CHAPTER 4:

RESEARCH DESIGN AND ANALYTIC STRATEGY

Research Design

Data for this dissertation are taken from a larger study funded under a National Institute of Justice grant awarded to Dr. Julie K. Horney titled, *Patterns of Violence: An Analysis of Individual Offenders*[12]. This study was initiated to study patterns of violence among a sample of incarcerated, male offenders in Lincoln, Nebraska. The major goal of this effort was to assess violence in a cohort of serious, incarcerated male offenders over a three year, retrospective time period prior to their most recent arrest.

According to Horney (2001), this study consists of three main components that largely reflect the major goals of data collection. First, interests in time invariant correlates to violence resulted in collection of background information on each respondent via self-report measures of key theoretical variables, including Grasmick et al.'s scale, and prison records. Second, data collection centered on time-varying characteristics to ascertain relationships between intra-individual change and stability in life-circumstances and intra-individual change and stability in violent behavior over a three year period. Although not particularly relevant to the current dissertation, a life events calendar was implemented to explore change and stability issues retrospectively. Roberts (2000) gives an excellent review of this methodological approach. Third, exploring situational aspects of violent and avoided

[12] 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. The National Institute of Justice bears no responsibility for the findings presented in this dissertation.

violent events over time was the final overarching goal of the study. While the topic of the current dissertation was not one of the broader goals of the original design, these data present a unique opportunity to assess questions concerning the psychometric properties of Grasmick et al.'s scale.

Participants

All newly admitted offenders to the Nebraska Department of Corrections between October 1997 and December 1998 served as the sampling frame for this study. Approximately 10 percent of the inmates solicited for the study refused to participate (Roberts, 2000). Horney (2001) made the decision to sample from a population of prison inmates for several reasons. First, unlike general population samples where violent events are rare, an inmate sample would maximize the prevalence of violence. Therefore, analyses would not be limited by marginal variation in violence variables. Second, she argues that an inmate sample will allow for a comparison of differential involvement in violence for those who, on average, have high criminal propensities. The latter point is important for this research because Hirschi and Gottfredson (1993) argue that such samples, i.e., high in criminal propensity, are important groups for testing their theory. For example, Hirschi and Gottfredson (1993: 48) stated, "general population samples, especially samples of adults, would be expected to have difficulty in producing adequate variation on the dependent variable." In other words, they would prefer a sampling strategy that would ensure a sufficient number of low self-control subjects. In the current study, we might expect respondents to have, on average, low self-control or, at least, sufficient variability for analyses.

Prospective study participants were recently convicted offenders transported to the Diagnostic and Evaluation Center (D & E) upon conviction. This is where they reside for approximately 60 to 180 days before being transported to a permanent placement at another state facility. Inmates are housed together at the D & E, a maximum-security facility, where they are subjected to "lock down" approximately seventeen hours a day. During this time, inmates are administered a number of intelligence and psychological batteries.

Before selection of participants was started, a sampling strategy was designed so that two out of every three inmates entering the center in Lincoln were randomly selected. With the exception of parole violators, all inmates sampled were invited to participate, regardless of their most recent offense that led to their current incarceration. Parole violators were excluded from this study because they could have spent a large portion of time behind bars during the period of interest for retrospective data collection, which was three years prior to a participant's current conviction (Horney, 2001; Roberts, 2000). Randomly selected inmates were approached and asked to participate in this study during their initial stay at the D & E in Lincoln, Nebraska.

Over the course of a year at the D & E center, data were collected on a total of 704 inmates. Demographic characteristics of this sample were comparable to a nationally representative sample of over 14,000 inmates from 277 state institutions. The Nebraska sample only differed substantially with respect to race; whereas, the Nebraska sample consisted of 59 % whites, the national sample was comprised of 69 % non- whites (Beck et al., 1993). The Nebraska sample, however, was slightly more

educated in that 62.3% reported having a high school diploma or GED relative to 59% of the national sample. The two samples were similar in age in that approximately 68% of both samples were below 35 years of age. The Nebraska sample was 59% (415) white, 19.9 % (140) African-American, 10.8% (76) Hispanic, 6.5% (46) Native American, and 3.6% (26) other. The average age of the sample was 30.53 with a standard deviation of 9.57. Only 24 (3.4%) inmates from the sample had college degrees, while 121 (17.2%) had attended college. A more thorough description of this sample will be discussed in the results chapter.

Interviewers and Administration of the Interview Instrument

Graduate students from the University of Nebraska at Omaha were trained to conduct interviews of inmates at the D & E center. In past studies, graduate students have been essential to the success of conducting and completing interviews at the D & E facility (Horney and Marshall, 1991); therefore, a similar approach was used here. Although both male and female interviewers were used, research has indicated that such demographic characteristics do not have any observed impact on survey responses (Bradburn and Danis, 1984; Brenner, 1982; Hindelang et al., 1981).

After being trained, the interview/research team of graduate students began conducting interviews. First, the research team provided facility administrators with a list of inmate names to be asked for interviews on a daily basis. Once this list was in the appropriate hands, facility workers would contact the housing unit via telephone to request that a particular inmate be released to come discuss the study upstairs. This contact procedure was used two times before considering an inmate as refusing to participate in the study (Roberts, 2000).

After inmates were selected into the sample as being eligible for participation, notified of the request from staff members, and the request was accepted by the inmate, a project interviewer walked him from his cell to one of the D& E visiting rooms where the project was explained. During this interval of time, interviewers would initiate conversation to start building rapport with the inmate (Wells, 1999). Arrangements to use visiting rooms were made possible when visiting hours were not in session; typically, interview times were scheduled for two time periods on Mondays and Wednesdays, in addition to evenings for the other days of the week. The design of the rooms allowed three interviews to be conducted concurrently (Roberts, 2000).

Upon arriving in the visiting room, a trained interviewer explained the purpose of the project to inmates. At this time, an informed consent approved by the University's that all personal information collected would remain confidential. Furthermore, inmates were told they would be rewarded five dollars for participation in a completed interview. The response rate turned out to be quite good, as approximately 90% of inmates solicited agreed to enroll in the study (Roberts, 2000; Wells, 1999).

A self-report instrument was used to gather a considerable amount of data from inmates agreeing to participate in this study, which took between two to five hours for inmates to complete. The instrument included several components. First, as mentioned already, one of the largest components included retrospective data collection so that short-term change and stability could be explored across individuals with varying criminal propensities. This particular effort was designed to capture

local life-circumstances, e.g., intimate partnership, drug use, gang membership, employment, military service, correctional supervision, incarceration, etc., and involvement in violence and other criminal activity on a month-to-month account dating back three years before inmates most recent arrests (Horney, 2001). Second, questions on violent and avoided violent events were a large part of the instrument as well. Specifically, a rather large portion of the instrument was designed to collect data on multiple incidents and situational aspects of those incidents over a month-to-month, three-year, retrospective period of time (Horney, 2001; Roberts, 2000; Wells, 1999).

A life-events calendar was used to facilitate retrospective data collection, as these calendars have been used, and empirically shown, to enhance the memory of events for respondents (Caspi et al., 1996). Although this method is, overall, probably the most essential aspect of the data collection instrument, data collected using this method are not utilized in the current dissertation. While a detailed discussion of this method is not warranted here, those interested in such are referred to Horney (2001) and Roberts (2000).

Additionally, the instrument included a self-report survey component of time-invariant factors. This component consisted of hundreds of variables intended to measure key theoretical concepts including: criminal propensity/predispositions such as early onset of criminal involvement and low self-control; high-risk for violence vignettes; family history variables concerning childhood and adolescence such as educational achievement of caregivers, role of religion, safety of the neighborhood

they grew up in, parental monitoring, supervision, and physical and verbal abuse directed towards them by caregivers while growing up.

Responses to all survey questions were recorded using a computerized survey instrument. Interviewers read survey questions to inmates from laptops and responses were entered directly into computers. In doing so, inmates sat at the same table next to the interviewer and computer so they could view responses being recorded. According to Roberts (2000: 50), this positioning of inmates during the interview "encouraged positive rapport between the respondent and the interviewer by alleviating any suspicions about what was being recorded."

Some benefits of administering a self-report survey in this fashion are apparent. First, interviewers can assist inmates in understanding questions. Second, reading questions aloud to inmates could minimize problems they may have with literacy and reading. Such advantages should help enhance the internal validity of the instrument. Third, this method of survey administration is more efficient for the research team. Particularly, this method minimizes the amount of time researchers would have used for entering and coding data. Furthermore, data entry errors associated with paper survey instruments now become much less of a concern (Maxfield and Babbie, 2001; Wells, 1999; Roberts, 2000).

While not part of the survey instrument itself, official records data were also collected to supplement, and in some cases validate, survey data. First, criminal records data, including prior arrests and criminal convictions, were obtained mainly from pre-sentence investigation reports. Specifically, criminal records data extracted from reports included: date of first arrest and conviction, number of times given

prison and jail sentences, total arrests and total violent arrests, and information related to each conviction. Second, psychological test data were collected from tests administered to inmates while residing in the D & E center. These included an IQ test as well as standardized scores from the Minnesota Multiphasic Personality Inventory (MMPI).

Measures

Most important for the current dissertation's purposes, the computerized self-report instrument includes the original 24-items of Grasmick et al.'s scale. These 24 items will be used for the majority of all analyses. Items have response categories ranging on a four point Likert scale. Responses were coded as 0 (strongly disagree), 1 (disagree), 2 (agree), and 3 (strongly agree). As in the original Grasmick et al. scale, larger numbers reflect low self-control. Descriptive statistics for the scale items are presented in Chapter Five.

Inmates' racial identifications are self-reported and used as a variable in this dissertation. Race was originally coded as 0 (White), 1 (African American), 2 (Hispanic), 3 (Native American), and 4 (Other); however, comparisons will only be made between blacks and whites[13]. For the current study, race is coded as 0 (White) and 1 (Black). Race is important to the current study for construct validity purposes.

[13] A couple of arguments can be made for investigating differences between only Blacks and Whites. First, many of the statistical methods that are employed in the current study use a Maximum Likelihood (ML) function for model estimation. This estimation procedure requires larger sample sizes for stable results; therefore, ruling out several groups from the sample, e.g., Hispanics and Native-Americans. Second, although Gottfredson and Hirschi (1990: 152-153) state that differences in self-control exist across racial groups, they do not explicitly state which racial and/or ethnic groups will have low self-control. They do, however, discuss difference in crime rates for Blacks and Whites, implying that variation in crime rates is due to variation in self-control across the two groups. Gottfredson and Hirschi's (1990) vague propositions would then prevent specific predictions about Hispanics and Native Americans. This is something, however, that future research and theory may want to consider.

As such, the race variable will be used to assess Grasmick et al.'s scale across racial groups to test specific hypotheses about the scale and its items.

Analytic Strategy

The analytic strategy component of this chapter has several sections. First, a general outline is presented showing how the analyses will progress. Second, a more detailed discussion will be presented on each quantitative method that will be used to test the core research questions that were proposed in Chapter Three.

General Outline of Analyses

The analyses begin with univariate and bivariate statistics. Descriptive sample statistics for individual demographic variables are presented first[14]. Next, descriptive statistics for all 24 items of Grasmick et al.'s scale are presented including means, standard deviations, and Pearson product-moment correlations between items. These statistics will allow for 1.) detection of variability for each item and 2.) a preliminary assessment of relationships between scale items to assess whether items are significantly correlated in the expected directions. Finally, independent sample t-tests are performed on the scale and its components to assess differences across race. This analysis is employed as a preliminary group differences analysis of scale validity that assesses Gottfredson and Hirschi's (1990) proposition that minority groups should have substantially lower self-control.

Next, reliability estimates using Cronbach's alpha coefficients will be calculated for Grasmick et al.'s scale items, as well as its components, to assess the coherence of items. This particular type of reliability estimation is used to stay

[14] While only one demographic variable is used in the analyses, I present other demographics for the study sample to give the reader a more thorough description of the inmate sample under investigation.

consistent with past studies. Furthermore, these analyses will be able to detect whether any items should be taken out of the scale to increase the reliability of the measure. Reliability estimates will be calculated for the total sample and for racial groups.

To address questions concerning the internal structure validity of Grasmick et al.'s scale, the next set of analyses will follow a three-step analytic approach. Step one and two employ conventional factor analytic methods to remain consistent with previous research testing the dimensionality of the scale. To remain consistent with previous studies, a Principal Components Analysis will first be performed on the total sample. This analysis will be similar to Grasmick and his colleagues' analysis reported in their initial study where they inferred unidimensionality. In addition, Principal Component Analyses will be conducted across racial groups. Second, conventional Confirmatory Factor Analyses are calculated using structural equation models (SEM's) using AMOS 4.0. Since recent empirical research has not reached an agreeable consensus on the conceptual definition of self-control nor the empirical dimensionality of the scale, three models will be calculated including a one-, six-, and second-order factor model. If any of these models fit the data acceptably well according to SEM model fit standards, the fitted model will be further tested for invariance across racial groups using a stacked model analysis. The final analysis will subject Grasmick et al.'s scale to a Rasch model. The next three sections will describe each of the three analytic methods in some detail. The goal of doing this is to differentiate the qualities of each of the three approaches, and highlight what the Rasch model has to offer that Principal Components Analysis and conventional

Confirmatory Factor Analysis techniques cannot. Before discussing each of analytic method, a discussion of the merits of categorical versus continuous data is needed.

Continuous Versus Categorical Data in Statistical Estimation

With the exception of the Rasch model, it should be noted that most statistical analyses, e.g., standard deviation, correlation, factor analysis, employed in the current dissertation typically assume the use of interval level measurement. In both psychology, and criminology, however, this assumption has often been relaxed and ordinal level data have been treated as continuous for reasons that will be discussed in this section. Although the interval level measurement assumption can be important, evidence has been compiled to show that divergence from using interval level measures, e.g., ordinal or categorical measures, will not produce fatal flaws in statistical estimations if other requirements are met.

Most importantly, approximation of normality is essential if ordinal level measures are to be used. If ordinal measures have highly skewed distributions the following can occur: standard error estimates can be unreliable, correlations can be underestimated, and model fit statistics can be inaccurate (Byrne, 2001). When the ordinal level data are not severely skewed, the above negative consequences are minimized. In this dissertation, the ordinal level items of Grasmick et al.'s scale do not show signs that indicate high degrees of skewness, thus, severe departures from normality are not problematic. As reported in Table 2, the largest skewness statistic for all items is a 1, whereas, the other items have substantially lower skewness statistics. A skewness statistic of 3 or higher has been suggested as a "rule of thumb" for departure from normality (Dennis W. Roncek, personal communication,

September, 1999), but some argue that skewness statistic higher than 1 can cause problems in certain instances (West, Finch, and Curran, 1995). According to these criteria, data used in the current dissertation are not violating this assumption.

If normality of ordinal variables is achieved, Bentler and Chou (1987: 88) argue that, "continuous methods can be used with little worry when a variable has four or more categories." As such, the Grasmick et al. self-control items used in this dissertation have four categories per item. In general, several negative consequences are alleviated when ordinal variables approximate normality. For example, the number of categories of ordinal variables will have minimal effects on model fit in SEM confirmatory factor models. In addition, factor loadings and correlations between factors are, at the most, moderately underestimated. Underestimation, however, can become quite severe when ordinal level variables have three or fewer categories and skewness statistics are greater than 1. As for the current dissertation, these concerns are not problematic, as Grasmick et al.'s scale items have four categories and skewness is 1 or less for each item. In addition to the potential for violating the normality assumption when using ordinal level data, there is also a concern that correlations will be attenuated when correlating ordinal level measures (See Bollen and Barb, 1981). This is important because many of the analyses presented in Chapter Five rely on correlation matrices, e.g., factor analysis. West and colleague's (1995) have stated that Pearson correlations can be higher when assessed between two interval level measures than when the same two variables are correlated after being collapsed into ordered categorical scales. They go on to suggest that attenuation is greatest when ordinal variables have a high degree of skewness. Once

Table 2. Skewness and kurtosis statistics for Grasmick et al.'s scale items

Items	Skewness	Kurtosis
Impulsivity		
I1	-.262	-.959
I2	.976	.552
I3	-.267	-.848
I4	-.212	-.988
Simple Tasks		
S1	.589	-.644
S2	.968	-.118
S3	-.126	-.875
S4	.541	-.735
Risk Seeking		
R1	-.752	-.034
R2	.051	-1.328
R3	.530	-1.093
R4	.551	-.799
Physical Activities		
P1	-.106	-.912
P2	-.811	.076
P3	-.485	-.595
P4	-.423	-.338
Self-centered		
Sc1	.362	-.864
Sc2	1.000	.281
Sc3	.917	.144
Sc4	.627	-.541
Temper		
T1	.447	-1.084
T2	.915	-.101
T3	.322	-1.181
T4	.105	-1.028

again, a high level of skewness is not problematic for the items used in this dissertation. Bollen and Barb (1981) analyzed simulated data to assess the differences in correlations when variables are both continuous and then when the same variables are both categorical. In contrast to West and colleagues (1995), they concluded that it may be justifiable to analyze categorical data as if they were continuous. Specifically, Bollen and Barb (1981) concluded that in practically all of their simulations the differences between the collapsed correlations (between categorical variables) and continuous correlations (interval level measures) were rather small. In consideration of the above points, a degree of confidence can be placed in the ordinal level data used in the analytic models in this dissertation[15].

Exploratory Factor Analysis

According to Gardner (2001: 237) "factor analysis is a generic term used to describe a family of techniques that have as their purpose the investigation of the relationships among a set of individual difference variables." Factor analysis has several purposes for analyzing items in a scale or items that may lead to the creation of a scale. Its most primary function has been to aid researchers in detecting how many latent variables may underlie a set of scale items or variables. In addition, it provides a strategy that helps explain variation between a relatively large number of items using few variables created from the analysis. Finally, these methods are beneficial to researchers wanting to define the substantive meaning of factors that

[15] It should also be noted that alternative ways for dealing with categorical data do exist. Few programs exist that allow for analysis of categorical data from a factor analysis framework. Comparisons between such methods and those used in the current analysis are beyond the scope of this dissertation. Nevertheless, one particular method used in this dissertation, i.e., Rasch modeling techniques, can readily deal with categorical data when assessing internal structure validity of a measure. As discussed later in this chapter, the Rasch rating scale analysis takes into account the ordinal scale items.

account for variation between many items. This is done by identifying items that covary with each other (Devillis, 1991).

Several analytic methods exist that could be listed under the category of exploratory factor analysis. The most common approaches have been Principal Components Analysis (PCA) and Principal Axis Factor Analysis (PAF). These methods have several features in common. Although the nature of the mathematics used for estimation slightly differ across techniques, both tend to provide very similar answers concerning the dimensionality underlying a given set of variables (See MacIntyre, 1990). It is not the current dissertation's goal to debate the relative merits of each of these techniques, as this is beyond its scope. However, differences between PCA and PAF will be discussed.

Disagreement exists concerning the similarities and differences between PCA and PAF. Cliff (1987: 349) states that:

> Some authorities insist that component analysis is the only suitable approach, and that the common-factors methods just superimpose a lot of extraneous mumbo jumbo, dealing with fundamentally unmeasurable things, the common factors. Feelings are, if anything, even stronger on the other side. Militant common-factorists insist that components analysis is at best a common factor analysis with some added error and at worst an unrecognizable hodgepodge of things from which nothing can be determined. Some even insist that the term factor analysis must not be used when a components analysis is performed.

The above statements represent how two camps are divided in terms of the differences between factor analysis, i.e., PAF, and components analysis, i.e., PCA.

Although the mathematical properties underlying PCA and PAF are similar, the theories underlying these techniques are quite different (Gardner, 2001). The main difference is that PCA is aimed at explaining total item variance, i.e., both error

variances and variance unique to the variables; whereas, PAF is designed to extract common variance, i.e., variance shared by the indicators excluding error. Although error is included in the components or factors extracted from PCA, the components or factors are not correlated with each other. PCA extracts the largest amount of shared variance among items for the first factor, the next largest amount for the second factor, and so on. Such a technique is typically used when little is known about the construct or items being investigated and when no a priori theoretical structure exists. PAF analyzes only the variance that is common to the indicators; therefore, error variance is not included (Carmines and Zeller, 1979; DeVellis, 1991; Gardner, 2001). Rotation can be used with PAF so that factors can be independent or correlated with each other[16]. Rotation is discussed later in this section.

The main difference between the two techniques is reflected in the correlation matrix used for each estimation technique. While the off-diagonal of the correlation matrices used in both methods represent the correlation between items, the diagonals of the correlation matrices used for each technique differ. The diagonal of a correlation matrix consist of unities, the variance of an indicator or item[17]. The diagonal values of the correlation matrix in PCA consist of 1's. In contrast, the diagonal of the correlation matrix analyzed in PAF consists of communalities rather than a particular variables variance. Communalities reflect the variance accounted for by the common factors[18]. Such a correlation matrix is also known as a reduced correlation matrix. This particular correlation matrix has a principal diagonal with

---

[16] Stevens (1996) argues that rotation can be performed on PCA results as well.
[17] The variance of a standardized variable is 1.
[18] Communalities are typically estimated for a particular variable by calculating an R^2. This indicates the multiple correlation of a particular variable with the other variables or the common variance for a particular items

values different from 1. In PAF, the number of factors extracted is less than the number of variables (DeVellis, 1991; Gardner, 2001); whereas, the number of components extracted in PCA will equal the number of variables in the analysis. Typically, a decision must be made regarding factor extraction in PAF which is usually guided by the results from PCA

While some have argued that the above differences are important (Pedhazur and Scmelkin, 1991), researchers have concluded that these two techniques produce similar results (Gardner, 2001). The current dissertation reports results from both PCA and PAF for comparison purposes. The reason for reporting both is that, as mentioned above, there are different opinions about these two techniques. Some argue that PCA and PAF can produce different results under certain conditions (Pedhazur and Schmelkin, 1991), while others suggest that these differences are minimal and choose PCA as a psychometrically valid and mathematically simpler approach (Stevens, 1996).

The typical exploratory factor analysis starts by calculating a covariance matrix from the items or variables. After calculation of the matrix, the factor analysis program attempts to extract factors that account mathematically for the covariation among items, thus, attempting to identify shared variance among them. Two widely used criteria guide factor extraction in exploratory factor analysis: Kaiser's eigenvalue rule (Nunnally, 1978) and Cattell's (1966) scree discontinuity plot. The Kaiser rule is based on extracting only factors that explain more variance than the average amount explained by one of the original items. Eigenvalues above one are interpreted as potentially meaningful factors or representative of the possible number

of dimensions that exist. The scree discontinuity plot is a factor extraction criterion consisting of graphed eigenvalues. Cattell (1966) recommends graphing the amount of variance explained by each successive factor extracted from the covariance matrix of items. A general rule endorsed by Cattell (1966) is that factors should be retained above what he labels the "elbow." The "elbow" is the point at which extracted factors begin to explain small amounts of variance. Factors above this point explain the most variance among items. Similar to the "elbow", another interpretation of the number of potential factors or dimensions is through assessing the largest drop in eigenvalues. These techniques have produced conflicting results when used to draw conclusion about Grasmick et al.'s scale, indicating the possibility of a one factor and six factor solution.

Factor rotation has been an important issue in EFA. According to DeVellis (1991: 100) "a goal of factor rotation is to find a set of arbitrary factors that provides the clearest conceptual picture of the relationships among the items by approximating simple structure." Orthogonal and oblique rotations are two types of factor rotation methods. These rotational methods can be distinguished by the outcome they produce. Orthogonal rotation implies rotating the factors at right angles which keep the factors independent of each other. This means that factors will not be allowed to correlate with each other; thus, the interpretation of any particular factor does not affect the interpretations of any other factor. Such a structure is obtained if subsets of items are associated with one, and only one, factor. Varimax is the most common orthogonal rotation method. In turn, oblique rotation allows factors to be correlated with each other. This correlation is allowed by rotating factors at more or less than

90 degree angles. The interpretation of the factors can overlap. This depends on the magnitude of the correlation between any two factors. Oblimin is the most common oblique rotation method (Devillis, 1991; Gardner, 2001; Pedhauzer and Schmelkin, 1991).

Factor rotation is inextricably entangled with theory. As such, the rotation method used in exploratory factor analysis depends largely on how the construct under investigation is understood. Theory should drive rotational decisions. To illustrate, different conceptions of self-control can lead to different ideas regarding the appropriateness of factor rotation. As noted in Chapter 3, some suggest that self-control consists of one general factor while others suggest that it consists of six factors that are correlated. Any tendency for a general factor in the data is minimized when an orthogonal (varimax) rotation is used (Pedhauzer and Schmelkin, 1991); therefore, employing this rotational method on Grasmick et al.'s scale items would assume that six uncorrelated factors best describes the data as opposed to a general self-control factor. In addition, this method would help also make decisions about whether a general factor exists. Typically, the first factor extracted in factor analysis tends to be a general factor that explains the largest amount of variance among items (Pedhauzer and Schmelkin, 1991). If a general one factor solution was expected then it could be argued that there would be no need for rotations. Finally, it could be that six correlated dimensions explain the data best. In this case an oblique (oblimin) rotation would be most appropriate.

Deciding the method of rotation is not an easy task. Pedhauzer and Schmelkin (1991: 615) suggest:

> From the perspective of construct validation, the decision whether to rotate factors orthogonally or obliquely reflects one's conception regarding the structure of the construct under consideration. It boils down to the question: Are aspects of a postulated multidimensional construct intercorrelated? The answer to this question is relegated to the status of an assumption when an orthogonal rotation is employed. This is grounds enough to question the wisdom of limiting oneself to orthogonal rotations, even when theoretical formulations lead one to expect factors to be not correlated. The preferred course of action is, in our opinion, to rotate both orthogonally and obliquely. When on the basis of the latter, it is concluded that the correlations among the factors are negligible, the interpretation of the simpler orthogonal solution becomes better.

Although factor rotation is a central issue in exploratory factor analysis, it will not be a central focus in the current dissertation, as correlations between factors will be assessed using confirmatory factor analysis methods.

In light of the above discussion, PCA and PAF have several limitations. First, both are typically used to assess covariation between items when no a priori theory is proposed by a researcher or the number of latent variables is not determined before conducting the analysis. Second, it leaves the task of defining the factor structure up to the statistical program itself; thus, a detailed model linking the latent to the observed variables or items is not specified in advance. Consequently, all latent variables will have a tendency to influence all observed variables, error terms are not allowed to correlate, and errors for observed variables are not taken into account (Bollen, 1989). Bollen (1989: 228) states that if "hypotheses about plausible structures exist, then exploratory factor analysis can frustrate attempts to test these ideas." In the case of Grasmick et al.'s scale, plausible hypothesized structures have been proposed. In the presence of such hypothesized structures, a different analytic technique is needed that can specify, a priori, the dimensionality or factor structure of items.

Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) requires specifying, a priori, items that theoretically should be indicators of underlying latent constructs. Thus, CFA is used when fairly clear ideas have already been advanced about what factors are thought to underlie items. In short, the CFA requires a detailed understanding of which factor structure(s) might be expected for a given set of items. In the current dissertation, such analyses are appropriate because several specifications have been advanced for Grasmick et al.'s scale and Gottfredson and Hirschi's (1990) construct of self-control.

CFA is usually estimated using structural equation modeling programs where measurement models are specified to assess whether the model fits the data or covariance structure. In doing so, the CFA can specify whether a latent variable is influencing a particular set of observed variables or several items, measurement errors for items are specified as variables and assumed to be uncorrelated, latent variables can covary, and factors can be specified so that they have no effects on items loading on other factors (Bollen, 1989; Kline, 1998). In general, a researcher wants to test the covariance structure hypothesis that $\Sigma = \Sigma(\theta)$. However, because Σ and $\Sigma(\theta)$ are population parameters, they are unknown, so the goal of the CFA is to assess the overall fit of the hypothesized covariance structure to the observed covariance structure. The fit of a model to the observed covariance among items is usually determined by a number of goodness of fit statistics, e.g., GFI, CFI, RMSEA, etc., as well as a chi-square statistic (Bollen, 1989; Kline, 1998; Hayduk, 1987). A Maximum Likelihood (ML) function is the typical fitting function used to estimate a

CFA (Bollen, 1989). This is a much more powerful analytic tool than EFA when attempting to confirm or disconfirm hypotheses concerning dimensionality.

According to Bollen (1989), the general CFA model is little more than the following equations:

$$X = \Lambda_x \xi + \delta \quad (4.1)$$

$$Y = \Lambda_y \eta + \varepsilon \quad (4.2)$$

Equations 4.1 and 4.2 have the same meaning; where X and Y are observed variables or items, ξ and η are latent variables, and δ and ε are variables representing measurement error. In these equations, observed indicators depend on one or more latent variables and errors of measurement vectors, errors are uncorrelated with latent variables, and coefficients representing effects of latent variables on observed variables are in the matrix of Λ_x or Λ_y depending on the equation used. Futhermore, δ or ε in either equation consists of two components in a CFA:

$$\delta = s + e \quad (4.3)$$

In equation 4.3, *s* reflects the specific variance affiliated with each indictor or item; whereas, *e* is the remaining random component in the specific variance associated with each observed indicators. Bollen (1989) refers to both δ and ε as random errors of measurement which are observed variables in a CFA model.

These equations can be applied to the logic of the current study. For a one factor confirmatory analysis, Xs or Ys would represent each of the 24 items of Grasmick et al.'s scale; ξ or η would indicate the specified latent variable of self-control; Λ_x or Λ_y would contain the coefficient describing the effects of self-control

on each of the 24 items; and δ or ε would be the random error of measurements associated with each of the 24 items.

Although the above CFA equations are straight forward, the equations for a second-order factor model are more difficult. The second-order model extends beyond basic CFA's in that associations between factors do not remain unanalyzed and are presumed to have a common, unmeasured cause that accounts for their intercorrelations (Kline, 1998). This model is based on a theoretical argument that suggests low self-control affects specific latent components of the construct (i.e., impulsivity, risk seeking, simple tasks, etc.) that in turn are measured by four item subsets that are directly observed. Equations for a second-order confirmatory factor analysis are explained below. The first-order model is given by Bollen (1989) and Joreskog and Sorbom (1989) as:

$$y = \Lambda_y \eta + \varepsilon \quad (4.4)$$

$$\eta = (\Gamma \xi + \zeta) + \varepsilon \quad (4.5)$$

Equation 4.4 describes the first-order loadings where η denotes lower order factors and ε represents measurement error. The first-order factor loadings of η on y are in Λ_y as discussed in the basic CFA equation in 4.1. These factors are Impulsivity (η_1), Simple Tasks (η_2), Risk Taking (η_3), Physical Activities (η_4), Self-Centeredness (η_5), and Temper (η_6). Each first-order factor has direct effects on each of the four indicators. Finally, the Λ_y is a 24 x 6 matrix with one indicator per factor (η) being scaled to one. Equation 4.5 describes the second-order loadings where ξ denotes the second-order factor, or the self-control construct. Second-order factor loadings of low

self-control (ξ) on η are in the Γ matrix and ζ is a vector of unique variables for η. Combining equation 4.4 and 4.5 gives the following equation:

$$y = \Lambda_y (\Gamma \xi + \zeta) + \varepsilon \quad (4.6)$$

Finally, equation 4.7 represents the covariance matrix for the second-order confirmatory solution:

$$\Sigma = \Lambda_y (\Gamma \xi \xi' \Gamma' + \zeta' \zeta) \Lambda_y + \varepsilon' \varepsilon \quad (4.7)$$

Although one-factor and second-order CFA's are important for answering questions concerning Grasmick et al.'s scale, such models cannot answer questions concerning (1) how item difficulty, Grasmick et al. scale items, interacts with person abilities, e.g., level of self-control (2) whether scale items can adequately discriminate among levels of self-control in a sample, and (3) whether each item of the scale contributes importantly to an underlying construct. The next section discusses the merits of a model that can answer these questions.

