving reliability estimation is more complex than the illustrative example above. Reliability estimation is guided by the early work of Charles Spearman (1904a) who proposed the true-score model or measurement error theory now known as classical test theory, the dominant theory guiding the estimation of reliability. Since, many books and monographs have been published on the topic resulting in revisions of its original form (Gulliksen, 1950;

Lord and Novick, 1968; McDonald, 1999; Nunnally and Bernstein, 1994). Although attempts have been made to extended classical test theory, reliability is explained here in its original form. The original form is explained because it is the point of departure for other theories and is still important to modern reliability estimation.

True score theory is the basic theory of measurement. Theoretically, any given measurement is an additive composite of two elements: true ability (true score) and random error. Social scientists, as do all scientists, strive to eliminate random error or noise in their measurements, but all measures will contain it to a certain degree. The true score model is based on the following equation (equations adopted from Trochim, 2001: 94-96):

$$X = T + E \qquad (2.1)$$

'X' in equation 2.1 equals the fallible score that a scientist observes. The observed score consists of a true score and an error component. 'T' in equation 2.1 indicates the score that would be obtained under perfect measurement, otherwise known as the true score. The true score is an unobtainable quantity that will never be directly observed. It can be thought of as the mean score if a person was measured an infinite number of times. Although true scores, in essence, are hypothetical entities they are central to the classical test tradition.

Since the true score will never be obtained, there will always be error in the measurement of a variable. Indicated by E in equation 2.1, error can be any influence that may affect measurement across a sample. For example, measurement error could be introduced by a subject's mood, test instructions given to subjects, testing conditions, or methods of instrument administration to name a few. Such

errors may have inconsistent effects across a sample. Furthermore, on one occasion of measurement error can be higher and the next be lower across individuals. Error may affect the variability around the mean score. Scientists try to minimize this variability so that a more precise estimation of the true score can be obtained. In sum, the observed score equals the true score plus error (Carmines and Zeller, 1979; McDonald, 1999; Nunnally, 1978)

Equation 2.1 has a parallel equation composed of the variance (spread or distribution) of a measure for a set of observations taken across individuals:

$$VAR\ (X) = VAR\ (T) + VAR\ (E) \qquad (2.2)$$

The variability of the observed measure is the sum of the variance due to the true score and random error variance (Tochrin, 2001). The reliability coefficient can be shown as the ratio of variance in true scores to the variance in observed scores in equation 2.3:

$$VAR\ (T) / VAR\ (X) \qquad (2.3)$$

The ratio of true score variance to observed score variance can be thought of as the proportion of truth in the measure or reliability of X as a measure of T. The true score variance is not observed; therefore, it can not be calculated directly. Equation 2.4 shows how the true score variance is calculated indirectly:

$$(VAR\ (X) - VAR\ (E)) / VAR\ (X) = VAR\ (T) / VAR\ (X) \qquad (2.4)$$

where the variance of the observed measure minus the variance of the error (standard error squared) is an estimate of the true score variance.

Several assumptions exist in the classical test model. First, it assumes that all errors are random and normally distributed. Second, true scores are uncorrelated with

errors. Third, different measures of a variable taken on the same person are statistically independent of each other. If these three assumptions are met, the above is a simple equation that will allow for the estimation of the true score, i.e., reliability. Furthermore, it can be assumed, given that measurement error is random, the mean of the measurement errors should equal zero. Finally, the true score will be equal to the mean of the observed scores over an indefinite number of repeated measures (Carmines and Zeller, 1979).

Reliability coefficients have a set range that is ultimately contingent on the error variance. The range of the reliability coefficient is 1 to 0, where 1 indicates perfect reliability and 0 indicates no reliability. This tells the proportion of the measures variability that is attributable to the true score:

$$\text{VAR (T)} / (\text{VAR (T)} + \text{VAR (E)}) \qquad (2.5)$$

If the measure is perfectly reliable there will be no measurement error, i.e., zero error variance, and the equation reduces to:

$$\text{VAR (T)} / \text{VAR (T)} = 1 \qquad (2.6)$$

No true score will exist if the measure is perfectly unreliable, thus, amounting to a measure that is all error variance as indicate in the following equation 2.7:

$$0 / \text{VAR (E)} = 0 \qquad (2.7)$$

The above discussion and equations were meant to illustrate true score theory at a basic level of conceptualization. More detailed reviews of the theory and its variations can be found elsewhere (Lord and Novick, 1968; McDonald, 1999; Nunnally, 1978; Nunnally and Bernstein, 1994). While alternative versions have

been proposed, they all share the same basic aim. The aim of constructing measures designed to measure the same phenomenon consistently.

The discussion of reliability thus far has centered on both definitions and the theory underlying its estimation. Several techniques exist to calculate reliability and choosing one is often contingent on the scientist's perception of its limitations and the measurement itself (Carmines and Zeller, 1979; Nunnally and Bernstein, 1994). So many techniques exist that scientists attempting to reach a formula for estimating reliability could get confused. Realizing this, the American Psychological Association (1985: 19) stated, "statements of reliability and reliability coefficients need to be regarded as generic." Regardless of its generic nature, two categories of reliability estimation have emerged in psychometrics: internal consistency and reproducibility. The most important and widely used internal consistency methods are coefficient alpha and split-half reliability. Reproducibility methods consist of test-retest reliability, inter-rater reliability, and intra-class correlation (ICC).

The coefficient alpha is the most popular internal consistency estimation technique (Carmines and Zeller, 1979) and probably the most frequently used in criminological research; therefore, this measure will be discussed in detail. Some criminologists argue that internal consistency approaches are not particularly appropriate for certain measures such as self-report delinquency scales due to conceptual concerns. As a result, some argue that test-retest reliability coefficients are preferable (Huizinga and Elliot, 1986; Thornberry and Krohn, 2000). The coefficient alpha is used when multiple test items are employed, typically Likert scale items, to measure an underlying construct. In basic terms, the coefficient alpha

reflects the degree to which all items in a scale measure the same underlying construct. This estimation technique is often employed to test the homogeneity/consistency of items or to conclude whether items show strong inter-item correlations. According to Thornberry and Krohn (2000: 46), "internal consistency simply means that multiple items measuring the same underlying concept should be highly intercorrelated."

The internal consistency estimate most commonly used is Cronbach's (1951) alpha. Alpha has an advantage over other internal consistency methods, such as the split halves method. One advantage is that alpha does not depend on how the items are divided then correlated among themselves. Coefficient alpha estimates the reliability of a measure without having to split items into random groups, requires only a single test administration, and provides a unique estimate of reliability for the test. The estimate that is provided can be defined as the expected correlation with an alternative form of the test containing the same number of items (Carmines and Zeller, 1979) or, as Nunnally (1978) has shown, the expected correlation between an actual test and a hypothetically different form of the same instrument. Finally, alpha represents a conservative estimate of the reliability of a measure, as the reliability of a scale should never be lower than the estimated alpha (Novick and Lewis, 1967).

Cronbach's alpha can be calculated from a correlation matrix using unique item variance and total scale variance or a correlation matrix of items using the average inter-item correlation. Equation 2.8 illustrates the variance-covariance matrix formula (Carmines and Zeller, 1979: 44):

$$\alpha = N / (N - 1) [1 - \Sigma\sigma^2 (Y_i) / \sigma_x^2] \quad (2.8)$$

where N is equivalent to the number of scale items, $\Sigma\sigma^2 (Y_i)$ equals the sum of the principal diagonal of the matrix, and σ_x^2 equals the variance of the sum of all element in the data matrix. Equation 2.9 illustrates the same formula using the correlation matrix (Carmines and Zeller, 1979: 44):

$$\alpha = N \bar{p} / [1 + \bar{p} (N - 1)] \quad (2.9)$$

where N equals the number of items and $\bar{p}$ equals the mean of the inter-item correlations. As can be seen from equations 2.8 and 2.9, alpha is contingent upon the number of items in a scale and the correlations and/or covariance among items. In both equations, alpha can range between 0 to 1. The closer the estimate is to 1 the more precise and reliable the measure will be and less measurement error will be present.

Although the alpha coefficient is one of the most common reliability estimation techniques, alternative methods are often preferable in some research situations. For example, criminologists have typically preferred the test-retest method of reliability estimation from the reproducibility methods. This method has even been chosen over internal consistency measures when calculating the reliability of particular measures such as self-report delinquency scales (Huizinga and Elliot, 1986; Thornberry and Krohn, 2000). Using a test-retest method makes more sense logically to some when considering all types of delinquency items that are not expected to be highly correlated with each other (Huizinga and Elliot, 1986).

The calculation of the test-retest method differs from the alpha coefficient. First, a group of respondents are administered a measurement instrument and the

same instrument is re-administered after a given time interval. The correlation between the two test scores serves as the reliability estimate. It is assumed that tests will correlate across time because they correspond to the identical true score (Carmines and Zeller, 1979; DeVellis, 1991; Nunnally and Bernstein, 1994). Equations 2.11 and 2.12 show formulas representing two administrations of the same measurement instrument to the same sample with an interval of time between test administrations.

$$X_1 = X_t + e_1 \quad (2.11)$$

$$X_2 = X_t + e_2 \quad (2.12)$$

Under the parallel measurement assumption the true scores are equal, the error variances are equal, the correlation between errors and true scores are 0, and if the correlations between errors are 0 it can be shown that the correlation between observed scores at time 1 and time 2 gives an estimate of reliability that varies between 0 and 1. Thus, estimates closer to 1 indicate higher reliability and less measurement error.

Although some criminologists argue for the use of test-retest reliability estimations in certain measurement situations (Huizinga and Elliot, 1986), this method has many limitations (See Carmines and Zeller, 1979; McDonald, 1999; Nunnally and Bernstein, 1994). First, researchers are often only able to obtain a measure of a construct at a single point in time, therefore, making it impossible to use this method. Second, its use can be impractical. For example a low correlation between time one and two of measurement may not indicate low reliability; instead, it could mean that the underlying concept, due to development, has changed and can no

longer be measured using the same instrument. The longer the time interval between measurements the more likely it has changed. In contrast, very short time intervals between testing could produce a reliability coefficient that is overestimated and, consequentially, incorrect. Such an inflated reliability estimate can be due to the fact that a respondent's memory of the first testing will influence his/her retest. In addition, subjects also tend to use repeated work habits and employ similar guess patterns that may impact test-retest reliability estimates (Nunnally and Bernstein, 1994). Such problems can make the estimates spuriously low or high, thus, resulting in a distorted reliability coefficient. Together, these limitations should be seen as cautionary indicators before using the test-retest estimation technique.

There is one last critical limitation of the test-retest method that has caused some of the most renowned psychometricians to suggest abandoning its use in most cases (Nunnally and Bernstein, 1994). A multiple item measure of a construct should reveal consistently high inter-item correlations; otherwise, it makes no logical sense to compute a scale to represent a construct from a set of items. The test-retest method can not take this into account. For example, correlations between some items might be zero at the first time of measurement, possibly implying weak internal consistency. Yet, each item could exhibit a strong correlation with itself over the two measurement periods. The test-retest method, therefore, would imply that reliability is substantial when the internal consistency of a measure is questionable (McDonald, 1999; Nunnally and Bernstein, 1994). In sum, high test-retest reliability can emerge despite weak internal consistency.

The goal of this section has been to describe some of the important methods used for estimating the reliability of a measure. Many methods exist, but the two described in this section are the most commonly used in criminology, with coefficient alpha being the one most frequently applied in the social sciences. The purpose of this section was to demonstrate how a theory, i.e., classical test theory, and its assumptions are applied to generate estimates of reliability for a measure of a construct. Furthermore, this section has explained theoretically and conceptually two frequently used reliability estimation techniques in the social sciences.

Measurement Validity

A measurement instrument must be more than reliable if it is to provide an accurate representation of a concept (Carmines and Zeller, 1979), it must be valid. The focus of validity is on whether the measurement instrument accurately reflects what it is supposed to measure. Validity is the extent that an instrument measures a concept as it has been defined and the degree to which the construct is the underlying cause of item covariation (DeVellis, 1991).

Several important points should be made about measurement validity before going into any detail. First, validity, as like reliability, is a matter of degree (Carmines and Zeller, 1979; Devellis; 1991; Nunnally and Bernstein, 1994) and, therefore, obtaining a completely valid measure is unachievable. Second, validity focuses on the critical relationship between a construct and its indicators (Carmines and Zeller, 1979). Third, according to Nunnally and Bernstein (1994: 84) "one validates the use to which a measuring instrument is put rather than the instrument itself," implying that a measuring instrument may achieve a certain degree of validity

for one purpose but not for another. For example, a test may be valid for selecting first-year college students, but not for selecting first-year graduate students. More pertinent to this dissertation, a self-control measure may be a valid indicator of self-control for college students but not for known criminal offenders.

