Validation of the Crime and Violence Scale (CVS) Against the Rasch Measurement Model Including Differences by Gender, Race and Age

Kendon J. Conrad[1]

Barth B. Riley[2]

Karen M. Conrad[1,3]

Ya-Fen Chan[2]

Michael L. Dennis[2]

(1) University of Illinois at Chicago, Chicago, Illinois

(2) Chestnut Health Systems, Normal, Illinois

(3) Program Metrics, LLC, Oak Park, Illinois

Corresponding author:
Kendon J. Conrad, PhD, MSPH
Professor of Health Policy and Administration
University of Illinois at Chicago
School of Public Health (M/C 923)
1603 West Taylor Street
Chicago, IL 60612-4310
Phone: 312.996.3185
Fax: 312.996.5356
Email: kjconrad@uic.edu

Running head: Crime and Violence Scale

ACKNOWLEDGEMENT: This development of this paper was supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHSA) via Westat under contract 270-2003-00006 to Dr. Dennis at Chestnut Health Systems in Bloomington, Illinois using data provided by the following grants and contracts from CSAT (TI-11320, TI-11317, TI-11321, TI-11323, TI-11324, TI-11422, TI-11424, TI-11423, TI-11894, TI-11874, TI-11888, TI-11892, TI-11871, TI-13309, TI-13356, TI-13305, TI-13340, TI-13344, TI-13322, TI-13323, TI-13345, TI-13308, TI-13354, TI-13313, TI-14254, TI-14376, TI-14311, TI-14196, TI-14214, TI-14261, TI-14090, TI-14189, TI-14252, TI-14283, TI-14355, TI-14272, TI-14103, TI-14267, TI-14315, TI-14188, TI-14271, TI-15686, TI-15671, TI-15486, TI-15545, TI-15672, TI-15475, TI-15678, TI-15447, TI-15461, TI-15433, TI-15481, TI-15514, TI-15478, TI-15413, TI-15483, TI-15670, TI-15674, TI-15479, TI-15682, TI-15467, TI-15511, TI-15562, TI-13601, TI-13190, TI-12541, TI-00567; Contract 207-98-7047, Contract 277-00-6500), the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (R01 AA 10368), the National Institute on Drug Abuse (NIDA) (R37 DA11323; R01 DA 018183), the Illinois Criminal Justice Information Authority (95-DB-VX-0017), and the Illinois Office of Alcoholism and Substance Abuse (PI 00567). The opinions are those of the authors and do not reflect official positions of the contributing project directors or government. We appreciate the editorial support provided by Jessica Mazza.

Validation of the Crime and Violence Scale (CVS) against the Rasch Measurement Model Including Differences by Gender, Race and Age

Abstract

In assessing criminality, researchers have used counts of crimes, arrests etc. because interval measures were not available. Additionally, crime seriousness varies depending on demographic factors. This study examined the *Crime and Violence Scale* (CVS) regarding: psychometric quality using item response theory (IRT); and invariance of the crime seriousness hierarchy for gender, age, and racial/ethnic groups on 7435 respondents. The CVS is a useful measure of criminality, though some items could be improved or dropped. Differential item functioning analysis revealed that crime seriousness varies by age and gender. IRT shows promise in assessing and adjusting for demographic variations in crime seriousness.

Key words: *Criminality measurement, item response theory, Rasch analysis, crime seriousness, Global Appraisal of Individual Needs (GAIN)*

Introduction

Although it may seem obvious that there is a hierarchy of seriousness in the construct of crime and violence, most measures consist of counts, such as the number of reported crimes, arrests, and convictions because hierarchical, linear, interval measures have not been available. Many authors have noted that merely counting the number of crimes committed gives a faulty estimate of criminality because this method weights all crimes equally, e.g., larceny treated as equal in seriousness to murder (Anderson and Newman 1998; Kwan et al. 2000; Wilkins 1980). This problem is further compounded in counts of the numbers of crimes or arrests since less serious crimes (e.g. shoplifting, prostitution) are often much more common than more serious crimes (e.g., assault, murder).