The Rasch Model

In general, tests of internal structure validity for criminological measurement scales have been limited to models that utilize conventional factor analytic methods described above. Such approaches have received considerable use as methods for summarizing covariances among variables, dominantly being used to create measures and test the internal structure validity of measures. Recently in criminology (Piquero et al., 2000), and rather routine in psychology and education (Bond and Fox, 2001), Rasch measurement models are being used for two purposes: (1) to create linear measures from categorical data and (2) scale validation. This section will discuss the following aspects of the Rasch model: (1) creation of the Rasch model, (2) calculation

of the Rasch model and its most fundamental tenants, (3) advantages the Rasch model has over traditional approaches, (4) an overview of important output produced from a Rasch analysis, and (5) common uses of Rasch measurement models .

The Rasch model was created by a Danish mathematician named Georg Rasch (1960). The Rasch model represents a group of measurement models used by researchers to develop variables or measures from dichotomous or categorical data (Bond and Fox, 2001; Wright and Masters, 1982). In addition, the Rasch model is used to test the internal structure validity of scales comprised of multiple items that are intended to measure only one attribute or construct, e.g., self-control (See Bond and Fox, 2001).

Georg Rasch (1960, 1980) developed the Rasch model when analyzing data from remedial tests administered to training recruits of the Danish army. Rasch administered a difficult and easy test to recruits in a pre-post test fashion. He started by plotting recruit abilities on a graph where the x-axis represented scores on the easy test and the y-axis represented scores on the difficult test. In doing so, he noticed that the plotted scores were approximating a straight line, but the line was not completely straight. A curvilinear relationship emerged (Mike L. Linacre, personal communication, May, 2003).

The curvilinear relationship that emerged indicated that those scoring 100% on the hard test also scored 100% on the easy test, and those scoring 0% on the easy tests scored 0% on the hard test. However, those scoring 25 % on the hard test scored 50 % on the easy test, and those scoring 50 % on the hard test scored 75 % on the easy test. In light of these findings, Rasch pursued the reason for the observed

relationship. He concluded that it must be a person's ability interacting with the difficulty level of the test, but ability could not be observed directly (Mike L. Linacre, personal communication, May, 2003).

As a mathematician, Rasch proposed a mathematical solution to his problem by using a logistic curve to represent how ability influences test scores[19]. In doing this, Rasch was able to calculate abilities of persons in logits and place them on the x-axis, equal interval distances apart. The y-axis represented test difficulty from easy to hard. If a good measure was to be produced the test would have to measure abilities on a straight line that ranged from $+\infty$ to $-\infty$. This mathematical solution led to what is known as the Rasch model, a mathematically derived model that estimates person ability independent of the difficulty of a test or instrument, where the relationship between ability and difficulty are modeled as a probabilistic function (Linacre, personal communication, May, 2003).

Unlike conventional factor analytic approaches, Rasch models are used to transform nonlinear measures into linear measures by using simple mathematical procedures (See Wright and Masters, 1982: 33-37). In doing so, the Rasch method converts raw scores from test items and persons into log-odds units or logits, thus, resulting in two parameters: ability of person n, (B_n), and the difficulty of item i, (D_i). The ability of a person represents his/her level on some latent trait, e.g., high or low self-control, relative to the difficulty of items that comprise the measurement instrument. Item difficulty is defined by the level of ability required to have a 50

[19] This curve should be familiar to criminologists because it is the distribution used for logistic regression analysis, where logits are the natural unit for the logistic curve.

percent chance of answering an item correctly or agreeing to a particular item in the case of Likert scaled responses (Bond and Fox, 2001; Wright and Masters 1982).

Person ability and item difficulty are placed on the same linear scale consisting of logit units that range from + ∞ to - ∞ (Bond and Fox, 2001; Wright and Masters, 1982). This allows the distance between person abilities, e.g., high or low self-control, and item difficulties, e.g., easy or hard to endorse self-control items, to be articulated and compared with relative ease. Luce and Tukey (1964) refer to this property as additive conjoint measurement.

By using a method that allows for the estimation of both person ability (or agreeability) and item difficulty, researchers are able to calculate the probability of agreeing to or correctly answering an item given a person's level of ability on an underlying trait. Thus, the Rasch model assumes that item responses are determined by a person's position on the underlying trait and the difficulty of the items that comprise the scale, both which are estimated by the Rasch model. This suggests that an interaction exists between persons and items. For example, when a difference between a person ability and item difficulty is zero, the probability of an agreeable response is 50 percent. The probability of an agreeable response is increased when a person's ability is higher than the item's difficulty. In contrast, the probability of a correct response is decreased when a person's ability is lower than an item's difficulty (Bond and Fox, 2001). Such probabilities can be produced by the Rasch model since both person ability and item difficulty are transformed into the same logit scale. The Rasch model then produces expected probabilities for how persons with varying levels of abilities should respond to items with varying difficulty levels.

These model expectations are then compared to the actual observed patterns in the data.

Initially, the Rasch model was intended for tests or instruments that consisted strictly of dichotomous items. This model is expressed as follows (Rasch, 1960/1980: See also Bond and Fox, 2001):

$$Ln\,[P_{ni} / 1 - P_{ni}] = B_n - D_i \qquad (4.8)$$

Equation 4.8 shows that the log odds of person *n* selecting the correct response to item *i* is expressed as the difference between person *n's* ability, i.e., *B*, and item *i's* difficulty, i.e., *D*. This equation can be extended to calculate the probability of choosing the correct answer for a given item as a function of person ability and item difficulty as follows (Bond and Fox, 2001):

$$P_{ni}\,(X_{ni} = 1 \mid B_n, D_i) = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}} \qquad (4.9)$$

Equation 4.9 shows that $P_{ni}\,(X_{ni} = 1 \mid B_n, D_i)$ represents the probability of giving a correct response, given person *n's* ability (B_n) and item *i's* difficulty level (D_i). The probability involves using the exponential contant *e*, which is 2.7183, raised to the difference between a person's ability and an item's difficulty divided by 1 plus the same value.

Since its inception, several extensions of the original Rasch model have been advanced that can take into account data that are not dichotomous. One of these has been the rating scale model (Andrich, 1978; Wright and Masters, 1982). The rating scale model is an expansion of the dichotomous model in that it is equipped to handle items that have more than two categories that are ordered (Bond and Fox, 2001). Specifically, the rating scale model was created to analyze several ordinal level items

that comprise a measure of a particular construct. As will be discussed, this approach can create interval level measures from ordinal level data using similar transformation techniques discussed previously. This technique is especially important to the current dissertation because Grasmick et al.'s scale items are ordinal.

The rating scale model has a similar equation to the original Rasch model with the exception of an additional parameter, a threshold parameter where each item threshold has its own difficulty estimate. The number of thresholds for an item is contingent on the number of categories. The number of thresholds is determined by the number of categories for an item minus 1. Thresholds are cumulative across response categories. For example, if an item has four categories (0 = strongly disagree, 1 = disagree, 2 = agree, 4 = strongly agree) then the number of thresholds will be three. Specifically, this parameter is modeled as a threshold at which a person has a 50 percent chance of choosing one category over another, e.g., the likelihood of a person choosing disagree over strongly disagree or the likelihood of a person choosing strongly agree over agree. Each threshold parameter is estimated once across the entire set of items in the rating scale; therefore, the threshold parameter doesn't vary by item. The rating scale model is expressed as follows (Bond and Fox, 2001):

$$Ln\,[P_{ni} / 1 - P_{ni\,(k\text{-}1)}] = B_n - D_i - F_k \qquad (4.10)$$

where the log odds of agreeing or endorsing category k relative to k – 1 is determined by the difference between D_i, person ability, and B_n, item difficulty, where F_k is the threshold parameter for category k.

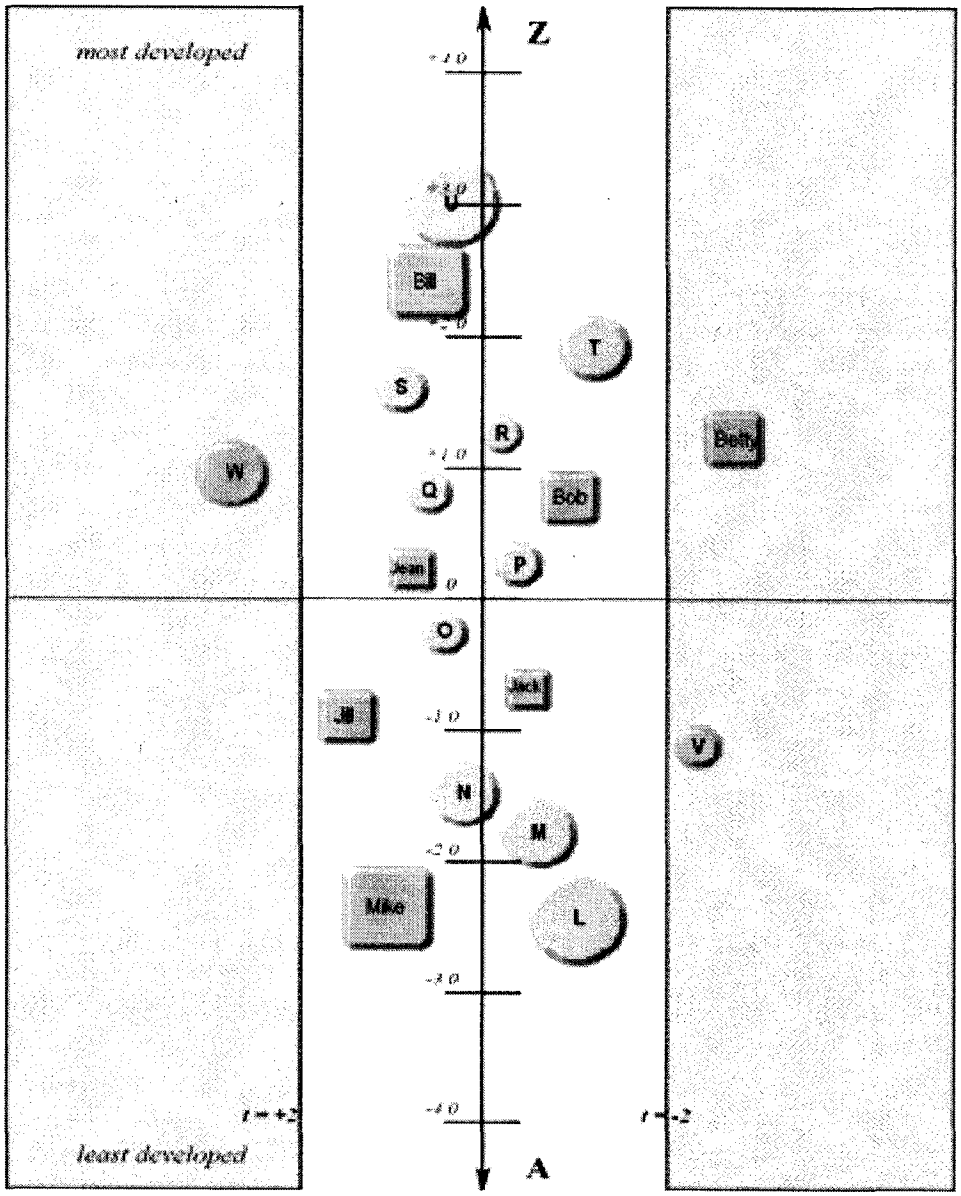
The log odds in equation 4.10 can easily be converted to probabilities using equation 4.11 (Bond and Fox, 2001):

$$P_{nil}(\mathrm{x} = 1 \mid B_n,\ D_i, F_l) = \frac{e^{(B_n - D - F_1)}}{1 + e^{(B_n - D_i - F_1)}} \qquad (4.11)$$

where P_{nil} is the probability of person n choosing category 1 (disagree) over category 0 (strongly disagree) on item i. The difficulty parameter for the first threshold is represented by F_1. Equation 4.11 is similar to equation 4.9 with the exception of a threshold parameter.

The above discussion has described the formulas for the Rasch model. What is lacking is a visual depiction of what the model is doing. Figure 1, which was adopted from Bond and Fox (2001), shows a conceptual drawing of how persons and items are considered on the same logit metric. Furthermore, it gives a visual explanation of what is important about a Rasch model. The logit scale is an interval scale that is located at the center of the diagram. It is vertical, with a mean of zero and can range from + ∞ to - ∞. Both items and persons are located on this diagram, where persons are indicated by squares and items by circles. Each person and item is located on the same interval, logit metric according to its estimated value. Thus, both items and persons are located on the same map and their estimates are read vertically. More positive persons on the vertical logit scale have higher ability and more positive items on the vertical logit scale are more difficult to endorse. For example, if this was a map of the Grasmick et al. measure,

Figure 1. A Visual display of the Rasch Model



individuals with more positive scores would have higher ability (or agreeability), in turn, indicating lower self-control. Bill would have the highest ability and Mike would have the lowest ability. As for items, those that are more positive would be items that are more difficult to agree on. For example, item U would be the most difficult and L would be the easiest. Given the difference between a person's ability (or agreeability) and a particular item's difficulty, it would be easy to estimate a person's probability of endorsing a particular item. A Rasch model would expect that Mike should have a lower probability of agreeing to item U than Bill, given that Bill has a higher ability and item U has a high difficulty. However, this may not be the case, as the observed data, or responses to items, may not fit the expected probabilities of a Rasch model. If not, the unidimensional expectation of the Rasch model is not supported. This leads to the issue of item fit, which is depicted by persons and items falling into the gray boxes on each side of the vertical logit scale. However, item fit will be discussed in regard to Figure 1 in the next section. Finally, measurement error for items and persons are indicated by the size of circles and squares, where larger circles and squares indicate more error for items and persons, respectively.

This section has outlined the model specifications of the most used Rasch models. Furthermore, it has described how the Rasch model places persons and items on the same logit scale to produce a linear measure. Finally, this section discussed the Rasch model in relation to a visual diagram that hopefully helps readers understand the model's conceptual properties. Next, I discuss the advantages of Rasch models over traditional approaches.

Advantages of the Rasch Model

A Rasch modeling approach differs substantially from traditional methods, e.g., factor analytic approaches, which are used for scale creation and validation. According to Schumaker and Linacre (1996: 470), "factor analysis is confused by ordinal variables and highly correlated factors. Rasch analysis excels at constructing linearity out of ordinality and at aiding the identification of the core construct inside a fog of collinearity." Four advantages the Rasch model has over traditional methods will be discussed.

First, respondent ability cannot be separated from test characteristics in traditional approaches, therefore, such methods are sample and item dependent. In traditional psychometric approaches a respondent's ability is defined only in terms of the choice of items that comprise a particular test; thus, respondent ability is dependent on the items used to measure a construct. As such, respondents will have lower ability scores on difficult tests and high ability scores on easier tests. The problem with this approach is that test statistics are dependent on the sample in which they are obtained. As discussed above, the Rasch model makes a correction for this problem by estimating item difficulty and person ability scores, i.e., by percentage of items agreed and then converting into logits, separately. By doing so, researchers can determine if items used in the test are too easy, too difficult, or well suited to measure the range of abilities in the sample. Item statistics are reported on the same scale as person ability estimates, allowing the researcher to know the expected relationships between items and persons. Specifically, the Rasch model takes into account the

interaction between item difficulty and person ability by generating expected probabilities of a correct or agreeable response for a particular item given the difficulty of that item and the person's ability.

Second, traditional approaches are largely test-oriented as opposed to item-based. Traditional approaches such as factor analysis only allow for conclusions to be drawn concerning the internal structure validity of a measure through assessing inter-item correlations. In contrast, a Rasch analysis allows researchers to assess each item's contribution to the underlying construct that is being measured by investigating item fit statistics. The Rasch model assumes that the items are measuring a unidimensional construct. If items do not fit the mathematical expectations of unidimensionality of the Rasch model then such items could be measuring a different construct or the items could be confusing to the respondent. Thus, items that do not fit diverge from the expected item/ability pattern defined by the Rasch model. Item fit will be discussed in detail later. In contrast, item fit, relative to person ability, can not be assessed using any of the traditional factor analytic approaches discussed earlier. In other words, EFA's and CFA's methods can not provide any information on how persons respond to items.

Third, the Rasch model creates linear, continuous measures from categorical data, e.g., dichotomous or ordinal variables, which range from $+\infty$ to $-\infty$. The Rasch model explicitly recognizes the categorical nature of items that comprise a scale. That is, the data are regarded as ordinal and not continuous. These measures are calculated by taking into account the difficulties of answering or endorsing items which traditional methods do not. The linear measures created by the Rasch model

are only meaningful if all items contribute to the measure of a single construct, thus, unidimensionality is mathematically implied by the Rasch model. As such, this approach is a construct validation tool that allows researchers to assess whether items create a unidimensional measure or not. As stated above, this can be detected by assessing fits statistics for each item. In stark contrast from the Rasch model, traditional factor analytic strategies assume that items comprising the scale are continuous level measures and they are not capable of creating continuous measures from categorical data.

Fourth, the most common approach to scale construction consists of administering a set of items to a sample of individuals to measure a particular construct (Fox and Bond, 2001; Fox and Jones, 1998). Once administered, it is common for researchers to sum responses to items for each person to create composite scores to represent person ability on the entity being measure. Using this approach assumes that items have equal weight and they each contribute equally to the underlying construct. Furthermore, after the item responses are summed for each individual they are treated as if they all fall on an equal interval scale. This approach is flawed because it does not take into account the difficulty level or endorsability of items comprising the scale, suggesting that all items are equally weighted and contribute the same amount in measuring a construct under investigation when in fact they may not. Fox and Jones (1998: 30) give the following example:

> For example, items measuring anxiety with respect to mundane events (such as asking a sales clerk for help) are weighted the same as those measuring anxiety in more extreme situations, such as speaking in front of a large crowd. It seems nonsensical to treat endorsement of both of these qualitatively different items as equal contributors to a total anxiety scale.

The Rasch model empirically tests the assumption of equal item weighting before responses to items are summed to create a trait score. Thus, items are not arbitrarily treated as equal contributors to represent the quantity of a trait, they are empirically assessed before hand by examining the overall responses to each item (Fox and Jones, 1998). This is done through separately calculating person ability and item difficulty parameters (Bond and Fox, 2001; Hambleton, Swaminathan, and Rogers, 1991; Wright and Masters, 1982). A psychometrically sound alternative to summating items is offered by the Rasch model when creating measures from scale items (Bond and Jones, 1998).

Finally, this section has shown that the Rasch model can answer questions that factor analytic methods cannot. Nevertheless, factor analysis should not be viewed completely separate from the Rasch model, as factor analysis can help determine whether or not to use a Rasch model. For example, Smith (1996) discussed when a researcher should progress from a factor analytic method to a Rasch model when analyzing scale items. For construct validity purposes, a Rasch analysis should be pursued when factor analytic methods reveal that one factor dominates the items comprising the measure or when scale items are dominated by highly correlated factors. In turn, when the correlation structure of scale items is dominated by distinct, uncorrelated factors factor analysis will suffice, as this would suggested that several unique constructs are being measured by the items that comprise a particular scale. In the case of Grasmick et al.'s scale, the majority of investigations have shown that factor analytic methods do not produce distinct, uncorrelated factors. A Rasch

analysis then should be used to separate its items out of the test to identify the dimensionality and scalability of the items.

A Rasch Analysis: What is Important to Report?

A Rasch analysis typically proceeds by estimating one of the models discussed earlier. The model chosen depends on the type of items comprising the scale that is under investigation. That is, the equation will change depending on whether scale or test items are comprised of dichotomous or ordinal data. WINSTEPS version 3.42 (Linacre and Wright, 1999-2001) is the program used to estimate a Rasch model in this dissertation. Specifically, a rating scale analysis will be conducted on Grasmick et al.'s self-control scale using WINSTEPS. Once the model has been estimated and the linear measure created, interpretation of the output progresses in several stages. This section will explain five aspects of output generated from a Rasch analysis that are important for answering research questions proposed in this dissertation. These aspects are as follows: (1) category functioning (2) item fit statistics (3) Rasch map that places person abilities and item difficulties on a common logistic scale (4) Item Characteristic Curve (ICC), and (5) Differential Item Functioning (DIF) across racial groups. These are the most commonly reported aspects of a rating scale analysis[20] (See Fox and Bond, 2001; Wright and Masters, 1982).

The first stage of a Rasch analysis should determine whether respondents or examinees are using response categories or the rating scale (e.g., 0 = strongly

[20] Aspects of a Rasch analysis are very similar across both dichotomous and rating scale models. The main difference is that a rating scale analysis must assess category functioning across items to assess whether response categories are being used as would be expected by the Rasch model. All other analyses discussed are common to both models.

disagree, 1 = disagree, 2 = agree, 3 = strongly agree) across items as anticipated by the researcher. The way in which rating scales are constructed can influence the quality of data obtained from the scale (Clark and Schober, 1992). Furthermore, measurement quality can be reduced and the fit of items to a Rasch model can be affected if categories are being used by respondents inappropriately or respondents have difficulty distinguishing between certain categories, e.g., agree vs. strongly agree. A Rasch analysis can be used to empirically assess these potential problems. If a category is being used by respondents in an inappropriate way or they lack understanding of a category, the researcher must consider restructuring the actual rating scale by collapsing problematic categories with adjacent categories. A Rasch analysis can establish how respondents actually used response categories. Bond and Fox (2001: 161) suggest that the goal is to "produce the rating scale that yields the highest quality measures for the construct of interest."

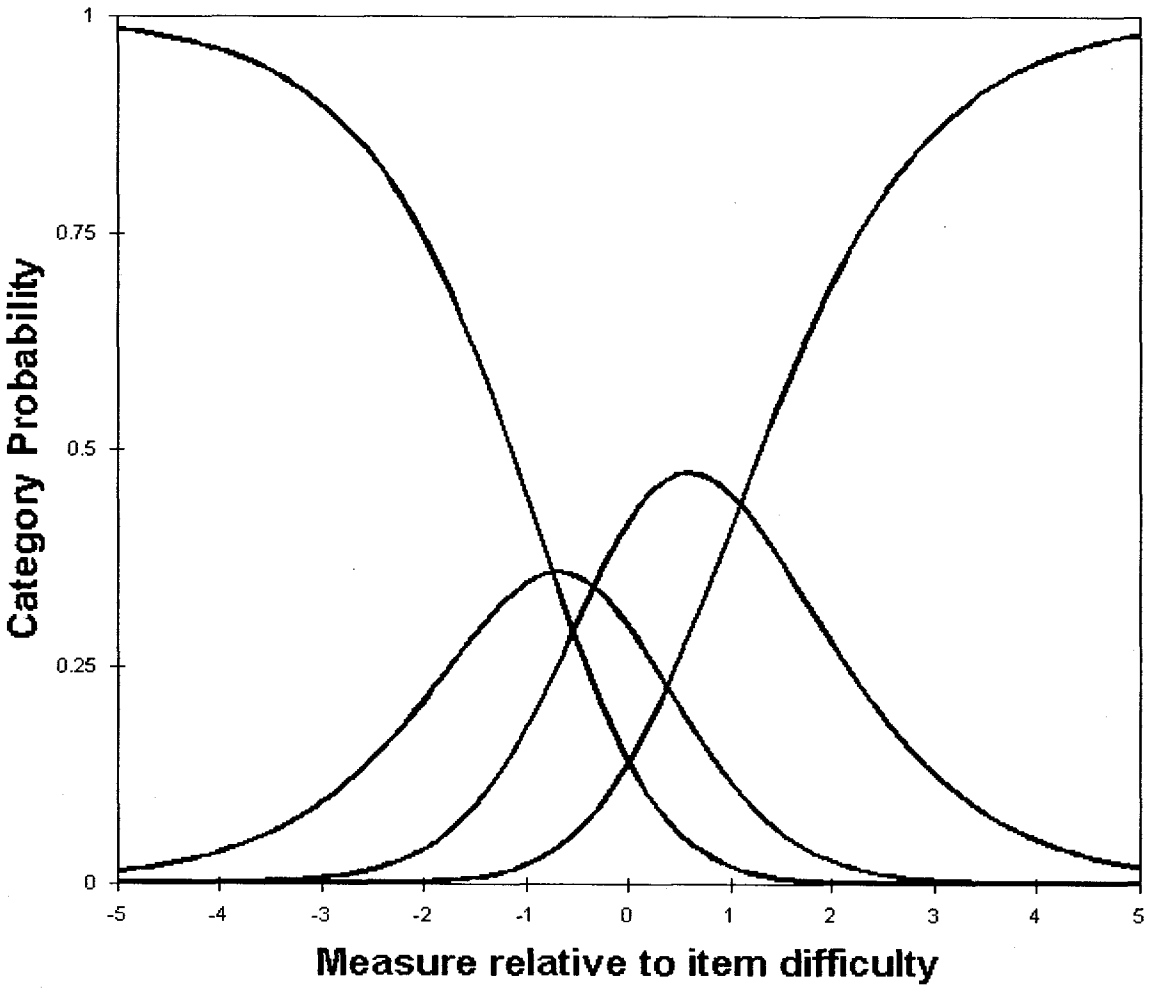
In WINSTEPS, several options exist for empirically assessing category functioning. In the first stage of a Rasch analysis, it is important to assess the category use statistics for each response category, i.e., category frequency and monotonicity of average measures (Bond and Fox, 2001; Linacre, 1995). A category frequency distribution can detect the number of respondents choosing a particular response category, which is summed for each category across all scale items. In examining this distribution, it is important to give attention to the distributions shape and the frequency of responses per category. The distribution should not be highly skewed and each category should have a sufficient number of responses. Linacre (1999) has suggested that 10 responses per category should be the minimal

recommended number. Low frequency categories are problematic because they do not have enough observations to estimate stable threshold parameters and they often reflect unnecessary or redundant categories; thus, these are the categories that should be collapsed into better functioning categories. (Bond and Fox, 2001).

To assess the monotonicity of average measures, the average ability measure, i.e., logit score, for persons endorsing a particular category across any item is investigated empirically. The mean of person ability measures are expected to increase in size as the variable increases, e.g., strongly disagree to agree. As such, a monotonic increase is expected where respondents with high ability endorse the higher categories across items. In turn, those who have lower abilities endorse lower categories. In the current dissertation, high ability will reflect lower self-control and is represented by larger positive logit scores. The rating scale is coded so that higher scores reflect lower self-control, i.e., strongly agree indicates low self-control and strongly disagree reflects high self-control. Therefore, the average person ability is expected to increase in size from the strongly disagree to strongly agree categories. Collapsing categories should be considered if this pattern does not emerge (Bond and Fox, 2001).

Another way to empirically assess category functioning is by plotting the probability of choosing any particular category across any item given a range of person ability estimates. Figure 2 shows a category probability plot. As shown, this analysis is reported on a graph that represents category probability distributions on the y-axis and differences between any person abilitiy and item difficulty on the x-axis. Each category in the rating scale has an estimated probability distribution that

Figure 2. Example of a category probability plot



spreads across the person ability estimates. These probability distributions should have a distinct peak reflecting that each category is the most probable response at some point along the linear measure that the Rasch model has estimated. A category is considered disordered or ill-functioning, e.g., a category that is confusing to respondents, if it has a flat probability distribution across the person ability continuum or linear measure. Such a confusing category would not distinguish between different ability levels well. That is, those with high and low agreeability would have a similar probability of endorsing the category that had a flat probability distribution. It is expected that a person with high ability (or high agreeability) has a higher probability of choosing strongly agree for any item as opposed to strongly disagree. If a category is not functioning well then this pattern would not be observed. Also, the monotonicity of thresholds[21] across the rating scale can be visually inspected by assessing the category probability distributions (Bond and Fox, 2001). Thresholds that do not increase monotonically across the response categories may also reflect a disordered or ill functioning category.

The second stage of a Rasch analysis requires an empirical assessment of the fit of items to the model's expectation of unidimensionality. This analysis is particularly important in the current dissertation as it will be used to test whether the Grasmick et al. scale items are contributing to a unidimensional construct or not. With regard to the Rasch model, Wright and Masters (1982: 114) state, "when an explicit measurement model is used, the internal validity of a test can be analyzed in terms of the statistical fit of each item to the model in a way that is independent of the

[21] Remember that a threshold in a rating scale is the point on the linear measure or person ability continuum where a person has a 50% chance of endorsing one category of another, e.g., agree to strongly agree.

sample distribution." Each respondent, depending on their ability, has a certain probability of endorsing an item that has a particular level of difficulty. If an item has internal validity the probabilistic expectation will be seen in the observed data. Thus, internal validity in a Rasch model is assessed by how well observations across items meet the probabilistic expectations of a Rasch model.

Fit statistics are used to assess the success of each item in meeting the unidimensional property of the Rasch model. An item may show signs of misfit for several reasons. For example, the item may be to complex, confusing, or actually measuring another construct. Two primary measures are used to asses item fit in a Rasch analysis: infit and outfit. Infit statistics assess misfitting responses to items near the person's ability level. Outfit statistics are concerned with misfitting responses to items farther away from a person's ability level. Both measures are mean squared residual statistics with an expected value of 1 and a range from 0 to + ∞. Mean squared residuals are basically the difference between observed responses to items and expected responses to items given a person's level of ability. Residuals for each person answering a particular item are squared and then summed across persons (Bond and Fox, 2001). Values greater than 1 indicate that responses to a particular item could be driven by something other factor than the construct under investigation, e.g., bad question, item is measuring another construct, or a confused respondent. As a "rule of thumb" for rating scale analyses, Wright and Linacre (1994) suggest that mean squared residuals ranging between 0.6 – 1.4 indicate acceptable item fit. A more common method of reporting fit statistics is the standardized normal transformation of the mean squared residual which is identical to

a t-statistic with a mean close to 0 and standard deviation of 1. Values greater than +2.0 and less than -2.0 indicate a statistically significant misfit ($p < .05$). Items with less than – 2.0 suggest substantially less variation in the item than predicted. Items with greater than a +2.0 indicate substantially more variation in the item than predicted. The Rasch model expects there to be a range of unpredictability around a person's level of ability; therefore, the infit statistics are most important for reporting purposes (Bond and Fox, 2001; Wright and Masters, 1982). Once indentified, misfitting items can be examined further to understand why they fall outside of the rasch model standards of unidimensionality.

Referring back to Figure 1, misfit items can be shown in a visual graph. As seen, some circles (items) fall in the white area and others in the two gray areas. Items having acceptable fit fall in the white area, having t-statistics between -2.0 and +2.0. Those falling in the gray area are considered misfitting, as they have t-statistics greater than 2 in absolute value. More specifically, items located in the gray area to the right (t-statistic of 2.0 or above) perform too erratically to the Rasch model, and items located in the gray area to the left (t-statistics of -2.0 or below) perform too predictably and overfit the Rasch model.

The third stage of a Rasch analysis requires an investigation of the person-item map or "ruler." As discussed earlier, person abilities and item difficulties are estimated by a Rasch model and then expressed on a common scale, i.e., logit scores ranging from $-\infty$ to $+\infty$. This allows for an examination of item functioning relative to the sample of respondents. The distribution of item difficulties can be compared to the distribution of person abilities graphically by creating a map in WINSTEPS.

Figure 3 shows an example of a Rasch person-item map created for a measure of visual ability among a sample of low vision patients. This map was adopted from Stelmack et al. (2004) for illustrative puposes only. According to Stelmack et al. (2004: 239), "the Rasch person-item map displays a ruler created from the measurements of persons' abilities to perform activities of daily living and the visual ability needed to perform each activity. The Rasch person-item map...orders the level of self-reported visual ability of the patients in our study (left side) and the difficulty of the activities (right side). Activities at the top of the scale are easier to perform. Activities become more difficult to perform further down the scale. Subjects with the least visual ability (at the top of the scale) have difficulty even with the easiest activities; subjects with more visual ability (at the bottom of the scale) have no difficulty performing any of the activities." This map is for illustrative purposes. While the format of this map is conceptually the same as the one that will be displayed in this dissertation, there are differeneces. For example, in Stelmack et al.'s (2004) person-item map, items with increasingly more positive logits indicate items that are easist to endorse. In this dissertation, items with more positive logits indicate items harder to endorse. The same goes for person logits as well. Therefore, this map should not necessarily be compare to the one estimated in this dissertation.