According to Pedhazur and Schmelkin (1991), a cursory review of the discussion of measurement validity indicates that the term validity can be used quite differently by researchers, suggesting that definitions of measurement validity can be encountered that are not always consistent with one another (Carmine and Zellers, 1979; Cronbach and Meehl, 1955; DeVellis, 1991; McDonald, 1999; Nunnally and Bernstein, 1994). A collaboration among many professional associations charged with the mission of developing standards for defining validity in educational and psychological measurement was largely unsuccessful. As an alternative, this collaborative effort resulted in recognition of characteristics of measurement validity in that it "refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from the test score. Test validation is the process of accumulating evidence to support such inferences." (American Psychological Association, 1985: 9). Pedhazur and Schmelkin (1991) suggest that inferences of test scores may have high or low accuracy depending on the purpose, the respondents, and the circumstances for which they are made.

Historically, the validity of a test score was determined by its ability to predict some criterion or an outcome measure external to the test itself (McDonald, 1999). In the early 1950's the American Psychological Association became collectively aware of the need to empirically define the validity of all new tests that were being

generated. In doing so, they decided to convene to set standards for measurement validity. Progress starting during the early 1950's was made by committees from the American Psychological Association and others to identify and define several aspects of measurement validity (Cronbach and Meehl, 1955). This effort resulted in four branches of validity known as predictive, concurrent, content, and construct validity; which can be collapsed into three major categories: content, criterion-related, and construct validity. Respectively, these major domains of measurement validation infer validity from the way the scale was constructed or its domain of content, its ability to predict events or some outcome, and its correlation with other measures (DeVellis, 1991).

It is important to recognize validation of a measure as a "unitary process" (American Psychological Association, 1985: 9). Thinking of this process as representing distinct dimensions of validity then would be a mistake (Pedhazur and Schmelkin, 1991). Some have argued that classifying validity types "leads to confusion, and in the face of confusion, oversimplification" (Dunnette and Borman, 1979: 483). Others have been more critical of classifying measurement validity by type. For example, Guion (1980: 386) stated that validity types are treated as "something of a holy trinity representing three different roads to psychometric salvation. If you can not demonstrate one kind of validity you have two more chances!" One preferred form of measurement validation that encompasses a three prong process has emerged as dominant in some social science disciplines. This approach is titled construct validity (Cronbach and Meehl, 1955; Loevinger, 1957),

and it lays the framework for which the current dissertation will empirically assess the quality of a self-control scale (Grasmick et al., 1993).

Construct validity was articulated by Cronbach and Meehl (1955) and later refined by Loevinger (1957). In 1955, Cronbach and Meehl (1955: 282) proposed that "construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured." Since then, construct validation has been woven into the theoretical cloth of the social sciences and is configured in a way that allows social scientists to investigate inferences about unobserved attributes, i.e., concepts, through observed variables, i.e., measures.

According to Cronbach and Meehl (1955:283) "a construct is some postulated attribute of people, assumed to be reflected in test performance." The construct validation process is important to the measurement of constructs. This process must be conceived of in a theoretical context, and it is the primary form of validation underlying the trait-related approach to psychometrics (Carmine and Zellers, 1979; DeVellis, 1991; Nunnally, 1978; Nunnally and Bernstein, 1994). A construct validation approach is used most often when a scientist believes that his or her measurement instrument measures a theoretically-derived construct. A theoretical framework allows researchers to generate testable hypotheses about a construct that will ultimately help in confirming or disconfirming claims about their measurement instruments (Cronbach and Meehl, 1955).

Theoretical guidance is essential when pursuing construct validation of a measure. Without a theoretical network linked to a construct it is nearly impossible

to validate a measure intended to represent the construct. Without a theory, it is difficult to generate scientific predictions (Carmines and Zeller, 1979). Attempts at construct validation, therefore, are only as good as the conceptualization of the theory that underlies a construct. Cronbach and Meehl (1955) proposed "nomological network" as a term to define theoretical relations a construct should have with other variables. Particularly, Cronbach and Meehl (1955: 290) define "nomological network" as an "interlocking system of laws which constitute a theory." It is important that agreement exists between researchers concerning the nomological network surrounding a particular construct; if not, validation could be impossible. This should not be misconstrued in a way that only formal, fully developed theories are pertinent to construct validation. As Cronbach and Meehl (1955: 291) indicated, "the logic of construct validation is involved whether the construct is highly systematized or loose, used in ramified theory or a few simple propositions, used in absolute propositions or probability statements..." Finally, Cronbach and Meehl (1955: 291) stated, "a rigorous chain of inferences is required to establish a test as a measure of a construct."

Three general steps are essential to construct validation. First, the theoretical relationships between constructs themselves must be specified. Second, empirical relationships between the measures of the concepts must be examined. Finally, results from empirical investigations must be interpreted in terms of how they clarify the validity of a construct (Carmines and Zeller, 1979). Put differently, Cronbach and Meehl (1955) stated, "the system involves propositions relating test to construct, construct to other constructs, and finally relating some of these constructs to

observables" (294). Similar to reliability, construct validity is not an all-or-nothing approach. Construct validity is rather a cumulative process where support for measurement validity is derived from many empirical studies that have subjected the same instrument to stages of the validation process. Construct validation, therefore, is not established by confirming a single prediction about the measure of a construct on different occasions, nor is it achieved by confirming many predictions in one study (Carmines and Zeller, 1979). Instead, it requires a pattern of persistent, favorable results involving different researchers using different samples and methodologies across many studies.

If negative evidence for the construct accumulates over empirical studies it can mean one of several things. First, the theoretical framework surrounding the construct is wrong. Second, the methods used to empirically test theoretical propositions regarding the construct are faulty or inappropriate. Third, the evidence is indicative of a lack of construct validity or the unreliability of some other variables in the analysis (Cronbach and Meehl, 1955).

The construct validation process can be described in more detail. Particularly, Pedhazur and Schmelkin (1991) describe three components of the assessment process underlying construct validation: logical analysis, internal structure analysis, and cross structure analysis. Although these components have been assigned different names by scholars, they ultimately retain the same meaning. For example, Loevinger (1957) labeled these components substantive, structural, and external, respectively.

Logical analysis, also known as face validity, alone can not confirm or disconfirm the construct validity of a measure, as it largely concerns the definition of

the construct and not the empirical aspects of analysis. Similar to content validity, this component is concerned with whether the measure adequately reflects the full domain of the construct as it is defined (Thornberry and Krohn, 2000). Although a construct's definition is the most important aspect of logical analysis, other aspects of logical analysis do exist. For example, logical analysis also includes paying close attention to item content and operationalization (i.e., are items consistent with the definition), method of measurement, when the instrument should be used, directions given to respondents, and scoring procedures for the instrument. Nevertheless, logical analysis alone is insufficient for disproving the validity of a measure (Cronbach, 1971).

Internal structure analysis consists primarily of quantitative techniques, driven by theory, that are used to empirically test the validity of a set of indicators or items representing a construct. Strategies employed to assess internal structure consist of a variety of exploratory and confirmatory techniques that include principal components analysis, structural equation models, and Rasch models. Some variant of factor analysis is the typical technique used to assess internal structure. Factor analysis derives the dimensionality or number of factors that underlie the correlations among a set of items to confirm whether or not data are consistent with theoretical expectations. It is necessary to show that data are consistent with the construct before accepting that a measure has good internal structure validity. In other words, if the construct is proposed as unidimensional, then the data should reflect this. Relations among indicators must be accounted for by the statistical model employed. It is

important that a construct be conceptually articulated so that appropriate measurement models can be used and comparisons can be made.

Evidence from internal structure analyses, although necessary, is not completely sufficient for determining the construct validity of a measure. The final stage of construct validation is cross-structure analysis (Pedhazur and Schmelkin, 1991). This stage concerns the extent to which the measure being validated is related in theoretically expected ways to other constructs and/or variables. As Cronbach and Meehl (1955) suggested, this means that the measure is correlated with variables in the "nomological network of interlocking laws which constitute a theory" (290). Similar to other psychometric endeavors, only after conducting extensive empirical studies, using diverse samples, and investigating a measure's relationship with many theoretically derived variables can a conclusion concerning the degree of measurement validity be made.

It is important to discuss one special case of cross-structure analysis referred to as the known group or group differences approach (Cronbach and Meehl, 1955; Pedhazur and Schmelkin, 1991). This approach will be used to conduct a series of analyses in the current dissertation; other cross structure statistical analyses are beyond the scope of this dissertation. According to Cronbach and Meehl (1955: 287), " if our understanding of a construct leads us to expect two groups to differ on the test, this expectation may be tested directly." Statistical evidence supporting such a hypothesis would lend support to a measure's validity. A simple example can be given by applying the group differences approach to a self-report delinquency scale. A researcher could group a sample of adolescents based on sex and make a theoretical

claim that boys will be more delinquent than girls as has been shown consistently in the past. Boys should have higher scores on a delinquency measure than girls if support for a delinquency measure's validity is to be asserted. If no differences are observed the measure could be tapping a very small domain of delinquency that is equally observed in boys and girls. According to Gottfredson and Hirchi's theoretical stance, a similar argument should be true for race and self control. As will be discussed in later chapters, a valid measure of self-control should show difference across whites and blacks.

In summary, the purpose of this section has been to develop a definition and understanding of measurement validity. First, this section has shed light on the process that must be invoked to draw conclusions about the validity of a measure. Second, this section has suggested that validity is measured in terms of degree, and it is not an all-or-none matter. Third, the degree of measurement validity is contingent on the results of an accumulation of empirical investigations across a number of studies, employing different samples, and using the same measure of a specific construct to test many proposed relationships embedded in its nomological network. Finally, this section has laid a general foundation for assessing the validity of the self-control scale employed in this dissertation. Now I turn to a discussion of the most frequent uses of psychometrics in criminological research.

The Use of Psychometrics in Criminology

The demand for reliable and valid measurement across scientific disciplines encourages the use and continuous development of psychometric methods. Although many appalling events in history are linked to the use of psychometrics, much

progress in the discipline has been made due to theoretical and statistical developments in measurement reliability and validity that emerged from early works of eager psychometricians. For example, modern psychometric methods such as Rasch measurement models were introduced and are now the "cutting-edge" of psychometric methodology (Andrich, 1988; Rasch, 1960; Rasch, 1980; Wright and Masters, 1982; Wright and Stone, 1979).

Similar to its role in the early development of psychometrics, psychology is at the forefront in leading the advancement of psychometric methodology and theory. These continual developments go beyond classical approaches such as factor analysis in ways that are discussed in Chapter Four. Further evidence of the importance of psychometrics today is reflected by the number of academic journals (e.g., *Psychometrika*, *Journal of Outcome Research*, and the *Journal of Applied Measurement)* and texts (Nunnally and Bernstein, 1994) devoted to objective psychological measurement and methods that come closer to achieving such goals.

While aware of measurement concerns in their respective disciplines, other social

scientists, including criminologists, have not fully reaped the benefits of advances made in psychometric theory and methodology. Criminologists, like other social scientists, face the challenging and critical task of theoretically conceptualizing constructs of interest, agreeing on the meaning of these constructs, measuring them (e.g., delinquency, peer delinquency, personality traits), and developing classification schemes (e.g., risk assessment of probationers, prisoners, etc.). Ideally, this process would produce consensus among criminologists on the nature and definition of

constructs and the measuring devices employed so that standardized measures can be used and empirical results can be compared across studies. Realistically, this is hardly ever the case (Gibson, Zhao, and Lovrich, 2002). Criminologists seldom undertake rigorous psychometric analyses of their data.

Some of the most extensive measurement investigations conducted by criminologists focus on one of their most important dependent variables, delinquency (Bendixen and Oleweus, 1999; Elliot and Ageton, 1980; Elliot, Huizinga, and Ageton, 1985; Hindelang, Hirschi, and Weis, 1981; Huizinga and Elliot, 1986; Piquero, MacIntosh, and Hickman, 2002). These studies have predominately focused on reliability and validity issues regarding self-report delinquency measures (See Junger-Tas and Marshall, 1999; Thornberry and Krohn, 2000). The same can not be said for studies exclusively devoted to understanding the psychometric properties of measures of constructs used to predict outcomes such as delinquency. Unfortunately, a cursory review of the empirical research shows that similar empirical attention has not been given to self-report measures involved in the etiology of delinquent and criminal behavior. Although exceptions exist (Arneklev et al., 1999; Gibson, Zhao, and Lovrich, 2002; Piquero, MacIntosh, and Hickman, 2000; Piquero, MacIntosh, and Hickman, 2002), they are uncommon.

Criminologists, as do psychologists, often use multiple-item scales to measure constructs. In general, attempts at validating these scales have largely been limited to classical test models of traditional psychometric theory such as exploratory and confirmatory factor analysis. Such assessments are largely limited and produce a host of concerns when assessing measurement quality (Piquero, MacIntosh, and Hickman,

2002) that will be discussed in Chapter Four. Again, there are several exceptions (See Piquero, MacIntosh, and Hickman, 2000; Piquero, MacIntosh, and Hickman, 2002; Raudenbush, Johnson, and Sampson, 2003).