However, hierarchical linear measures are beginning to be developed. For example, Piquero et al. (2002) recently used item response theory (IRT) measurement, which provides an empirically derived hierarchy, to analyze one of the most commonly used delinquency measures, i.e., the Self-Reported Delinquency Scale (Elliott et al. 1985). The Cronbach's alpha for the original 24-item scale was .76, whereas the reliability of the 9-item version that fit the IRT/Rasch model was .58. Piquero et al., (2002) noted that more remains to be done in developing self-reported delinquency measures. In the 2003 workshop on "Measurement Problems in Criminal Justice Research," one of the key issues of concern was "improving the reliability and validity of self-report surveys, rather than simply assessing these characteristics" (Pepper and Petrie 2003, p. 8). Thornberry and Krohn (2003) argued for the need to better understand self-administration, e.g., response errors across self-report and administrative surveys. They also called for the development of instruments to better measure serious offenses, since early self-report scales tended to ignore serious criminal and delinquent events and concentrated almost exclusively on

minor forms of delinquency. Perhaps the major concern in assessing crime seriousness is its variability depending on demographic factors since there is consensus that it is socially defined (Sellin and Wolfgang 1964; Rossi, Waite, Bose, and Berk 1974; Rossi, Simpson, and Miller 1985).

Research Goals

The purpose of this study was to examine the Crime and Violence Scale (CVS) in terms of: (1) psychometric quality using the IRT/Rasch measurement model, which provides a hierarchical linear measure (Bond and Fox 2007; Rasch 1960), as the standard; and (2) the invariance of the crime seriousness hierarchy for demographic factors, specifically: males vs. females, youth vs. adults, and racial/ethnic groups. If the items of the CVS were not invariant, then changes to the measures, e.g., dropping items and developing new ones, would be considered. The CVS is part of the Global Appraisal of Individual Needs (GAIN) long and short forms (Dennis et al. 2006) and, by the end of 2008, was already in use by over 850 agencies in 47 states. Thus understanding its properties and/or potentially improving it have great potential to impact the field.

Background

One of the earliest attempts to create a linear, interval psychological measure was by Louis Thurstone on the construct of crime (Thurstone 1927). Using his method of paired comparisons, Thurstone administered a list of 19 crimes to 266 students at the University of Chicago. Respondents ranked the crimes in seriousness by comparing all possible pairs (n=171 pairs). Thurstone's scaling method produced a linear, interval hierarchy ranging from the least serious, such as vagrancy, to more serious, such as larceny and assault-battery, to even more serious, such as arson and kidnapping to the most serious, such as rape and homicide. Although

the method was never widely used because of its heavy burden on respondents, Thurstone's crime study was recently replicated twice with remarkably similar results (Kwan et al. 2002; Stone 2000).

Thurstone scaling provides a linear interval ruler of the crimes themselves (Kwan et al. 2002), but not of the persons, i.e., no mathematical relationship between the crimes and persons is possible. In other words, Thurstone scaling is less than ideal since it provides a quantitative estimate of the relative seriousness of the crimes, but not the relative amount of criminality of the persons who committed them.

While replications of Thurstone's work have yielded quite similar results, cultural differences in ratings of the seriousness of particular crimes have been found. An example of this cultural sensitivity is the fact that prostitution is not a crime in some cultures whereas in others it is one of the most serious crimes. Rossi, Waite, Bose, and Berk (1974) used correlation and regression methods to analyze differences in perceived crime seriousness by age, educational attainment, gender and race. They concluded that while norms concerning crime seriousness were widely diffused throughout subgroups of society, educational attainment was the best predictor of agreement vs. disagreement with common norms.

As another example, Kwan et al. (2000) found a large difference in the perceived high seriousness of "drug offense" by Hong Kong residents compared to the low severity of the legal penalty, which was based in British law. Using Thurstone scaling, these authors found that women rated violent crimes against persons, such as rape and assault, as more serious than did men. They also found large differences in ratings based on age and educational attainment. They concluded that "crime seriousness is an evaluation mediated by the social structural context in which it is embedded" (p. 630).