Creating a visual map of the two distributions is important for several reasons. Most importantly, a map can be used to determine the extent to which item positions, i.e., easy to hard agreeability, match the range of person abilities. Several problems could exist if the distribution of scale items does not resemble the person ability distribution on the logit ruler (Bond and Fox, 2001; Wright and Masters, 1982).

Figure 3. Example of a Rasch person-item map: A measure of visual ability

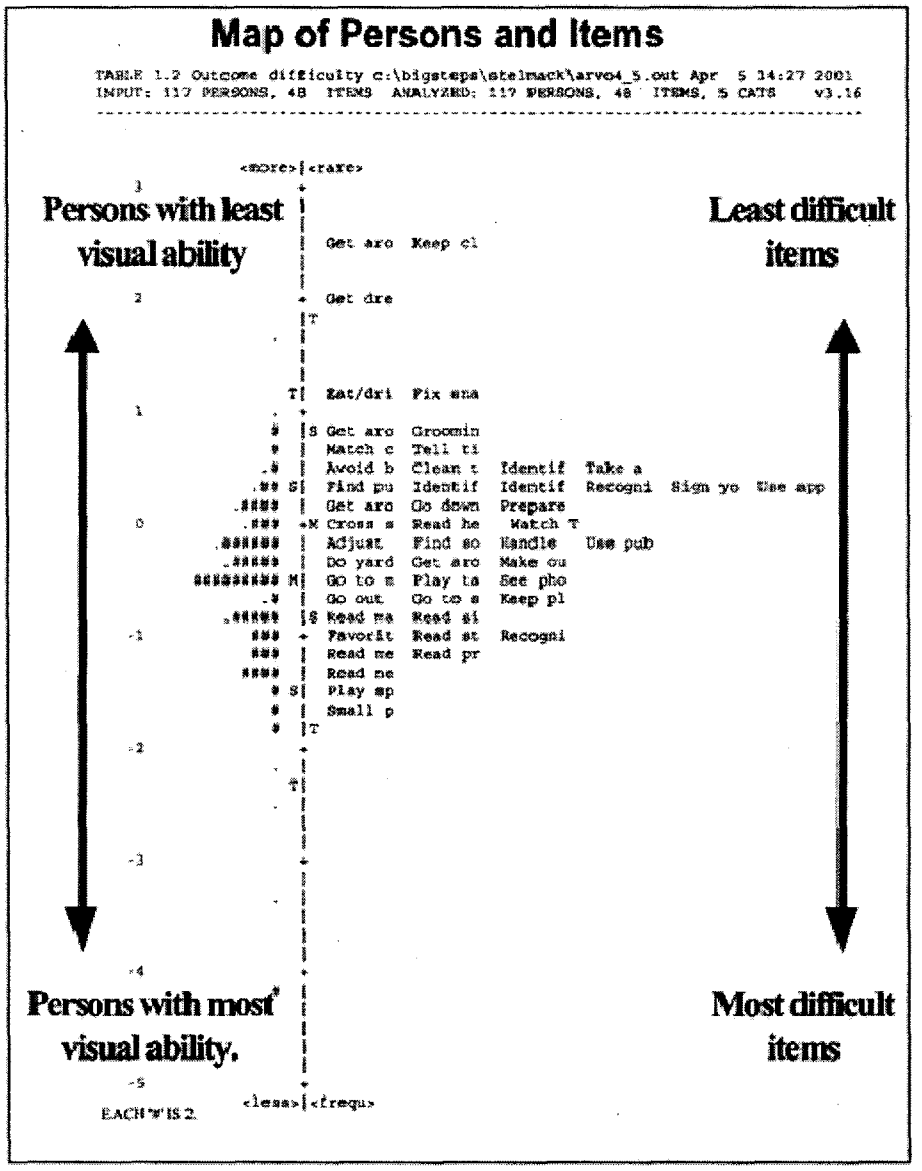


Figure 2.
Map of persons and items.

Note: Adopted from Stelmack et al. (2004)

Items could be too difficult for the sample to endorse or items could be too easy for the sample to endorse; thus, resulting in a set of items that do not accurately depict person abilities or producing a floor or ceiling effect. This would imply that scale items are not appropriate for measuring the range of abilities in the sample. This is particularly important for the current dissertation because until now it is unknown whether the Grasmick et al. scale items are appropriate for a sample of incarcerated offenders. It could be that these items are too easy to endorse for these offenders because they, on average, may have very low levels of self-control. This would imply that new items be created that can discriminate between levels of self-control for a sample of people that, on average, have low levels of self-control.

The fourth stage of a Rasch analysis consists of investigating the Item Characteristic Curve (ICC) plotted against person ability estimates. The ICC is the basic logistic curve that ranges from $-\infty$ to $+\infty$. The ICC reflects the expected number of scale items endorsed as a function of person ability estimates. The observed distribution can be plotted against the expected distribution to detect incongruence across person abilities (Linacre, personal communication). This stage of the analysis is important to this dissertation because it explores a central research question, that is, whether attitudinal self-control instruments designed to measure self-control are less effective for low self-control individuals?

The fifth stage of a Rasch analysis should investigate item functioning across different subgroups. The idea is that items should function in the same manner across groups defined on some variable. This research defines groups by race. Therefore, the goal of this analysis is to empirically assess whether items comprising the

Grasmick et al. self-control scale function differently or the same across white and black incarcerated male offenders. This analysis is referred to as a Differential Item Functioning Analysis (DIF). This procedure requires item difficulties to be estimated for each groups separately. Item functioning should remain invariant across groups, thus, no statistically significant differences in item difficulties should be observed across these groups for the measure to show invariance. Item "bias" is a concern if statistically significant differences are observed for items across groups (Bond and Fox, 2001). A standard z score is used when estimating statistical significance of item differences across groups as follows (Hickman et al., 2004):

$$Z = \frac{(d_1 - d_2)}{(se_1^2 + se_2^2)^{1/2}} \qquad (4.12)$$

where *d* indicates item difficulty and *se* indicates the standard error associated with the item being compared across groups.

Some Uses of the Rasch Model

Rasch measurement models have a rich history dating back several decades in educational and achievement testing (Rasch; 1960; Rasch, 1980). The importance of the Rasch model has been recognized across several disciplines, resulting in the use of this modern psychometric method for test construction and scale validation. The Rasch model has been used for the development and validation of many scales such as client satisfaction with public education (King and Bond, 2003), computer anxiety (King and Bond, 1996), self-esteem (McRae, 1991), decision making for nurses (Fox, 1994), academic responsibility (Green, 1996), attitudes towards rape (Fox and Jones, 1998), children's exposure to violence (Kindlon, Wright, Raudenbush, and Earls, 1996; Selner-O'Hagan, Kindlon, Buka, Raubenbush, and Earls, 1998), fear of crime

(Wright and Masters, 1982), and fear of falling among elderly (Velozo and Peterson, 2001), to name a few. The following paragraphs discuss the details of several uses of a Rasch rating scale analysis from studies cited above. Although not exhaustive, discussion of the following studies is provided to show uses of the Rasch model that are, in several cases, similar to the one used in this dissertation.

In assessing the measurement properties of an exposure to violence instrument, i.e., MY ETV, Selner-O'Hagan et al. (1998) performed a Rasch rating scale analysis on responses from subjects ages 9 to 24 participating in the Project on Human Development in Chicago Neighborhoods. According to Selner-O'Hagan et al. (1998: 218):

> The Rasch model constructs a linear measure for ETV by using response data as realizations of the probabilities of item endorsement given the level of that items extremeity [item difficulty] and the exposure of the person responding to the item. The measure of a person's ETV and the item clibrations are on the log-odds metric, which can vary from minus infinity to plus infinity, and are expressed in unit called logits."

My ETV is an 18 item scale measuring exposure to violent events that range in level of severity, e.g., witnessing a murder to witnessing someone being hit. Standardized item fit statistics were estimated and results showed that items generally fit a unidimensional construct. Two items, hearing gunfire and being hit or punched, showed poor fit using standardized item fit statistics, with a cut-off for acceptable fit being 2.00 in absolute value. That is, on these two items, some subjects with high ability, high exposure to violence, showed low frequency for these items and some with low ability, or low exposure to violence showed high frequency for these items. They also assessed a person-item map to investigate the distribution of items and

persons on a logit scale. Doing this allowed them to explore how well items (i.e., difficulty of endorsing items) covered the range of person abilities (i.e., exposure to violence). They found that the items did well for covering the range of abilities in the sample. That is, they were not to difficult or too easy to endorse.

Support for the construct validity of My ETV was also provided in a way that is different from the approach taken in this dissertation. Selner-O'Hagan et al. (1998) not only were interested in the unidimensionality of their ETV measure, but they hypothesized that items would have differing degrees of seriousness or item difficulty. Therefore, they assessed seriousness of items in relation to difficulty of endorsement. For example, they expected individuals to have a more difficult time endorsing items such as witnessing a murder or rape vs. witnessing a person being hit. This is another possible way to assess construct validity from a Rasch measurement approach. They found support for their hypothesized ordering of items.

In assessing the measurement properties of a fear of falling measure among elderly, i.e., University of Illinois Chicago fear of falling measure (UIC-FFM), Velozo and Peterson (2001) performed a Rasch rating scale analysis on responses from over 100 community dwelling elderly subjects. The UIC-FFM is a 19 item scale that measures common activities designed to represent an increasing level of concern about falling among older adults. First, a category probability plot showed that individuals were using categories as expected. In other words, those having high fear of falling were more likely to use the rating "very worried" than they were the other categories. Second, A Rasch rating scale analysis showed that three items did not fit the unidimnsional expectations of the model. That is, the three items causing

the least worry (i.e., getting dressed, getting out of bed, and getting on/off the toilet) were statistically erratic in that they had standardized fit statistics 2.0 in absolute value. Specifically, those with generally lower levels of fear (lower ability) worried about falling when doing these activities and those with higher levels of fear (high ability) did not worry about falling when doing these activites. Third, a Rasch person-item map was used to assess person measures of fear (or ability) and item measures on same linear continuum, a logit scale of ruler. Velozo and Peterson (2001) found that the items were well matched to the sample under study. That is, items were not too difficult or too easy to endorse given this sample's distribution of abilities or general fear of falling.

In assessing the measurement properties of a computer anxiety scale, i.e., the Computer-Anxiety Index (CAIN), King and Bond (1996) performed a Rasch rating scale analysis on responses from over 300 eleven-to twelve-year-old elementary school students. The CAIN is a 26 item Likert-type test that measures an underlying trait called "computer anxiety." This instrument has items such as "I enjoy using computers" and "I sometimes feel that computers are smarter than I am." King and Bond (1996: 55) state that, "the extent to which the CAIN data conform to the estimations based on the Rasch model can be interpreted as evidence for the test's adherence to the basic assumptions that all items...measure the same thing." Discussing the Rasch model, King and Bond (1996: 55) state:

> It is a mathematical formulation that predicts the probabilities of the responses of a sample of subjects, who vary in the exhibition of a trait, to a test that has items of varying endorsability designed to detect the presence of that trait; in the case of the CAIN, the presumed trait is computer anxiety. Rasch analysis is an Item Response Theory model and presumes that, in this case, it is the difference between the anxiety level of any person and the degree of anxiety

> detected by any item (its "difficulty" level) which determines the likelihood that any person will endorse any item.

With respect to unidimensionality and the Rasch model, King and Bond (1996: 55) state:

> Rasch first determines the extent to which the items in a test conform to the unidimensionality premise that all items should measure the same latent trait (variable) and as a consequence, that the individual item scores can be added together arithmetically to produce a meaningful total score.

King and Bond (1996) found six of their twenty-six items to have substantial misfit, therefore, the twenty remaining items conformed to the unidimensional expectation of the Rasch model. The remaining twenty items could be used to measure a single trait for their sample of elementary school students. They also reported a Rasch person-item map of person ability and item diffculty distributions on the same logit scale. They found that few items detect anxiety levels at the low end of computer anxiety, as most items were more diffcult to endorse for individuals having low ability. Using the person-item map, King and Bond (1996: 60) concluded that, "researchers might be interested in attempting to develop new items to detect more precisely lower levels of computer anxiety and thereby provide further discrimination between persons exhibiting extremely low computer anxiety."

So far, several uses of the Rasch model have been discussed from different studies. One area that has not been discussed is the use of the Rasch model in educational testing. Although this particular use of the model is beyond the current dissertations scope, the model's use in an educational setting is no different from how it is used in other settings. For example, creators of applied educational examinations, e.g., medical board examinations, entrance examinations, etc, attempt

to find items that are measuring a unidmensional construct where items vary in difficulty so that they cover the range of abilities tested.

Test development using the Rasch approach has been a success in many applied educational settings. For example, Masters (1997) shows that Rasch analysis has been beneficial for developing many high stakes medical certification exams in the United States. Also, Rasch analysis has been used by the National Board of Medical Examiners, National Board of Osteopathic Medical Examiners, American Society of Clinical Pathologists, to name a few. Furthermore, the Rasch model has been successfully applied to medical school competency tests including the Australian Medical School Admissions Test. Masters (1997) also found that the Rasch model has been used successfully to create a variety of school assessment tests in public school systems such as Vancouver, Washington and Portland Oregon.

Unlike some disciplines, criminology has not begun to realize the full potential of the Rasch model for creating measures and testing scale validity. To date, only few criminologists have applied the Rasch model to scales commonly used in criminological research (Hickman, Piquero, and Piquero, 2004; Piquero, MacIntosh, and Hickman, 2002; Piquero et al., 2000). With construct validity being the major goal, criminologists have tested whether items comprising a scale fit the unidimensional expectation of the Rasch model and if items comprising a scale function differently across different sub-groups of respondents.

Recent applications of the Rasch model in criminology have begun to change how the measurement qualities of commonly used criminological scales are perceived. For example, using data from the National Youth Survey (NYS), Piquero

et al. (2002) applied the Rasch model to one of the most used delinquency scales, i.e., Self Report Delinquency Scale (SRD). They found that several scale items functioned poorly across groups, as several items were biased. Furthermore, they recommended that researchers using the scale in the future may want to consider dropping several of the misfitting items and reword items to make them more sensitive to age, gender, and race.

Using data collected from Philadelphia police officers, Hickman, Piquero, and Piquero (2004) applied the Rasch model to a commonly used police cynicism scale. They found that many of the items did not fit the unidimensionality expectation of the Rasch model, scale items exhibited a lack of invariance indicating gender and race bias for items, and the scale could be improved by dropping several confusing items. Using data on responses from college students, Piquero and his colleagues (2000) found a similar pattern when they applied the Rasch model to the Grasmick et al. self-control scale in that several items did not fit the unidimensional expectations of the Rasch model and that item bias was present across different groups. The above criminological examples suggests that the Rasch model can produce new insights concerning the quality of measures that are used in criminology, as it has in other disciplines; however, criminologists are only now beginning to understand the use and importance of the Rasch model.

Summary

This chapter has provided a discussion on the data and methods that are used to test research questions presented in Chapter Three. First, this chapter discussed the research ideas steering data collection efforts for the overall project on patterns of

violence, characteristics of the sample of offenders being used for this dissertation, interview methods used to obtain data from offenders, and measures used in this dissertation. Second, a detailed discussion of the analytic procedures used in this dissertation followed. The goal was to describe each method that will be used to assess the internal structure validity of Grasmick et al.'s self-control scale, including: exploratory factor analysis, confirmatory factor analysis, and the Rasch model. Each method was discussed in the context of how they will be applied to the Grasmick et al. scale items. In addition, the Rasch model was discussed in detail to show how it diverges from traditional ways of assessing internal structure validity, the advantages of the model, how and what to interpret from output generated from a Rasch analysis, and common uses of this model.

Before moving to Chapter Five, the important research questions of this dissertation should be repeated. As stated in Chapter Three, 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?*
2. *Does Grasmick et al.'s scale show observed differences across racial groups for a sample of incarcerated offenders?*
3. *Is Grasmick et al.'s scale unidimensional?*
4. *Is Grasmick et al.'s scale multidimensional?*
5. *Can Grasmick et al.'s scale items discriminate among levels of ability for a sample of incarcerated offenders?*

6. *Do respondents' levels of ability on Grasmick et al.'s scale affect survey responses?*
7. *Are Grasmick et al.'s scale items invariant across racial groups?*

CHAPTER 5:

RESULTS

Univariate and Bivariate Analyses

Before moving to the analyses that address the main research questions, it is important to describe the sample of offenders used for this dissertation. Table 3 reports descriptive and univariate statistics for selected variables. Due to missing data, this dissertation uses a sample of 651 offenders from the original sample. Missing cases did not significantly differ from other cases on main variables such as race, education, and self-report offending. As shown in Table 3, the first group of variables is demographics and the second is general, self-report criminal offending variables.

Demographic variables include race/ethnicity, education, and the number of reported marriages. As for race and ethnicity, 18.7% (122) of these offenders are black, 13.1% (85) are Hispanic/Mexican, 58.1% are white, 6.5% (42) are Native American, none of the offenders are Asian, and 3.7% (24) are categorized as other. The majority of this offender sample has less than a high school education or a high school education (including GED's). Specifically, 39.1% (254) has less than a high school education, 39.8% (259) are either high school graduates or have their GED, 17.2% (112) have some college and 3.8% (15) are college graduates or received post graduate study. As for marriage, 56.5% (368) of offenders report never being married, 32.6% (212) report being married once, 8.9% report being married twice, 1.5% (10) report being married three times, .3% (2) report being married four times, and .2% (1)

reported being married five times. The mean for number of times married is .57 with a standard deviation of .76.

The second part of Table 3 reports self-report criminal offending variables that include onset and frequency of offending. As for age of onset, this sample of offenders were asked: How old were you when you were first involved in crime (not necessarily caught or arrested)?, How old were you when you were first arrested-that is, officially charged by the police (an adult or juvenile arrest, other than a traffic violation)?, and How old were you when you were first convicted of a criminal offense (an adult or juvenile conviction, other than a traffic violation)? Offending frequency questions are the following: Altogether in your life, how many times have you been arrested (don't count traffic violations) and how many times have you been convicted of a felony?

Table 3 shows that age of 1st involvement ranges from 4 to 63, the mean age is 14.76, and the standard deviation is 6.63. Age of 1st arrest ranges from 5 to 63, the mean age is 17.3, and the standard deviation is 6.71. Age of 1st conviction ranges from 7 to 63, the mean age is 18.65, and the standard deviation is 7.14. This offender sample accumulated many arrests. Over 38% (249) of these offenders report being arrested over eleven or more times, 14.9% (97) report seven to ten times, 21.2% (138) report four to six times, 18.1% (118) report two to three times, and only 7.5% (49) report being arrested only one time. Furthermore, 2.6% (17) report having eleven or more felony convictions, 4.5% (29) report seven to ten times, 13.1% (85) report four to six times, 38.2% (249) report two to three times, 39.3% report one time, and 2.3% (15) report never being convicted of a felony.

Table 3. Descriptive statistics for the offender sample (n = 651)

Variable	%	Mean	SD	Min	Max
Race/ethnicity					
Black	18.7%				
Hispanic/Mexican	13.1%				
White	58.1%				
Native American	6.5%				
Other	3.7%				
Education					
Less than HS	39.1%				
HS graduate or GED	39.8%				
Some college	17.2%				
College graduate>	3.8%				
Times married		.57	.76	0	5
0	56.5%				
1	32.6%				
2	8.9%				
3	1.5%				
4	.3%				
5	.2%				
Age of 1st involvement		14.76	6.63	4	63
Age of 1st arrest		17.30	6.71	5	63
Age of 1st conviction		18.85	7.14	7	63
# arrests					
1 time	7.5%				
2-3 times	18.1%				
4-6 times	21.2%				
7-10 times	14.9%				
11 or more times	38.2%				
# felony convictions					
Never	2.3%				
1 time	39.3%				
2-3 times	38.2%				
4-6 times	13.1%				
7-10 times	4.5%				
11 or more times	2.6%				

Appendices A through C show univariate and bivariate statistics for Grasmick et al.'s 24 self-control items. Appendix A shows frequency distributions for each item, Appendix B shows means and standard deviations, and Appendix C shows a Pearson'sproduct-moment correlation matrix of the 24 items. First, the majority of correlations between items are positive and statistically significant ($p < .05$). Unlike most items, I2 and P4 did not consistently show statistically significant correlations with other scale items. Second, scale items in the same domain, e.g., Temper, are correlated stronger with each other than with items from other domains.

Results from Reliability Analyses

Table 4 shows Cronbach's reliability coefficients for Grasmick et al.'s 24 item self-control scale. Reliability coefficients were estimated for the full, White, and Black offender samples. The 24 item scale has high reliability or internal consistency. Specifically, Cronbach's alpha was .87 for the full sample of offenders, .85 for Black offenders, and .89 for White offenders. The coefficients indicate that this scale is internally consistent and would be expected to correlate highly with an alternative form of the Grasmick et al. measure for the entire sample and across racial groups.

Table 4 also shows reliability coefficients for the subscales reflecting each of the six elements of self-control. Overall, each subscale has adequate reliability when determined for the full sample and across racial groups; however, some important observations should be noted. First, the Impulsivity subscale has the lowest reliability of all six subscales: .60 for the full sample, .62 for the White sample, and .50 for the Black sample. Second, the Physical Activities subscale has the second lowest

Table 4. Cronbach's reliability analysis of the Grasmick et al.'s self-control scale and its six dimensions.

Scales	(n = 651) Full sample	(n = 122) Black Sample	(n = 378) White Sample
Self-Control (24 items)	.87	.85	.89
Impulsivity	.60	.50	.62
Simple Tasks	.79	.79	.80
Risk Seeking	.78	.72	.80
Physical Activities	.67	.60	.72
Self-Centered	.75	.71	.77
Temper	.81	.72	.85

reliability: .67 for the full sample, .72 for the White sample, and .60 for the Black sample. Although possibly due to a lower sample size, the reliability estimates for the 24 item scale and each subscale were consistently smaller for the Black offender sample. In sum, the 24item scale has good internal consistency and most of the subscales do also, with the exception of the Impulsivity and Physical Activity subscales.

Results from Independent Samples T-tests

As stated in earlier chapters, Gottfredson and Hirschi (1990) propose that minority groups will have lower levels of self-control. Based on their assertion, it is argued from a construct validation approach that a valid measure of self-control should capture these differences. Table 4 reports analyses that serve as a preliminary investigation of the cross-structure validity of the Grasmick et al. scale.

Table 5 shows results from a series of independent sample t-tests that were used to investigate differences on Grasmick et al.'s self-control measure, and its subscales, for the Black and White offender samples. From a construct validity perspective, it is expected that Blacks, on average, will have lower self-control than Whites if Grasmick et al.'s measure is valid. Results do not support this expectation, which, in turn, could be interpreted as an initial strike against the scale's validity. In fact, some of the results indicate the exact opposite. Specifically, White offenders (mean = 30.77, SD = 11.61), on average, score significantly higher on Grasmick et al.'s scale than Black offenders (mean = 27.74, SD = 10.72), indicating that White offenders have lower self-control than Black offenders. Differences also emerge across racial groups for two of the subscales, Impulsivity and Risk Seeking. White

Table 5. Independent samples t-tests assessing racial group differences on Grasmick et al.'s self- control scale and its six dimensions.

Variables	(n = 122) Black Sample		(n = 378) White Sample		t-statistic
	Mean	SD	Mean	SD	
Self-Control (24 items)	27.74	10.72	30.77	11.61	-2.55* (.01)
Impulsivity	4.61	2.35	5.58	2.51	-3.75* (.00)
Simple Tasks	4.85	3.12	3.97	2.84	1.37 (.17)
Risk Seeking	3.98	2.77	5.65	3.10	-5.62* (.00)
Physical Activities	7.45	2.38	7.91	2.47	-1.80 (.07)
Self-Centered	3.17	2.56	3.31	2.58	-.52 (.60)
Temper	4.13	2.88	4.34	3.36	-.68 (.53)

$p < .05$ (two-tailed); Note: Due to significant F-statistics ($p < .05$) on Levene's test for equality of variances, t-statistics assessing racial differences on impulsivity and temper were calculated under the assumption of unequal variance. Probabilities are below t-values.

offenders (mean = 5.58, SD = 2.51), on average, score significantly higher on Grasmick et al.'s Impulsivity subscale than Black offenders (mean = 4.61, SD = 2.35), indicating that White offenders are more impulsive than Black offenders. In addition, White offenders (mean = 5.65, SD = 3.10), on average, score significantly higher on Grasmick et al.'s Risk Seeking subscale than Black offenders . (mean = 3.98, SD = 2.77), indicating that white offenders are higher in risk seeking than black offenders. Finally, statistically significant differences were not found across Black and White offender samples for the following subscales: Simple Tasks, Physical Activities, Self-centeredness, and Temper. A more detailed interpretation of these results will be discussed in Chapter 6. Next, the results from the internal structure analyses of Grasmick et al.'s scale are discussed.

Results from Principal Components Analyses

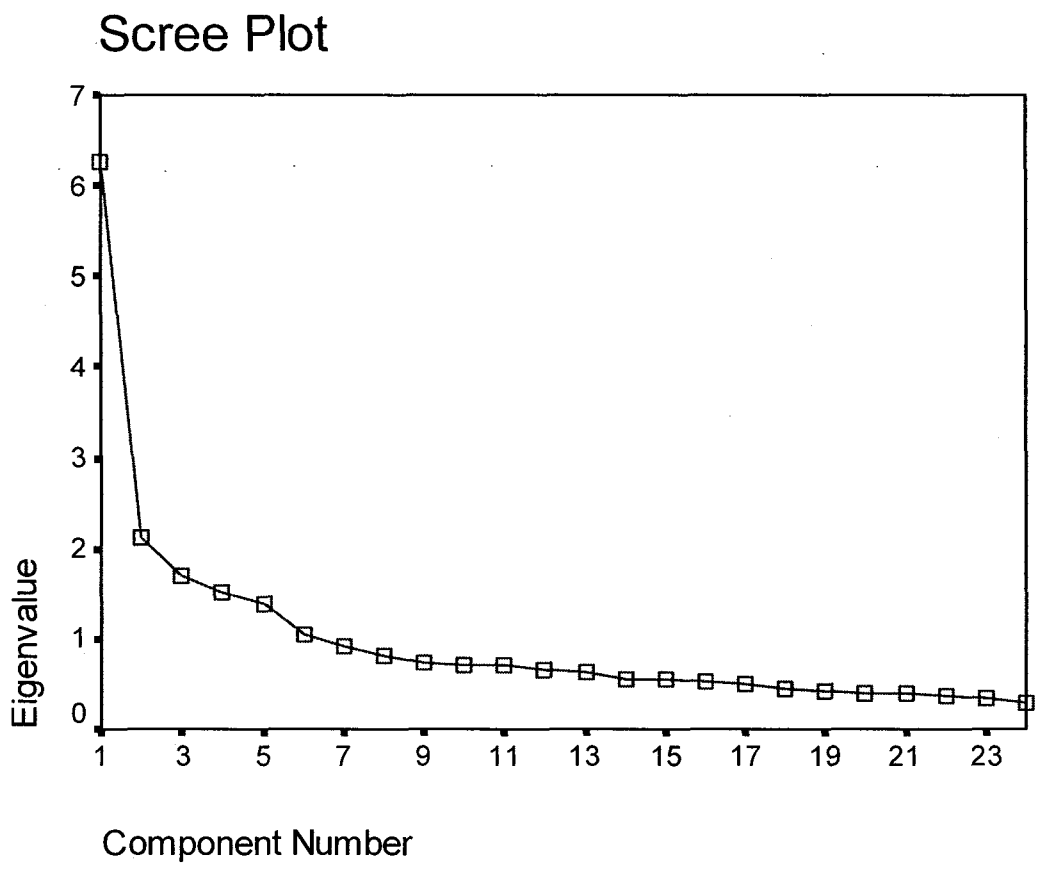
Following Gramsick et al. (1993), an internal structure analysis of their scale begins by performing a series of Principal Components Analyses (PCA). A PCA is conducted for the full sample of offenders, Black sample, and White sample. These analyses attempt to confirm whether Grasmick et al.'s. (1993) findings are replicated with a sample of male offenders and across racial groups.

Table 6 reports results from a PCA for the full sample of offenders (n = 651). As seen, 22 of the 24 items load high on the first component. Just like regression coefficients and correlations, there is debate on what constitutes a high loading in factor analysis. Following several researchers (see Gardner, 2001: 257), this dissertation will use .30 as a cut-off. This cut-off is recommended with sample sizes larger than one-hundred (Gardner, 2001: 257). Loadings are between .36 and .66,

Table 6. Principal Components Analysis of Grasmick et al.'s 24 self-control items: Results for the full sample (n = 651).

Variable	Loadings on component 1
Impulsivity	
I1	.54
I2	.29
I3	.52
I4	.53
Simple Tasks	
S1	.42
S2	.47
S3	.57
S4	.46
Risk Seeking	
R1	.42
R2	.55
R3	.62
R4	.59
Physical Activities	
P1	.40
P2	.36
P3	.42
P4	.26
Self-Centered	
Sc1	.55
Sc2	.52
Sc3	.51
Sc4	.63
Temper	
T1	.54
T2	.66
T3	.60
T4	.58

Figure 4. Scree plot for the principal components analysis of Grasmick et al.'s 24 self-control items: Results for the full offender sample (n = 651)



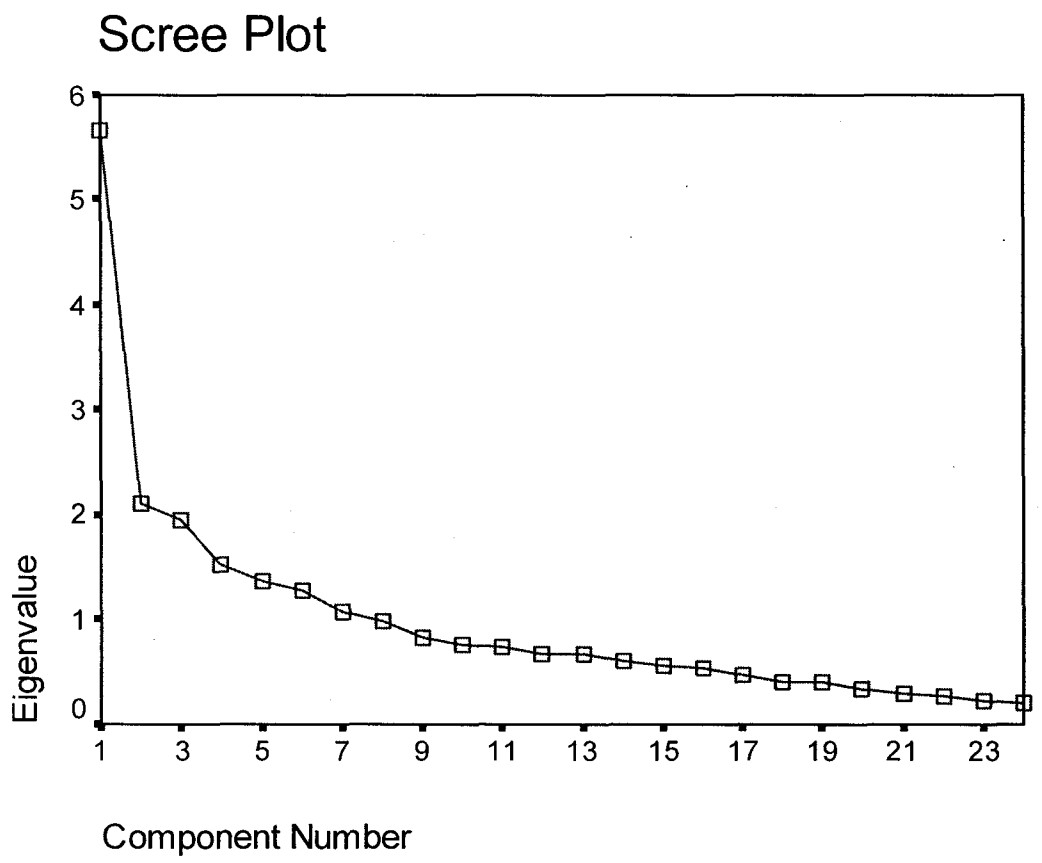
with two exceptions (I2), which has a loading of .29, and P4, which has a loading of .26. As shown in Figure 4, the scree plot from the PCA for the full sample reveals a conflicting pattern of results. Six of the extracted components have eigenvalues greater than 1. Specifically, component 1 explains 26.07 percent of variance between items, component 2 explains 8.89 percent of variance, component 3 explains 7.07 percent of variance, component 4 explains 6.36 percent of variance, component 5 explains 5.77 percent of variance, and component 6 explains 4.32 percent of variance. According to the Kaiser rule (Nunnally, 1967), a six factor solution is appropriate. In contrast, the largest and most obvious break between eigenvalues is the difference between the first (eigenvalue = 6.25) and second (eigenvalue = 2.14) values, indicating that only one component above the "elbow" is extracted. According to Cattell (1966), the scree plot results imply that only one meaningful factor exists.