Criminologists do realize the importance of obtaining accurate and precise measures and are often aware of the problems that can inhibit achieving such measurement (Junger-Tas and Marshall, 1999; Thornberry and Krohn, 2001). For example, Huizinga and Elliott (1986: 293) stated, "few issues are as critical to the study of crime and delinquency as the question of the reliability and validity of our measures of this phenomenon." Not only did they realize the importance of measurement, they go on to assess psychometric properties of self-report delinquency scales. Extensive reviews have been composed on the psychometric properties of self-report delinquency measures and biases of official records data that can serve as a learning device when considering measures of constructs that criminologists use to predict delinquency (Junger-Tas and Marshall, 1999). Nevertheless, reviews and studies designed to test important measurement issues related to constructs of interest to criminologists appear to be few and far between.

Criminologists face many struggles in attempting to measure delinquency, as did psychologists and psychometricians when attempting to measure intelligence in the early 20th century. There has been an important discourse dating back to the early 1960's concerning the advantages and disadvantages of self-report measures of delinquency relative to official records. Criminologists have taken sides and showed skepticism about which measurement method was more accurate and which one most suited their ideological or theoretical preference. Particularly, Gibbons (1979: 84)

was one of the most unconvinced of the new method by saying, "the burst of energy devoted to the self-report studies of delinquency has apparently been exhausted. This work constituted a criminological fad that has waned, probably because such studies have not fulfilled there promise." According to Huizinga and Elliot (1986: 294), "this [resistance and skepticism] resulted in part because there was limited information available on the reliability and validity of self-report measures and in part because these measures appeared to generate different findings regarding the volume and distribution of crime..." Although resistance to the self-report method was encountered (Gibbons, 1979), criminologists eventually started to realize the potential of this methodology.

Most delinquency research before the 1960's was dominated by official police records which many criminologists recognized were not suitable for the task (Merton, 1938; Sutherland, 1939). Such measures did not tap "hidden delinquency." This resulted in a picture of crime that portrayed lower-class youth, African-Americans, and males as the most common criminals. The inception of Short and Nye's (1957, 1958) self-report methodology led to a new way for criminologists to study crime and revealed startling inconsistencies with police records in that (1) delinquent behavior was common to most youth not just particular groups and (2) much crime went undetected by authorities. Although such measurement inconsistencies have been topics of debate, the self-report method for measuring delinquency has become a fixture in modern criminology. In fact, self-report delinquency scales have been employed in large, longitudinal studies to generate most of what is currently known about delinquency and its antecedents (Elliot, Huizinga, and Mernard, 1989;

Esbensen and Osgood, 1999; Loeber, Farrington, Stouthamer-Loeber, Moffitt, Caspi, 1998; Moffitt, Caspi, Dickson, Silva, Stanton, 1996). Self-report delinquency scales have evolved from a controversial method for measuring delinquency to standard practice in criminological research. The sophistication of self-report studies has advanced remarkably in the past five decades (Thornberry and Krohn, 2000). This evolution, however, has not been smooth.

One major area of work influencing the acceptance of self-report delinquency scales has been quantitative in nature. Delinquency scales have been subjected to rigorous psychometric investigations in multiple studies using national and local samples to assess both reliability and validity (Farrington, Loeber, Stouthamer-Loeber, and Van Kammen, 1996; Hindelang, Hirschi, and Weis, 1981; Huizinga and Elliot, 1986; Piquero, MacIntosh, and Hickman, 2002). These scales have probably undergone more psychometric assessment than any other measures used in criminology. These assessments have focused on internal consistency, test-retest reliability, and several aspects of validity.

The self-report method for measuring delinquency has acceptable psychometric properties. First, results from both internal consistency and test-retest estimates have indicated that scale reliability is strong, thus, leading some to conclude that, "if self-report measurement is flawed, it is not here, but in the validity" (Hindelang et al., 1981: 84). Second, the collective results on content, construct, and criterion-related validity have produced favorable evidence supporting the validity of the self-report method. Particularly, evidence from both construct and criterion validity assessments has been the strongest (Farrington et al., 1996; Hindelang,

Hirschi, and Weis, 1981; Huizinga and Elliot, 1986; Thornberry and Krohn, 2000). These investigations have attempted to answer whether such scales work as intended by correlating them with other variables and external criteria of the same constructs. The extent to which these scales have predictive validity are shown in their ability to predict occurrences of arrest and convictions.

With respect to self-report delinquency measures, Thornberry and Krohn (2000: 2001) stated, "the self-report method for measuring this rather sensitive topic-undetected criminal behavior-appears to be reasonably valid...On the other hand, despite this general conclusion, there are several substantial issues concerning the validity of self-report measures." For example, studies of the validity of self-report delinquency measures have shown differential validity across groups such as race. While studies show that most people who have been arrested do report their offenses in self-report scales, there are considerable differences for self-reports of African American males relative to others (See Junger-Tas and Marshall, 1999). Self-report measures of delinquency and official records do not correlate highly for African-American male adolescents (Hindelang, Hirschi, and Weis, 1981; Huizinga and Elliott, 1986); however, this finding has not been consistent across studies (Farrington et al., 1996).

Another differential validity issue for self-report delinquency scales centers on types of self-report offenses. Some studies show that the accuracy of self-reporting for more serious types of offenses may be questionable. These include questions on hard drugs and serious forms of delinquency such as violence (Huizinga and Elliott, 1986). As implied by Hindelang and colleagues (1981), and directly stated by

Hirschi and Gottfredson (1993: 48), "self-report measures, whether of dependent or independent variables, appear to be less valid the greater the delinquency of those to whom they apply." Furthermore, some researchers have discussed differential validity of self-reports across age, indicating that self-reports from adults are lower than those from juveniles (Junger-Tas and Marshall, 1999) In sum, these findings question the validity of self-report delinquency measures when comparisons are made across gender, race, age, seriousness of offense, and criminal involvement of respondents. Nevertheless, without these psychometric assessments such measurement validity issues would go unnoticed and attempts at improvement would most likely not be pursued.

With respect to self-report delinquency measures, comments from Hindelang, Hirschi, and Weis (1981: 114) are probably the most sensible:

> [T]he self-report method appears to behave reasonably well when judged by standard criteria available to social scientists. By these criteria, the difficulty of self-report instruments currently in use would appear to be surmountable; the method of self-reports does not appear from these studies to be fundamentally flawed. Reliability measures are impressive and the majority of studies produce validity coefficients in the moderate to strong range.

While these comments may be reasonable when summarizing the evidence for the reliability and validity of self-report delinquency measures, the same breadth of analysis, investigation, and evidence is lacking for measures of important theoretical constructs used by criminologists as independent variables. Like delinquency scales, independent variables used to predict and explain variance in delinquency are often multiple item measures that should equally be scrutinized through logical, internal structure, and cross-structure analyses as well as reliability analyses. This is

important so that agreement among researchers can be achieved and comparisons of findings across studies can be useful.

Neither theorists nor researchers should be held responsible for the lack of measurement quality. Operationalization and measurement of a construct can only be as good as the theoretical conceptualization that drives both. Refinement efforts from both theorists and researchers are needed if agreement is going to be reached concerning the conceptual, operational, and empirical aspects of a measure. It is important for both theorists and researchers to devote the same attention to their explanatory measures as they have to their outcome measures, e.g., delinquency.

If measures are not created with acceptable degrees of reliability and validity conclusions drawn about the effects of explanatory variables on outcomes are placed in jeopardy because it is not certain what the measure's variance represents. Furthermore, measures of independent variables can be confounded with items measuring different constructs than intended. When reliability and validity analyses are conducted on independent variables in criminological studies, a reliability coefficient and/or principal components analysis are the common procedures used, although exceptions do exist (Gibson et al., 2002; Hickman et al., 2004; Piquero et al, 1999; Piquero et al., 2002; Raudenbush et al., 2003). Such analyses are typically reported in a footnote or methods section of a manuscript as standard practice and not mentioned again. As will be shown in this dissertation, empirical scrutiny is critical for generating measures that meet psychometric standards and such scrutiny will assist in doing away with or refining measures having poor measurement quality.

Several criminological constructs lack independent variable measures that have known psychometric properties. These include constructs within general strain theory such as removal of positive stimuli, presence of negative stimuli, and the disjunction between expected and observed goals (Agnew, 1992); self-control which supposedly consists of six element that coalesce in individuals (Gottfredson and Hirschi, 1990); and social integration and perceptions of collective efficacy, to name a few (Gibson et al., 2002). It is not the purpose of this dissertation to debate the relative merits and psychometric properties of measures of different constructs from criminological theories. It is important, however, to point out that measures of constructs perceived to be important in the etiology of criminal and/or delinquent behavior lack the type of psychometric investigation that has been conducted on self-report delinquency scales.

The purpose of this dissertation is to subject one widely used self-report measure of an important theoretical construct in the etiology of delinquency and criminal behavior to an extensive psychometric analysis. This measure, i.e., the Grasmick et al. self-control scale, is one rare example in criminology where several recent psychometric assessments have been undertaken on an independent variable (Arneklev et al., 1999; Grasmick et al., 1993; Longshore et al., 1996; Piquero and Rosay, 1998; Vazsonyi et al., 2001). Despite the attention awarded to this self-control measure, no conclusive evidence has emerged on its psychometric properties. Since Gottfredson and Hirschi (1990) and Hirschi and Gottfredson (1994) argue that delinquency is closely linked with the nature of self-control, several important ideas can be taken from the conceptualization, operationalization, and measurement of self-

reported delinquency and applied to the investigation and creation of self-control measures. Some of these ideas, such as differential validity, will be elaborated on in the next chapter when discussing the construct and measurement of self-control.

Summary

This chapter emphasized several themes relating to the current dissertation. One of the most important themes is that measurement is a fundamental aspect of the social science research process. Whether conducting applied or basic research, social scientists must be critical of the measures they are using and must understand the reliability and validity of them. Furthermore, theorists must clearly conceptualize constructs deemed as important because this process often guides how constructs are operationalized, measured, and subjected to statistical analysis.

The importance of psychometrics has also been a key theme of this chapter. It has been one goal of this chapter to introduce psychometrics and its evolution. Furthermore, the theory, process, and several analytic frameworks underlying psychometrics were discussed because it is the foundation and guiding framework for investigating the self-control measure in this dissertation.

Finally, the influence of psychometrics in criminology was briefly discussed. Criminologists have used psychometric analysis to understand the measurement properties of self-report delinquency measures, but have not given equal attention to measures of constructs viewed as important in the etiology of delinquency and criminal behavior. This reluctance, as noted already, can produce damaging consequence when interpreting effects of independent variables on dependent

variables. In sum, this chapter has provided a background and framework to the importance of this dissertation. The construct of self-control is discussed next.

CHAPTER 3:

SELF-CONTROL AND THE PSYCHOMETRIC PROPERTIES OF THE GRASMICK ET AL. SCALE

Over a decade has passed since the publication of Gottfredson and Hirschi's (1990) book titled *A General Theory of Crime*. Their theory remains at the center of criminological discourse. This discourse has resulted in persistent theoretical and empirical scrutiny of Gottfredson and Hirschi's key statements. The roots of this intense scrutiny that permeates criminological and criminal justice literature lie in Gottfredson and Hirschi's controversial, yet parsimonious and well-argued, constellation of propositions.

Their propositions concerning the etiology of criminal behavior practically dismiss most criminological theories as incorrect. Gottfredson and Hirschi (1990) argue that traditional theories of delinquent and criminal behavior generate unreasonably multifarious explanations for why people commit crime. They further believe that such theories generally propose spurious relationships between social and behavioral domains of life. This, they argue, is a result of ignoring the nature of crime. Theorists and researchers alike have remained attentive to Gottfredson and Hirschi's formulation for a number of reasons including a.) its parsimonious character, with one main explanatory construct, i.e., self-control, b.) its breadth of explanatory power over the life-course; and c.) the prestige of the authors' past work.

In perhaps one of the most controversial statements made in criminology in the last decade, Gottfredson and Hirschi (1990) argue that their general theory of crime can account for all types of criminal, deviant, and reckless behaviors. This

shows the level of generality they ascribe to their theory. In their own words, the generality proposition of the theory:

> covers common delinquency (theft and assault), serious crime (burglary and murder), reckless behaviors (speeding), school and employment difficulties (truancy, tardiness, in-school misbehavior, job instability), promiscuous sexual behaviors, drug use, and family violence (spouse abuse or child abuse), all of which have negative long-term consequences. No special motivation for any of these acts is assumed. They all provide immediate, obvious benefits to the actor (as indeed, do all purposeful acts). They typically entail no certain or meaningful short-term costs. They all, however, invoke a substantial long-term costs to the actor" (Hirschi and Gottfredson, 1994: 16).

Thus, Gottfredson and Hirschi (1990) base their theory on the postulate that crime, among other deviant and reckless behaviors, provides easily accomplished, instantaneous gratification. Hence, those who commit crime will also engage in acts analogous to law- breaking behaviors. Such people have a disposition that dictates their engagement in all behaviors that provide immediate satisfaction, pleasure, and gratification.

Gottfredson and Hirschi (1990), then, would argue that there is an underlying factor accounting for involvement in all sorts of behaviors. This factor manifests itself across a variety of life's domains in ways that are "not conducive to the achievements of long-term goals and aspirations...that can impede educational and occupational achievement, destroy interpersonal relations, and undermine physical health and well being" (Gottfredson and Hirschi, 1990: 96). As such, the correlations between crime, drug use, unstable employment, failure in marriage, and having delinquent peers are all manifestations of a latent tendency to pursue short-term, immediate pleasure at the expense of long-term consequences (Evans, Cullen, Burton, Dunaway, and Benson, 1997; Hirschi and Gottfredson, 1994). They call this latent

tendency low self-control. Hirschi and Gottfredson (1994: 1-2) link self-control and crime in the following manner:

> Criminal acts are a subset of acts in which the actor ignores the long-term negative consequences that flow from the act itself (e.g., the health consequences of drug use), from the social or familial environment (e.g., a spouses reaction to infidelity), or from the state (e.g., the criminal justice response to robbery). All acts that share this feature, including criminal acts, are therefore likely to be engaged in by individuals unusually insensitive to long-term consequences. The immediacy of the benefits of crime implies that they are obvious to the actor, that no special skill or learning is required. The property of individuals that explains variation in the likelihood of engaging in such acts we call "self-control.

In contrast, those with high self-control are the opposite from those possessing low self-control. As such, Gottfredson and Hirschi (1990: 118) argue that these individuals "are less likely under all circumstances throughout life to commit crime." People who possess self-control are substantially less likely to engage in acts for short-term pleasure even in settings that have marginal social or legal surveillance. For example, they do not steal, drive recklessly, or do drugs even when opportunities, absent from the possibility of legal or social sanctions, are present.

Hirschi and Gottfredson (1994) argue that two primary sources of evidence indicate a latent trait of low self-control causes deviant behavior. First, evidence has shown that a number of heterogeneous criminal, deviant, and reckless acts have a consistent statistical association with one another, occur in a vast array of situations, and have different sets of necessary conditions. Hirschi and Gottfredson (1994) argue that it is reasonable to suspect that the commonalities among these acts reside in the persons committing them. Therefore, according to them, there is no specialization in crime but rather general involvement, and it can be explained by low self-control.

Second, individual differences remain stable over time, that is, those likely to commit the acts mentioned above are also more likely to commit such acts later in time. Therefore, Hirschi and Gottfredson (1994) argue that it would be sensible to attribute the correlations between behaviors over time to a persistent underlying trait. This argument concurs with a population heterogeneity perspective. This perspective attributes the covariation between crime, deviance, and criminal behavior at two points in time to an underlying trait, rendering the correlation between past and future behavior as spurious. For Gottfredson and Hirschi (1990) this underlying trait is low self-control. In this light, Gottfredson and Hirschi's (1990) theory does not only attempt to explain juvenile delinquency, but rather, it offers an explanation for stability in crime and general deviance throughout the life-course.

It is clear that Hirschi and Gottfredson (1994: 2-3) view low self-control as a time-stable trait. Less clear, however, is whether they view low self-control as a general propensity or criminality. They argue that propensity and criminality are both terms rooted in psychological positivism that are directly opposite from their own conception of low self-control. Hirschi and Gottfredson (1993) suggest that the term propensity is the equivalent to a criminal predisposition, which is contrary to control theory. Furthermore, Hirschi and Gottfredson (1993: 49) state, "there may be in our theory an enduring predisposition to consider the long-term consequences of one's acts, but there is no personality trait [propensity] predisposing people toward crime." Similarly, Gottfredson and Hirschi (1990: 88) view criminality as a "positive tendency to crime that is contrary to the classical model [classical theory]." Furthermore, Gottfredson and Hirschi (1990: 88) state, "whereas self-control suggests

that people differ in the extent to which they are restrained from criminal acts, criminality suggests that people differ in the extent to which they are compelled to crime." Contrary to this distinction, Gottfredson and Hirschi (1990: 109) use self-control and criminality interchangeably when discussing personality and criminality. It appears that Gottfredson and Hirschi want to separate their concept of low self-control from propensity and criminality by arguing that self-control is not a personality trait or a predisposition that compels people towards crime. Whether self-control reflects a propensity, criminality, or something else is beyond the scope of the current study; however, this issues will be revisited in Chapter Six.

Regardless of Gottfredson and Hirschi's (1990, 1994) position concerning the above discussion, they do make one thing clear. Self-control (or lack there of), they argue, is stable throughout life (Gottfredson and Hirschi, 1990, 1994). Low levels of self-control increase the probability of virtually all types of criminal and deviant acts that bring pleasure, gratification, and fulfillment in the short-term. Although they attribute generality and stability of criminal and deviant behavior to a trait that resides in an individual, they argue criminal and deviant behaviors will be probabilistic and contingent on opportunities. Although different people may have the same level of the trait, expressions of specific types of criminal and/or deviant acts can reflect variation in opportunities to commit them. While opportunity is important, their theory accords self-control the most explanatory power.

Gottfredson and Hirschi (1990) propose that a person's level of self-control is formed in early childhood. In their opinion, the development of self-control originates from a dynamic process of education and socialization of a child from birth

through pre- adolescence, largely attributing low self-control to inadequate parenting styles. For Gottfredson and Hirschi (1990), weak direct parental controls in childhood are largely responsible for the inability of individuals to delay gratification and the reasons why people pursue behaviors that produce short-term satisfaction. Specifically, Gottfredson and Hirschi (1990: 97) suggest that parents must do three tasks to instill self-control in their children: "(1) monitor their child's behavior; (2) recognize deviant behavior when it occurs; and (3) punish such behavior." Attachment is a key mechanism which determines the quality of parent-child interaction. Parents who are attached to their children will monitor, recognize, and punish naughty, unruly, and disobedient behaviors. To them (Gottfredson and Hirschi, 1990), parental affection towards the child is the motivating factor that will satisfy the three conditions. Conversely, children will develop low levels of self-control if parents are not affectionate; unsuccessful at recognizing misbehavior; do not monitor misbehavior once noticed; and do not appropriately punish the behavior when exhibited by the child.

Gottfredson and Hirschi (1990) devote a lengthy discussion to explaining how their theory may account for the effect of race on crime. Racial disparities in offending rates have been consistently observed and widely acknowledged. As Gottfredson and Hirschi (1990:194) point out, "there is substantial agreement that there are large, relatively stable differences in crime and delinquency rates across race and ethnic groups." Important to the current effort is the mechanism that Gottfredson and Hirschi (1990) put forth to account for these racial disparities. According to Gottfredson and Hirschi, past theories trying to explain these racial differences are

incorrect; differences can be largely explained by inadequate childrearing and, consequently, differences in self-control. Specifically, they argue, "differences in self-control probably far outweigh difference in supervision in accounting for racial or ethnic variation [in crime]" (Gottfredson and Hirschi, 1990: 149). Statements made by Gottfredson and Hirschi concerning race are particularly important to the efforts of this dissertation's investigation of the construct validity of Grasmick et al.'s self-control scale. For them, substantial racial differences in self-control are expected, implying that blacks, as well as other minority racial groups, will have substantially lower levels of self-control than whites. They argue that these differences are due to differences in socialization across racial groups.

Numerous studies have now been published that, when considered collectively, show moderate, yet consistent, support for the proposition that low self-control predicts involvement in a wide range of criminal, deviant, and reckless behaviors (Burton, Cullen, Evans, Alarid, and Dunaway, 1998; Evans, Cullen, Burton, Dunaway, and Benson, 1997; Forde and Kennedy, 1997; Gibbs and Giever, 1995; Gibbs, Giever, and Martin, 1998; Gibson and Wright, 2001; Grasmick et al., 1993; LaGrange and Silverman, 1999; Longshore and Turner, 1998; Nagin and Paternoster, 1993; Paternoster and Brame, 1998; Piquero, Gibson, and Tibbetts, 2002; Piquero and Tibbetts, 1996; Polakowski, 1994; Tremblay, Boulerice, Arsenault, and Junger, 1995) . Many of these studies also indicate that the effects of low self-control hold in the presence of competing theoretically derived variables; across different groups consisting of college students, adolescents, offenders, community samples, different countries; and when using both cross-sectional and longitudinal data.

Pratt and Cullen (2000) recently completed a meta-analysis on the empirical status of Gottfredson and Hirschi's theory. In summarizing the results of 21 empirical studies, they show that low self-control has an average effect size of approximately .27[1]. According to Pratt and Cullen (2000: 952) this effect size qualifies low self-control as "one of the strongest known correlates of crime." Nevertheless, Pratt and Cullen (2000) question whether low self-control is the sole cause of a range of deviant and criminal acts as other variables still have important mean effects in their meta-analysis, namely social learning variables.

Studies testing the relational proposition that low self-control is the cause of a wide array of behaviors have been the cornerstone of support for Gottfredson and Hirschi's (1990) theory. Low self-control has been shown to affect the following: gambling (Arneklev et al., 1993); binge-drinking (Piquero et al., 2002); using force or fraud in the pursuit of self-interest (Grasmick et al., 1993); drunk driving or intention to drive while drunk (Keane et al., 1993; Nagin and Paternoster, 1993; Piquero and Tibbetts, 1996); intentions to commit larceny and sexual assault (Nagin and Paternoster, 1993, 1994); cutting class and alcohol use among undergraduates (Gibbs and Giever, 1995; Gibbs et al., 1998); academic dishonesty (Cochran, Wood, Sellers, Wilkerson, and Chamlin, 1998); drug use among adolescents (Winfree and Bernat, 1998); offending behaviors among a sample of criminal offenders (Longshore and Turner, 1998; Longshore et al., 1996); speeding, driving without a seat belt, and

[1] The standardized correlation coefficient *r* was used to estimate effect sizes found in the 21 studies included in their meta-analysis. According to Pratt and Cullen (2000: 940), this estimation procedure was chosen because "... its ease of interpretation, and because formulae are available for converting other test statistics (e.g., F, t, chi-square) into *r*."

smoking (Forde and Kennedy, 1997); intimate violence (Sellers, 1999); involvement in accidents (Junger and Tremblay, 1999); and victimization (Schreck, 1999).

Other key propositions embedded in Gottfredson and Hirschi's (1990) theory have received less empirical attention. As noted earlier, Gottfredson and Hirschi (1990) state that early in a child's life parents and/or caregivers will have a direct impact on the development of self-control. Once developed, self-control (or a lack there of) will be a stable trait throughout life that, in the presence of opportunity, will explain variation in the persistence of criminal and deviant behavior, versatility in deviance, and predict other negative social outcomes.

Some studies have generated preliminary empirical support for the above claims made by Gottfredson and Hirschi (1990, 1994). First, Hay (2001) found that a lack of parental monitoring and discipline predicts low levels of self-control. Furthermore, Gibbs et al. (1998) found that parental management has an indirect impact on delinquency through self-control. Second, levels of self- control have been shown to be relatively stable over short periods of time, e.g. one academic semester, (Arneklev, Cochran, and Gainey, 1998) as well as longer periods of time, e.g., 5 years (Piquero and Turner, 2002). Third, moderate support has been observed for an interaction between low self-control and opportunity in predicting deviant and criminal outcomes (Grasmick et al.1993; LaGrange and Silverman, 1999; Longshore and Turner, 1998). Finally, several studies have shown that low self-control is related to negative social consequences beyond deviance and criminal behaviors. For example, Wright and colleagues (1999) found that a lack of self-control in childhood predicted disrupted social bonds, e.g., lack of educational attainment, unemployment,

and poor intimate relationships later in life. Furthermore, Gibson and colleagues (2000) found similar results in that low self-control predicted lack of school commitment, lack of cohesiveness with parents, limited goals and aspirations, and involvement with delinquent peers. Both the Wright et al. (1999) and Gibson et al. (2000) studies, however, found that low self-control did not substantially reduce the impact of other social and psychosocial variables on delinquency.

Overall, considerable evidence shows support for several propositions proposed by Gottfredson and Hirschi; however, each study has its own limitations. This dissertation will not subject all this evidence to critical examination. The purpose of the current work is to address one particular theoretical as well as empirical dilemma concerning Gottfredson and Hirschi's (1990, 1994) theory. This dilemma concerns their key construct of self-control, its conceptualization, operationalization, and the adequacy of a self-report scale most commonly used to measure self-control, i.e., Grasmick et al's scale. The quality of this self-control scale is crucial to the body of empirical evidence supporting Gottfredson and Hirschi's theoretical claims, as this scale has been used in numerous studies to draw conclusion concerning the explanatory power of self-control.

Self-control theory has attracted numerous criticism that include: the theory is too general by attempting to explain a broad range of deviant behaviors; it is based on a misconception of the age-crime relationship; it ascribes too much explanatory power to self-control; it overlooks the distinction between prevalence and incidence of criminal involvement and the possibility that the predictors of participation may not be the same as those for frequency of offending; and it is tautological (Hirschi and

Gottfredson, 1994). While these are all important criticisms of Gottfredson and Hirschi's theory, one particular critisizm should be singled out since it illustrates the difficulties with conceptualizing, operationalizing, and measuring self-control. Akers (1991) has accused Gottfredson and Hirschi's (1990) theory as being tautological. Akers (1991: 204) states:

> it would appear to be tautological to explain the propensity to commit crime by low self-control. They are one and of the same, and such assertions about them are true by definition. The assertion means that low self-control causes low-self control. Similarly, since no operational definition of self-control is given, we cannot know that a person has low self-control (stable propensity to commit crime) unless he or she commits crimes or analogous behaviors. The statement that low self-control is a cause of crime, then, is also a tautology.