Table 7 reports results from a PCA for the Black offender sample (n = 122). Similar to the full sample analysis, 23 of the 24 items load strongly on the first component. Loadings range between .39 and .66, with one exception (I2), which has a loading of .15. As shown in Figure 5, the scree plot from the PCA for the Black sample, once again, reveals a conflicting pattern of results, with only minor differences from the full sample. Seven of the extracted components have eigenvalues greater than 1. Specifically, component 1 explains 23.56 percent of variance between the items, component 2 explains 8.82 percent of variance, component 3 explains 8.19 percent of variance, component 4 explains 6.32 percent of variance, component 5 explains 5.73 percent of variance, component 6 explains 5.32 percent of variance, and component 7 explains 4.52 percent of variance.

Table 7. Principal Components Analysis of Grasmick et al.'s 24 self-control items: Results for the Black sample (n = 122).

Variable	Loadings on component 1
Impulsivity	
I1	.48
I2	.15
I3	.53
I4	.60
Simple Tasks	
S1	.56
S2	.56
S3	.61
S4	.59
Risk Seeking	
R1	.40
R2	.45
R3	.47
R4	.52
Physical Activities	
P1	.44
P2	.39
P3	.41
P4	.35
Self-Centered	
Sc1	.44
Sc2	.40
Sc3	.41
Sc4	.47
Temper	
T1	.44
T2	.65
T3	.59
T4	.44

Figure 5. Scree plot for the principal components analysis of Grasmick et al.'s 24 self-control items: Results for the Black sample (n = 122)



Interestingly, a seventh component was extracted from the Black sample which consisted of items I1 (.44) and I2 (.38); however, the same items load higher on the first component extracted. According to the Kaiser rule (Nunnally, 1967), a seven factor solution could be appropriate for the Black sample.

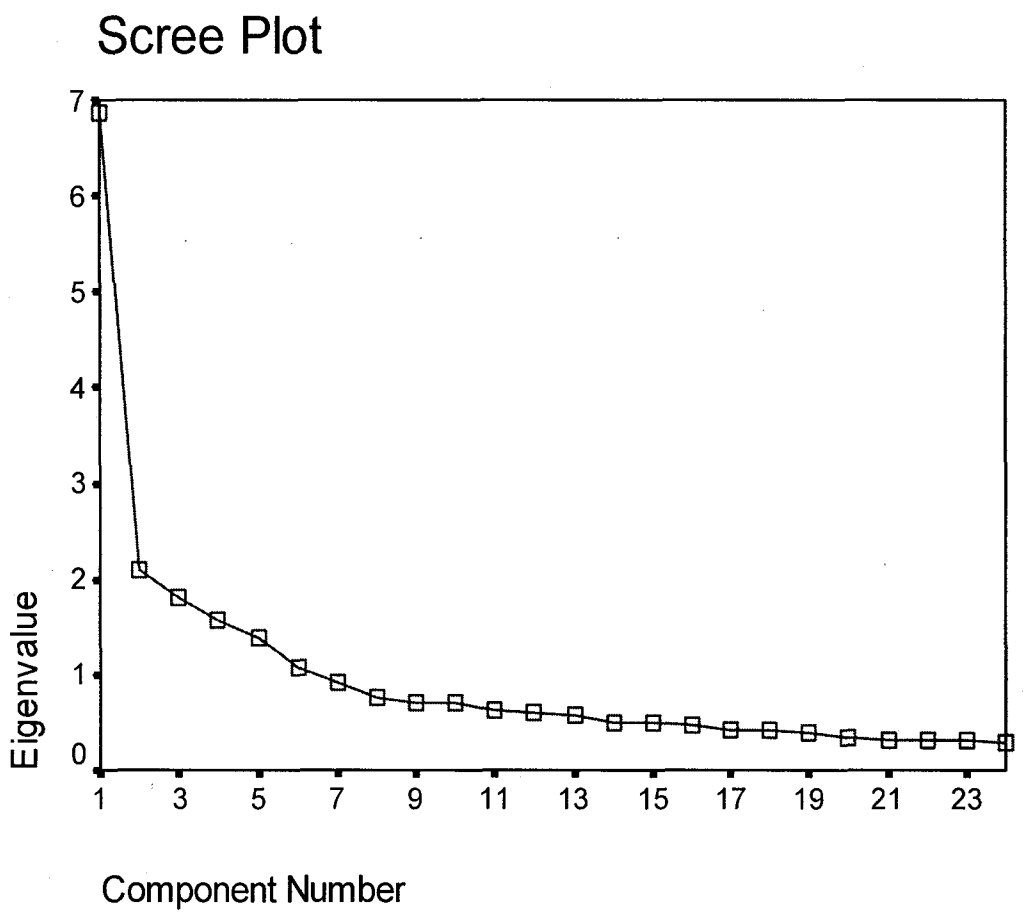
Figure 5 also shows that the largest and most obvious break between eigenvalues is the difference between the first (eigenvalue = 5.65) and second (eigenvalue = 2.11) values, indicating that only one component above the "elbow" is extracted. According to Cattell (1966), the scree plot results for the Black sample imply that only one meaningful factor exists.

Table 8 reports results from a PCA for the White sample of offenders (n = 378). As seen, 22 of the 24 items loaded strongly on the first component. The loadings range between .35 and .69, with two exceptions (I2) and (P4), which have loadings of .28 and .26, respectively. Figure 6 shows a scree plot for the White offender sample and reveals, once again, a conflicting pattern of results. Six of the extracted components have eigenvalues greater than 1. Specifically, component 1 explains 28.67 percent of variance between the items, component 2 explains 8.78 percent of variance, component 3 explains 7.56 percent of variance, component 4 explains 6.57 percent of variance, component 5 explains 5.82 percent of variance, and component 6 explains 4.49 percent of variance. According to the Kaiser rule (Nunnally, 1967), a six factor solution is appropriate. In contrast, the largest break between eigenvalues is the difference between the first (eigenvalue = 6.88) and second (eigenvalue = 2.10) values, indicating that only one component above the

Table 8. Principal components analysis of Grasmick et al's 24 self-control items: Results for the White sample (n = 378).

Variable	Loadings on component 1
Impulsivity	
I1	.61
I2	.27
I3	.53
I4	.49
Simple Tasks	
S1	.46
S2	.52
S3	.61
S4	.50
Risk Seeking	
R1	.43
R2	.55
R3	.64
R4	.62
Physical Activities	
P1	.41
P2	.35
P3	.44
P4	.26
Self-Centered	
Sc1	.59
Sc2	.53
Sc3	.53
Sc4	.65
Temper	
T1	.60
T2	.69
T3	.63
T4	.63

Figure 6. Scree plot of the principal components analysis of Grasmick et al.'s 24 self-control items: Results for the White sample (n = 378)



"elbow" is extracted. According to Cattell (1966), the scree plot results for the White sample imply that only one meaningful factor exists.

Results from Principal Axis Factor Analyses

Because some researchers argue that PCA is not a valid form of exploratory factor analysis, results from Principal Axis Factoring (PAF) analyses are reported for comparative purposes. Results are reported for the full, White, and Black samples. Results were similar to those from the PCA analyses.

Table 9 reports results from a PAF analysis for the full sample of offenders (n = 651). As seen, all items load on the first factor. The loadings range between .24 and .64. Two of the twenty-four items have small loadings on the first factor, these were I2 (.26) and P4 (.24). Figure 7 shows a scree plot for the full offender sample and reveals that six factors are extracted that have eigenvalues greater than 1. Specifically, factor 1 explains 26.07 percent of variance between items, factor 2 explains 8.89 percent of variance, factor 3 explains 7.07 percent of variance, factor 4 explains 6.36 percent of variance, factor 5 explains 5.77 percent of variance, and factor 6 explains 4.32 percent of variance. According to the Kaiser rule (Nunnally, 1967), a six factor solution is appropriate. In contrast, the largest and most obvious break between eigenvalues is the difference between the first (eigenvalue = 6.25) and second (eigenvalue = 2.14) values, indicating that only one factor above the "elbow" is extracted. According to Cattell (1966), the scree plot results imply that only one meaningful factor exists.

Table 10 reports results from a PAF analysis for the Black sample of offenders (n = 122). As seen, all items load on the first factor. Factor loadings were

Table 9. Principal Axis Factor analysis of Grasmick et al.'s 24 self-control items: Results for the full sample (n = 651).

Variable	Loadings on factor 1
Impulsivity	
I1	.51
I2	.26
I3	.49
I4	.50
Simple Tasks	
S1	.41
S2	.46
S3	.53
S4	.45
Risk Seeking	
R1	.39
R2	.56
R3	.60
R4	.56
Physical Activities	
P1	.37
P2	.35
P3	.40
P4	.24
Self-Centered	
Sc1	.52
Sc2	.48
Sc3	.49
Sc4	.61
Temper	
T1	.52
T2	.64
T3	.59
T4	.56

Figure 7. Scree plot from the principal axis factor analysis of Grasmick et al.'s 24 self-control items: Results for the full sample (n = 651)

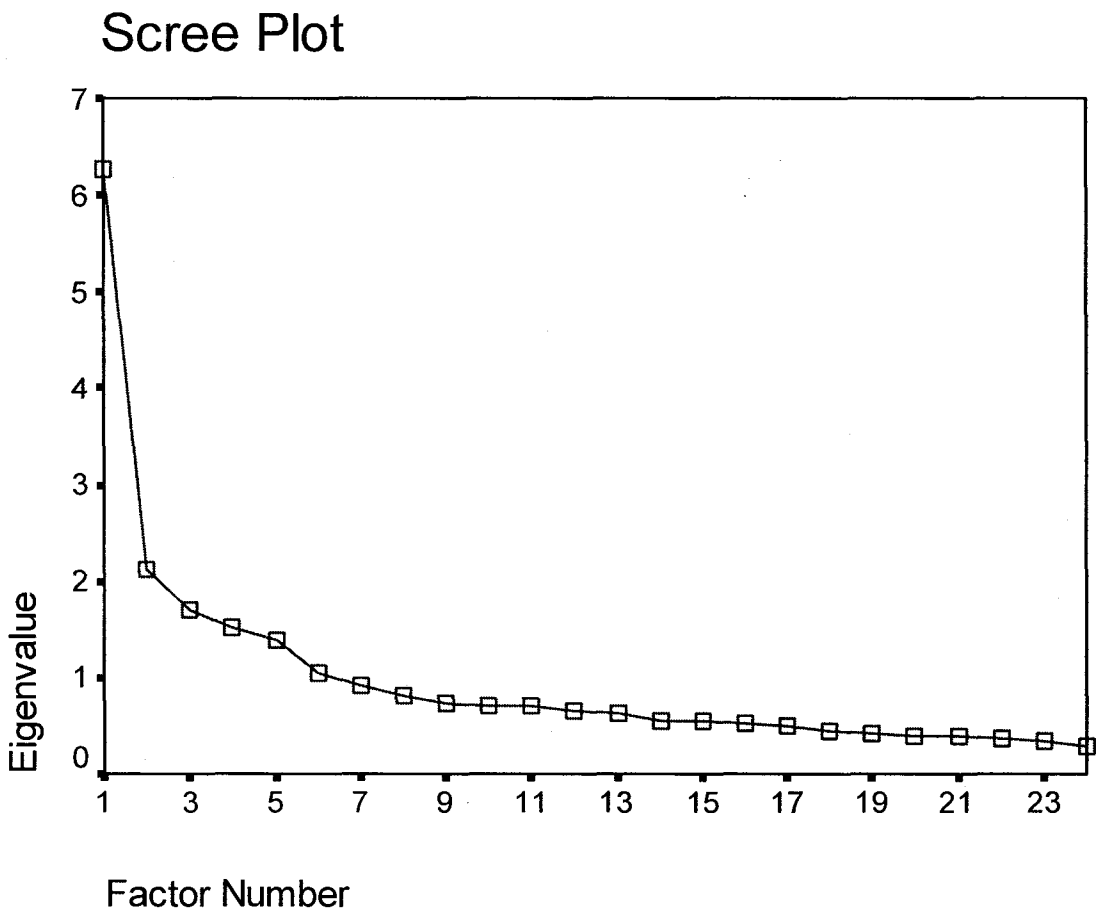
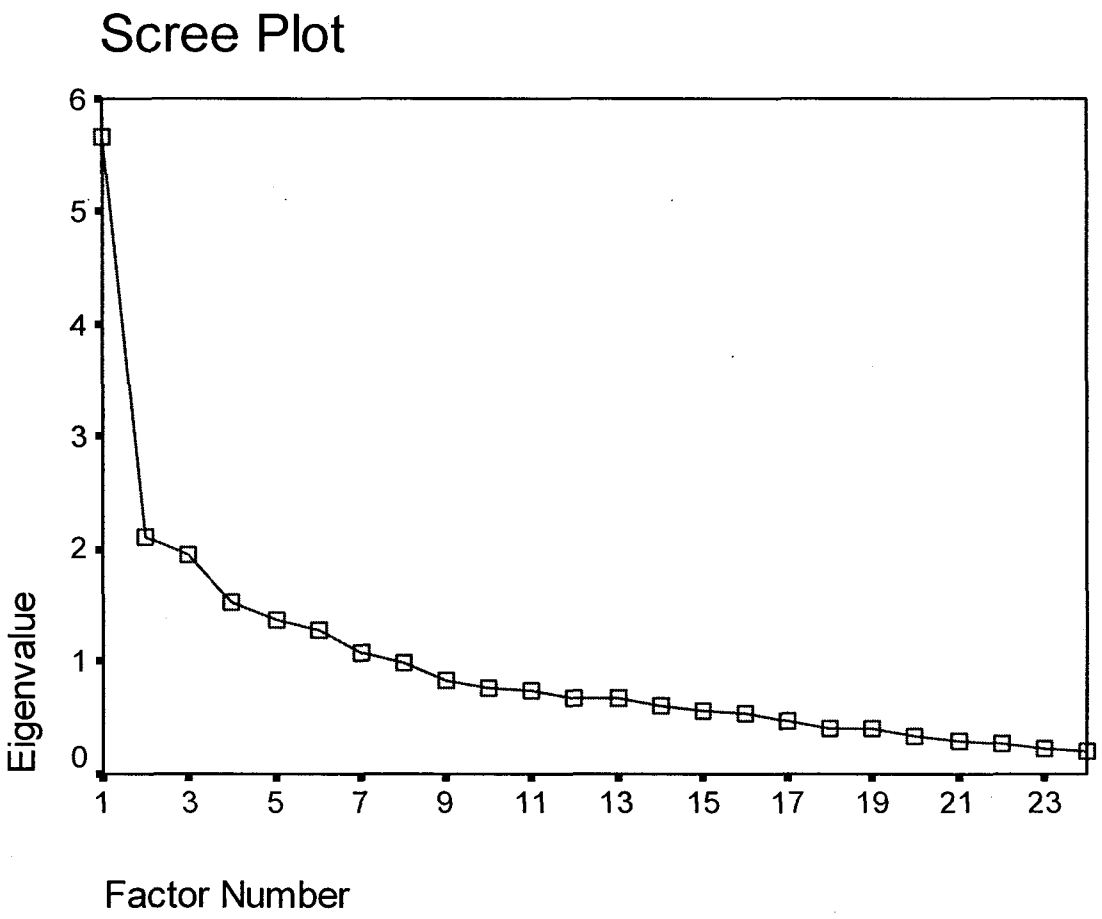


Table 10. Principal Axis Factor Analysis of Grasmick et al.'s 24 self-control items: Results for the Black sample (n = 122).

Variable	Loadings on factor 1
Impulsivity	
I1	.47
I2	.13
I3	.49
I4	.57
Simple Tasks	
S1	.55
S2	.54
S3	.59
S4	.57
Risk Seeking	
R1	.37
R2	.47
R3	.45
R4	.48
Physical Activities	
P1	.41
P2	.36
P3	.38
P4	.32
Self-Centered	
Sc1	.42
Sc2	.38
Sc3	.39
Sc4	.46
Temper	
T1	.42
T2	.63
T3	.57
T4	.42

Figure 8. Scree plot from the principal axis factor analysis of Grasmick et al.'s 24 self-control items: Results for the Black sample (n = 122)



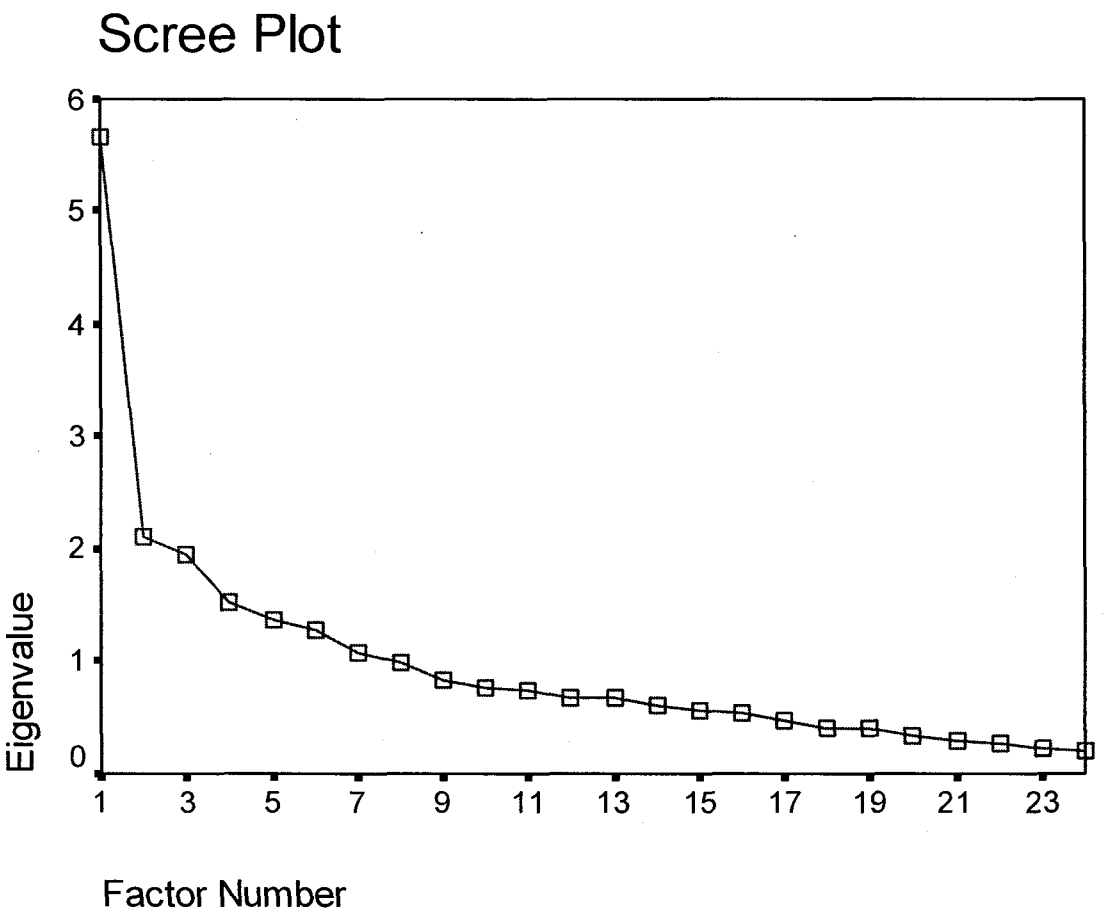
high ranging between .32 and .64, with the exception of I2, which has a loading of .13. As shown in Figure 8, the scree plot for the Black offender sample reveals that seven factors are extracted having eigenvalues greater than 1. Specifically, factor 1 explains 23.56 percent of variance between items, factor 2 explains 8.82 percent of variance, factor 3 explains 8.19 percent of variance, factor 4 explains 6.32 percent of variance, factor 5 explains 5.73 percent of variance, factor 6 explains 5.32 percent of variance, and factor 7 explains 4.52 percent of variance. According to the Kaiser rule (Nunnally, 1967), a seven factor solution is appropriate for the Black sample. In contrast, the largest and most obvious break between eigenvalues was the difference between the first (eigenvalue = 5.65) and second (eigenvalue = 2.12) values, indicating that only one factor above the "elbow" is extracted. According to Cattell (1966), the scree plot results imply that only one meaningful factor exists.

Table 11 reports results from a PAF analysis for the White sample of offenders (n = 378). As seen, all items load on the first factor. The loadings are between .24 and .67. Two of the twenty-four items have smaller loadings on the first factor, these were I2 (.26) and P4 (.24). As shown in Figure 9, the scree plot for the White sample reveals that six factors were extracted that have eigenvalues greater than 1. Specifically, factor 1 explains 28.67 percent of variance between items, factor 2 explains 8.79 percent of variance, factor 3 explains 7.56 percent of variance, factor 4 explains 6.58 percent of variance, factor 5 explains 5.82 percent of variance, and factor 6 explains 4.49 percent of variance. According to the Kaiser rule (Nunnally, 1967), a six factor solution is appropriate. In contrast, the largest and most obvious break between eigenvalues is the difference between the first (eigenvalue = 6.88) and

Table 11. Principal Axis Factor Analysis of Grasmick et al.'s 24 self-control items: Results for the White sample (n = 378).

Variable	Loadings on factor 1
Impulsivity	
I1	.57
I2	.25
I3	.51
I4	.47
Simple Tasks	
S1	.45
S2	.51
S3	.58
S4	.48
Risk Seeking	
R1	.40
R2	.55
R3	.63
R4	.60
Physical Activities	
P1	.38
P2	.33
P3	.42
P4	.24
Self-Centered	
Sc1	.56
Sc2	.50
Sc3	.51
Sc4	.64
Temper	
T1	.59
T2	.67
T3	.62
T4	.61

Figure 9. Scree plot from the principal axis factor analysis of Grasmick et al.'s 24 self-control items: Results for the White sample (n = 378)



second (eigenvalue = 2.10) values, indicating that only one factor above the "elbow" is extracted. According to Cattell (1966), the scree plot results suggest that only one meaningful factor exists.

So far, results from the PCA and PAF analyses reveal similar findings. That is, based on standard criteria for component and factor extraction, two possible factor structures exist for the Grasmick et al. scale. First, a one factor solution is possible for the full, Black, and White Samples. Such a solution would be consistent with what Grasmick and his colleagues initially discovered. Second, however, is that the eigenvalue rule shows the possibility of six factors or dimensions. For the Black offender sample, however, both sets of exploratory analyses reveal that a one factor or seven factor solution is possible. Due to the exploratory nature of these analyses, it is still unknown whether a unidimensional or multidimensional structure is better for explaining the 24 items of Grasmick et al.'s self-control scale for a sample of incarcerated male offenders. Next, a series of Confirmatory Factor Analyses is conducted to test the fit of theoretically driven unidimensional and multidimensional models for the Grasmick et al. self-control items.

Results from Confirmatory Factor Analyses

In this section results are reported from three different CFA models which are estimated using the AMOS 4.0 structural equation modeling program. Unlike the exploratory results reported above, each confirmatory model explicitly tests a unique theoretical conceptualization of the self-control construct. As seen in Figure 10, a one factor model will be estimated that allows all 24 items to load on only one factor. Next, Figure 11 shows the six factor model that will be estimated, where each item

Figure 10. A one factor model for Grasmick et al.'s self-control items

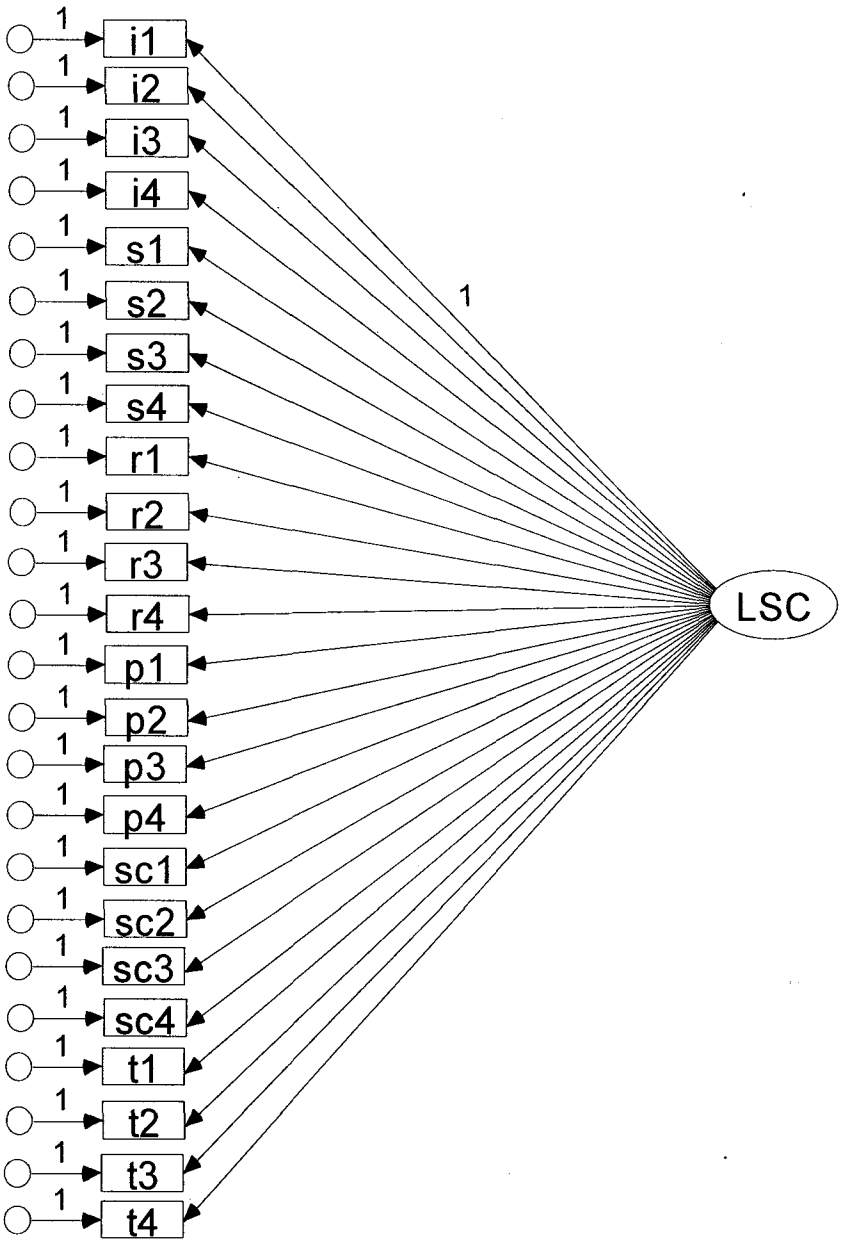


Figure 11. A six factor model for Grasmick et al.'s self-control items

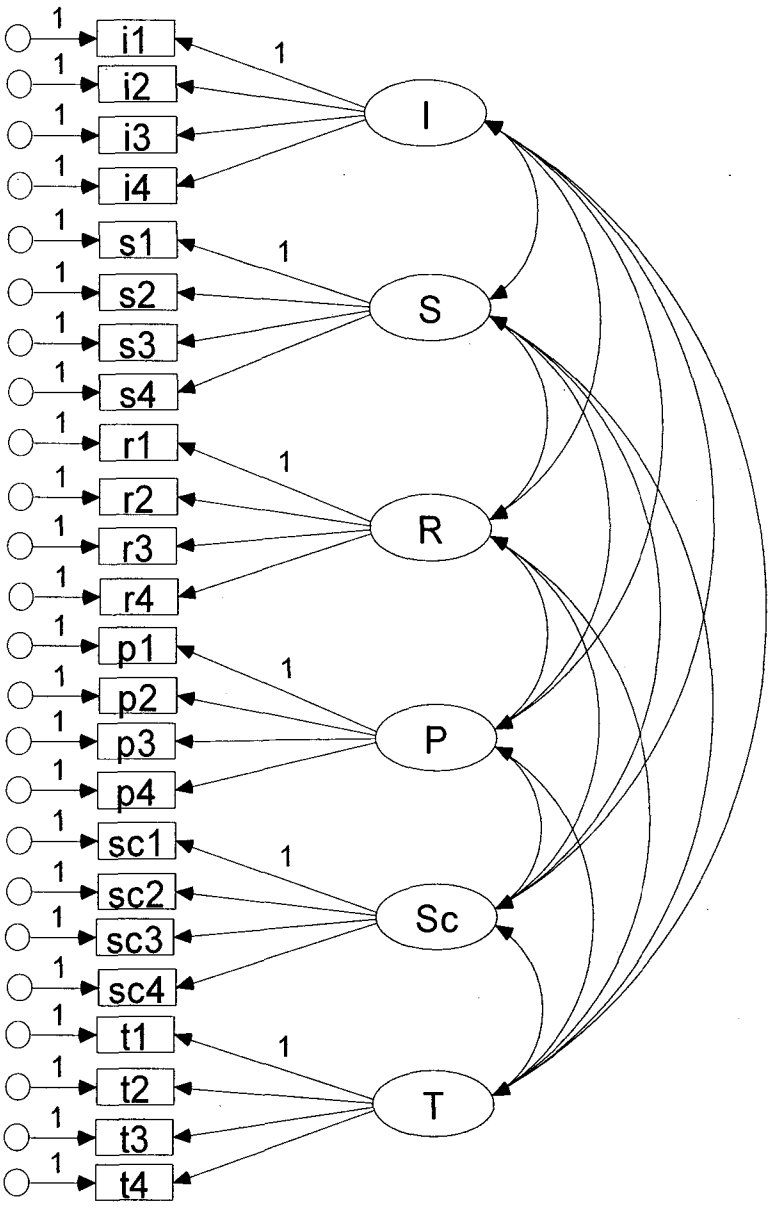
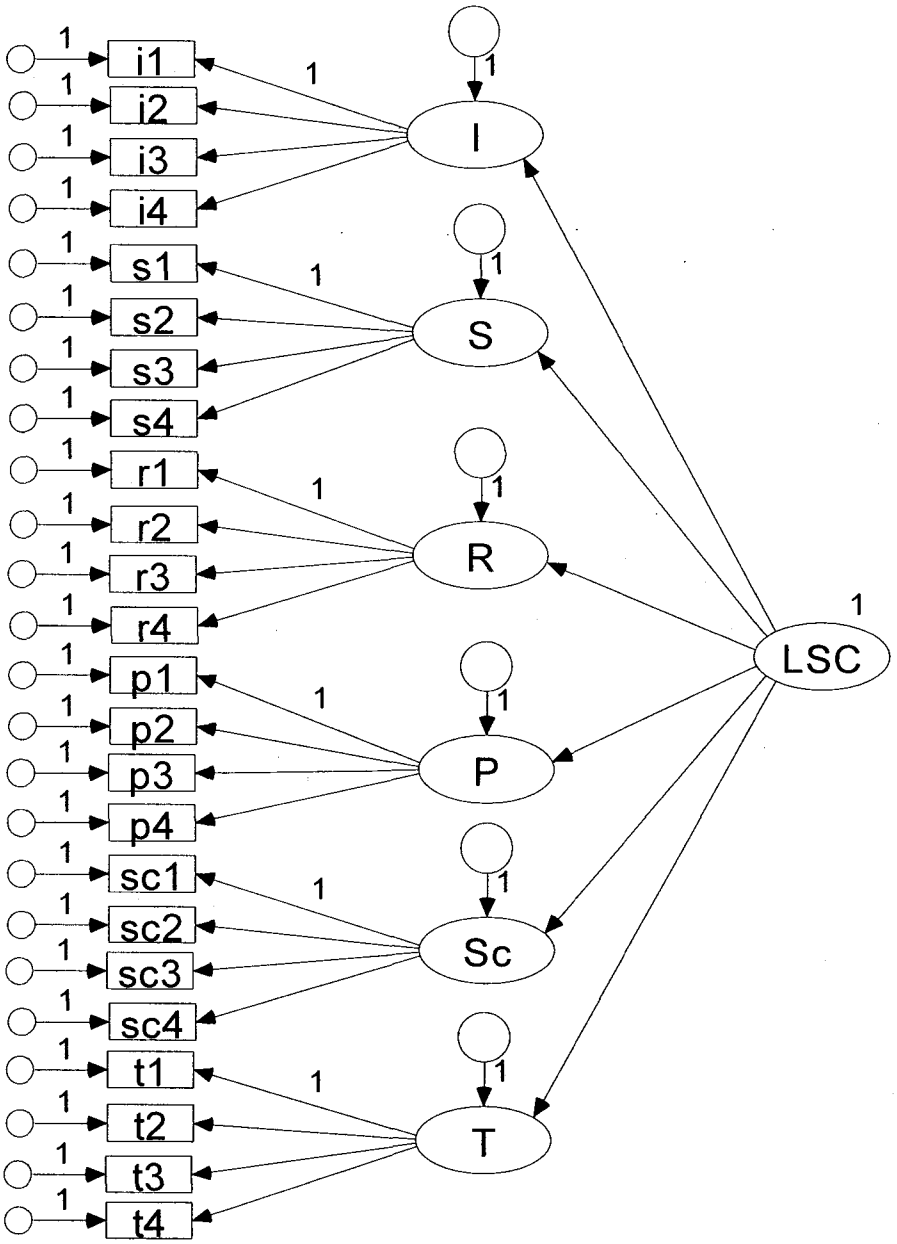


Figure 12. A second-order model for Grasmick et al.'s self-control items



loads on its respective factor and each factor is allowed to correlate with the other factors. Finally, Figure 12 shows the second-order model that will be estimated, where six separate factors are allowed to load onto a second-order factor of self-control. Results from all three models are discussed next, focusing particularly on factor loadings and the overall fit of each model.