Akers implied that Gottfredson and Hirschi's logic is flawed since they contend that crime and low self-control are indistinguishable. This is problematic for Akers because Gottfredson and Hirschi also advocate the use of behavioral indicators to measure low self-control. The result, therefore, would closely resemble an empirical tautology because the independent and dependent variables resemble each other too closely. Thus, Akers (1991: 204) writes, "to avoid the tautology problem, independent indicators of self-control are needed."

Accusations of tautology do not bother Hirschi and Gottfredson (1994), as for them, it shows the strength of their theory. They argue that the character of the actor is reflected in the character of the act; therefore, crimes and behaviors analogous to crime are both consequences and indicators of low self-control Simply stated, their theory implies unrestrained people behave in unrestrained ways. Nevertheless, whether Gottfredson and Hirschi's (1990) logic is flawed or ingenious, this particular dilemma has led to many questions of how to operationalize and measure self-control.

The next section will describe the elements of the self-control construct as originally put forth by Gottfredson and Hirschi (1990). Following a description of the construct, there will be a discussion about the conceptual disagreement among scholars concerning self-control. Finally, advantages and disadvantages of different operational definitions used in past studies are discussed.

Conceptualization and Operationalization of Self-control

In describing their central construct, Gottfredson and Hirschi (1990: 89) provide a generally meticulous account of the "elements of self-control." They identify six elements which, they advocate, mirror the nature of criminal acts and largely define one's degree of self-control. Those lacking self-control will have a "concrete here and now orientation", "lack diligence, tenacity, or persistence in a course of action", are "adventuresome, active, and physical, are indifferent, or insensitive to the suffering and needs of others", and "tend to have minimal tolerance for frustration and little ability to respond to conflict through verbal rather than physical means" (Gottfredson and Hirschi, 1990: 89-90).

Gottfredson and Hirschi (1990: 89-90) link each element to the criminal act. First, the 'here and now' orientation reflects the immediate gratification provided by crime, and those with low self-control have an inclination to respond to tangible stimuli in the immediate environment. Second, lacking diligence, tenacity, or persistence reflects the easy and simple gratification provided by crime, and those with low self-control tend to want immediate rewards without much effort. Third, being adventuresome, active, and physical is reflective of the excitement, risk, and thrill attached to the criminal act. Those having low self-control will be risk-seekers

as well as prefer physical activity. Fourth, being insensitive or indifferent reflects the lack of relevance of the discomfort or pain the victims of criminal acts may experience. Those with low self-control have a tendency to be unkind and lack empathy, therefore, are insensitive towards people on whom they directly or indirectly inflict pain or discomfort. Finally, possessing a marginal tolerance for frustration reflects not the pleasure of the criminal act, but rather the relief from temporary irritation. Those with low self-control will have a minimal tolerance for frustration, and they have a tendency to respond to a situation of conflict with physical rather than verbal means.

In sum, Gottfredson and Hirschi (1990: 90) argue that, "people who lack self-control will tend to be impulsive, insensitive, physical (as opposed to mental), risk-taking, short-sighted, and nonverbal..." In addition, these individuals will also possess a volatile temper indicative of their low tolerance for frustration. Furthermore, they note that, "there is a considerable tendency for these traits to come together in the same people" (Gottfredson and Hirschi, 1990: 90-91).

Gottfredson and Hirschi's conceptual definition of self-control, as well as the operational procedures that have been followed, have sparked a rather interesting debate among criminologists. This debate has led to an interpretive divide. First, a division exists among criminologists concerning the appropriate conceptualization of the self-control construct. Second, operationalization of self-control has led to an unsettled dilemma among criminologists when choosing indicators that are most appropriate to reflect self-control (Hirschi and Gottfredson, 1993; Stylianou, 2002)

With regards to conceptualization, some interpret self-control as being unidimensional and others argue that it is a multidimensional construct. Unidimensionality implies that one trait or attribute is being measured, in this case, the attribute is self-control. According to Trochim (2001: 136), it is easy to think of a dimension as a ruler or number line. Unidimensionality would then mean one line can be used to reflect higher or lower levels of self-control. For example, weight is a concept that is unidimensional. For the current study, this would mean that all elements specified by Gottfredson and Hirschi are one and the same; indistinguishable, and therefore do not represent different attributes as they can all be captured on one ruler to indicate more or less self-control. In contrast, it is not possible to measure a multidimensional construct on one ruler or a single number line (Trochim, 2001: 135). For example, intelligence consists of multiple dimension such a math and verbal ability. A person may have strong verbal ability and weak math ability. As will be shown, some argue the same could be true for self-control. For example, self-control could be multidimensional in that different elements indicate different constructs; therefore, it would be impossible to depict a person's level of self-control using one number line because multiple measures could be confounded in one.

Grasmick et al. (1993) conducted one of the first empirical tests of Gottfredson and Hirschi's theory. In doing so, Grasmick et al. (1993: 9) explicitly interpreted the conceptualization of self-control as a unidimensional construct that is evident in their following statement:

> A factor analysis of valid and reliable indicators of the six components is expected to fit a one factor model, justifying the creation of a single scale

> called low self-control. In effect, this is a very crucial premise in Gottfredson and Hirschi's theory. A single, unidimensional personality trait is expected to predict involvement in all varieties of crime as well as academic performance, labor force outcomes, success in marriage, various "imprudent" behaviors such as smoking and drinking, and even the likelihood of being involved in accidents. Evidence that such a trait exists is the most elementary step in a research agenda to test the wealth of hypotheses Gottfredson and Hirschi have presented.

Since, others have pursued measurement of self-control under a conceptual framework of unidimensionality. For example, Nagin and Paternoster (1993: 478) note that, "the construct was intended to be unidimensional," therefore, implying that they conceptually interpret Gottfredson and Hirschi's construct as reflecting one entity. A similar line of thought was followed by Arneklev and colleagues (1993: 232) when they examined the same scale based on "Gottfredson and Hirschi's assertion that low self-control is a unidimensional construct" (232). Furthermore, Piquero and Rosay's (1998: 157) conceptual interpretation is apparent when they stated, "evidence for a solution that has more than one factor would not be consistent with Gottfredson and Hirschi's claim." Although some researchers interpretation of Gottfredson and Hirschi's construct of self-control imply unidimensionality, others interpret the original formulation differently.

In contrast from those cited above, several researchers have interpreted the original conceptualization of self-control as multidimensional. This is most likely due to Gottfredson and Hirschi's identification of several elements embodied in their construct. A multidimensional interpretation implies that more than one attribute is being measured. On a conceptual level, this implies that the elements of self-control identified by Gottfredson and Hirschi (1990) can be related but yet are distinct from one another. While some suggest that evidence of multidimensionality would be

damaging to the intended conceptualization of the self-control construct (Longshore et al., 1996), others would interpret such evidence as support for Gottfredson and Hirschi's claims (Arneklev et al., 1999; Vazsonyi et al., 2001).

Changing their conceptual interpretation in a later publication, Arneklev and colleagues (1999) interpret Gottfredson and Hirschi's construct as multidimensional. They argued that the elements were distinct yet were accounted for by an underlying trait. Arneklev and colleagues (1999) argued that Gottfredson and Hirschi specify six dimensions of self-control so how can the characteristics be anything but multidimensional. What is questionable, according to Arneklev and his colleagues (1999), is whether or not these six elements account for a final, higher-order construct. While the 1999 conceptualization departs from Arneklev and his colleagues (1993) earlier interpretation, they still imply there is an underlying construct of self-control, but six elements should be identifiable in the construct.

Vazsonyi et al. (2001) also argued that Gottfredson and Hirschi conclusively outline self-control as a multidimensional trait. They go on to argue, however, that this is not in total contrast to a unidimensional interpretation when they stated that "a multidimensional measure of self-control still can and does imply that these elements together form a single latent trait of self-control" (Vazsonyi et al., 2001: 98).

The conceptual confusion that has resulted from interpretations of Gottfredson and Hirschi's description of self-control's elements can be partially attributed to Gottfredson and Hirschi themselves. Although they clearly describe the elements of their construct, the dimensionality of their construct remains ambivalent, except to state that these six elements have a tendency to come together in the same people

(Gottfredson and Hirschi, 1990: 91). This does complicate efforts to validate scales designed to specifically test self-control because there is no consensus on how to conceptually interpret the construct. The most reasonable line of action would then be to empirically test all conceptualizations of the construct of self-control.

An appropriate operational definition is the second source of controversy regarding self-control as a construct (Akers, 1991; Gibbs and Giever, 1995; Hirschi and Gottfredson, 1993; Hirschi and Gottfredson, 1994; Stylianou, 2002). An operational definition implies a process that articulately defines how a construct will be measured Maxfield and Babbie, 2001: 106). In doing so, this process moves closer to measurement by considering a pool of questions, statements, or behaviors that will be considered to represent the construct as well as the method(s) that will be used to collect data (e.g., self-report, observational, etc.) (Maxfield and Babbie, 2001: 106). Currently, no agreed-upon operationalization of Gottfredson and Hirschi's self-control construct exists. The controversy surrounds two different operationalizations: attitudinal and behavioral. For Hirschi and Gottfredson (1993, 1994) behavior-based operationalizations are to be preferred.

Hirschi and Gottfredson (1993:49) explicitly stated, "behavioral measures of self-control seem preferable to self-reports" and "multiple measures [items] are desirable." They seem to prefer such measures because they oppose the inclination to interpret the concept of self-control as a personality predisposition. In contrast, Akers (1991) has warned against such operationalizations due to a tautology issue of not having indicators of self-control that are independent of outcomes that it should

predict. Nevertheless, Hirschi and Gottfredson (1994) argue that behavioral indicators can be identified independent of crime. They propose the following:

> "With respect to crime, we would propose such items as whining, pushing and shoving (as a child); smoking and drinking and excessive television watching and accident frequency (as a teenager); difficulties in interpersonal relations, employment instability, automobile accidents, drinking and smoking (as an adult)" (Hirschi and Gottfredson, 1994: 9).

Such behavioral operationalizations become problematic to researchers for several reasons. First, such indicators are not only outcomes in Gottfredson and Hirschi's theory, but they are being promoted as actual measures of self-control. On the one hand, Gottfredson and Hirschi (1990) argue that the above indicators can be used as behavioral indicators to measure self-control. On the other hand, they argue that these are also outcomes of low self-control. Not only can this be conceived as presenting a threat of empirical tautology, but it poses a problem to researchers when attempting to disentangle causes from effects. Stylianou (2001) points out that when using such behavioral indicators causes and effects will possibly become entangled. Causal hypotheses require distinctions between the independent and dependent variables, in this case, elements and manifestations of low self control. She argued, "when modeling low self-control as a cause of crime and analogous behavior, one cannot use crime and analogous behavior as measures of low self-control" (Stylianou, 2001: 536).

Operational definitions of self-control based on behavior may have serious limitations for understanding relationships between low self-control and its manifestations. Mainly, limitations are apparent in the interpretation of effects of low self-control. As such, no consensus exists on whether to interpret the results as

support that low self-control predicts negative outcomes or whether the effects observed indicate versatility in deviant and criminal behavior.

Another problem is the use of behavioral definitions to operationalize self-control in childhood to predict teenage and adult criminal/deviant behavior. When testing the relationship between self-control in childhood and future behavior, Paternoster and Brame (1998) employed data from the well-known Cambridge Youth Study. They constructed an operational definition of self-control consisting of, "proneness of the boy to act out, rating of the boy's daring or adventurousness, and teachers ratings on laziness, concentration skills, and disciplinary difficulty" to predict future misbehavior (Paternoster and Brame, 1998: 642), concluding that self-control in childhood predicts future deviant and criminal behavior. However, such a link could be interpreted as heterotypic behavioral continuity and not that childhood low self-control predicts adult criminal outcomes[2]. Finally, Gibbs and Giever (1995) have pointed out other possible flaws in behavioral measures. They state that "crime and analogous behaviors as measures of self-control can be expected to contain substantial error because they reflect several underlying variables or constructs" (Gibbs and Giever, 1995: 249).

A few studies have used directly observable behavioral indicators to measure self-control. For example, Keane and colleagues (1993) used direct observation (i.e., failure to wear a seatbelt) as well as self-report behavioral items (i.e., drinking,

[2] Heterotypic continuity implies that misbehavior may manifest in different forms from childhood to adulthood, but is caused by the same underlying, unobserved characteristics. As such, this would imply a population heterorgeneity position in that the observed correlation between misbehavior in childhood and adulthood is due to unmeasured differences across persons established early in life (Nagin and Paternoster, 2000). Therefore, the link between childhood behavioral measures of low self-control and adult offending outcomes could represent heterotypic behavioral continuity in that the relationship is caused by some other trait that is not observed, once again an empirical tautology.

perceived risk of being stopped by police, etc.) in operationalizing self-control to predict driving under the influence. Most behavioral operationalizations of self-control, however, have relied on self-report[3]. In operationalizing self-control, Zager (1994: 75) used a self-report index consisting of "six self-report delinquency items, including alcohol use, marijuana use, making obscene phone calls, avoiding payment, strong arming students, and joyriding." Similarly, Evans and his colleagues (1997) used an operational definition of lack of self-control that included self-report behavioral items consisting of violating the speed limit, drunk driving, illegal gambling, and using drugs. In sum, many of these self-report behaviors are deviant and criminal acts that Gottfredson and Hirschi (1990) would argue are predicted by low self-control, however, they are used in some studies to measure low self-control as well.

The above operational definitions exemplify the use of behavior, whether self-reported or directly observed, to represent self-control of children, adolescents, and adults. These behavioral definitions are more consistent with Gottfredson and Hirschi's operational preference than attitudinal/trait-based operationalizations. Hirschi and Gottfredson (1993) prefer, however, directly observable behaviors in operationalizing self-control. They put less faith in the self-report methodology. Importantly, they argue that "the level of self-control itself affects survey responses...self-report measures, whether of dependent or independent variables,

[3] The distinction between observed and self-report behavioral measures of self-control should be made as Hirschi and Gottfredson (1993) imply that differences between the two do exist. For them, all self-report measures should be used with caution because survey responses are affected by an individual's self-control, whether answering a question about behavior or attitude. Hirschi and Gottfredson would have us believe that behavioral measures independent of self-report, i.e., direct observation, are preferred.

appear to be less valid the greater the delinquency of those whom they are applied" (Hirschi and Gottfredson, 1993: 48). Hirschi and Gottfredson (1993) imply that this will be true for any self-report measure whether it be attitudes or behavior. While they do not argue for abandoning operational definitions that employ self-report methods to test self-control theory, they do suggest that differences among respondents should be considered in research design and measurement when testing their theory (Hirschi and Gottfredson, 1993: 48).

The other, probably more favored, operational definition among criminologists has been attitudinal and/or personality based self-report items designed to represent the construct of self-control (Grasmick et al., 1993; Gibbs and Giever, 1995; Stylianou, 2002). Some argue this operational method is a way to overcome the tautology issue (Stylianou, 2002), while others argue that this approach implies psychological positivism that is incongruent with the self-control construct (Hirschi and Gottfredson, 1993). Nevertheless, many support such an operational definition for several reasons.

Gibbs and his colleagues (1998: 95) suggest that a variable used to explain behavior "can be most clearly grasped and tested when it is defined as something broader or different than behavior." An advantage to such an approach with respect to self-control is that it "leaves no space for tautology: Conceptually, attitudes and behaviors are mutually exclusive categories" (Stylianou, 2002: 538). Furthermore, Gibbs and Giever (1995) argue that self-inventory, personality-based operational definitions, which would include Grasmick et al.'s scale, are constructed specifically based upon elements of self-control described by Gottfredson and Hirschi. Such

operational definitions, they say, have advantages over behavioral ones for two reasons: 1.) they are more useful in tapping more cognitive aspects of self-control and 2.) allow for a more comprehensive coverage of domains of self-control because items can be developed to capture typical modes of behavior that relate to everyday life (Gibbs and Giever, 1995: 249). In contrast, behavioral measures are restricted by time, money, and access in the cross section of daily life they cover (Gibbs and Giever, 1995; Nunnally, 1978).

Few self-report attitudinal and/or personality based operational definitions have been developed specifically to test Gottfredson and Hirschi's construct of self-control (Gibbs and Giever, 1995; Grasmick et al., 1993). While Gibbs and Giever (1995) created such a measure, its creation was intended to be relevant only to college students and has not received much empirical attention beyond their own exploratory scrutiny. Similarly, Grasmick et al (1993) created a 24-item attitudinal/personality scale based on their interpretation of Gottfredson and Hirschi's conceptual definition of self-control. Grasmick et al. (1993) employed this 24-item scale in one of the first investigations to test key propositions in self-control theory. This particular operational definition has been used widely in tests of Gottfredson and Hirschi's (1990) theory. For example, Pratt and Cullen (2000) show that at least 12 studies have used Grasmick et al.'s (1993) scale in pursuing empirical tests of self-control theory. The following section will discuss the creation of this scale in detail.

Creation of Grasmick et al.'s Scale

In creating their self-control scale, Grasmick and his colleagues (1993) gave close attention to how Gottfredson and Hirschi (1990) conceptually define elements

of self-control. In doing so, they derived an operational definition to reflect its conceptual properties. This process required justification for the items under each component to create a scale[4]. Such logical analysis is the first step in any construct validation process.

From Gottfredson and Hirschi's (1990) theory, Grasmick and his colleagues (1993) identified six components of self-control that they interpret as a "personality trait" that should be unidimensional. The components are: impulsivity, preference for simple rather than complex tasks, risk-seeking , preference for physical rather than cerebral activities, self-centered-orientation, volatile temper linked to a low tolerance for frustration. This gave Grasmick and his colleagues a starting point for identifying items that correspond to each component (or element) of self-control[5].

Grasmick and his colleagues used a combination of many items in pre-testing college students to identify a final 24 items. This resulted in four items for each of the six components. Grasmick and his colleagues pretest found sufficient variation within items and items tended to be unidimensional in their factor structure. Table 1 lists the original items. Items were originally scored on a four point Likert scale ranging from (1) strongly disagree, (2) disagree somewhat, (3) agree somewhat, and (4) strongly agree According to Grasmick et al. (1993), agreeing to many of these

---

[4] This process is common to all social science research endeavors, whether using existing variables from secondary data to create scales or constructing items to represent a particular construct. Grasmick et al.'s effort is unique in that it is one of the few attempts to create items to measure a specific criminological construct that resembles the way psychologists construct scales.

[5] Initially, they considered using the self-control subscale of the California Psychology Inventory (CPI) (Gough, 1975). Although some CPI items reflect domains of self-control, Grasmick and his colleagues discovered that several items lacked face validity in regards to Gottfredson and Hirschi's description of self-control. In addition, the CPI subscale did not contain items that tapped preference for simple tasks or preference for physical activities. Grasmick and his colleagues decided to create their own items, influenced by the CPI subscale, to formulate an operational definition that matched Gottfredson and Hirschi's (1990) elements as closely as possible.

Table 1. Grasmick et al.'s (1993) self-control items

Item
Impulsivity
I1: I often act on the spur of the moment without stopping to think.
I2: I don't devote much thought and effort to preparing for the future.(reverse coded)
I3: I often do whatever brings me pleasure here and now, even at the cost of some distant goal.
I4: I'm more concerned with what happens to me in the short run than in the long run.
Simple Tasks
S1: I frequently try to avoid projects that I know will be difficult.
S2: When things get complicated, I tend to quit or withdraw.
S3: The things in life that are easiest to do bring me the most pleasure.
S4: I dislike really hard tasks that stretch my abilities to the limit.
Risk Seeking
R1: I like to test myself every now and then by doing something a little risky.
R2: Sometimes I will take a risk just for the fun of it.
R3: I sometimes find it exciting to do things for which I might get in trouble.
R4: Excitement and adventure are more important to me than security.
Physical activities
P1: If I had a choice, I would almost always rather do something physical than something mental.
P2: I almost always feel better when I am on the move than when I am sitting and thinking.
P3: I like to get out and do things more than I like to read and contemplate ideas.
P4: I seem to have more energy and a greater need for activity than most other people my age.
Self-centered
Sc1: I try to look out for myself first, even if it means making things difficult for other people.
Sc2: I'm not very sympathetic to other people when they are having problems.
Sc3: If things I do upset people, it's their problem not mine.
Sc4: I will try to get the things I want even when I know it's causing problems for other people.
Temper
T1: I lose my temper pretty easily.
T2: Often, when I'm angry at people I feel more like hurting them than talking to them about why I am angry.
T3: When I'm really angry, other people better stay out of my way.
T4: When I have a serious disagreement with someone, it's usually hard for me to talk calmly about it without getting upset.

items would indicate low self-control or, in other words, higher scores would mean a lack of self-control.

Grasmick et al.'s (1993) effort was the first attempt to create a self-control measure distinctively operationalized to embody self-control as described by Gottfredson and Hirschi (1990). Following closely Gottfredson and Hirschi's (1990) description of self-control, they were able to identify items that appeared to represent each element. While achieving face and content validity is an important part of any logical analysis, they failed to discuss other important aspects of a logical analysis. For example, Grasmick and his colleagues (1993) did not give other researchers any advice for which populations the instrument is appropriate, e.g., college students, juveniles, incarcerated populations, if directions were explicitly given to respondents, and which scoring procedures should be used for scale construction.

The failure to discuss the conditions under which the instrument is appropriate is a question that should and can be pursued through empirical testing. It is not yet clear whether Grasmick et al.'s (1993) scale can be equally applied to different samples to discriminate between levels of self-control (or a lack there of). Perhaps, their scale items are more suitable for low- risk samples, such as college students, rather than high-risk samples, such as serious criminal offenders. The scale items could be too easy or too endorsable for a sample of respondents who, on average, were likely to have lower self-control. This could result in the inability of Grasmick et al.'s scale to accurately measure levels of self-control among such respondents. In contrast, the scale items could be well-suited for a community or college student

sample who, on average, were likely to have higher self-control than a sample of criminal offenders.

Although Grasmick and his colleagues (1993) advise readers not to accept their work as a definitive operationalization of self-control, their scale remains the measuring instrument of choice for researchers attempting to quantify self-control (See Delisi, Hochstetler, and Murphy, 2003). To support its continued use there must be evidence showing the scale is empirically reliable and valid, and that it is applicable to different samples of subjects. Researchers are only now beginning to investigate the psychometric properties of the scale across different samples using multiple reliability and validation techniques. The next sections will review these studies.

Psychometric Properties of Grasmick et al.'s Scale

A review of the recent research indicates that several studies have used selected items from Grasmick et al.'s scale (Burton et al., 1998; Gibson and Wright, 2000; Winfree and Bernat, 1998); other studies employ Grasmick et al.'s scale in combination with other items/constructs to assess self-control (LaGrange and Silverman, 1999); and some studies use differing methods to test Grasmick et al.'s measure, such as, telephone interviews (Forde and Kennedy, 1997). These studies did not exclusively focus on the conceptual and measurement issues of self-control using Grasmick et al.'s scale.

Fewer studies have solely examined the psychometric properties of Grasmick et al.'s scale (Arneklev et al., 1999; Delisi et al., 2003; Grasmick et al., 1993;

Longshore et al., 1998; Piquero et al., 2000; Piquero and Rosay, 1998; Vazsonyi et al., 2001). In reviewing these studies, evidence for the scale's reliability and construct validity will be presented separately. With respect to construct validity, evidence will be presented concerning internal structure (or dimensionality of the scale) analyses.

Reliability of Grasmick et al.'s Scale

As discussed in Chapter Two, Cronbach's alpha is the most common internal consistency measure for estimating reliability in social science research. Most studies employing Grasmick et al.'s scale typically report this measure. Assuming that the self-control construct is unidimensional, the alpha for this scale should be quite high, or at least modest, ranging from .7 to .9. Some researchers, however, argue that a reliability coefficient of .8 may not be nearly large enough to make decisions about individuals, but in the initial stages of scale construction a modest reliability will suffice, e.g., .7 (Nunnally and Bernstein, 1994). Estimates larger than the above criteria would be evidence that the scale has minimal measurement error and that items are highly correlated, as they should if the scale is measuring one trait or attribute. Lower internal consistency scores may be a function of the scale's multidimensionality. This remains an empirical question that reliability estimates alone cannot answer.

Several of the seven studies testing the psychometric properties of Grasmick et al.'s scale reported Cronbach's alpha coefficients for the total 24-item scale (Delisi et al., 2003; Grasmick et al., 1993; Longshore et al., 1996; Piquero and Rosay, 1998). Furthermore, some studies reported alpha for each four item subset representing each

self-control element (Piquero and Rosay, 1998; Vazonyi et al., 2001). Finally, two of the seven studies did not report reliability coefficients (Arneklev et al., 1999; Piquero et al. 2000).

Using data from a simple random sample of 395 adult respondents who completed the Oklahoma City Survey, Grasmick and his colleagues (1993) conducted the first reliability analysis of their scale. They concluded that by dropping one item (the last item under the physical activities component) from the scale they could increase reliability from .80 to .81. Although they made the adjustment, this adjustment did not substantially improve the internal consistency of the scale.

Two studies using the same data set emerged in 1996 and 1998 revealing the psychometric complexity of Grasmick et al.'s scale. It should be noted that the version of the scale in these studies diverges slightly from its original form in a few ways. First, Longshore et al. (1996) modified the original response scale and added an additional category to make it a five-point Likert scale: never (0), rarely (1), sometimes (2), often (3), and almost always (4). Second, item wording was changed and often reversed to detect any bias from yes-saying. These data came from a multi-site evaluation of Treatment Alternatives to Street Crime (TASC) programs to identify drug using adult and juvenile offenders in the criminal justice system to gauge their treatment needs, place them in treatment, and monitor progress that is made. The sample consisted of the first 623 offenders providing all relevant data during intake between 1991 through 1992. Most respondents had lengthy criminal histories, and the sample had variability in sex, race, and age.