Table 12 shows results for the first model testing the hypothesis that Grasmick et al.'s scale items form a unidimensional construct in that all items load on one, and only one, factor. Initial support for the unidimensionality of Grasmick et al.'s scale would be gained if all items have statistically significant, positive loadings that are strong, and the model fits the data well. As can be seen, all factor loadings are positive and statistically significant ($p < .05$), which gives support to the notion that all items load on one factor. Although each item had a statistically significant loading, some items had larger loadings others. Most items have loadings between .32 and .65. Once again, I2 and P4 have smaller loadings, .27 and .22 respectively, compared to other items. The self-control factor explains a small to moderate amount of variance for each item, ranging from 10% (P2) to 42 % (T2). The self-control factor is unsuccessful in explaining more than 42% of the variation for any item. Although these results are somewhat encouraging, they do not indicate that the one factor model fits the data well.

The next step is to assess the overall fit of the one factor confirmatory model. Several indices are used to assess the fit of a confirmatory model in SEM. The most common fit indices include the χ^2 statistic, ratio of χ^2/df, goodness of fit index (GFI), adjusted goodness of fit index (AGFI), comparative fit index (CFI), normative fit

Table 12. Confirmatory Factor Analysis- One Factor Model

Variable	Factor 1 loading (se)
Impulsivity	
I1	--- ---
I2	.28* (.07)
I3	.49* (.09)
I4	.49* (.10)
Simple Tasks	
S1	.37* (.09)
S2	.42* (.09)
S3	.52* (.10)
S4	.41* (.09)
Risk Seeking	
R1	.40* (.09)
R2	.51* (.11)
R3	.60* (.11)
R4	.56* (.10)
Physical Activities	
P1	.36* (.09)
P2	.32* (.07)
P3	.37* (.08)
P4	.22* (.07)
Self-centered	
Sc1	.52* (.10)
Sc2	.48* (.09)
Sc3	.48* (.08)
Sc4	.61* (.09)
Temper	
T1	.52* (.11)
T2	.65* (.11)
T3	.59* (.12)
T4	.56* (.11)

$\chi^2/df = 2163.93/252$.

Note: *p < .05; Factor loadings are reported as standardized regression coefficients

index (NFI), and the root mean square error of approximation (RMSEA). A discussion of each index is presented before interpreting the overall fit of the one factor model.

Also known as a test of exact fit, the χ^2 statistic reflects the discrepancy between the unrestricted variance-covariance matrix and the restricted matrix. This statistic is used to test a null hypothesis that states the restricted and unrestricted models variances and covariances are equal. The probability associated with the χ^2 statistic indicates the probability of obtaining a χ^2 value that exceeds the χ^2 value when the null hypothesis is true (Byrne, 2001). A higher probability associated with χ^2 would represent a closer fit between the hypothesized measurement model and a perfect model. In contrast, a χ^2 statistic that has a probability level of .05 or less indicates that the hypothesized model does not achieve an adequate fit or is not equal to the unrestricted sample covariance matrix. Unlike standard regression models, a good model fit is indicated by a non-significant χ^2 value where the χ^2 statistic would approximate degrees of freedom (*df*) and not the opposite where χ^2 is large relative to *df* (Bollen, 1989).

Two important limitations have been identified for the χ^2 statistic as a fit index. First, results of this test are highly sensitive to sample size. Large samples almost always lead to rejection of the null hypothesis, even when differences between observed and model implied covariances are actually small. In contrast, badly specified models might be accepted when small samples are used. Second, the χ^2 statistic does not have an upper bound which, in turn, prevents a standardized interpretation of the estimated value (Hayduk, 1987: 167; Kline, 1998). The $\chi 2/df$

ratio has been recommended as an alternative to the χ^2 statistic because dividing it by its degrees of freedom reduces sensitivity to sample size. A value less than 3 has been recommended for a good model fit (Kline, 1998). Due to the above limitations, researchers have argued for the use of multiple fit statistics that provide more useful information about the degree of model fit, are less sensitive to sample size, and can take into account model complexity (Byrne, 2001; Kline, 1998).

The GFI is one alternative for assessing the fit of a SEM model. The GFI is analogous to a squared multiple correlation and its value represents the proportion of covariance explained by model implied covariance or the relative amount of variance and covariance in S that is explained by Σ. The range of the GFI is from 0 to 1 and a value of .90 or higher has been used to indicate a well fitting model (Hayduk, 1987; Kline, 1998). Similarly, the AGFI is a commonly used goodness of fit index that takes into account model complexity and the degrees of freedom of a specified model. The AGFI corrects for the fact that complex models often fit the same data better than simpler ones. Although interpretation of the AGFI value is consistent with the GFI, it corrects downward the value of the GFI when the number of parameters in a model are increased (Kline, 1998). An AGFI value of .90 or above indicates satisfactory or adequate model fit.

Incremental fit indexes are also commonly used indicators of model fit. The most common of these fit measures are the NFI and CFI (Bentler, 1990; Bentler and Bonnett, 1980). These fit indexes compare a hypothesized model to a null model, i.e., a model that assumes observed variables are uncorrelated, so that the proportion of improvement can be detected relative to a null model. With a range from 0 to 1, a

value of .90 or larger represents a well fitting model for both the NFI and CFI. Although they are interpreted similarly, the NFI underestimates fit in smaller samples and the CFI is less affected by sample size (Bentler, 1990).

Finally, RMSEA has been recognized as one of the most informative criteria for assessing the fit of a covariance structure model. RMSEA asks the question, "How well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?" (Browne and Cudeck, 1993: 137-138). RMSEA accounts for the error of approximation in the population (Byrne, 2001). Values less than .05 indicate good fit and those greater than .10 indicate poor fit (Browne and Cudeck, 1993).

Model fit is a highly debated issue in the SEM literature. Researchers do not agree on only one fit index that determines the adequacy of a model. As Kline (1998) suggested, a favorable value of one fit index can not by itself indicate good fit. It is more ideal to investigate the values of multiple fit indexes. In doing so, the more criteria that the model satisfies, the more satisfactory is its fit. I now return to the overall fit of the one factor model for Grasmick et al.'s items.

Although Grasmick et al.'s self-control items all have positive and significant loadings on one factor, the overall fit of the one factor model is not too encouraging. Table 13 reports fit indexes for the one factor model, indicating that this model poorly fits the data; therefore, not supporting the hypothesis that Grasmick et al.'s items represent a unidimensional construct. The ratio of $\chi 2/df$ is 8.59, a value that is larger than the critical value of 3 recommended by some researchers (Kline, 1998). The values for the GFI and AGFI are .74 and .69, respectively. These values are both

Table 13. Fit statistics for each Confirmatory Factor Analysis

Fit Statistic	One Factor Model	Six Factor Model	Second Order Model
X^2	2163.93*	639.03*	674.79*
df	252	237	246
X^2/ *df*	8.59	2.70	2.74
GFI	.74	.92	.92
AGFI	.69	.90	.90
CFI	.59	.91	.91
NFI	.56	.87	.86
RMSEA	.11	.05	.05

*$p < .05$

lower than the accepted value of .90. The values for the NFI and CFI are.59 and .56, respectively, which again are both substantially lower than the critical value of .90. The value for the RMSEA is equal to .11, well above the criteria of .05 indicating a poor fit.

Table 14 shows results from the second confirmatory model testing the hypothesis that Grasmick et al.'s scale items represent six separate, but correlated, factors: Impulsivity, Simple Tasks, Risk Seeking, Physical Activities, Self-Centered, and Temper. Items are only allowed to load on their corresponding factors. For example, items measuring impulsivity are specified to load on the Impulsivity dimension. As shown in Table 14, all factor loadings are positive and statistically significant ($p < .05$). Most item loadings are strong, ranging from .45 to .82. The smallest loading is item I2 on the Impulsivity factor, which has a standardized loading of .31.

Although not reported, each of the six dimensions has a positive and statistically significant ($p < .05$) correlation with the other dimensions. For example, individuals higher in Iimpulsivity are also more likely to score higher on other dimensions. Correlations among the six separate subscales are between .20 (for simple tasks and risk seeking) and .64 (for impulsivity and risk seeking). More specifically, the correlation between Impulsivity and Physical Activity is .42, Impulsivity and Self-Centeredness is .62, and Impulsivity and Temper is .55. The correlation between Simple Tasks and Physical Activity is .28, Simple Tasks and Self-Centeredness is .44, and Simple Tasks and Temper is .38. The correlation between Risk Seeking and Physical Activity is .35, Risk Seeking and Self-

Table 14. Confirmatory Factor Analysis- Six Factor Model

Variable	Factor 1 loading(se)	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Impulsivity						
I1	--- ---					
I2	.31* (.07)					
I3	.60* (.09)					
I4	.61* (.10)					
Simple Tasks						
S1		--- ---				
S2		.76* (.06)				
S3		.57* (.06)				
S4		.73* (.06)				
Risk Seeking						
R1			--- ---			
R2			.84* (.12)			
R3			.76* (.11)			
R4			.61* (.10)			
Physical activities						
P1				--- ---		
P2				.65* (.10)		
P3				.70* (.11)		
P4				.49* (.09)		
Self-centered						
Sc1					--- ---	
Sc2					.58* (.07)	
Sc3					.67* (.07)	
Sc4					.78* (.08)	
Temper						
T1						--- ---
T2						.78* (.06)
T3						.75* (.07)
T4						.72* (.07)

$\overline{\chi^2}/df = 639.03/237.$

Note: *p < .05; Factor loadings are reported as standardized regression coefficients

Centeredness is .49, and Risk Seeking and Temper is .48. The correlations between Physical Activity and Self-Centeredness is .36 and Physical Activity and Temper is .37. Finally, Self-Senteredness and Temper has a correlation of .56. These correlations indicate that each subscale has a moderate to strong correlation with the other dimensions of the scale and that these dimensions are not distinct from one another.

Table 14 shows that Grasmick et al.'s self-control items all have positive and significant loadings on their respective factors, but the question of model fit, however, remains to be answered. That is, does the six factor model fit the data any better than the one factor model? Refering back to Table 13, most fit indexes indicate that the six factor model fits the data substantially better than the one factor model; supporting the idea that Grasmick et al's scale is measuring six dimensions that are correlated. Although $\chi 2$ was statistically significant ($p < .05$), several fit statistics provide support for the six factor model. The ratio of $\chi 2/df$ is 2.69, a value that is smaller than the critical value of 3 recommended by some researchers (Kline, 1997). The values for the GFI and AGFI are .92 and .90, respectively. These values both exceed the recommended value of .90 or above for good model fit. The values for the NFI and CFI are.87 and .91, respectively, which are both a substantial improvement from the one factor model. The value for the RMSEA is equal to .05, which meets the criteria of .05 or below.

Table 15 shows results from a second-order model, which uses seven latent variables, including the six separate dimensions of self-control, and the overall latent trait of self-control. This model tests the hypothesis that Grasmick et al.'s scale items

Table 15. Confirmatory Factor Analysis- Second Order Model

Variable	Factor 1 loading(se)	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	LSC
Impulsivity							.86* (.04)
I1	--- ---						
I2	.31* (.07)						
I3	.59* (.09)						
I4	.59* (.09)						
Simple Tasks							.52* (.04)
S1		--- ---					
S2		.76* (.06)					
S3		.58* (.06)					
S4		.73* (.06)					
Risk Seeking							.68* (.03)
R1			--- ---				
R2			.82* (.13)				
R3			.77* (.12)				
R4			.62* (.10)				
Physical activities							.51* (.03)
P1				--- ---			
P2				.65* (.10)			
P3				.70* (.11)			
P4				.45* (.09)			
Self-centered							.75* (.04)
Sc1					--- ---		
Sc2					.58* (.07)		
Sc3					.67* (.07)		
Sc4					.75* (.08)		
Temper							.71* (.04)
T1						--- ---	
T2						.76* (.06)	
T3						.75* (.07)	
T4						.72* (.07)	

$\chi^2/df = 674.79/246$.

Note: *p < .05; Factor loadings are reported as standardized regression coefficients

measure six separate factors that can be explained by a second-order factor, i.e., self-control. Consistent with a six factor confirmatory model, results for the second-order model indicate that all items have positive and statistically significant loadings ($p < .05$) on their respective factors. Most item loadings are strong, ranging from .45 to .82. Similar to a six factor model, the weakest loading is item I2 on the Impulsivity factor, which has a standardized loading of .31. Loadings for each of the six dimensions on the self-control latent factor are shown in the second-order component of the model. Each of the six lower order factors has strong, positive, and statistically significant ($p < .05$) loadings on the self-control factor. The Impulsivity factor has a standardized loading of .86, the Simple Task factor has a standardized loading of .52, the Risk Seeking factor has a standardized loading of .68, the Physical Activity factor has a standardized loading of .51, the Self-Centeredness factor has a standardized loading of .75, and the Temper factor has a standardized loading of .71.

Although loadings for all six dimensions on a second-order, self-control factor are statistically significant, the explained variance in six lower order dimensions range between .27 and .75. Specifically, a second-order factor explains 75% of the variance in the Impulsivity dimension, 27% in the Simple Tasks dimension, 46% in the Risk Seeking dimension, 26% in the Physical Activities dimension, 57% in the Self-Centeredness dimension, and 50% in the Temper dimension. This indicates that a fair amount of variance is not explained by the higher order factor of self-control.

While the overall fit of the second-order model is a substantial improvement over the one factor model, the second-order and six factor models fit the data almost identically. Once again, Table 13 shows that most fit indices indicate that the second-

order model fits well also. The ratio of $\chi 2/df$ is 2.74, a value that is smaller than the critical value of 3 recommended by some researchers (Kline, 1998). The values for the GFI and AGFI are .92 and .90, respectively. These values both meet the critical value of .90 for good model fit. The values for the NFI and CFI are .86 and .91, respectively, which are both a substantial improvement from the one factor model. The value for the RMSEA is equal to .05, which meets the criteria of .05 or below.

The fit indices for the six-factor and second-order models suggest a good fit to these offender data. While both the six-factor and second-order models fit these offender data well, it is unknown if the second-order model is a significant improvement over the six-factor model. Following others (Taub, 2001; Vazsonyi and Crosswhite, 2004), additional analyses were conducted to see if the change in parameters in the second-order model would result in an improvement beyond the six-factor model. The change in Chi-square and degrees of freedom was used to evaluate the competing models. Results from this analysis indicate that there is an increase in χ^2 and degrees of freedom (35.76(3), $p < .05$), suggesting that the data fit the second-order model significantly worse than the six- factor model. These results suggest the six-factor model provides the most parsimonious fit to the data. Neverthless, both the six-factor and second-order models indicate that Grasmick et al.'s instrument is measuring a multidimensional construct.

The CFA analyses presented so far reveal several interesting findings. First, the unidmensional, one factor model is not a good fit for the 24 Grasmick et al. self-control items. Second, according to several fit indices, both the six factor and second-order models fit the data well. Third, a Chi-square difference test indicated

that the second-order model fits the data significantly worse than the six-factor model. Overall, it can be concluded from these analyses that the Grasmick et al. scale items are measuring a multi-dimensional construct and not a unidimensional construct.

While both the six factor and second-order measurement models fit the offender data well, it is not known if these two models are invariant across Black and White offender groups. Tests for the invariance of a factorial structure are often referred to as multi-group models (Byrne, 2001). Multi-group models allow researchers to investigate whether the factorial structure of an instrument is equivalent across different groups. For CFA, the most common set of parameters investigated for group invariance are the factor loadings and of less importance are tests of the invariance of error terms (Byrne, 2001). In conducting such an analysis, it is important to first estimate a baseline model with no constraints on the factor loadings or coefficients. Coefficients are estimated freely for both groups at the same time. The fit of this model provides one set of baseline values, i.e., fit statistics, against which constrained models are compared. The constrained model requires factor loadings or coefficients across groups to be equal or invariant. In other words, the factor loadings are forced to be the same and not vary across groups, in this instance Black and White offender groups. The constrained model is of primary importance because it provides the basis of comparison with previously fitted models. In testing for invariance, it is typical to assess the difference between the constrained model's $\chi 2$ value and the $\chi 2$ of the unconstrained model where no equality constraints are imposed. The difference in $\chi 2$ values is distributed as $\chi 2$, with the difference in

degrees of freedom associated with the $\chi 2$ values serving as the degrees of freedom which is also is equal to the number of constraints. A statistically significant $\chi 2$ is evidence to support the notion that equality constraints do not hold across groups, indicating that the factor loadings are not equal or different across groups. In contrast, an insignificant $\chi 2$ is evidence to support the invariance of the factorial structure across racial groups (Byrne, 2001).

Both the six factor and second-order models are subjected to multi-group CFAs across racial groups, i.e., White and Black offender samples. The baseline, multi-group model for the six factor designation is estimated first, where both groups are estimated simultaneously and factor loadings are free to vary. This model reveals an insignificant $\chi 2$ value 797.71 with 495 degrees of freedom. This estimate provides the baseline value against which all subsequent tests for invariance are compared. Furthermore, and following the reporting practices of Byrne (2001), CFI was .92 and RMSEA was .04, indicating that the six factor model is a good fit for the Black and White offender samples. Having established the baseline model, it is possible to proceed to testing the factorial invariance across samples. In testing the invariance hypothesis, constraints are placed on all factor loadings and covariances among the factors in a six factor model. The constrained model had an insignificant $\chi 2$ of 846.85 with 507 degree of freedom, CFI was.91, and RMSEA was .04. Of primary importance, however, is the $\chi 2$ value because it provides the basis for comparison with the unconstrained, six factor, multi-group model. In testing for the invariance of the constrained model, the $\chi 2$ value of 846.85 (507 *df*) is compared with that of the initial model with no constraints imposed. This is done by taking the difference in $\chi 2$

and *df* between the two models. The comparison yields a $\chi 2$ difference of 49.14 with 33 degrees of freedom ($p > .05$), indicating that the factorial structure of the six factor model is invariant across Black and White offender samples.

The baseline, multi-group model for the second-order factor structure has a $\chi 2$ value 842.27 with 492 degrees of freedom, and CFI and RMSEA are .91 and .04, respectively. These fit statistics indicate that the second-order model is also a good fit for both the White and Black offender samples. In testing the invariance hypothesis, constraints are placed on all factor loadings in the second-order model. The constrained model has an insignificant $\chi 2$ of 869.53 with 516 degrees of freedom, CFI is .91, and RMSEA is .037. The comparison between the unconstrained and constrained second order model yields a $\chi 2$ difference of 27.26 with 24 degrees of freedom ($p > .05$), indicating that the factorial structure of the second-order model is also invariant across Black and White offender samples.

To summarize, the multi-group analyses for both a six factor and second-order models have been shown to be equal across racial groups. A six factor model for the Grasmick et al scale is invariant for Black and White offender samples. A second-order model for the Grasmick et al. scale is also invariant for Black and White offender samples. Overall, the CFA findings reported in this section do not support the idea that the Grasmick et al. scale is measuring a unidimensional construct and that multiple dimensions are being measured by the scale. Furthermore, the multidimensional solutions, i.e., six factor and second order models, not only fit the data well but were both invariant across racial groups, indicating that both factor structures fit Blacks and Whites equally well. Although both models fit these

offender data well across racial groups, an earlier analysis did show that the six factor model fit these data a little better than the second-order model. Results from a Rasch rating scale analysis of Grasmick et al.'s scale items are presented next.

Results from a Rasch Rating Scale Analysis

Category Functioning Analysis

The first step before estimating a Rasch model is to assess category functioning. This is done to understand if response categories are being used appropriately given ability levels of respondents on the trait being measured. Table 16 reports several estimates assessing how the sample of offenders use the categories of Grasmick et al.'s scale items. These estimates include observed counts, average ability scores, and thresholds.

A category frequency distribution of Grasmick et al.'s four category rating scheme, i.e., observed counts, is reported in column 2 of Table 16. It is important to give attention to the distributions shape and the frequency of responses per category. The distribution does not appear to be highly skewed. As noted in Chapter Four, low frequency categories, i.e., say 10 counts or less, are problematic because they do not have enough observations to estimate stable threshold parameters and they often reflect unnecessary or redundant categories. None of the response categories for Grasmick et al.'s items have low frequencies. Across all items, the "strongly disagree" category is chosen 4,563 (29%) times, the "disagree" category is chosen 4,332 (28%) times, the "agree" category is chosen 4,520 times (29%), and the "strongly disagree" category is chosen 2,185 (14%) times.

The average measure, i.e., average ability measure, for persons endorsing a particular category across any item is reported in column 3 of Table 16. The mean of the person ability measure is expected to increase in size as the variable increases, e.g., from strongly disagree to strongly agree. A monotonic increase is expected where respondents with higher ability are more likely to endorse the higher categories across any item. In turn, those who have lower abilities, on average, endorse lower categories on any item. As would be expected for an appropriately functioning category scheme, the average ability measures of respondents' increase as response category usage increased from strongly disagree to strongly agree. The average measure for category 0 "strongly disagree" is -1.17, meaning that the average ability estimate for offenders answering "strongly disagree" across any item is -1.17 logits. The logit-person measure is the natural logarithmic transformation of person raw scores, where more negative scores indicate lower ability, i.e., higher self-control. For offenders who answered 1 "disagree" on any item, the average ability estimate is -.54, meaning that the average ability estimate for offenders answering "disagree" across any item was -.54 logits. The average measure for category 2 "agree" is .09, meaning that the average ability estimate for offenders answering "agree" across any item was .09 logits. Finally, the average measure for category 3 "strongly agree" is .61, meaning that the average ability estimate for offenders answering "strongly agree" across any item is .61 logits (i.e., these persons are more agreeable on average than the offenders who answered 0,1, or 2).

Table 16. Category functioning of Grasmick et al.'s four category rating scale: Observed counts, average measures, and thresholds.

Category	Observed Count	Average Ability	Thresholds
Strongly Disagree (0)	4563 (29%)	-1.17	None
Disagree (1)	4332 (28%)	-.54	-.79
Agree (2)	4520 (29%)	.09	-.30
Strongly Disagree (3)	2185 (14%)	.61	1.09

Column 4 of Table 16 reports thresholds across categories. As noted in Chapter 4, thresholds are the difficulties estimated for choosing one response category over another. Similar to the average measures, thresholds should increase monotonically across the rating scale and those that do not are considered disordered. As illustrated in column 4, the rating scale used for the Grasmick et al. self-control items meets this criterion. As expected, thresholds increase monotonically from -.79 to 1.09 across the rating scale.

In sum, findings discussed above indicate that the average measures and the thresholds function as expected, they increase monotonically across the rating scale from "strongly disagree" to "strongly agree." This is preliminary evidence that the response categories employed for the Grasmick et al. scale items are used appropriately by the sample of offenders and functioning well according to Rasch expectations.

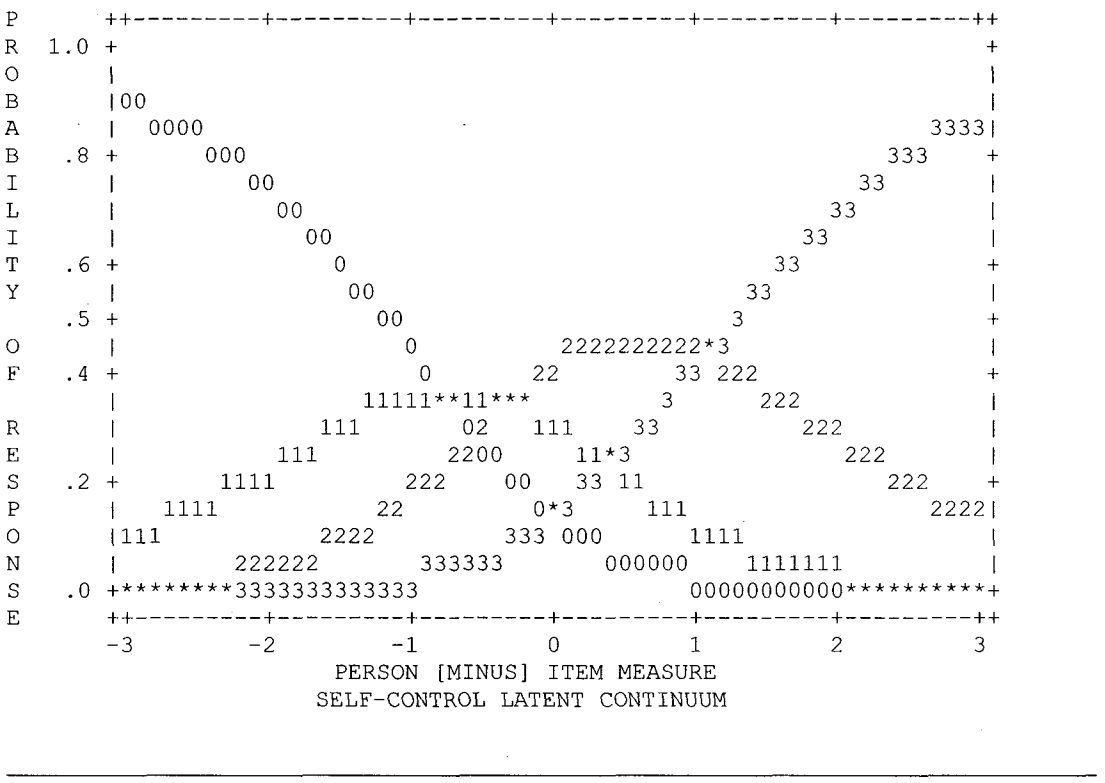
Another way to empirically assess category functioning is by plotting the category probability curves. Figure 13 shows the relationship between the latent trait and the probability of selecting response category k relative to category k - 1. The probability curves reported in Figure 7 display the probability of responding to any particular category (y-axis), given the differences in estimates between any person ability and any item difficulty (x-axis). To explain further, the probability curves in Figure 12 reflect each of the rating scale categories. Specifically, the curve depicted by 0's is the probability curve for the "strongly disagree" category, the curve depicted by 1's is the probability curve for the "disagree category," the curve depicted by 2's is

the probability curve for the "agree" category," and the curve depicted by 3's is the probability curve for the "strongly agree" category.

In Figure 13, the x-axis is labeled as the "person [minus] item measure" and the "self control latent continuum." The latent continuum on the x-axis ranges from positive to negative infinity, but it is standard practice in Rasch analyses to report the latent continuum ranging between -3 and +3 (Bond and Fox, 2002). There are two overlapping ways to explain what the x-axis actually represents. First, due to how items were coded, higher positive scores on the latent continuum indicate higher ability, i.e., lower self-control, when taking into account item difficulty. In turn, more negative scores on the latent continuum reflect higher self-control after taking into account item difficulty. Second, the continuum can be described as representing the difference in logits between any person ability and any item difficulty, thus taking both measures into account when predicting the probability of responding to a certain category. For example, a -3 on the continuum indicates a person's ability is 3 logits lower than the difficulty of the item, whereas, a + 3 on the continuum would indicate a persons ability is 3 logits higher than the difficulty of an item.

Each probability curve should have a distinct peak reflecting that each category is the most probable response at some point along the linear measure. This is true for the four category scheme used for the Grasmick et al. self-control items. For example, while any response is possible, those having a – 3 on the continuum are more likely to choose response category 0 "strongly disagree." Furthermore, the probability of choosing a 0 on any item decreases as scores along the continuum increase. For example, the probability of selecting response category 0 for those with

Figure 13. Category probability curves for Grasmick et al.'s four-category scheme: The relationship between the latent trait and the probability of selecting response category K.



a + 3 score is substantially lower. In turn, those who have + 3 scores are more likely to choose response category 3 "strongly agree;" on the other hand, the probability of choosing a 3 on any item decreases as scores decrease along the continuum. In addition, there will be no category inversions where a higher category, e.g., category 3, is more likely at a lower point than a lower category. Figure 13 shows that each probability curve has a distinct peak along the continuum, none of the probability curves are flat across the latent continuum, and there are no category inversions. Taken together, these findings indicate that categories are not disordered or ill-functioning, i.e., not confusing to respondents.

In sum, findings from the category function analyses indicate that the average measures and thresholds function as expected; they increase monotonically across the rating scale from "strongly disagree" to "strongly agree." In addition, the probability category curves are orderly across the latent continuum, and each of the curves has a distinct peak along the continuum. These results provide evidence that the response categories for the Grasmick et al. self-control scale items are being used appropriately by the sample of offenders and functioning well according to Rasch expectations. Next, an item fit analysis is discussed to assess the unidimensionality of the 24 Grasmick et al. self-control items.

Item Fit Analysis

The next step of a Rasch rating scale analysis for the Grasmick at et al. self-control scale is to evaluate the fit of the 24 items to the Rasch model. Fit statistics are used to assess the success of each item in meeting the unidimensional expectations of the model. All items should fit the Rasch model's expectations if Grasmick et al.'s

self-control items are measuring one trait. This would be the case if all items had standardized infit and outfit statistics that are between the values of -2.00 and +2.00. Items that have values greater than 2 in absolute value indicate that responses to those items are significantly more varied than predicted by the Rasch model. In turn, items that have values less than -2 indicate that responses to those items are significantly less variable than those predicted by the Rasch model.