While results from both studies drew different conclusions concerning the empirical dimensionality of the scale (discussed in the internal structure section), they did show similarities in reliability. Longshore et al. (1996) reported a Cronbach's alpha of .80 for the Gramsick et al. scale, whereas, Piquero and Rosay (1998) reported a Cronbach's alpha of .71. Unlike Longshore et al. (1996), Piquero and Rosay (1998) reported gender specific alpha's, .72 for males and .68 for females. Both studies reported estimates for each component of the scale, which were low compared to acceptable standards. Longshore et al.'s (1996) estimates were .65 for Impulsivity/Self-centeredness, .48 for Simple Tasks, .58 for Risk Seeking, .35 for Physical Activities, and .71 for Temper. Piquero and Rosay's (1998) estimates were .45 for Impulsivity (.46 for males and .43 for females), .44 for Simple Tasks (.47 for males and .28 for females), .58 for Risk Seeking (.58 for males and .56 for females), .37 for Physical Activities (.40 for males and .31 for females), .68 for Temper (.71 for males and .59 for females), and .57 for Self-centeredness (.59 for males and .49 for females). The difference was that Longshore et al. (1996) reported alpha's on only five of the six components; they combined Impulsivity and Self-centeredness items due to the results of their internal structure analysis that will be discussed in the next section.

Delisi et al. (2003) used data collected from 208 male parolees residing in work release facilities in a Midwestern state that had been previously released from prison and were currently serving provisional parole sentences. They reported that Cronbach's alpha for the total scale was .91. They also computed reliability estimates for each component showing coefficients of .79 for Impulsivity, .81 for Simple Tasks,

.79 for Risk Seeking, .72 for Physical Tasks, .81 for Self-centeredness, and .86 for Temperament. Evidence showing that the scale's components had lower alpha estimates than the total scale does not give more support for unidimensionality, as this difference could be a function of the number of items in each estimate.

In the largest study undertaken to investigate the psychometric properties of Grasmick et al.'s scale, Vazonyi et al. (2001) gathered data on over 8,000 adolescents from four different countries including schools in Hungary (n = 871), Netherlands (n = 1,315), Switzerland (n = 4,018), and the United States (2,213). While total scale reliability estimates are not reported, they do report them for self-control subscales for both the total and country samples. Specifically, Cronbach's alpha was .50 for Impulsivity ranging from .45 to .62; .68 for Simple Tasks ranging from .61 to .73; .79 for Risk Seeking ranging from .69 to .84; .63 for Physical Activities ranging from .55 to .74; .60 for Self-centeredness ranging from .45 to .68; and .76 for Temper ranging from .68 to .76.

Internal Structure Analyses of Grasmick et al.'s Scale

There is considerable disagreement on the conceptual interpretation of Gottfredson and Hirschi's self-control construct. Some interpret self-control as being unidimensional while others see it as being multidimensional. Consequently, no consensus exists on the number of factors that should emerge to support construct validity of Grasmick et al.'s scale. Grasmick and his colleagues (1993) do argue, however, that a factor analysis of valid and reliable indicators of self-control should produce a unidimensional structure. From a construct validity perspective, this disagreement is troublesome because researchers have no clearly defined theoretical

model to pursue in empirical tests of this scale. As a result, researchers pursuing internal structure tests of Grasmick et al.'s scale have employed different models including both unidimensional and multidimensional solutions using both exploratory (EFA) and confirmatory factor analysis (CFA).

Grasmick et al. (1993) were the first to assess the dimensionality of their scale. First, they performed a principal components exploratory factor analysis (EFA) with one-, five-, and six-factor solutions. Then, based on associated evaluative criteria, e.g., Kaiser rule and Scree plots, they could not "find strong evidence that combinations of items into subgroups produces readily interpretable multidimensionality" (Grasmick et al., 1993: 17). In contrast, their analysis led them to conclude that, "the strongest case can be made for a one-factor unidimensional model" (Grasmick et al., 1993: 17). Their decision to infer unidimensionality was largely based on results of a Scree Discontinuity plot that showed the largest break in eigenvalues was between the first and second factor. Some suggest that the largest break will determine how many factors are present, but this rule is a very descriptive and preliminary first step that does not confirm dimensionality (Carmines and Zeller, 1979; Nunnally, 1967). In contrast, the Kaiser Rule states that eigenvalues greater than 1.0 imply how many factors are present in the data. Grasmick and colleague's (1993) results showed that six factors had eigenvalues greater than 1.0. Several studies have shown similar results using the same method across different samples (Arneklev et al., 1998; Nagin and Paternoster, 1993; Piquero and Tibbetts, 1996; Piquero et al., 2002; Delisi et al., 2003), concluding that the largest "break" between

eigenvalues is between the first and second factor with six factors having eigenvalues greater than 1.0.

Although these studies show consistency, this alone does not indicate that the scale is either unidimensional or multidimensional. Depending on researchers' understanding of the original conceptualization of self-control, findings from these studies have been interpreted both as suggesting multiple factors as well as one factor. Furthermore, EFA's, e.g., principal components analysis[6], reduce multiple variables (or items) without an imposed theoretical structure, and they try to extract the most variance possible from the first factor. EFA leaves the task of defining the factors up to the factor analysis program, therefore being inadequate for construct validity purposes (Devillis, 1991). Due to limitations of EFA, results from the above studies are descriptive and not capable of confirming a multidimensional or unidimensional structure. In sum, the results from EFA's imply that the Grasmick et al.'s scale could be either. More recently, others have used confirmatory models that are more appropriate for construct validation. These tests have led to quite different conclusions than Grasmick et al.'s (1993) original analysis (Arneklev et al., 1999; Delisi et al., 2003; Longshore et al., 1996; Piquero et al., 2000; Piquero and Rosay, 1998; Vazonyi, 2001).

Longshore et al. (1996) and Piquero and Rosay (1998) found results that differed from those of the original study of Grasmick et al.'s scale. These studies both used the same data from a sample of drug using offenders and found that the

[6] Principal components analysis is one of several exploratory factoring methods used for initial investigations. Principal components analysis has been the most common method of EFA used in testing Grasmick et al.'s scale. A general overview of this method will be presented and compared to other EFA's in the next chapter.

scale fit two different models, both a unidimensional and multidimensional structure with slight modifications to the scale, e.g., dropping items from the analysis. In assessing the internal structure of the scale, Longshore et al. (1996) did not find initial support for a single underlying factor, and the scale did not appear to function equally across subgroups defined by race, sex, and age. They modified the scale by dropping two items and allowing several error terms to correlate in a confirmatory measurement model, still concluding that a one-factor solution did not adequately fit the data. Next, they assessed a five-factor solution, combining two of the components, i.e., Impulsivity and Self-centeredness. This solution also provided a poor fit to the data until they allowed four error terms to correlate and one item to load on two different factors. Their modified five-factor solution provided a better fit to the data especially for juveniles (CFI = .89), males (CFI = .92), Caucasians (CFI = 93), African-Americans (CFI = .92), and adults (CFI = .91). This solution, however, provided a poor fit for women (CFI = .80)[7]. Most importantly, their results questioned the unidimensionality of the scale for a criminal population.

Given several concerns they had about Longshore et al.'s analysis, Piquero and Rosay (1998) reanalyzed the same offender data. They hypothesized that the scale could conform to a one-factor solution, could be equally reliable and valid across gender, and could produce a good fit to the data without allowing error terms to correlate. Indeed, their confirmatory measurement model showed that a unidimensional model fit the data for both males and females. While they were able

[7] CFI is the Comparative Fit Index, which varies between 0 and 1. A score exceeding .90 is recommended for a good fit, indicating that 90% of the covariation in the data is accounted for by the model (Bentler, 1992). This is one of several available goodness of fit indicators used to assess model fit in SEM measurement models. These will be discussed in detail in the results section.

to make this conclusion without allowing error terms to correlate, they did drop several items from the scale reducing it to only 19 items as they were not able to derive a unidimensional solution using all scale items. For example, the physical activities component was reduced to two items, and the impulsivity, simple tasks, and self-centeredness components were all reduced to three items each. Although Piquero and Rosay (1998) are confident that the results from their study supported scale unidimensionality, others disagree and conclude that their results are analogous to a second-order factor analysis where one overarching factor accounts for the relationships among lower level factors such as temper, risk seeking, etc. (Longshore et al., 1998). This criticism is based on Piquero and Rosay's (1998) averaging of the scores within each component, i.e., each subscale, and their use of the final six composite scores as indicators in a one-factor measurement model. In sum, the results produced by the two studies do not provide a clear, unambiguous understanding of the scale's dimensionality. In both cases, modifications were made to the scale so that the results from the analyses could conform to either a multidimensional or unidimensional structure.

Arneklev et al. (1999) employed a second-order, confirmatory factor model to test the internal structure of Grasmick et al.'s scale. They explicitly argued that theory guided their analysis. In doing so, they suggested that a valid measure of self-control should have six distinct dimensions that load on a higher-order factor of self-control; this reasoning is similar to Longshore et al.'s (1998) interpretation of Piquero and Rosay. Using a simple random sample of adults and a convenience sample of college students, they concluded that the second-order factor model fit the data well.

For the adult sample, the results showed that the coefficients between both the indicators and the six dimensions, and the six dimensions and self-control are sufficiently large. Although each of the six dimensions was significantly related to the second-order self-control factor, they found that impulsivity had the highest loading. Additionally, the physical activities dimension loaded less strongly on self-control than any other dimension. Overall, Arneklev and his colleagues (1999) concluded that the second-order factor model for the adult sample provided support for Gottfredson and Hirschi's (1993) theory. As such, the goodness of fit statistic (GFI = .89) had an acceptable magnitude, indicating that the proposed theoretical model fit the data well.

Arneklev et al. (1999) showed similar results for their college sample and all factor loadings were sufficiently large. The loading for impulsivity was the largest while the physical activity dimension was relatively small compared to other dimensions. The GFI was .88 leading them to conclude that the magnitude was sufficient for the model to fit the data. In comparing analyses from both samples, it appears that all dimensions had similar loadings on low self-control with the exception of temper. The factor loading for temper on self-control was substantially stronger for the college sample (.43) than the adult sample (.28). Considering results from both samples, Arneklev et al. (1999) concluded that evidence of six distinct dimensions exist, but evidence also indicated that all six dimensions loaded on a higher-order construct that they called self-control. Although Arneklev et al. (1999) concluded that the data fit the model well, some of the second-order loadings were stronger than others.

Vazsonyi and his colleagues (2002) used both exploratory and confirmatory factor analysis in their study on Grasmick et al.'s scale, which they conducted on adolescents in four countries. In an *a priori* fashion, they interpreted Gottfredson and Hirschi's (1990) conceptualization of self-control to be multidimensional, consisting of six separate dimensions. They first calculated exploratory factor analysis models for the total sample as well as for groups by sex, age, and country. Vazsonyi et al. (2002) argued that their preliminary results indicated the existence of six factors and that the scale is not unidimensional. Second, they use all 24 items to conduct a series of more rigorous confirmatory models including a one-factor and six-factor model to confirm their exploratory efforts. Using several fit statistics (e.g., CFI = .65, GFI = .82, and RMSEA = .09 for the total sample), they concluded a one-factor model was not a good fit to their data. Each item, however, did show statistically significant loadings. In contrast, they showed that a six-factor solution fit the data much better (e.g., CFI = .91, GFI = .95, and RMSEA = .05 for the total sample), even across groups by age, sex, and country. In a final attempt to improve the six-order factor model, they allowed for two correlated error terms and dropped two items from the scale to achieve a consistent, overall, improved fit (e.g., CFI = .93, GFI = .96, and RMSEA = .04 for the total sample) which did not vary much across groups. Unlike others (Arneklev et al., 1999; Delisi et al., 2003; Piquero et al., 2000), Vazsonyi and his colleagues did not attempt to test a second-order factor model; therefore, it is unknown if what they are calling six separate factors coalesce into a latent self-control factor.

Delisi et al. (2003) tested the dimensionality of Grasmick et al.'s scale using a sample of male offenders residing in work release facilities in a Midwestern state. Aware of the lack of clarity surrounding the self-control construct, they employed confirmatory factor models to test one-factor, six-factor, and second-order factor structures. In doing so, they allowed: 1.) all items to load on a latent variable, i.e., self-control, when testing the one-factor solution, 2.) each item to load on its respective factor for the six factor model, and 3.) items to load on their respective dimensions and then have each dimension load on the higher-order factor, i.e, self-control, for the second-order factor model. While all loadings were statistically significant in all models, they concluded that all models fit the data poorly. These conclusions were drawn using numerous fit statistics. They rejected the six-factor model that had a GFI of .85 and the second-order model that had a GFI of .84; whereas, others have interpreted similar estimates as being acceptable (Arneklev et al., 1999). Most troubling for the unidimensionality hypothesis was their results showing that the one-factor solution had the worst fit among the confirmatory factor models (GFI = .65, AGFI = .59, RMR = .11, and $x^2/df = 4.27$).