Table 17 reports item fit statistics for the 24 items of Grasmick et al.'s self-control scale. The second column shows the item difficulty estimates, the third and fourth columns show the standardized infit and outfit statistics associated with each item, and the fifth column identifies misfitting items by a plus sign. As shown in Column 2, the most easily endorsable or agreeable item is P2 (logit = -1.44) and the item most difficult to endorse is Sc3 (logit = .87). A discussion of item difficulty will be expanded on when explaining the person/item map. Columns 3 through 5 in Table 17 show that 11 of the 24 items have statistically significant misfits; that is, 11 items have infit and/or outfit statistics greater than +2 and/or less than -2, indicating that responses to these items are not consistent with predictions of the Rasch model. The misfitting items are as follows: Impulsivity: I2 (Infit = 2.2, Outfit = 3.5), I3 (Infit = -2.8, Outfit = -2.4); Simple Tasks: S1 (Infit = 2.6, Outfit = 2.5), S2 (Infit = 2.7), S3 (Infit = -3.4, Outfit = -3.4); Risk Seeking: R2 (Infit = 2.9, Outfit = 2.4); Physical Activities: P4 (Outfit = 2.4); Self-Centerdness: Sc4 (Infit = -5.1, Outfit = -3.9); Temper: T1 (Infit = 3.3, Outfit = 2.6), T2 (Infit = -3.0, Outfit = -3.7), T3 (Outfit = 2.2). The remaining 13 items of the Grasmick et al. self control scale fit the unidimensional expectations of the Rasch model well.

Table 17. Item Fit Statistics for Grasmick et al.'s 24 Self-Control Items for the Full Sample (n = 651)

Item	Measure	Infit (zstd)	Outfit (zstd)	Misfit
I1	-.45	-.4	-.6	
I2	.79	2.2	3.5	+
I3	-.30	-2.8	-2.4	+
I4	-.38	-.9	-.4	
S1	.37	2.6	2.5	+
S2	.79	2.7	1.8	+
S3	-.35	-3.4	-3.4	+
S4	.38	1.4	.9	
R1	-.97	.7	1.1	
R2	-.07	2.9	2.4	+
R3	.44	1.2	.6	
R4	.38	-.7	-1.1	
P1	-.45	1.5	1.7	
P2	-1.44	1.3	1.6	
P3	-1.19	-.3	.1	
P4	-1.02	.9	2.4	+
Sc1	.12	-.9	-.2	
Sc2	.84	-.7	-.8	
Sc3	.87	-1.7	-.5	
Sc4	.65	-5.1	-3.9	+
T1	.25	3.3	2.6	+
T2	.75	-3.0	-3.7	+
T3	.09	1.8	2.2	+
T4	-.11	-1.8	-1.5	

As discussed in Chapter Four, the Rasch model is a confirmatory model that tests the unidimensionality of a multiple item scale. Given the results in Table 17, the items do not encompass a unidimensional scale that permits the use of a single summary score that can meaningfully indicate a respondent's level on an underlying trait, i.e., self-control. The item fit analysis from the Rasch model, coupled with the poorly fitting one factor CFA model, provide two independent sets of results that show the Grasmick et al. scale is not measuring a unidimensional construct.

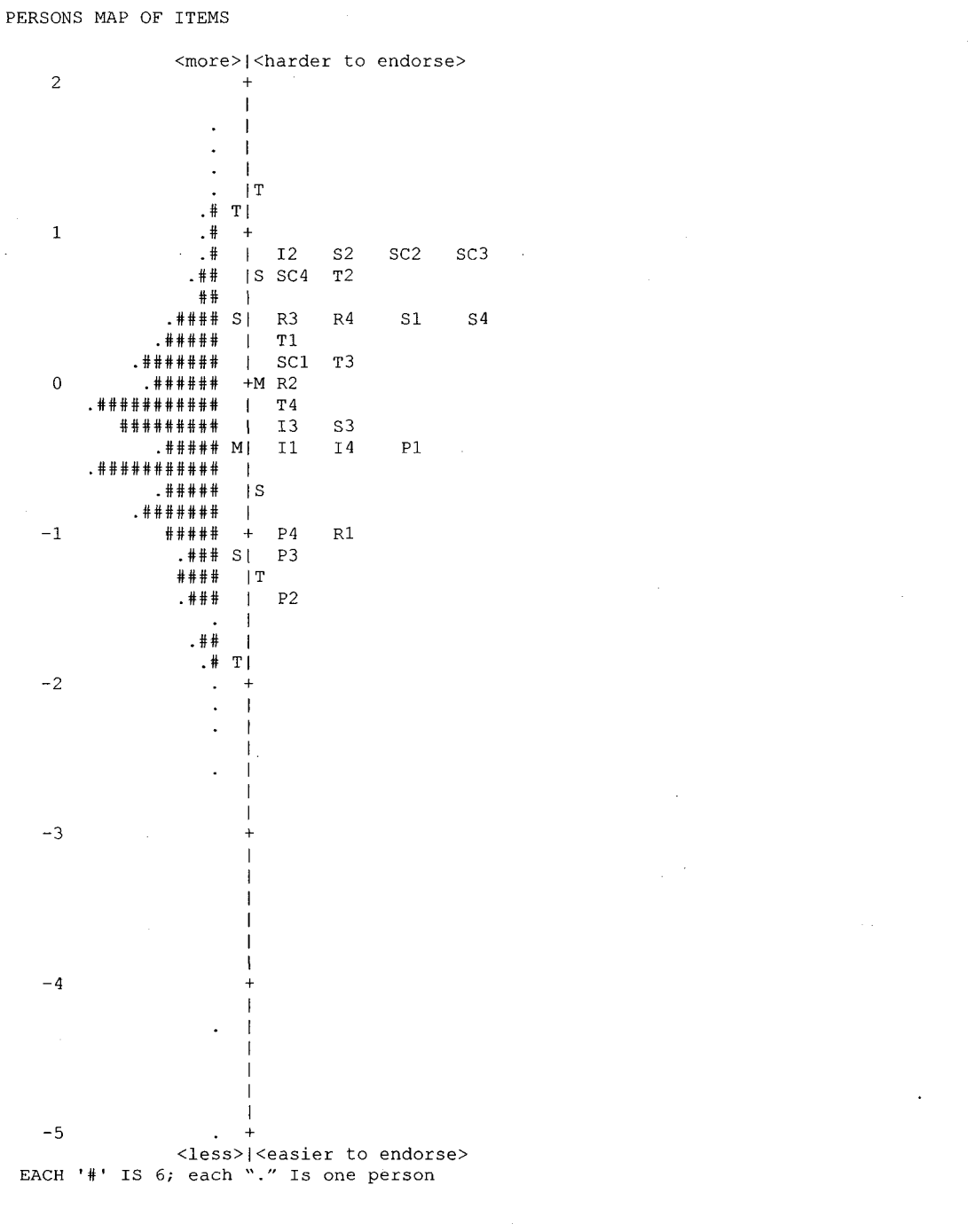
A Rasch Person/Item Map

At the core of a Rasch analysis is the map or "ruler" that visually displays the person and item measures in tandem. Both person positions and item difficulties are expressed on the same logit ruler, allowing for an examination of item functioning relative to the sample of respondents. Although the person/item map does not allow for an assessment of the unidmensionality of a scale, creating a visual map is important. As mentioned earlier, a person/item map is used to assess whether items are too difficult or easy for a sample to endorse given the range of person abilities. As for the Grasmick et al. scale, it is important to know whether items are too difficult or to easy given the offender sample's range of abilities on the latent trait. Since the Grasmick et al. scale items were not created with offender samples in mind, it could be that these items are too easy to endorse for these offenders because they, on average, may have high abilities. Such a finding would imply that new items should be created that can discriminate levels of ability for a sample of people that, on average, has low self-control.

Figure 14 shows the person/item map for the full offender sample. The distribution of person positions is on the left side of the vertical line and items are on the right side. The numbers to the far left represent the logit scores for the map of persons and items. On the left side of the vertical line, each "#" represents about six to ten persons in this figure and a "." represents one person. Those at the upper end of the scale, i.e., larger logit scores, agree with more items and agree more strongly, reflecting lower self-control; and those with more negative logits agree with less items, reflecting higher self-control. Items are represented on the left side of the vertical line. Items at the upper end of the scale, i.e., larger logit scores, are more difficult to endorse; whereas, items at the lower end of the scale are easier to endorse. Several letters are reported on the vertical line that divides persons and items. "M" marks the person and item means (average logit scores), "S" is one standard deviation from the mean, and "T" is two standard deviations away from the mean.

If the Grasmick et al. self-control items were too easy for the offender sample, items would be expected to have a mean logit score substantially lower than the mean logit score for the person ability measure. Furthermore, the expectation would be that items are distributed at the bottom of the ruler and persons distributed at the top of the ruler, Figure 14 shows that this is not the case. In fact, the mean item difficulty score is larger than the mean person ability score, although the difference is rather small. Also, the person ability and item difficulty distributions are not that much different along the logit ruler. It should be noted that items cover a range of logits from -1.44 (P2) to .87 (Sc3), narrower than the range of person abilities. This indicates that items are not measuring the entire range of abilities in the offender sample, especially at the

Figure 14. Rasch person/item map

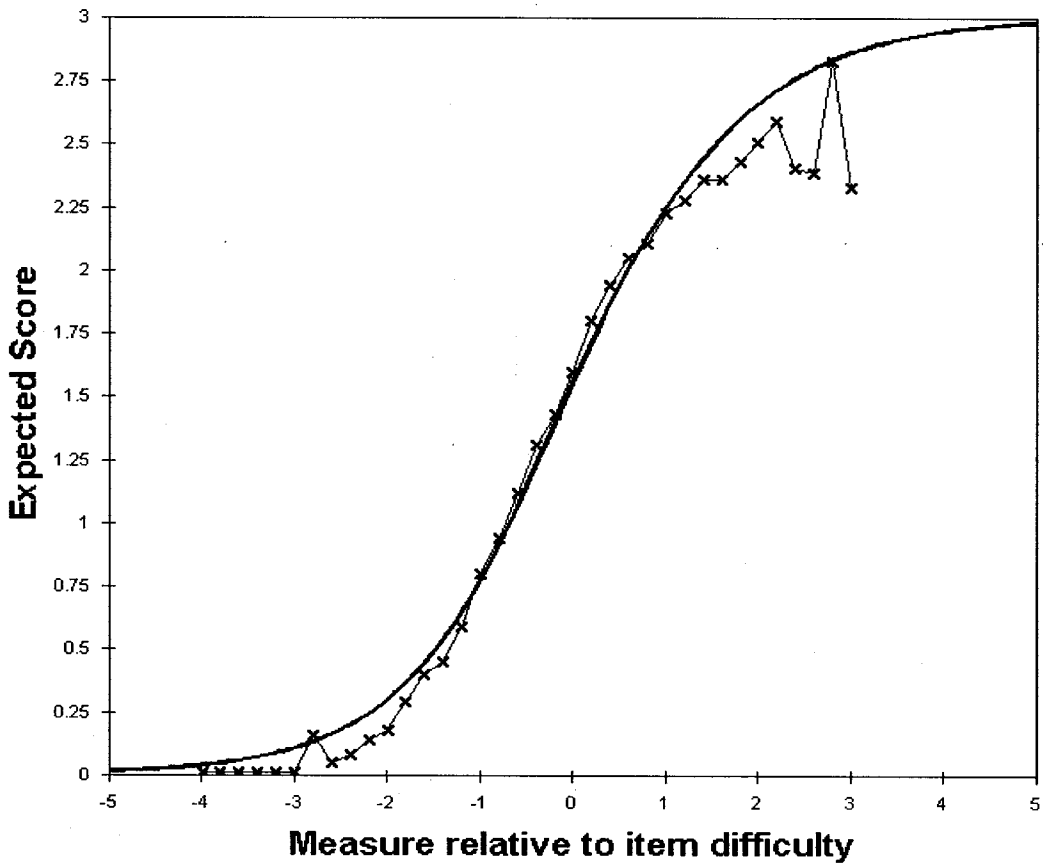


extreme ends where some individuals are very high and low in ability. Also, it should be noted that at several points on the item logit scale there are several items at the same position, indicating that some of the items are redundant. For example, items I1 (logit = -.45) and I4 (logit = -.45) are indistinguishable. The same can be said for items I2 (logit = .79) and S2 (logit = .79), and again with R4 (logit = .38) and S4 (logit = .38). This is also shown in Table 17 which reports the logits for item difficulty.

Assessment of the Item Characteristic Curve

Figure 15 shows the ICC graph for the full offenders sample. This graph shows "measure to item difficulty." The x-axis has the same meaning as explained earlier in Figure 13 and it can be thought of as representing the underlying trait being measured. Those with more positive logits on the continuum, i.e., x-axis, are lower in self-control and those having more negative logits are considered higher in self-control. The y-axis represents the expected average score on an item for persons at each measure relative to the item. The y-axis ranges between 0 and 3, where 0 equals "strongly disagree" and 3 equals "strongly agree." Two plots are observed. The solid line is the ICC that the Rasch model predicts, the average expected ICC. The dotted line is the ICC according to the data, the average empirical ICC. This is a summarized average across all items. The observed average score for persons at each measure relative to the item is represented by the dotted curve. Figure 15 shows that the expected and observed ICC are close to one another, indicating that the offender data follow what was predicted by the Rasch model. Although these curves overlap one another almost identically, it is important to point out that this is not true for the

Figure 15. Item Characteristic Curve for the Grasmick et al. self-control measure



high end of the observed curve. Importantly, some overall high performers on the measure, i.e., those who would be labeled as having lower self control, have chosen some unexpectedly low categories causing the incongruence in the higher tail of the curves (Linacre, personal communication).

Why is this finding important? Hirschi and Gottfredson (1993) have cautioned against using self-report attitudinal measures of self-control. They argue that an individual's level of self-control may affect how he/she responds to such items. Those individuals with the lowest levels of self-control will be less likely to respond to items as would be expected. The ICC graph in Figure 15 is consistent with Hirschi and Gottfredson's concerns about attitudinal self-control measures. That is, attitudinal measures may not be suitable for those with the lowest levels of self-control. Although not a statistical hypothesis test, the ICC curve provides preliminary evidence for Hirschi and Gottfredson proposition.

DIF analysis across Racial Groups

DIF is a condition when a scale item or several items function differently for respondents from one group to another. In other words, do respondents, with similar levels on a latent trait, but who belong to different populations, have a different probability of agreeing to or endorsing an item or items? The idea is that scale items should not measure people differently, there should be no bias. A conceptual example of DIF can be given using a ruler. A ruler provides a measure of height in inches. Height scores are invariant regardless of the ruler used. A ruler can be used to measure anyone's height and is not restricted for measuring the height of certain groups; therefore, a ruler would not show differential functioning in measuring height

across different people. A ruler is not biased in measuring the height of men or women or Blacks or Whites. DIF is used to test such an idea when applied to multiple item scales. In other words, do items of a scale measure people the same or differently depending on the group of people being measured by the items?

The current dissertation uses DIF analysis to assess item functioning of the Gramsick et al. items across Black and White offender samples. This analysis assesses whether or not items vary significantly in difficulty or agreeability across samples while controlling for the level of ability. Do Black and White offenders with similar levels of ability have different probabilities of endorsing the 24 items? For scale validity purposes, Black and White offenders should have the same probability of endorsing the same items if item difficulties are invariant across offender samples. Item bias would exist if items significantly vary in their difficulty levels across groups.

Table 18 shows a DIF analysis of the Grasmick et al. items across offender samples. Column 1 lists the items; columns 2 and 3 show the estimated item difficulties for the two offender samples; and column 4 shows z-tests to assess the difference of individual item functioning across samples. Similar to a t-test, Z values that are above a 1.96 in absolute value represent statistical significance ($p < .05$), indicating that a particular item significantly differs in difficulty across samples. Table 18 shows that 10 of the 24 items exhibit DIF. Generally, these findings indicate that 10 of the self-control items significantly vary in their difficulty levels or agreeability across the two samples Specifically, this shows that many items are biased or easier/harder to endorse for the Black and White offenders, while

Table 18. Differential Item Function (DIF) analysis for the Grasmick et al. scale: An assessment across Black and White samples.

Items	Black Sample Measure	White Sample Measure	z-test
I1	-.20	-.64	3.52*
I2	1.03	.71	2.14*
I3	-.43	-.35	-.63
I4	-.34	-.38	.34
S1	-.10	.62	-5.61*
S2	.73	.92	-1.34
S3	-.59	-.27	-2.59*
S4	.29	.53	-1.79
R1	-.93	-1.17	1.78
R2	.31	-.27	4.47*
R3	.71	.32	2.79*
R4	.96	.27	4.68*
P1	-.35	-.57	1.74
P2	-1.45	-1.54	.65
P3	-1.15	-1.28	.93
P4	-1.40	-.94	-3.37*
Sc1	-.06	.27	-2.63*
Sc2	.89	1.01	-.83
Sc3	.96	.90	.42
Sc4	.51	.67	-1.18
T1	.13	.28	-1.14
T2	.79	.84	-.32
T3	.14	.12	.17
T4	-.36	-.05	-2.45*

*$p < .05$

controlling for their level on the latent trait, i.e., ability. If scale items were invariant or unbiased, like a ruler, none of these items should have shown differences in difficulty across samples.

Two of the "Impulsivity" items show DIF across samples. Black offenders find it more difficult to agree to items I1 (black logit = -.20; white logit = -.64) and I2 (black logit = -1.03; white logit = .71) when compared to White offenders with similar self-control levels. Two of the "Simple Task" items show differences across offender samples. White offenders find it more difficult to agree to items S1 (black logit = -.10; white logit = .62) and S3 (logit = -.27; logit = -.59) compared to White offenders with similar self-control. Three of four "Risk Seeking" items show DIF across offender samples. Black offenders find it more difficult to agree to items R2 (black logit = .31; white logit = -.27), R3 (black logit = .71; white logit = .32), and R4 (black logit = .96; white logit = .27) when compared to white offenders with the similar ability levels. One of four "Physical Activities" items showed DIF across offender samples. It is more difficult for White offenders to agree to P4 (black logit = -1.40; white logit = -.94) when compared to Black offenders with similar ability levels. One of four "Self-Centered" items shows a difference across Black and White offender samples. It is more difficult for White offenders to agree to Sc1 (black logit = -.06; white logit = .27) when compared to Black offenders with similar self-control. One of four "Temper" items shows DIF. It is more difficult for White offenders to agree to T4 (black logit = -.36; white logit = -.05) when compared to Black offenders with similar levels of self-control.

Many of Grasmick et al.'s items were shown to be biased or differences across racial groups of offenders. If scale items were invariant, or unbiased, none of these items should have exhibited statistically significant differences in difficulty. The DIF analysis has provided evidence for the lack of invariance of items across racial groups. This poses another threat to the validity of Grasmick et al.'s scale. Failure to produce invariance prohibits comparisons of groups on this measure because some have a higher probability of endorsing some items than other items. This important finding is discussed in more detail when summarizing the results in Chapter 6. A general discussion and summary of results from this dissertation are now presented.

CHAPTER 6:

DISCUSSION AND CONCLUSIONS

It has now been over a decade since the publication of Gottfredson and Hirschi's (1990) *General Theory of Crime*, i.e., self-control theory. While there is little doubt that their theory is one of the more controversial and debated theories in criminology (Akers, 1991; Geis, 2000); there is no question that it continues to be at the center of criminological explanations for criminal offending, delinquency, deviance, and antisocial behavior (see Pratt and Cullen, 2000). Self-control theory's continued presence as an important explanation for crime is not a surprise, as a large body of supportive scientific evidence has accumulated in its favor over the last decade.

The body of empirical research testing hypotheses on the wide ranging effects of low self-control has generally drawn the conclusion that individuals low in self-control will have many negative experiences throughout life because of their dispositions. They will be victims of crime, engage in antisocial and criminal acts frequently, be unsuccessful in school, associate with delinquent peers, have difficulty in dating relationships and marriages, and have difficulty garnering prosocial bonds in society, only to name a few. Favorable evidence in support of self-control theory has been reported for offenders, adolescents, across gender, across race, and for people residing in different countries (Forde and Kennedy, 1997; Pratt and Cullen, 2000; Schreck, 1999; Stewart, Elifson, and Sterk, 2004; Tittle, Ward, and Grasmick, 2003; Vazsonyi et al., 2000).

In summarizing this growing body of research, Pratt and Cullen (2000) assessed the empirical status of self-control theory by performing a meta-analysis on twenty-eight published studies and drew several conclusions. First, Pratt and Cullen (2000:952) concluded that their findings "would rank self-control as one of the strongest known correlates of crime" and "future research that omits self-control from its empirical analyses risks being misspecified." Second, Pratt and Cullen (2000: 953) indicated that "on an absolute level, therefore, it appears that low self-control must be considered an important predictor of criminal behavior and the general theory warrants a measure of acceptance."

Acknowledging Pratt and Cullen's analyses, Marcus (2004) stated that, "it would appear that a settlement has been reached, allowing for a well-informed evaluation of its [self-control theory's] empirical status." Marcus (2004), however, went on to suggest that this is not the case and that any well-informed evaluation of the empirical status of self-control theory would be premature, as a major issue has gone largely unaddressed. Minimal attention has been devoted to the measurement of self-control theory's major construct – self-control. Marcus (2004) went on to say that we are far from a well informed empirical evaluation because the evidence to date supporting self-control theory is questionable due to the conceptual, operational, and measurement problems surrounding Gottfredson and Hirschi's most important construct, i.e., self-control.

Many of Marcus's concerns resonate throughout this dissertation. While it has not been the current dissertation's goal to offer a new definition and measure of Gottfredson and Hirschi's self-control construct, a goal has been to assess the

psychometric properties of the most commonly used measure of self-control, the 24 item Grasmick et al scale.

As discussed in Chapter Three, many studies have used Grasmick et al.'s scale as the definitive measure of self-control while neglecting to have thoroughly examined its measurement qualities. In doing so, researchers have often resorted to summing responses to Grasmick et al.'s scale items for each person in their sample and using the summated score to indicate a person's level of the trait, implying that the intrument was measuring one thing, self-control or lack there of. Unfortunately, this crude treatment of scale items has been justified based on evidence of good internal consistency and evidence from exploratory factor analyses which has shown all items to load on one factor.

Studies that have examined the measurement qualities of this scale have shown that, as a whole, the scale has good internal consistency, but results pertaining to its internal structure validity have been questionable. The latter has led to the following question: Should responses to Grasmick et al.'s scale items be summed to represent respondents' levels of self-control if items did not conform to a unidimensional structure? Particularly, with regard to internal structure validity, Grasmick et al.'s scale has been shown to reflect one underlying trait, six distinct but correlated dimensions, and six dimensions that can be explained by a higher order trait. As discussed in Chapter Three, these conclusions have been drawn using community, college, parolee, and drug treatment offender samples. Given such inconsistent results it would be premature to treat Grasmick et al's scale items as measuring one thing. This dissertation has empirically confronted this problem.

This chapter has been divided into two primary sections. First, a summary of results will be provided as they are related to each research question from this dissertation. In addition, current results will be compared to those that have been previously published on Grasmick et al.'s self-control scale. Second, limitations of the current study's design will be identified and discussed. Finally, given the results and limitations, directions for future research are explored which should help identify and guide further investigations of the measurement of self-control.

Summary of Findings

Using a sample of serious, incarcerated male offenders, this dissertation addressed several measurement concerns regarding Grasmick et al.'s scale. As part of a larger National Institute of Justice funded study, the Grasmick et al. scale was administered to a random sample of newly admitted male offenders to the Nebraska Department of Corrections between October 1997 and December 1998. This section will summarize results from this offender sample as they are related to each research question proposed in this dissertation.

Is Grasmick et al.'s scale a reliable measure for a sample of incarcerated offenders?

Grasmick et al.'s scale was shown to have strong reliability. Particularly, a series of Cronbach's reliability analyses conducted on the 24 item scale revealed high internal consistency for the full sample of offenders ($\alpha = .87$), as well as, the black ($\alpha = .85$) and white ($\alpha = .89$) offender samples. Furthermore, deletion of any of the items did not improve scale reliability.

Results from this dissertation confirm what has been observed in past studies that have evaluated the reliability of Grasmick et al.'s scale. The scale appears to be measuring a common trait, has little measurement error, and is highly consistent. This is the general agreement across samples ranging from college students, adolescents, international respondents, community residents, and offenders. This dissertation, however, has been the first to assess the reliability of the scale when administered to a sample of male inmates. Furthermore, analyses in this dissertation were the first to investigate the reliability of Grasmick et al.'s scale across Black and White male inmates.

Given evidence from prior studies that suggested Grasmick et al.'s scale was measuring six dimensions (i.e., impulsivity, risk seeking, self centeredness, temper, preference for simple tasks, and preference for physical activities), separate reliability analyses were conducted for each dimension for the full offender sample, as well as, the Black and White samples. Generally, most Cronbach's coefficients for the six subscales indicated high internal consistency. For the full sample, reliability estimates ranged from .61 for the Impulsivity subscale to .81 for the Temper subscale. As for the Black offender sample, reliability estimates ranged from .50 for the Impulsivity subscale and .79 for the Simple Tasks subscale. As for the White offender sample, reliability estimates ranged from .62 for the Impulsivity subscale to .85 for the Temper subscale.

As mentioned, most subscales showed high internal consistency. Nevertheless, some noteworthy differences emerged. First, reliability estimates for each subscale were consistently lower for the black offender sample. These

differences were rather small and were possibly due to the smaller sample size of black offenders in the study. On the other hand, the observed differences for the black sample could be meaningful, indicating that subscales were not as reliable when administered to blacks. Second, the Impulsivity and Physical Activities subscales had smaller alpha coefficients relative to the other subscales. Particularly, the Impulsivity subscale had the most measurement error and/or lack of internal consistency. Item I2 ("I devote much thought and effort to preparing for the future") could have been deleted to improve the reliability of the Impulsivity subscale, but the improvement would have been marginal. This item was problematic throughout the entire set of analyses conducted in this dissertation.

To date, few studies have conducted separate reliability analyses for each subscale of the Grasmick et al. measure. Using criminal samples, Piquero and Rosay (1998) and DeLisi et al. (2003), respectively, found different and similar subscale reliability estimates from those in the current dissertation. Particularly, Piquero and Rosay (1998) had subscale reliability coefficients of .45 for Impulsivity, .44 for Simple Tasks, .58 for Risk Seeking, .37 for Physical Activities, .68 for Temper and .57 for Self-centeredness. DeLisi et al. (2004) found stronger coefficients for all subscales. They reported coefficients of .79 for Impulsivity, .81 for Simple Tasks, .79 for Risk Seeking, .72 for Physical Activities, .86 for Temper, and .81 for Self-centeredness. Similar to the current study, both of these studies found the Impulsivity and Physical Activities subscales to exhibit lower reliability.

As discussed in Chapter Two, reliability has been a necessary but not sufficient for achieving valid measurement. A finding of good reliability for Grasmick et al.'s scale did not mean that it was an accurate portrayal of self-control

nor did it mean that researchers were justified in treating the scale as unidimensional. Unfortunately, several studies using Grasmick et al.'s scale to investigate the effects of self-control on different outcomes have relied on little more than a reliability coefficient as justification for treating it as measuring a unidimensional construct. The next set of results to be summarized address validity concerns of Grasmick et al.'s scale.

Does Grasmick et al.'s scale show observed differences across racial groups for a sample of incarcerated offenders?

Gottfredson and Hirschi (1990) stated that differences exist in offending behaviors across racial groups. Theoretically, they argued that racial variation in offending can be explained by racial differences in self-control, suggesting that blacks and whites should vary significantly in their levels of self-control. Although it was not the current dissertation's goal to explore whether racial differences in offending could be explained by differences in self-control, it was one of its goals to investigate whether an instrument used to measure self-control was able to detect the racial differences in self- control that Gottfredson and Hirschi suggested in their theory.

From a construct validity framework, a valid measure of self-control should show observed differences for minority groups when compared to whites. Particularly, blacks, on average, should have lower self-control than whites, which could then be interpreted as preliminary evidence for the cross structure validity of Grasmick et al.'s scale. A series of independent sample t-tests were estimated to assess average differences for the full scale and each subscale across black and white offender groups. Several statistically significant differences emerged. White

offenders, on average, scored higher on the full scale than black offenders. In addition, white offenders, on average, scored higher than black offenders on the Impulsivity and Risk seeking subscales. Black and white offenders did not significantly differ on any of the other dimensions.

Although differences were observed across racial groups, these differences were opposite of what would be anticipated from a construct validity framework. That is, assuming Gottfredson and Hirschi are theoretically correct, a valid measure of self-control should have shown blacks to have higher scores than whites across all subscales and the full scale. To date, no study has approached validation of Grasmick et al.'s scale using a cross-structure analysis as the one employed in this dissertation.

Results from the independent sample t-tests could be interpreted in several ways. First, these findings can be interpreted as an initial strike against the validity of Grasmick et al.'s scale. This could have been due to the fact that a measure which was specifically designed to test Gottfredson and Hirschi's (1990) concept of self-control was not empirically consistent with the theoretical expectation that minorities should have lower levels of self-control than whites. On the other hand, Grasmick et al.'s scale may not be measuring self-control, but something actually different (Marcus, 2002). This point is discussed latter in this chapter.

Second, and aside from construct validity, it could be that Gottfredson and Hirschi's theoretical expectation driving this cross-structure analysis was incorrect. That is, blacks should not theoretically be expected to have lower self-control than whites. If so, future theoretical developments should offer competing explanations that articulate why minorities and whites should be expectd to have similar levels of

self-control. Third, the observed differences between white and black offenders could be explained by the fact that several of Grasmick et al.'s self-control items were biased across racial groups. These findings emerged from the Rasch analysis and will be discussed later in this section.

Finally, differences in blacks and whites on their levels of self-control could be real and due to criminal justice system processing. That is, black and white offenders could be screened differently so that more serious white offenders and less serious black offenders are convicted and incarcerated.

Is Grasmick et al.'s scale unidimensional?

One of the most debated topics concerning Grasmick et al.'s instrument is whether its items conform to a unidimensional structure that reflects the measurement of one trait, self-control. As discussed in Chapter three, several studies have concluded that its items conform to a unidimensional structure (Grasmick et al., 1993; Longshore et al., 1996; Nagin and Paternoster, 1993; Piquero and Rosay, 1998), while others have rejected this idea in support of a multidimensional structure (Delisi et al., 2003; Vazsonyi et al., 2002).

Even in the presence of conflicting results, the majority of research, past and present, has continued to treat Grasmick et al.'s scale items as measuring a one-dimensional trait. As mentioned, many of these studies have typically conducted reliability analyses and exploratory factor analyses on scale items to justify the summing of responses to items to represent individuals' levels of self-control. It is no surprise that researchers continue to treat Grasmick et al.'s scale this way, as these are the analytic techniques used by the creators of the scale when they assessed its

internal structure validity (Grasmick et al.1993). This continued practice has been followed even by researchers who have discovered, through very rigorous psychometric assessment, that Grasmick et al.'s scale is not measuring a unidimensional trait (see Piquero et al., 2000; Piquero, Gomez-Smith, and Langton, 2004).

Many researchers have followed in Grasmick and colleagues (1993) foot steps by treating their self-control scale as measuring one construct; but, as shown by the results from this dissertation, researchers may be following the wrong foot steps. That is, three different analytic procedures were used in this dissertation to explore and confirm that Grasmick et al.'s scale items are not measuring one trait, at least for a sample of incarcerated offenders. Particularly, this dissertation used exploratory factor analysis, confirmatory factor analysis, and a Rasch rating scale analysis to show that Grasmick et al.'s scale items do not conform to a unidimensional structure.

First, using exploratory techniques, principal components and principal axis factor analyses, a conflicting picture of the dimensionality of Grasmick et al.'s scale emerged. That is, results showed that all scale items loaded on one factor, but evidence also indicated that multiple factors were being measured. This pattern emerged for the full, black, and white offender samples. Following the practice of many studies, it would have been convenient to stop here and conclude that Grasmick et al. items are generally conforming to a unidimensional structure since all items clearly loaded on the first component or factor. However, such a practice is naive, as evidence from additional analyses showed that multiple factors exist. As discussed

below, additional analyses found that interpreting the exploratory models as evidence of unidimensionality would have been a large mistake.

Second, Gramsick et al.'s scale items were subjected to a strict test of unidimensionality. Unlike the exploratory analyses, a confirmatory factor analysis was estimated where scale items could not freely load on multiple factors. Items were forced to load on only one factor. In doing this, a clear picture emerged. That is, while items had positive and statistically significant loadings, each fit index showed that a one factor solution was definitely not a good fit for Grasmick et al.'s items. These findings were similar to what other researchers have reported using the same analytic strategy (DeLisi et al., 2003).

Finally, to further investigate the unidimensionality of Grasmick et al.'s scale a Rasch rating scale analysis was conducted. As discussed at length in Chapter four, a Rasch model is a confirmatory technique used to investigate whether scale items are contributing to the measurement of one trait. Particularly, given persons abilities on a latent trait and item difficulty, a Rasch model predicts how a person should respond if the item is contributing to measuring a unidimensional trait. That is, if observed responses to an item significantly diverge from Rasch expectation then the item is not contributing to the measurement of a unidimensional trait. This model allowed for an assessment of whether each scale item diverged from the Rasch expectation of unidimensionality.

A Rasch analysis rejected the hypothesis that the 24 item Grasmick et al. self-control scale was suitably unidimensional to warrant the use of a single summary score to represent respondents' levels of self-control. Particularly, 11 of the 24 items

did not meet a Rasch model's expectations for unidimensionality. These items included two of impulsivity items (I2 "I don't devote much thought and effort to preparing for the future" and I3 "I often do whatever brings me pleasure here and now, even at the cost of some distant goal"), three simple tasks items (S1 "I frequently try to avoid projects that I know will be difficult", S2 "When things get complicated, I tend to quit or withdraw", and S3 "The things in life that are easiest to do bring me the most pleasure."), one risk seeking item (R2 "Sometimes I will take a risk just for the fun of it"), one physical activities item (P4 "If I had a choice, I would almost always rather do something physical than something mental", one self-centeredness item (Sc4 "I will try to get the things I want even when I know it's causing problems for other people"), and three temper items (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", and T3 "When I'm really angry, other people better stay out of my way"). In sum, a Rasch analysis indicated that these 11 items be dropped for valid measurement of a unidimensional trait.

Results from a Rasch analysis were similar to those reported by Piquero and colleagues (2000). Using a sample of college students, they found that 10 of the 24 items diverged significantly from the unidimensional expectation of a Rasch model. Although similar findings emerged, some differences were also found. First, Piquero and colleagues found several items to have statistically significant misfit that the current analysis did not. Specifically, Physical Activity 1, Self-centered 2, and Self-centered 3 were found to have statistically significant misfit in the Piquero et al. (2000) study. Second, the current Rasch analysis found several items to have

statistically significant misfit that Piquero et al. (2000) found to fit well. Specifically, Simple Task 2, Risk Seeking 2, Self-centered 4, and Temper 1 were found to have statistically significant misfit in the current analysis, but were not shown to have misfit in the Piquero et al. (2000) study. While it is beyond the current dissertation's goal to explain why the two studies found dissimilar patterns in item fit, it was the current dissertation's goal to assess whether Grasmick et al.'s scale items meet the unidimensional expectations of a Rasch model. In general, the most important finding here is that both studies, the current dissertation and Piquero et al. (2000), reached the same general conclusion using two different samples. That is, the 24 item Grasmick et al. self-control scale is not sufficiently unidimensional to warrant the use of a summated score to represent a person's level of self-control.

The absence of a unidimensional structure for Grasmick et al.'s scale is an important finding. In creation of their scale, Grasmick and colleagues interpreted Gottfredson and Hirschi' concept of self-control as a unitary construct and, as such, predicted that their scale should measure one thing, self-control. That is, they stated that, "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." Testing the factor structure of their own scale they determine that it did fit a one factor model that justified the creation of a single scale. The current results seriously challenge their findings. Overall, evidence of the scale's lack of unidimnesionality represents another strike against the validity of Grasmick et al.'s scale.

In light of the current results, those using the Grasmick et al. scale as a unitary measure of self-control should more thoroughly investigate its internal structure before conveniently summing responses to items and calling the resulting score a measure of self-control. Furthermore, researchers that have done this should revisit their data and pay closer attention to the scale's internal structure; this includes Grasmick and his colleagues (1993). Additionally, studies that have interpreted the scale as being unidimensional using only exploratory factor analytic methods should revisit the scale and employ both traditional confirmatory models and a Rasch model to assess whether or not results from this dissertation can be replicated. If similar findings then emerge, we must seriously begin to question what and how many things are being measured by Grasmick et al.'s instrument.

Is Grasmick et al.'s scale multidimensional?

There was sufficient evidence from this dissertation to state that Grasmick et al.'s scale is not measuring one trait, at least for a sample of incarcerated male offenders. This leads to the next question. What is the dimensionality of Grasmick et al.'s scale? Several researchers have interpreted Gottfredson and Hirschi's (1990) concept of self-control as being multidimensional (Arneklev et al., 1999; Longshore et al., 1996; Vazsonyi et al., 2001). That is, six unique, but correlated, elements define self-control, i.e., impulsivity, risk seeking, preference for simple tasks, preference for physical activity, self-centeredness, and temper. Researchers have interpreted the concept of self-control as such because Gottfredson and Hirschi explicitly stated that self-control is made up of these six elements.

The current dissertation found support for the multidimensionality hypothesis that Grasmick et al.'s scale is measuring several traits. Two measurement models were tested to reach this conclusion. First, a confirmatory factor analysis that specified a six dimensional structure was tested where each dimension was allowed to covary with the others, and items were only allowed to load on their respective dimensions. Model fit statistic indicated that this model fit the offender data very well. Moreover, this model fit equally well for black and white offenders. That is, the six factor model was invariant across both samples of offenders. Furthermore, the six-factor model, after performing a chi-square difference test, was found to have the most parsimonious fit when compare to the second order-model. Second, a second-order model fit the data very well. Each of the six dimensions identified by Gottfredson and Hirschi (1990) were measured by the Grasmick et al self-control scale, and these dimensions could be partially explained by a higher order construct. The word partially should be emphasized, as the higher order factor, titled self-control, was not completely successful in explaining variation in the lower order factors. For example, the second order factor explained 75 % of the variance in the Impulsivity factor, but only explained 27 % of variance in the Simple Task factor. As like the six-factor model, the second-order model was invariant across Black and White offender samples. In sum, the confirmatory factor analyses reported in this dissertation revealed more support for the idea that Grasmick et al.'s self- control scale was measuring elements of what Gottfredson and Hirschi called self-control and not one trait. In addition, multidimensionality was validated across racial groups of

offenders. These findings were consistent with results reported by other researchers (Arneklev et al., 1999; Vazsonyi et al., 2001).

While some have suggested 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). Nevertheless, if six unique, but correlated, traits were being measured by Grasmick et al.'s self-control scale then empirical research using this scale to predict deviant and criminal outcomes should treat it as such. That is, each element of self-control should be considered as a predictor; if not, it will be unclear how each element contributes to explaining a particular outcome. As already mentioned, the majority of studies using Grasmick et al.'s scale to measure self-control have done the exact opposite by summing responses to all 24 items and treating the resulting score as a person's level of self-control while not acknowledging that the scale could be measuring several different traits. Such treatment of Grasmick et al.'s scale could mask important contributions of each element in predicting outcomes hypothesized by self-control theory. It could be that some elements are more important for predicting some deviant and criminal outcomes than others. This has been shown in several studies already (Longshore et al. 1996; DeLisi et al., 2003). For example, Delisi and his colleagues (2003) found that the temper dimension of Grasmick et al.'s scale was the only significant predictor of delinquency.

Can Grasmick et al.'s scale items discriminate between levels of ability for a sample of incarcerated offenders?

To date, most studies that have assessed the psychometric properties of Grasmick et al.'s self-control scale have investigated its dimensionality and largely neglected other important issues pertaining to the utility of its items (Arneklev et al., 1999; DeLisi et al., 2003; Longshore et al., 1996; Piquero and Rosay, 1998; Vazsonyi et al., 2001). More specifically, the majority of these studies have not investigated whether scale items are well suited for measuring the range of self-control within different samples of respondents. This was particularly important for the current dissertation because until now it was unknown whether the Grasmick et al. scale items were appropriate for a sample of incarcerated offenders, as they were originally tested on a community sample (Grasmick et al., 1993). It was anticipated that these items would be too easy to endorse for incarcerated offenders because offenders, on average, may have very low levels of self-control. If all items were too easy (or difficult) for the offender sample, the scale would not provide very accurate scores for that population.

Using the person/item map produced by the Rasch rating scale analysis, it was shown that Grasmick et al.'s self control items ranged in difficulty in a way that sufficiently matched the range of offender abilities. This is good news for the Grasmick et al.' s items, as it was originally anticipated that items would be to easily endorsed (or agreed on) by a sample of incarcerated offenders, thus not providing much discrimination or accurate information for persons who on average had high ability (lower self-control). However, results showed that items were sufficiently

capable of discriminating among the range of offender abilities. That is, item difficulties spread the range of offender abilities and were not all too easy or difficult to endorse.

Only one other study has assessed the distribution of person abilities and item difficulties when using Grasmick et al.'s scale items to measure self-control. In doing so, Piquero et al. (2000) administered these items to a sample of college students attending a large, public university on the east coast. In contrast to the current findings, Piquero et al. (2000: 918) found that "many items do not provide much information about the sample [college students] because they are too difficult for many to endorse." That is, many items were difficult to endorse for a sample of people that had low levels of ability (or higher self-control). These contrasting findings are interesting and important for several reasons. First, these analyses showed that Grasmick et al.'s self-control items were more suitable for offenders than college students, as many of the items weree too difficult for college students to endorse, but match the offender abilities reasonably well. Although beyond the scope of this dissertation, future studies should re-evaluate items and construct alternatives that can equally discriminate among samples of people that have different ability distributions so that the same scale items could provide accurate information regardless of the distribution of abilities. Second, contrasting results that emerged from the current study and the Piquero et al. (2000) study brought into question the ability of Gramsick et al.'s scale items to provide accurate information for different populations. Grasmick et al.'s items have been used to measure self-control for samples that potentially have varying levels of ability such as community residents,

drug treatment offender, paroled offenders, adolescents, and youngsters living in other countries. Therefore, researchers may want to revisit these samples and assess item difficulty relative to person ability, as items may or may not provide accurate information for other populations. Grasmick et al.'s items may provide more accurate information for samples that are expected to have lower self-control (or higher ability) and provide less accurate information for those expected to have high self-control (or lower abilities).

Do respondents' levels of ability on Grasmick et al.'s scale affect survey responses?

Although results from this dissertation have raised suspicion as to the unidimensionality of Grasmick et al.'s scale, it did show some support for a proposition Gottfredson and Hirschi put forth. They stated that self-control will affect the validity of survey responses in that responses will be less valid for individuals with higher criminality and/or lower self-control (Hirschi and Gottfredson, 1993; Piquero et al., 2000). Gottfredson and Hirschi suggested that attitudinal self-report measures of self-control may not be suitable for accurately capturing levels of self-control, especially for those that have a very low self-control.

Gottfredson and Hirschi's (1993) idea was explored in the current dissertation by subjecting Grasmick et al.'s scale to a Rasch model and then examining the Item Characteristic Curve. The ICC found that male offenders having very low self-control (or high ability scores) were more likely to have average item responses that were unexpectedly lower than what the Rasch model had predicted. Remember that the Rasch model predicts what a person's score should be on a measuring instrument

given their level of ability on the trait being measured. In this case the observed data matched the predictions of the Rasch model with the exception of those having high ability had unexpectedly lower average scores on Grasmick et al.'s scale items. Generally, the current findings echoed what Piquero et al (2000) found when they administered Grasmick et al.'s scale to a sample of college students, ability affected survey responses to Grasmick et al.'s self-control scale.

These findings led to several questions. First, what should be done to more accurately capture responses for individuals that have very low self-control? Hirschi and Gottfredson (1993) have argued that a solution would be to collect behavioral measures of self-control. While they state that such indicators would be preferred, they suggested that behavioral indicators collected independent of respondents will be most successful, as self-report behavioral indicators can also be influenced by low self-control. Personally acknowledging that such a solution could take sunstantial resources and be a financial burden to researchers that do not have funds to collect direct observational data, an alternative to Hirschi and Gottfredson's solution would be to employ methods that would enhance the accuracy of survey responses from those having lower levels of self-control. In doing so, researchers could consider the nature of self-control and how it may guide item creation and scale formatting. For example, efforts could be made to construct items that are clear and not confusing for individuals that already prefer physical tasks over more cognitive/mental tasks such as reading. Also, another line of exploration would be to construct scales that are not overly lengthy, as impulsive people may not have much patience in completing tasks that take relatively longer periods of time to complete.

Second, research should focus on the mechanisms that influence lower self-control individuals to respond to survey statements or questions in a less valid manner. Piquero and colleagues (2000) argued that numerous factors may be driving such responses that include: cognitive factors, reading ability, social learning factors, or social structural differences. While they list these factors as potential explanations they did not articulate how such variables would differentially apply to low and high self-control individuals in a way that would affect survey responses. Another explanation could be that the very nature of what is being measured by the instrument could influence responses to items. Specifically, individuals that are very impulsive, prefer simple tasks, largely self-centered, and would be unlikely to delay gratification could be less likely to give accurate responses to a self-control instrument or any instrument for that matter. For example, such a respondent could be thinking short term in that they want to finish filling out the questionnaire as quickly as possible so they can move to a more exciting task or anything other than a task that requires them to use cognitive skills to respond. If so, they may quickly respond without truly giving much thought to the questions being asked. Furthermore, since such individuals are not concerned with long-term outcomes and are highly self-centered individuals they will likely have problems with mentally calculating how thoughtful responses to questions may benefit others who are conducting research.

Are Grasmick et al.'s scale items invariant across racial groups?

Few studies have assessed the invariant properties of Grasmick et al.'s self-control scale across groups that vary by demographic characteristics (Arneklev et al., 1999; Longshore et al., 1996; Piquero and Rosay, 1998; Piquero et al. 2000;

Vazsonyi et al., 2001). The current study was the first to thoroughly investigate scale invariance across racial groups using multiple methods. While it was anticipated that racial groups should differ on mean levels of self-control, the factor structure and items of Grasmick et al.'s scale should both be invariant for valid measurement to occur.

As discussed, both multidimensional factor structures for Grasmick et al.'s scale were shown to be invariant across black and white offender samples. Also, a traditional confirmatory factor analysis was not able to shed light on item bias across racial groups. Assessing item bias was important for two primary reasons. First, if items were biased then mean level comparisons of scores across different racial groups would be questionable. It could appear that racial differences exist, but in reality these differences were due to the differential likelihood of agreeing to items for some groups. Second, a biased measuring instrument will not measure different groups equally.

Results from a DIF analysis of Grasmick et al.'s self-control items showed that several items were biased across racial groups of offenders. Particularly, while controlling for level of ability, white offenders were more likely to give agreeable responses, i.e., responses that indicated lower self-control, to several items than black offenders. In turn, black offenders found it more difficult to give agreeable responses to these items. When black and white offenders with similar levels of self-control were confronted with the same scale item the black offender was less likely than the white offender to give an agreeable response. This was the case for 10 of the 24 items. Particularly, two Impulsivity items (i.e., I1 "I often act on the spur of the moment

without stopping to think." and I2 "I don't devote much thought and effort to preparing for the future."), two Simple Task items (S1 "I frequently try to avoid projects that I know will be difficult." and S3 "The things in life that are easiest to do bring me the most pleasure"), three Risk Seeking items (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" and R4 "Excitement and adventure are more important to me than security."), one Physical Tasks item (i.e., P4 "I seem to have more energy and a greater need for activity than most other people my age."), one Self-Centeredness item (i.e., Sc1 "I try to look out for myself first, even if it means making things difficult for other people"), and one Temper item (i.e., T4 "When I have a serious disagreement with someone, it's usually hard for me to talk calmly about it without getting upset.") from the Grasmick et al scale showed DIF. It was more difficult for black offenders to agree on all 10 items that showed DIF.

Results from the DIF showed that scale items are biased and, consequently, racial group comparisons for the Grasmick et al. measure will not necessarily reveal meaningful results. In other words, the finding that white offenders, on average, had lower self-control than black offenders was likely due to item bias and not truly meaningful differences. In addition, the findings that black offenders, on average, were lower in impulsivity and risk seeking are also likely due to item bias. Given these results, observed mean differences in scores for white and black offenders could be explained by the fact that several of Grasmick et al.'s self-control items were biased across racial groups.

Once again, another strike against the validity of Grasmick et al.'s scale has been uncovered, as a valid instrument should measure white and black offenders equally. This is the beginning of a potentially larger problem with the Grasmick et al. measure. That is, items may be biased across different groups such as sex, age, etc. This is an empirical question that could be answered using a Rasch model. Researchers should revisit their data and explore this potential problem when using this scale.

Limitations and Future Research

Although this dissertation has replicated and provided new psychometric information on Grasmick et al.'s self-control measure, it is not without limitations. Identification of these limitations will provide a starting point for future research. The discussion in this section centers on two important themes: specific limitations to the current investigation of Grasmick et al.'s self-control instrument and general limitations to using this instrument to measure self-control. From this, an agenda will emerge for future research on Grasmick et al.'s measure and the measurement of self-control in general.

Reliability and Grasmick et al.'s Scale: What Should Be Done Next?

One primary goal of this dissertation was to assess the reliability of Grasmick et al.'s self-control instrument. In doing so, results showed that this measure generally has good internal consistency. These results confirmed what other studies have found using the same method, Cronbach's alpha. Although replication was achieved, it could be argued that Cronbach's alpha does not discern repeatability of a measuring instrument. Specifically, it only estimates the correlation among items and

gives an estimate of the expected magnitude of the correlation between a measure and an alternative form of the measure. Given that only one time of measurement was required to estimate reliability with Cronbachs's alpha, actual repeatability of the measure was not assessed. Repeatability could not be assessed, as multiple measurements using Grasmick et al.'s instrument were not taken on the offender sample used in this study.

Multiple methods of reliability assessment were not used in the current study to confirm the reliability of Grasmick et al.'s measure nor have they been used in other studies that have assessed this measure's reliability. Therefore, the reliability of Grasmick et al.'s measure could still be questioned. Future research should devote empirical attention to this issue. For example, one starting point would be to take several measurements on a sample or samples (e.g., offenders, college students, adolescents) using Grasmick et al.'s instrument. At the least, two times of measurement would be required while varying time periods between measurements. Measurements could be taken in short intervals (e.g., 2 to 3 week interval), as well as, long intervals (e.g., 1 to 5 year intervals) for the same sample of people. Varying the time intervals between measurements would serve two purposes. First, from a reliability perspective, a short time interval could be used to capture repeatability or consistency of measurement in a way that is different from Cronbach alpha's approach. If scores were highly correlated between time 1 and time 2 this would reveal more evidence for the reliability of Grasmick et al.'s scale. Given that Grasmick et al.'s instrument was found to measure six traits, it would be worthwhile to do the same for each of these components, e.g., Impulsivity, Risk Seeking, etc.

Second, from a theoretical perspective, a longer interval between measurements could be used to test the stability of the six elements being measured by Grasmick et al.'s scale. This investigation would be important not only for reliability purposes but also for validity. That is, Gottfredson and Hirschi (1990) indicated that self-control (or lack of) is highly stable over long periods of time. If their theory was correct, then a valid measuring instrument should exhibit stability in scores for a group of individuals over a relatively long time span.

Revisiting and Replication

A sample of incarcerated male offenders was used in the current study to assess the psychometric properties of Grasmick et al.'s self-control measure. Although using such a sample, as discussed, had several benefits for the purpose of measurement validation, it also placed restrictions on generalizations that could be made regarding the measure's reliability and validity. Future research may want to replicate the analyses performed in this dissertation with different samples, especially with other offender samples, as other offender populations could be quite different from the one studied herein. In addition, researchers that have used this scale to measure self-control, especially those that have not performed extensive psychometric analyses, are advised to revisit their data to take a closer look at its properties. Besides the current dissertation, only one published study has used a Rasch model to investigate the psychometric properties of Grasmick et al.'s measure and both found consistent evidence that did not support the idea of summing responses to the 24 items to measure one thing. Actually, both studies found that Grasmick et al.'s instrument would be better suited as a unidimensional measure by

dropping several of the misfitting items from the scale. Researchers are encouraged to use multiple techniques to investigate the internal structure of the Grasmick et al. scale and for that matter, all multiple item scales used in their research. This would be important because different methods, such as a Rasch model, may reveal important information that other methods do not.

Limits Placed on Validity: What else can be done?

A primary goal of this dissertation was to investigate the validity of Grasmick et al.'s self-control instrument. As such, an intensive effort was made to empirically assess the internal structure of the instrument and its invariance across two racial groups. To a lesser degree, this dissertation has investigated the cross structure validity of the scale by assessing differences across racial groups. As discussed in Chapter two, construct validation has three components: face validity, internal structure validity, and cross structure validity. This dissertation has provided a thorough empirical assessment of Grasmick et al.'s measure using one of these components, internal structure validity. Consequently, the current study was limited in that it did not assess this measures face validity nor was its cross structure validity thoroughly investigated. These two limitations are discussed next and directions for future research are put forth.

Challenges to the Face Validity of Grasmick et al.'s Scale

In chapter three, Gottfredson and Hirschi's concept of self-control was discussed and Grasmick et al.'s (1993) process of moving from conceptualizing to operationalizing self-control was detailed. This dissertation did not attempt to challenge the face validity of Grasmick et al.'s operationalization. Consequently,

face validity was assumed. Many studies have taken the same approach by assuming Grasmick et al.'s operationalization is a valid reflection of Gottfredson and Hirschi's concept of self-control (Delisi et al., 2003; Longshore et al., 1996; Piquero and Rosay, 1998). Furthermore, these same studies that have performed internal structure analyses on Grasmick et al.'s scale have concluded that self-control is either unidimensional or multidimensional. This could be completely inaccurate conclusions that have been drawn. The reason for this is that items used to measure self-control by Grasmick and colleagues (1993) may not reflect the definition of self-control that was intended by Gottfredson and Hirschi (1990).

Recently, Marcus (2004) has questioned the face validity of Grasmick et al.'s scale based on what he believes most accurately reflects Gottfredson and Hirschi's definition of self-control. Opposed to the idea that six elements define the construct of self-control, Marcus (2004) argued that Gottfredson and Hirschi's (1994: 3) construct is simply "the tendency to avoid acts whose long-term costs exceed their monetary advantages." In light of this purely behavioral definition of self-control, Marcus (2004) stated that researchers have continued to use the Gramick et al. scale that is based on self-reflective attitudes that measure a set of heterogeneous traits.

Marcus (2004) attacks the face validity of Grasmick et al.'s scale on several fronts. First, Marcus stated that Grasmick et al.'s 24 items reflect six elements or personality traits discussed by Gottfredson and Hirschi and these traits overlap with a large body of literature on the structure of personality. In fact, he argued that each of the six elements measured by this scale is linked to the Five Factor Model (FFM) of personality (see Digman, 1990) that includes neuroticism, extraversion, openness to

experience, agreeableness, and concientiousness. Ultimately, this is an empirical question. That is, if Grasmick et al.'s scale is a unique measure of Gottfredson and Hirschi's construct of self-control, and not merely a measure that is tapping several personality traits that overlap with personality measures, then it should show divergent validity when compared to a widely used and psychometrically sound instrument of the FFM. Future research could explore this idea by empirically investigating the correlations between Grasmick et al.'s subscales and the most widely used, and most well-researched, instrument to measure the FFM of personality, the NEO-personality inventory which consists of 240 Likert items where 48 items are used to measure each of the five factors of the FFM (Block, 1995; Costa and McCrae, 1992).

Adding to Marcus's observations, the traits being measured by Grasmick et al.'s scale appear to also overlap with symptoms of disruptive disorders and other constructs that have been identified in psychology. For example, the three elements of Attention Deficit Hyperactivity Disorder (ADHD) are impulsivity, inattention, and hyperactivity. Russell Barkley (1997), a renowned researcher and clinician on the topic of ADHD, has put forth a theory that integrates the constructs of self-control and ADHD. Some have also argued that self-control overlaps with psychopathy (Wiebe, 2003). Specifically, Wiebe (2003:299) points out that psychopathy is made up of several similar traits that are represented in Grasmick et al.'s scale including self-centeredness, self-gratification, and a lack of concern for others. In sum, it remains to be seen if Grasmick et al.'s instrument is uniquely measuring self-control or if it represents a set of indicators that measure several aspects of personality or

other psychological variables. For validation purposes, future research is needed on the linkage between Grasmick et al.'s scale and personality measures.

Marcus (2004) also argued that the six elements measured by Grasmick et al.'s scale are inconsistent with the idea of control, an essential feature of self-control theory. Regardless of whether Grasmick et al.'s scale is unidimensional or multidimensional, he argued that elements such as risk-seeking and preference for easy tasks measured by this instrument reflect a motivational basis of behavioral choice and do not reflect behavioral constraint. Given this conceptual flaw, he argued that Grasmick et al.'s measure, and those similar to it, is incompatible with the main premise of self-control theory. That is, the lack of concern for long term consequences explains involvement in crime and not differential attractiveness to short-term goals (Marcus, 2004).

Similar to Marcus's line of reasoning, the six elements measured by Grasmick et al.'s scale are also inconsistent with definitions of self-control in psychological research. For example, Barkley (1997: 51) defined self-control as "any response or chain of responses by the individual which serve to alter the probability of their subsequent response to an event and, in so doing, function to alter the probability of later consequences related to that event." Moreover, Barkley (1997: 52) argues that, "for self-control to occur, the individual must have developed a preference for the long term over the short term outcomes of behavior." Barkley's definition appears to be consistent with Marcus's interpretation of Gottfredson and Hirschi's concept of self-control. Furthermore, according to Marcus (2004), operationalizing self-control with a set of attitudinal indicators is, by definition, not based on the evaluation of

behavior which the theory was deduced. In sum, Marcus (2004) makes a strong argument against the face validity of Grasmick et al.'s measure and has opened up avenues of research on its measurement.

The above discussion of face validity has potentially altered the meaning of Grasmick et al.'s scale and what it could potentially be measuring. Researchers are encouraged to create behavioral measures of self-control that are based on the above ideas and empirically compare them to Grasmick et al.'s measure to assess divergent and convergent validity. Some have already done comparisons by assessing the predictive validity of behavioral versus attitudinal measures, e.g., Grasmick et al.'s scale, of self-control. For example, Tittle and colleagues (2003: 362-363) found that, "behaviorally based scales of self-control produce no advantage over attitudinal, cognitive based ones in the prediction of deviant/criminal behavior." Furthermore, Marcus (2004) has created a self-control measure that consists of 67 strictly behavioral statements that are designed to assess the frequency of prior conduct that have long-term negative consequences. His measure has not been implemented in the mainstream of criminological research testing self-control theory. Future research should not only investigate how Marcus's scale correlates with Grasmick et al.'s measure, but should assess the predictive validity of his measure and compare it to the predictive validity of Grasmick et al.'s measure. As pointed out by Arneklev et al. (1999), the Grasmick et al. scale is only one step towards a measure of self-control in criminology; therefore, others are encouraged to develop alternative measures of this construct.

Limitations on the Cross-Structure Validity of Grasmick et al.'s Scale

Cross structure validation was not a major empirical goal of this dissertation. That is, Grasmick et al.'s self-control scale was not used to predict outcomes that Gottfredson and Hirschi suggested could be explained by low self-control. This is a limitation of the current study. If Grasmick et al.'s scale is measuring self-control then it should be able to empirically explain variation in outcomes that are proposed by self-control theory. While their measure has been shown to predict several theoretically derived variables in past research, this is yet to be determined with the sample of incarcerated male offenders used in this dissertation. Future research using these data could investigate how each dimension of Grasmick et al.'s scale predicts involvement and frequency of participation in different deviant and criminal behaviors. Some have already found that not every dimension of the scale is important in explaining criminal outcomes (Delisi et al., 2003). Others have found particular dimensions of Grasmick et al.'s scale, e.g., Impulsivity and Risk Seeking, to be more important than the overall self-control scale when predicting criminal outcomes (Longshore et al., 1996).

As discussed in earlier chapters, many studies have established the link between Grasmick et al.'s self-control scale and deviant/criminal outcomes, showing that this measure has good cross structure validity. Similar to past research, it is likely to be the case that Grasmick et al.'s measure is correlated with deviant and criminal outcomes for the sample of offenders used in this study, but this still remains an empirical question. Even if cross structure validity was achieved, as has been done

in past studies, it would not necessarily mean that the scale has achieved more validity as a measure of self-control. It could potentially mean that a number of personality traits are important predictors of crime. This is because the definition of self-control and face validity of Grasmick et al.'s instrument has been questioned in this section.

Although not exhaustive, this section has laid out a reasonable research agenda derived from limitations of the current study. Hopefully many of the un-addressed analyses identified above will be pursued in years to come, especially as criminologists become more concerned with measurement problems facing their discipline. On this score, it's suggested that even though criminologists should continue to investigate the psychometric properties of Grasmick et al.'s self-control scale they are encouraged to engage in rigorous psychometric analyses of other measuring instruments commonly used in their discipline. It is believed that such pursuits of measurement reliability and validity will lead to advances in producing sound, criminological research.

REFERENCES

Agnew, Robert. 1992. Foundation for a general strain theory of crime and delinquency. Criminology 30: 47-87.

Akers, Ronald J. 1991. Self-control as a general theory of crime. Journal of Quantitative Criminology 7: 201-211.

American Psychological Association. 1985. Standards for Educational and Psychological Testing. Washington, DC: Author.

Arneklev, Bruce J., John K. Cochran, and Randy R. Gainey. 1998. Testing Gottfredson and Hirschi's low self-control stability hypothesis: An exploratory study. American Journal of Criminal Justice 23: 107-127.

Arneklev, Bruce J., Harold G. Grasmick, Charles R. Tittle, and Robert J. Bursik. 1993. Low self-control and imprudent behaviors. Journal of Quantitative Criminology 9: 225-247.

Arneklev, Bruce J., Harold G. Grasmick, and Robert J. Bursik. 1999. Evaluating the dimensionality and invariance of "low self-control." Journal of Quantitative Criminology 15: 307-331.

Andrich, David. 1978. Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement 38: 665-680.

Andrich, David. 1988. Rasch Models for Measurement. Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills: Sage.

Barkley, Russell A. 1997. ADHD and the Nature of Self-Control. New York, New York: Guilford Press.

Bendixen, Mons and Dan Olweus. 1999. Measurement of antisocial behavior in early adolescence and adolescence: Psychometric properties and substantive findings. Criminal Behaviour and Mental Health 9: 323-54.

Bentler, Peter M. 1990. Comparative fit indexes in structural equation models. Psychological Bulletin 107: 238-246.

Bentler, Peter and Douglas G. Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin 88: 588-606.

Bentler, Peter M. and Chih-Ping Chou. 1987. Practical issues in structural equation modeling. Sociological Methods and Research 68: 78-117.

Binet Alfred and T. Simon. 1911. A Method of Measuring the Development of the Intelligence of Young Children. Lincoln, Illinois: Courier Company.