From Delisi et al.'s (2003) confirmatory analyses a model building effort was undertaken. They found that a modified six-factor model was able to fit their data well. This particular model, however, was fitted in the absence of theory and driven by model modifications like other studies in the past (Longshore et al., 1996; Piquero and Rosay, 1998). Modifications consisted of dropping three items because they

loaded highly on other dimensions[8]; however, other modifications were not made explicit. For example, they do not clearly state whether they allowed for error terms to be correlated. In sum, Delisi et al. (2003) rejected both the one-factor and second-order factor models, arguing that inadequate fit statistics led them to these conclusions.

In one of the most advanced empirical statements concerning Grasmick et al.'s scale, Piquero and his colleagues (2000) employed a Rasch measurement model to investigate the psychometric properties of the scale, administered to a sample of college students. The Rasch model is a confirmatory model that tests for scale unidimensionality, but it diverges from traditional internal structure analyses discussed thus far in many important ways. While the details of this model are articulated in Chapter Four, several of its basic advantages are discussed here before describing Piquero and his colleagues (2000) substantive findings.

First, the Rasch model produces distribution-free estimates in that the values do not depend on the distribution of the trait or attitude, i.e., self-control, across samples as does conventional exploratory and confirmatory factor models. This is important because results from Rasch models can be compared across samples when the same scale is employed, while results from factor analysis models are questionable for comparative purposes (Piquero et al., 2000; Bond and Fox, 2001). Also, this means that the Rasch model, unlike conventional factor analysis methods, is not test based.

---

[8] Delisi et al. (2003) dropped the following items: "I act on the spur of the moment without stopping to think," Excitement and adventure are more important to me then security," and "I try to avoid project that I know will be difficult."

Second, a Rasch model separates person ability and item difficulty estimates, placing them both on the same logit ruler for comparative purposes. Importantly, this function allows for comparisons of item difficulty in relation to people's level of ability, i.e., self-control, on the same interval-level scale. By taking into account the interaction between persons and scale items, the Rasch model overcomes the test-based approach of conventional confirmatory factor analysis. As such, the ability estimates do not depend on the difficulty of items in the scale. A Rasch model allows the researcher to detect the difficulty of endorsing items in relation to the range of self-control in the sample. This is important in relation to Grasmick et al.'s scale because Delisi et al. (2003: 247) posed the questions, "Does self-control work differently for different populations?" and "Is self-control equally as salient among low-risk samples, such as university students, and high-risk samples, such as prison inmates?" While the analyses conducted by Delisi et al. (2003) could not address these questions, a Rasch analysis can begin to answer these questions by separating person abilities from item difficulties.

Third, the Rasch model is mathematically defined to assess unidimensionality; therefore, researchers fit the data to the model and not the model to the data as in conventional confirmatory factor analysis. In doing so, each scale item is examined to assess its fit to the model. Finally, Rasch models create interval-level measures from ordinal items; whereas, factor analysis mistakes ordinal responses for continuous responses violating assumptions inherent in factor analysis (see Piquero et al., 2000; Wright and Masters, 1982).

Before estimating a Rasch model, Piquero and his colleagues (2000) investigated conventional exploratory and confirmatory factor models. Their exploratory analysis closely resembled results from previous studies (Grasmick et al. 1993; Nagin and Paternoster, 1994; Piquero and Tibbetts, 1996). Furthermore, they tested three confirmatory factor models including a one-factor model, six-factor model, and a second-order factor model that have all been tested in other studies. Their one-factor model produced statistically significant loadings for all items. A variety of fit statistics, however, indicated the one-factor solution did not fit their data well. A six-factor model had a questionable fit, and the second-order factor model produced an adequate fit. In sum, their analyses resembled findings from other studies in that they provide no conclusive interpretation of the internal structure of Grasmick et al.'s scale.

Piquero et al. (2000) reported five tables of results from their Rasch model analysis. First, they were interested in whether respondents used item response categories as the designer of the scale intended. This analysis can be conducted since ability can be separated from item responses. Therefore, calculations can be made so that probabilities of endorsing a certain category can be determined given a person's underlying level of self-control. Examinees use of response categories were orderly, as those with low levels of self-control had a higher probability of agreeing to each item (selecting response categories that reflect low self-control) than those with high self-control. Second, they were interested in how well scale items fit the unidimensional Rasch model. In doing so, they found that many items had poor fit to

the unidimensional expectation of the model[9]. Particularly, they found that 11 of the 24 items showed statistically significant misfit when all items were considered as a unidimensional measure; thus rejecting the hypothesis that the scale is unidimensional.

Third, Piquero et al. (2000) investigated a person/item logit ruler created by the Rasch model and determined that several items were too difficult for their college sample to endorse. Produced by a Rasch analysis, a logit ruler, or person/item map, allows researchers to assess the distributions of ability and item difficulty on the same metric to determine if items are to difficult to endorse relative to the distribution of the sample's ability. This is discussed in more detail in Chapter four. Most of their sample had very low ability indicating high levels of self-control, which would be expected with a sample of college students. While the Grasmick et al. items do not discriminate well among a college sample with disproportionately high levels of self-control, it remains to be seen whether or not these items can discriminate well among a sample of criminal offenders. Items from Grasmick et al.'s scale could be too easily endorsable for a sample of incarcerated offenders to the extent that items cannot effectively discriminate levels of self-control between them.

Finally, Piquero et al. (2000) conducted a Differential Item Function (DIF) analysis to assess item responses across high and low self-control groups[10]. Low-self

---

[9] A Rasch model analysis provides item fit statistics to assess how well each item conforms to the model's unidimensional expectations. According to Bond and Fox (2001: 26) "fit indices help the investigator to ascertain whether the assumption of unidimensionality holds up empirically. Items that do not fit the unidimensional construct are those that diverge unacceptably from the expected ability/difficulty pattern." Divergence from model expectations is often determined by investigating standardized item statistics, similar to t-statistics in linear regression.

[10] According to Bond and Fox (2000: 170-171), "DIF models the invariance of item difficulty estimates by comparing items across two or more samples" requiring that "...item difficulties be estimated for each separate sample, and that the item calibrations be plotted against each other." In

control individuals were found in some instances to respond to items differently than those having high-self control. Particularly, the low self-control group was more (or less) willing to agree with some items than would have been expected by the Rasch model.

Summary and Research Questions

This section summarizes the main points of this chapter, which provide the background for the research questions of this dissertation. First, studies estimating the reliability of Grasmick et al.'s scale have employed diverse samples ranging from convicted offenders, adult community members, adolescents residing in different countries, and college students. Only one of these studies reports a Cronbach's alpha above .9 for the total scale (Delisi et al., 2003), however, other studies do indicate that the scale has modest to high internal consistency. As noted in Chapter Two, this is necessary, but not sufficient, for demonstrating scale validity. Furthermore, the scale items appear to cohere more closely in some samples than others, while some studies show very low reliability for subscales. Although these studies generally support the internal consistency of Grasmick et al.'s scale, other reliability tests have not been used. Nevertheless, psychometricians are often skeptical of other methods of estimation, e.g., test-retest, and often prefer alpha reliability coefficients (Nunnally and Bernstein, 1994).

Based on the findings from studies that have empirically investigated the Grasmick et al.'s scale, there is no clear internal structure that emerges. As initially

---

other words, a DIF analysis compares the item characteristic function of two or more groups (Hambleton et al., 1991: 110). As stated by Hambleton et al. (1991:110), an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right" or endorsing the item.

noted, no agreement on a conceptual definition of self-control exists against which to weigh the empirical evidence. Several researchers interpret the construct, a priori, to have different conceptual properties. With this in mind, researchers have proceeded with internal structure analyses from different conceptual frameworks and have tested multiple models to determine which structure (e.g., one-factor, six-factor, and second-order factor models) is most fitting for Grasmick et al.'s scale.

With out an agreed-upon definition of a construct, it can be very difficult to achieve internal structure validity. While this remains an important issue, one firm statement can be made about Grasmick et al.'s scale. The designers of the scale originally intended for the scale to be unidimensional regardless of the conceptual definitions extracted from Gottfredson and Hirschi's (1990) theory by other researchers. Grasmick et al. (1993) explicitly stated that a factor analysis of valid and reliable indicators of self-control should produce a unidimensional measure. They found support for their hypothesis, but used an inappropriate analytic method to make such an inference, i.e., EFA. Since then, several more rigorous examinations of the scale have refuted their claims, rejecting the original findings produced by Grasmick et al. In contrast, the scale has most often been shown to be multidimensional reflecting either six-factors or a second-order factor structure. Most of these solutions have achieved good fits, however, such models were often fit by making modifications to the factor structure through dropping items and allowing error terms to correlate.

The internal structure of Grasmick et al.'s scale for offending populations is unclear. Three studies have investigated the dimensionality of Grasmick et al.'s scale

using criminal samples (Longshore et al., 1996; Piquero and Rosay, 1998; Delisi et al., 2003)[11]. Findings from these studies are inconsistent in that they have produced results supporting both unidimensional and multidimensional structures, even when the same data are used. Thus, this particular divide warrants more empirical attention with a different criminal sample.

While the Rasch model has been applied to Grasmick et al.'s scale once, no study has applied this model to data collected from a criminal sample. Thus, using the Rasch model on a criminal sample is important for several reasons. First, it will help confirm or disconfirm whether the scale's items form a unidimensional construct. Second, such a model can detect whether the items are able to distinguish levels of self-control for a criminal sample. Third, it is unknown if levels of self-control affect responses to survey items in a criminal sample, a Rasch model can shed light on this question.

Finally, conflicting results have emerged as to the scale's dimensionality and validity across demographic groups. Specifically, little is known about how Grasmick et al.'s scale operates across racial groups of offenders. For example, conflicting results have been found for whether the scale works equally well for blacks, Hispanics, and whites. Gottfredson and Hirschi (1990) do not clearly specify the factor structure that a valid measure of self-control should posses, let alone how such a structure would hold up across different races. One thing they do state, however, is that minority groups will have lower levels of self-control than whites because they

[11] Two data sets were used for the three studies, Longshore et al. (1996) and Piquero and Rosay (1998) used the same data to find different results.

are disproportionately involved in more crime. This has yet to be explored from a construct validity framework.

Drawing from the preceding arguments, the current dissertation will assess the psychometric properties of Grasmick et al.'s scale for a large sample of incarcerated male offenders by answering the following questions:

1. *Is Grasmick et al.'s scale a reliable measure for a sample of incarcerated offenders?* To stay consistent with past studies, this question will be answered by using Cronbach's reliability coefficient for the total scale as well as each of its components. Estimates will be obtained for the total sample as well as for groups disaggregated by race.
2. *Does Grasmick et al.'s scale show observed differences across racial groups for a sample of incarcerated offenders?* This particular type of validity analysis was noted in Chapter Two when discussing cross-structure analyses in a construct validity framework. As such, support for the validity of a scale is gained if the scale can distinguish between groups according to what theory would predict. In this case, Gottfredson and Hirsch (1990) imply that blacks will have lower self-control than whites. From a construct validity perspective, Grasmick et al.'s scale should exhibit these differences across racial groups.
3. *Is Grasmick et al.'s scale unidimensional?* Although mixed results have appeared, Grasmick and his colleagues (1993) do imply that their scale should reflect a unidimensional, one-factor structure. They argue that this is implied in Gottfredson and Hirschi's conception of self-control.

Staying consistent with past studies, the current effort will employ conventional EFA and CFA analyses, as well as, a Rasch model to answer this question.

4. *Is Grasmick et al.'s scale multidimensional?* The two most common multidimensional models supported thus far have been a.) a six-factor model where six dimensions are distinctly identified, but yet are correlated and b.) a second-order factor model where six dimensions are distinctly identified, however, they are best explained by a second-order factor, i.e., self-control. Two conventional CFA models will be calculated to test the fit of both.
5. *Can Grasmick et al.'s scale items discriminate among levels of ability for a sample of incarcerated offenders?* Currently, this question has not been subjected to empirical scrutiny. Some researchers, however, have entertained this idea by implying that the scale may not be equally salient for populations expected to have low self-control compared to those expected to have more self-control. This will be done by observing the distribution of item difficulties relative to the distribution of person abilities on a person/item logit ruler produced by a Rasch analysis.
6. *Do respondents' levels of ability on Grasmick et al.'s scale affect survey responses?* Hirschi and Gottfredson (1993) have argued that this will most likely be the case when self-report methods are used to measure independent or dependent variables. It could be argued that low self-control individuals will have less valid and consistent responses to items

on a self-report self-control scale. This question will be explored using a Rasch measurement model's Item Characteristic Curve.

7. *Are Grasmick et al.'s scale items invariant across racial groups?* From a construct validity perspective, items of a scale should not have different meanings for different groups of individuals. In other words, the Grasmick et al. scale items should not function differently across racial groups. While Black and Whites should vary in their levels of self-control, items should not show significantly different levels of difficulty across these two groups. If items are not invariant across groups item bias could be present. Grasmick et al. scale items should show invariance across racial groups, thus, supporting the validity of the measure. A Rasch model will be estimated to answer this question.