Blalock, Hubert M. 1968. The measurement problem. In Hubert M. Blalock and A. Blalock (eds.). Methodology in Social Science Research. New York: McGraw- Hill.

Blalock, Hubert M. 1982. Conceptualization and Measurement in the Social Sciences. Beverly Hills: Sage

Blalock, Hubert M. 1984. Basic Dilemmas in the Social Sciences. Beverly Hills: Sage.

Block, Jack. 1995. A contrarian view on the five-factor approach to personality description. Psychological Bulletin 114: 187-215.

Bollen, Kenneth A. 1989. Structural Equations with Latent Variables. New York: John Wiley and Sons, Inc.

Bollen, Kenneth A. and Kenney H. Barb. 1981. Pearson's R and Coarsely Categorized Measures. American Sociological Review 46: 232-239.

Bond, Trevor G. and Christine M. Fox. 2001. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Browne, Michael. W and Robert Cudeck. 1993. Alternative ways of assessing model fit. In Kenneth Bllen and Scott Long (eds.). Testing structural equation modeling (pp. 445-455). Newbury Park, CA: Sage.

Brownfield, David and Ann Marie Sorenson. 1993. Self-control and juvenile delinquency: Theoretical issues and an empirical assessment of selected elements of a general theory of crime. Deviant Behavior 14: 243-264.

Burton, Velmer, S., Francis T. Cullen, David T. Evans, Leanne Fiftal Alarid, and Gregory R. Dunaway. 1998. Gender, self-control, and crime. Journal of Research in Crime and Delinquency 35: 123-147.

Byrne, Barbara M. 2001. Structural equation modeling with AMOS: Basic concepts, applications, and programming. Mahwah, NJ: Lawernce Earlbaum Associates.

Carmines, Edward G. and Richard A. Zeller. 1979. Reliability and Validity Assessment. Thousand Oaks: Sage.

Caspi, Avshalom T., Terrie E. Moffitt, A. Thornton, D. Freedman, J.W. Amell, H. Harrington, J. Smijers, and P. A. Silva. 1996. The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. International Journal of Methods in Psychiatric Research 6: 101-114.

Cattell, Raymond.B. 1966. The scree test for the number of factors. Multivariate Behavioral Research 1: 245-276.

Clark, H.H. and M.F. Schober. 1992. Asking questions and influencing answers. In J.M. Tanur (Ed.), Questions about Questions: Inquires into the Cognitive Bases of Surveys. New York: Russell Sage.

Cochran, John K., Peter B. Wood, Christine S. Sellers, Wendy Wilkerson, and Mitchell B. Chamlin. 1998. Academic dishonesty and low self-control: An empirical test of a general theory of crime. Deviant Behavior 19: 227-255.

Costa, Paul T., and Robert R. McCrae. 1992. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Odessa, Florida: Psychological Assessment Resources.

Cronbach, Lee J. 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16: 297-334.

Cronbach, Lee J. 1971. Test validation. In Edward L. Thorndike (ed.). Educational Measurement. Washington, DC: American Council on Education.

Cronbach, Lee J. and Paul E. Meehl. 1955. Construct validity in psychological tests. Psychological Bulletin 52: 281-302.

Darwin, Charles. 1871. The Descent of Man. London: John Murray.

Delisi, Matt, Andy Hochstetler, and Daniel S. Murphy. 2003. Justice Quarterly 20: 241-263.

DeVellis, Robert F. 1991. Scale Development. Newbury Park, California: Sage Publications, Inc.

Digman, John M. 1990. Personality structure: Emergence of the five factor model. Annual Review of Psychology 41: 417-440.

Dolna DNA learning Center, Cold Spring Harbor Laboratory. Image Archive on the American Eugenics Movement Retrieve July 2003, from http://www.eugenicsarchive.org/

Duncan, Otis Dudley. 1984. Note on Social Measurement: Historical and Critical. New York: Russell Sage Foundation.

Dunnette, M.D. and W.C. Borman. 1979. Personnel selection and classification systems. Annual Review of Psychology 30: 477-525.

Durkheim, Emile. 1938. The Rules of Sociological Method. New York, New York: The Free Press.

Elliot, Delbert S. and Suzanne S. Ageton. 1980. Reconciling race and class differences in self-reported and official estimates of delinquency. American Sociological Review 45: 95-110.

Elliot, Delbert S., David Huizinga, and Suzanne S. Ageton. 1985. Explaining Delinquency and Drug Use. Beverley Hills, California: Sage Publication.

Elliot, Delbert S., David Huizinga, and Scott Menard. 1989. Multiple Problem Youth: Delinquency, Substance Abuse, and Metnal Health Problems. New York: Springer-Verlag.

Esbensen, Finn-Age and Wayne D. Osgood. 1999. Gang resistance education and training (GREAT): Results from the national evaluation. Journal of Research in Crime and Delinquency 36: 194-225.

Evans, David T., Francis T. Cullen, Velmer S. Burton, Gregory R. Dunaway, and Michael L. Benson. 1997. The social consequences of self-control: Testing the general theory of crime. Criminology 35: 475-504.

Farrington, David P., Rolf Loeber, Magda Stouthamer-Loeber, Welmoet B. Van Kammen, and Laura Schmidt. 1996. Self-reported delinquency and a combined delinquency seriousness scale based on boys, mothers, and teachers: Concurrent and predictive validity for African-Americans and Caucasions. Criminology 34: 493-517.

Fordam University, Department of Psychometrics. Retrieved July 2003, from http://www.fordham.edu/aps/whatpsy.html

Forde, David R. and Leslie W. Kennedy. 1997. Risky lifestyles, routine activities, and the general theory of crime. Justice Quarterly 14: 265-294.

Fox, Christine M. 1994. A practical knowledge instrument: Psychometric characteristics and validity of an instrument for nurses. Dissertation Abstracts International, 56(06), 2210A.

Fox, Christine M. and James A. Jones. 1998. Uses of Rasch modeling in counseling psychology research. Journal of Counseling Psychology 45: 30-45.

Gardner, Robert C. 2001. Psychological Statistics Using SPSS for Windows. Upper Saddle River, New Jersey: Prentice Hall.

Geis, Gilbert. 2000. On the absence of self-control as the basis for a general theory of crime: a critique. Theoretical Criminology 4: 35-53.

Gibbons, Don C. 1979. The Criminological Enterprise: Theories and Perspectives. Upper Saddle River, New Jersey: Prentice Hall.

Gibbs, John J. and Dennis Giever. 1995. Self-control and its manifestations among university students: An empirical test of Gottfredson and Hirschi's general theory. Justice Quarterly 12: 231-255.

Gibbs, John J., Dennis Giever, and Jamie S. Martin. 1998. Parental management and self-control: An empirical test of Gottfredson and Hirschi's general theory. Journal of Research on Crime and Delinquency 35: 40-70.

Gibson, Chris and John Wright. 2001. Low self-control and co-worker delinquency: A research note. Journal of Criminal Justice 29: 483-492.

Gibson, Chris, John Wright, and Stephen Tibbetts. 2000. Testing the generality of the general theory of crime: The effects of low self-control on social development. Journal of Crime and Justice 23:109-134.

Gibson, Chris, Jihong Zhao, and Nicholas P. Lovrich. 2002. Sociological measurement confusion, paradigmatic imperfection, and etiological nirvana: Striking a pragmatic balance in pursuing science. Justice Quarterly 19: 793-808.

Gilbert, J. A. 1894. Researches on the mental and physical development of school children. Studies of the Yale Psychological Laboratory 2: 40-100.

Gottfredson, Michael R. and Travis Hirschi. 1990. A General Theory of Crime. Stanford, California: Stanford University Press.

Gould, Stephen Jay. 1996. The Mismeasure of Man. New York, New York: W.W. Norton and Company, Inc.

Grasmick, Harold G., Charles R. Tittle, Robert J. Bursik, and Bruce J. Arneklev. 1993. Testing the core empirical implications of Gottfredson and Hirschi's general theory of crime. Journal of Research in Crime and Delinquency 30: 5-29.

Green, K.E. 1996. Applications of the Rasch model to evaluation of survey data quality. New Directions for Evaluation 70: 81-92.

Guion, Robert M. 1980. On trinitarian doctrines of validity. Professional Psychology 11: 385-398.

Gulliksen, Harold. 1950. Theory of Mental Tests. New York: Wiley.

Hayduk L. A. 1987. Structural equation modeling with LISREL: Essentials and advances. Baltimore, MD: Johns Hopkins Press.

Hambleton, Ronald K., H. Swaminathan, and H. Jane Rogers. 1991. Fundamentals of Item Response Theory. Newbury Park, CA: Sage.

Hay, Carter. 2001. Parenting, self-control, and delinquency: A test of self-control theory. Criminology 39: 707-736.

Hayduk, Leslie A. 1987. Structural equation modeling with LISREL: Essential and advances. Baltimore, MD: Johns Hopkins University Press.

Hickman, Matthew, J., Nichole L. Piquero, and Alex R. Piquero. 2004. The validity of Niederhoffer's cynicism scale. Journal of Criminal Justice 32:1-13.

Hindelang, Michael J., Travis Hirschi, and Joseph G. Weis. 1981. Measuring Delinquency. Beverly Hills, CA: Sage.

Hirschi, Travis. 1969. Causes of Delinquency. California: University of California Press, Berkley.

Hirschi, Travis and Michael R. Gottfredson. 1993. Commentary: Testing the general theory of crime. Journal of Research in Crime and Delinquency. 30: 47-54.

Hirschi, Travis and Michael R. Gottfredson. 1994. The generality of deviance. In Travis Hirschi and Michael R. Gottfredson (eds.), The Generality of Deviance. New Brunswick, New Jersey: Transaction Publishers.

Horney, Julie. 2001. Criminal events and criminal careers: An integrative approach to the study of violence. In Robert F. Meier, Leslie W. Kennedy, and Vincent F. Sacco (eds.), The Process and Structure of Crime: Criminal Events and Crime Analysis. New Brunswick, New Jersey: Transaction Publishers.

Huizinga, David and Delbert Elliot. 1986. Reassessing the reliability and validity of self-report delinquency measures. Journal of Quantitative Criminology 2: 293-327.

Institute for Objective Measurement, Inc. 2000. Definition of measurement. Retrieved July 2000, from http://www.rasch.org/define.htm

Joreskog, Karl G. and D. Sorbom. 1989. LISREL 7: User's Reference Guide. Mooresville, Indiana: Scientific Software.

Junger, Marianne and Richard E. Tremblay. 1999. Self-control, accidents, and crime. Criminal Justice and behavior 26: 485-501.

Junger-Tas, Josine and Ineke Haen Marshall. 1999. The self-report methodology in crime research. In Michael Tonry (ed.). Crime and Justice 25. Chicago, Illinois: University of Chicago Press.

Keane, Carl, Paul S. Maxim, and Jame J. Teevan. 1993. Drinking and driving, self-control, and gender: Testing the general theory of crime. Journal of Research in Crime and Delinquency 30: 30-46.

Kline, Rex B. 1998. Principles and practice of structural equation modeling. New York, New York: The Guilford Press.

King, John A. and Trevor Bond. 1996. A Rasch analysis of a measure of computer anxiety. Journal of Educational Computing Research 14: 49-65.

King, John A. and Trevor G. Bond. 2003. Measuring client satisfaction with public education I: Meeting competing demand in establishing state-wide benchmarks. Journal of Applied Measurement 4: 111-123.

Kindlon, D. J., Benjamin D. Wright, Stephen W. Raudenbush, and Felton Earls. 1996. The measurement of children's exposure to violence: A Rasch analysis. International Journal of Methods in Psychiatric Research 6: 187-194.

Kline, Rex B. 1998. Principles and Practices of Structural Equation Modeling. New York: The Guilford Press.

LaGrange, Teresa C. and Robert A. Silverman. 1999. Low self-control and opportunity: Testing the general theory of crime as an explanation for gender differences in delinquency. Criminology 37: 41-72.

Linacre, Mike J. 1999. Investigating rating scale category utility. Journal of Outcome Measurement 3: 103-122.

Linacre, Mike J. 1995. Categorical misfit statistics. Rasch Measurement Transaction 9: 450-451.

Linacre, Mike J. and Benjamin D. Wright. (1999-2001). WINSTEPS. University of Chicago: MESA press.

Loeber, Rolf, David P. Farrington, Magda Stouthamer-Loeber, Terrie E. Moffitt, and Avshalom Caspi. 1998. The development of male offending: Key findings

from the first decade of the Pittsburgh Youth Study. Studies on Crime and Crime Prevention 7: 141-71.

Loevinger, Jane. 1957. Objective tests as instruments of psychological theory. Psychological Reports 3: 635-694.

Longshore, Douglas, Susan Turner, and Judith A. Stein. 1996. Self-control in a criminal sample: An examination of construct validity. Criminology 34: 209-228.

Longshore, Douglas and Susan Turner. 1998. Reliability and validity of a self-control measure: Rejoinder. Criminology 36: 175-182.

Lord, F.M. and M. R. Novick. 1968. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley.

Luce, R. Duncan and John W. Tukey. 1964. Simultaneous conjoint measurement: A new kind of fundamental measurement. Journal of Mathematical Psychology 1: 1-27.

MacIntyre, P.D. 1990. Issues and recommendations in the use of factor analysis. The Western Journal of Graduate Research 2: 59-73.

Marcus, Bernd. 2004. Self-control in the general theory of crime: Theoretical implications of a measurement problem. Theoretical Criminology 8: 33-35.

Masters, Geoffery, N. 1997. Where has Rasch measurement proved effective? Rasch Measurement Transactions 11: 568.

Maxfield, Michael G. and Earl Babbie. Research Methods for Criminal Justice and Criminology. (3rd ed.). Belmont, California: Wadsworth.

McDonald, Roderick P. 1999. Test Theory: A Unified Treatment. Mahwah, New Jersey: Lawrence Erlbaum Associates.

McRae, James A. 1991. Rasch measurement and differences between women and men in self-esteem. Social Science Research 20: 421-436.

Merton, Robert K. 1938. Social structure and anomie. American Sociological Review. 3: 672-682.

Moffitt, Terrie E., Avshalom Caspi, Nigel Dickson, Phil Silva, and Warren Stanton. 1996. Childhood-onset versus adolescent-onset antisocial conduct problems in males: Natural history from ages 3 to 18 years. Development and Psychopathology 8: 399-424.

Nagin, Daniel S. and Raymond Paternoster. 1993. Enduring individual differences and rational choice theories of crime. Law and Society Review 27: 467-496.

Nagin, Daniel S. and Raymond Paternoster. 2000. Population heterogeneity and state dependence: State of the evidence and directions for future research. Journal of Quantitative Criminology 16: 117-144.

Novick, Melvin R. and Charles Lewis. 1967. Coefficient alpha and composite measurements. Psychometrika 32: 1-13.

Nunnally, Jum C. 1978. Psychometric Theory. (2nd ed.). New York: McGraw-Hill.

Nunnally, Jum C. and Ira H. Bernstein. 1994. Psychometric Theory. (3rd ed.). New York: McGraw Hill.

Paternoster, Raymond and Robert Brame. 1998. The structural similarity of processes generating criminal and analogous behavior. Criminology 36: 633-70.

Pedhazur, Elazar J. Liora Pedhazur Schmelkin. 1991. Measurement, Design, and Analysis: An Integrated Approach. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Piquero, Alex R., Chris Gibson, and Stephen Tibbetts. 2002. Does low self-control account for the relationship between binge-drinking and alcohol-related behaviors: comparison across gender. Criminal Behavior and Mental Health 12: 135-154.

Piquero, Alex R., Randall MacIntosh, and Matthew Hickman. 2000. Does self-control affect survey response? Applying exploratory, confirmatory, and item response theory analysis to Grasmick et al.'s self-control scale. Criminology 38: 897-929.

Piquero, Alex R., Randall MacIntosh, and Matthew Hickman. 2002. The validity of a self-report delinquency scale: Comparison across gender, age, race. Sociological Methods and Research 30: 492-529.

Piquero, Alex R. and Andre B. Rosay. 1998. The reliability and validity of Grasmick etal.'s self-control scale: A comment on Longshore et al. Criminology 36: 157-173.

Piquero, Alex R. and Stephen Tibbetts. 1996. Specifying the direct and indirect effects of low self-control and situational factors in offender's decision making. Justice Quarterly 13: 481-510.

Polakowski, Michael. 1994. Linking self- and social control with deviance: Illuminating the structure underlying a general theory of crime and its relation to criminal activity. Journal of Quantitative Criminology. 10: 41-78.

Pratt, Travis C. and Francis T. Cullen. 2000. The empirical status of Gottfredson and Hirschi's general theory of crime: A meta-analysis. Criminology 38: 931-964.

Raine, Adrian. 1993. The Psychopathology of Crime: Criminal Behavior as a Clinical Disorder. San Diego, California: Academic Press.

Rasch, Georg. 1960. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.

Rasch, Georg. 1980. Probabilistic models for some intelligence and attainment tests. (Expanded ed.), Chicago: University of Chicago Press.

Raudenbush, Stephen, Chris Johnson, and Robert J. Sampson. 2003. A multivariate, multilevel Rasch model with application to self-report criminal behavior. Sociological Methodology 33:169-211.

Roberts, Jennifer. 2000. The person in context: The impact of early onset criminality on high risk for violence situations. (Doctoral dissertation, University of Nebraska at Omaha). Dissertation Abstracts International, UMI number: 9975583.

Rozeboom. 1966. Foundations of the Theory of Prediction. Homewood, Illinois: Dorsey Press.

Rust, John and Susan Golombok. 1999. Modern Psychometrics: The Science of Psychological Assessment. (2nd ed.). New York: Routledge.

Schreck, Christopher J. 1999. Criminal Victimization and low self-control: An extension and test of a general theory of crime. Justice Quarterly 16: 633-654.

Sellers, Christine S. 1999. Self-control and intimate violence: An examination of the scope and specification of the general theory of crime. Criminology 37: 375-404.

Short, James F. and F. I. Nye. 1957. Reported behavior as a criterion of deviant behavior. Social Problems 5: 207-213.

Short, James F. and F. I. Nye. 1958. Extent of unrecorded juvenile delinquency: Tentative conclusions. Journal of Criminal Law and Criminology 49: 296-302.

Spearman, Charles. 1904a. "General intelligence," objectively determined and measured. American Journal of Psychology. 15: 201-293.

Spearman, Charles. 1904b. The proof and measurement of association between two things. American Journal of Psychology 15: 201-293.

Stelmack, Joan, Janet P. Szlyk, Thomas Stelmack, Judith Babcock-Parziale, Paulette Demers-Turco, Tracy Williams, and Robert W. Massof. 2004. Use of Rasch person-item map in exploratory data analysis: A clinical example. Journal of Rehabilitation Research and Development 41: 233-242.

Stevens, J. 1996. Applied Multivariate Statistics for the Social Sciences (3rd ed.). Hillsdale, New Jersey: Lawrence Erlbaum.

Stewart, Eric A., Kirk W. Elifson, and Claire E. Sterk. 2004. Integrating the general theory of crime into an explanation of violent victimization among female offenders. Justice Quarterly 21: 159-181.

Stylianou, Stelios. 2002. The relationship between elements and manifestations of low self-control in a general theory of crime: Two comments and a test. Deviant Behavior 23: 531-557.

Sutherland, Edwin H. 1939. Principles of Criminology. (3rd ed.). Philadelphia: J. B> Lippincott.

Taub, Gordon E. 2001. A confirmatory analysis of the wechsler adult intelligence scale-third edition: is the verbal/performance discrepancy justified? Practical Assessment, Research & Evaluation 7

Terman, Lewis M. 1906. Genius and stupidity: A study of some of the intellectual processes of seven "bright" and seven "stupid" boys. Pedagogical Seminary 13: 307-373.

Terman, Lewis M. 1916. The Measurement of Intelligence. Boston, Houghton Mifflin.

Thornberry, Terence P. and Marvin D. Krohn. 2000. The self-report method for measuring delinquency and crime. In David Duffee (ed.). Measurement and Analysis of Crime and Justice. Washington, DC: National Institute of Justice.

Thurstone, Louis L. 1924. The Nature of Intelligence. London: Kegan Paul, Trench, Trubner and Company.

Thurstone, Louis L. 1947. Multiple-factor Analysis. Chicago: Chicago University Press.

Tittle, Charles R., David A. Ward, and Harold G. Grasmick. 2003. Self-control and crime/deviance: Cognitive vs. behavioral measures. Journal of Quantitative Criminology 19: 333-365.

Tremblay, Richard E., Bernard Boulerice, Louis Arsenault, and Marianne Junger. 1995. Does low self-control during childhood explain the association between delinquency and accidents in early adolescence. Criminal Behaviour and Mental Health 5: 439-451.

Trochim, William M.K. 2001. The Research Methods Knowledge Base. Cincinnati, Ohio: Atomic Dog Publishing.

Turner, Michael and Alex Piquero. 2002. The stability of self-control. Journal of Criminal Justice 30: 457-472.

Vazsonyi, Alexander T., and Jennifer M. Crosswhite. 2004. A test of Gottfredson and Hirschi's general theory of crime in African American adolescents. 41: 407-432.

Vazsonyi, Alexander T., Lloyd E. Pickering, Marianne Junger, and Dick Hessing. 2001. An empirical test of a general theory of crime: A four-nation comparative study of self-control and the prediction of deviance. Journal of Research in Crime and Delinquency 38: 91-131.

Velozo, Craig A. and Elizabeth W. Peterson. 2001. Developing meaningful fear of falling measures for community dwelling elderly. American Journal of Physical Medical Rehabilitation 80: 662-673.

Wells , William. 1999. The situational role of firearms in violent encounters. (Doctoral dissertation, University of Nebraska at Omaha). Dissertation Abstracts International, UMI number: 9955184.

West, Steven G., Finch, John F., and Patrick J. Curran. 1995. Structural equation models with nonnormal variables: Problems and remedies. In Rick H. Hoyle (ed.), Structural Equation Modeling: Concepts, Issues, and Applications. Thousand Oaks, California: Sage.

Wiebe, Richard P. 2003. Reconciling psychopathy and low self-control. Justice Quarterly 20: 297-331.

Winfree, Thomas L. and Frances P. Bernat. 1998. Social learning, self-control, and substance abuse by eighth grade students: A tale of two cities. Journal of Drug Issues 28: 539-558.

Wood, Peter B., Betty Pfefferbaum, and Bruce J. Arneklev. 1993. Risk-taking and self-control: Social psychological correlates of delinquency. Journal of Criminal Justice 16: 111-130.

Wright, Ben D. and Geofferey N. Masters. 1982. Rating Scale Analysis. Chicago, Illinois: MESA Press.

Wright, Ben D. and Mike J. Linacre. 1994. Reasonable mean-square fit values. Rasch Measurement Transactions 8: 370.

Wright, Ben. D. and Mark H. Stone. 1979. Best Test Design. Chicago, Illinois: MESA Press.

Wright, Bradley R., Avshalom Caspi, Terrie E. Moffitt, and Phil A. Silva. 1999. Low self-control, social bonds, and crime: Social causation, social selection, or both? Criminology 37: 175-194.

Zager, Mary A. 1994. Gender and crime. In Travis Hirschi and Michael R. Gottfredson (Eds.). The Generality of Deviance. New Brunswick, New Jersey: Transaction Publishers.

Appendix A. Frequency distributions of Grasmick et al.'s 24 self-control items

Variable	Strongly Disagree	Disagree	Agree	Strongly Agree
Impulsivity				
I1	121 (18.6%)	144 (22.1%)	269 (41.3%)	117 (18.0%)
I2	29 (4.5%)	274 (42.1%)	66 (10.1%)	29 (4.5%)
I3	124 (19.0%)	160 (24.6%)	298 (45.8%)	69 (10.6%)
I4	128 (19.7%)	151 (23.2%)	266 (40.9%)	106 (16.3%)
Simple Tasks				
S1	234 (35.9%)	232 (35.6%)	127 (19.5%)	58 (8.9%)
S2	334 (51.3%)	180 (27.6%)	97 (14.9%)	40 (6.1%)
S3	110 (16.9%)	191 (29.3%)	255 (39.2%)	95 (14.6%)
S4	239 (36.7%)	219 (33.6%)	141 (21.7%)	52 (8.0%)
Risk Seeking				
R1	69 (10.6%)	78 (12.0%)	331 (50.8%)	173 (26.6%)
R2	209 (32.1%)	121 (18.6%)	227 (34.9%)	94 (14.4%)
R3	301 (46.2%)	125 (19.2%)	170 (26.1%)	55 (8.4%)
R4	249 (38.2%)	204 (31.3%)	141 (21.7%)	57 (8.8%)
Physical Activities				
P1	95 (14.6%)	201 (30.9%)	234 (35.9%)	121 (18.6%)
P2	24 (3.7%)	89 (13.7%)	266 (40.9%)	272 (41.8%)
P3	23 (3.5%)	140 (21.5%)	259 (39.8%)	229 (35.2%)
P4	28 (4.3%)	141 (21.7%)	309 (47.5%)	173 (26.6%)
Self-Centered				
Sc1	184 (28.3%)	236 (36.3%)	159 (24.4%)	72 (11.1%)
Sc2	314 (48.2%)	228 (35.0%)	79 (12.1%)	30 (4.6%)
Sc3	311 (47.8%)	235 (36.1%)	85 (13.1%)	20 (3.1%)
Sc4	272 (41.8%)	231 (35.5%)	127 (19.5%)	21 (3.2%)
Temper				
T1	250 (38.4%)	168 (25.8%)	152 (23.3%)	81 (12.1%)
T2	312 (47.9%)	206 (31.6%)	96 (14.7%)	37 (5.7%)
T3	221 (33.9%)	173 (26.6%)	160 (24.6%)	97 (14.9%)
T4	153 (23.5%)	205 (23.5%)	206 (31.6%)	87 (13.4%)

Appendix B. Univariate statistics for Grasmick et al.'s 24 self-control items.

Variable	Mean	SD	Min	Max
Impulsivity				
I1	1.59	.99	0	3
I2	.76	.81	0	3
I3	1.48	.92	0	3
I4	1.54	.98	0	3
Simple Tasks				
S1	1.01	.96	0	3
S2	.76	.92	0	3
S3	1.51	.94	0	3
S4	1.01	.95	0	3
Risk Seeking				
R1	1.93	.90	0	3
R2	1.32	1.07	0	3
R3	.97	1.03	0	3
R4	1.01	.96	0	3
Physical Activities				
P1	1.59	.95	0	3
P2	2.21	.81	0	3
P3	2.07	.84	0	3
P4	1.96	.81	0	3
Self-Centered				
Sc1	1.18	.97	0	3
Sc2	.73	.85	0	3
Sc3	.71	.81	0	3
Sc4	.84	.85	0	3
Temper				
T1	1.10	1.05	0	3
T2	.78	.90	0	3
T3	1.20	1.06	0	3
T4	1.35	.98	0	3

Appendix C. Pearson correlations for Grasmick et al.'s 24 self-control items.

Items	I1	I2	I3	I4	S1	S2	S3	S4	R1	R2	R3	R4
I1	1.00											
I2	.22*	1.00										
I3	.34*	.17*	1.00									
I4	.32*	.15*	.41*	1.00								
S1	.11*	.03	.16*	.29*	1.00							
S2	.16*	.10*	.21*	.32*	.59*	1.00						
S3	.29*	.09*	.30*	.30*	.37*	.41*	1.00					
S4	.17*	.05	.16*	.27*	.54*	.54*	.43*	1.00				
R1	.22*	.06	.25*	.20*	-.03	-.00	.17*	.03	1.00			
R2	.31*	.15*	.30*	.26*	.01	.05	.21*	.03	.54*	1.00		
R3	.32*	.23*	.30*	.28*	.15*	.17*	.27*	.16*	.37*	.64*	1.00	
R4	.26*	.14*	.27*	.32*	.19*	.20*	.26*	.19*	.29*	.50*	.47*	1.00
P1	.21*	.06	.12*	.15*	.12*	.07	.17*	.16*	.18*	.16*	.12*	.26*
P2	.20*	-.00	.12*	.18*	.08*	.09*	.20*	.15*	.17*	.22*	.20*	.19*
P3	.21*	.08*	.11*	.18*	.14*	.16*	.22*	.16*	.12*	.12*	.17*	.20*
P4	.15*	-.05	.08*	.13*	.03	.01	.09*	.04	.15*	.12*	.07	.13*
Sc1	.21*	.11*	.23*	.29*	.28*	.29*	.29*	.27*	.18*	.16*	.22*	.31*
Sc2	.21*	.23*	.20*	.20*	.23*	.25*	.21*	.28*	.13*	.15*	.22*	.27*
Sc3	.27*	.18*	.23*	.13*	.08*	.15*	.19*	.15*	.08*	.22*	.29*	.22*
Sc4	.32*	.15*	.30*	.22*	.18*	.17*	.34*	.18*	.20*	.30*	.39*	.34*
T1	.30*	.13*	.22*	.18*	.20*	.19*	.24*	.21*	.17*	.23*	.27*	.16*
T2	.31*	.17*	.23*	.26*	.19*	.21*	.30*	.25*	.29*	.29*	.37*	.29*
T3	.23*	.15*	.22*	.20*	.11*	.17*	.23*	.15*	.22*	.27*	.32*	.29*
T4	.25*	.16*	.23*	.21	.18*	.25*	.24*	.18*	.23*	.23*	.24*	.22*

Appendix C (cont'd). Pearson correlations for Grasmick et al.'s 24 self-control items.

Items	P1	P2	P3	P4	Sc1	Sc2	Sc3	Sc4	T1	T2	T3	T4
I1	.21*	.20*	.21*	.15*	.21*	.21*	.27*	.32*	.30*	.31*	.23*	.25*
I2	.06	-.00	.08*	-.05	.11*	.23*	.18*	.15*	.13*	.17*	.15*	.16*
I3	.12*	.12*	.11*	.08*	.22*	.20*	.23*	.30*	.22*	.23*	.22*	.23*
I4	.15*	.18*	.18*	.13*	.29*	.20*	.13*	.22*	.18*	.26*	.20*	.21*
S1	.12*	.08*	.14*	.03	.28*	.23*	.08*	.18*	.20*	.19*	.11*	.18*
S2	.07	.09*	.16*	.01	.29*	.25*	.15*	.17*	.19*	.21*	.17*	.25*
S3	.17*	.20*	.22*	.09*	.29*	.21*	.19*	.34*	.24*	.30*	.23*	.24*
S4	.16*	.15*	.16*	.04	.27*	.28*	.15*	.18*	.21*	.25*	.15*	.18*
R1	.18*	.17*	.12*	.15*	.18*	.13*	.08*	.20*	.17*	.23*	.22*	.23*
R2	.16*	.22*	.12*	.12*	.16*	.15*	.22*	.30*	.23*	.29*	.27*	.23*
R3	.12*	.20*	.17*	.07	.22*	.22*	.29*	.39*	.27*	.37*	.32*	.24*
R4	.26*	.19*	.20*	.13*	.31*	.27*	.22*	.34*	.16*	.29*	.29*	.22*
P1	1.00	.31*	.35*	.30*	.14*	.16*	.17*	.25*	.18*	.18*	.18*	.19*
P2		1.00	.48*	.30*	.10*	.05	.05	.15*	.16*	.14*	.15*	.15*
P3			1.00	.28*	.15*	.17*	.17*	.20*	.19*	.22*	.20*	.18*
P4				1.00	.11*	.06	.11*	.12*	.11*	.11*	.13*	.12*
Sc1					1.00	.41*	.35*	.47*	.17*	.31*	.26*	.23*
Sc2						1.00	.44*	.36*	.16*	.30*	.27*	.23*
Sc3							1.00	.53*	.17*	.29*	.28*	.24*
Sc4								1.00	.22*	.40*	.33*	.26*
T1									1.00	.51*	.48*	.51*
T2										1.00	.57*	.52*
T3											1.00	.57*
T4												1.00