While methods like Thurstone's, sometimes referred to as normative rankings (Sellin and Wolfgang 1964), are most common, Cohen (1988) reported on an alternative method of ranking the seriousness of crime where actual victim injury rates were combined with jury awards in personal injury accident cases to estimate pain, suffering, and fear. Crime-related death rates were combined with estimates of the value of life to arrive at monetary values for the risk of death. These estimates were combined with out-of-pocket costs (such as medical costs and lost wages) to arrive at total dollar estimates of the cost of individual crimes to victims. These dollar estimates were then used to rank the seriousness of crimes. Such econometric methods are useful in understanding seriousness of crimes, but their psychometric application in assessing individuals' criminality has not been demonstrated.

Methods for Measuring Crime

Some ways to measure criminality are to count the number of crimes in public records, or count arrests, or the amount of time spent in jail. Of course, there are problems with these methods. Most violent and criminal behaviors are not prosecuted or adjudicated so official records are lacking in assessing this construct. The number of crimes committed does not take into account the seriousness of those crimes. The arrest record does not take into account how well one is able to avoid arrest. Time spent in jail does not take the ability to avoid jail time into account. All of these may vary depending on local and regional statutes, sociodemographic characteristics, and so on. See Pepper and Petrie, eds. (2003) for a thorough discussion of problems measuring crime and violence.

This leads unavoidably to estimating crime and violence through self-report. Philosophically, this is consistent with the notion that crime seriousness is not an objective attribute but is subjectively perceived among citizens (Black 1979). In fact, as Piquero, et al.,

(2002) noted, so much youth crime escapes official detection that self-report delinquency scales have formed the basis of much of our understanding of delinquency today.

Of course, crime is a sensitive issue, and some may be reluctant to report some behaviors out of fear of self-incrimination. Rather than directly ask about crimes, the principal method of assessing adults' criminality has been to measure psychological and other characteristics of persons that predict crime. For example, the Psychological Inventory of Criminal Thinking Styles (PICTS) is designed to assess eight thinking styles hypothesized to support and maintain a criminal lifestyle (Walters 2002), but it is not a measure of criminality itself. Another personality assessment, the Hare Psychopathy Checklist (Hare 2003) obtains professional ratings, e.g., psychologist or social worker, on two factors. Factor 1 is labeled "selfish, callous and remorseless use of others." Factor 2 is labeled as "chronically unstable, antisocial and socially deviant lifestyle." Walters, et al. (1991) developed the Lifestyle Criminality Screening Form (LCSF), a 14-item screening instrument designed to identify life-style criminality that is divided into four primary sections (irresponsibility, self-indulgence, interpersonal intrusiveness, and social rule breaking).. It does not assess types of crimes, but asks for their number, thereby treating all crimes as equal in weight.

IRT Measurement. Item response theory (IRT) measurement models enable the placement of individuals on the ruler in relation to the crimes (Piquero et al. 2002). The relationship between the person's ability and an item's difficulty is estimated mathematically using probability estimators (Rasch 1960; Wright and Stone 1979). This is called the Rasch or 1-parameter IRT model. Where applicable, additional parameters can be added to measure item slope (2 parameter IRT model) and guessing (3 parameter IRT model). Unlike Thurstone scaling which only scales crimes, the IRT/Rasch method places both persons and crimes on the same

ruler, and it is the only method that provides a linear, interval ruler like those used in the physical sciences (Embretson and Reise 2000; Wright and Stone 1979). These methods can also focus on the assessment of differential item functioning (DIF) on the linear measure between subgroups of individuals that may be the result of real differences in prevalence, cultural perception or measurement bias (Conrad et al. 2007). Therefore, this method employs a similar key assumption to that proposed by Ramchand, et al. (2009): *if for some group an offense, A, is less severe than another, B, then members of that group will be more likely to engage in A before they engage in B, rather than the reverse*. While the Ramchand et al. (2009) method examines temporal precedence, the IRT/Rasch method uses self-reported prevalence to estimate likelihood/probability.

Frequency or endorsement probability works very well to define difficulty in education. Likewise, after seeing hundreds of articles in health and related areas using the Rasch model (Conrad & Smith, 2004), we can say that it also works well to define illness severity. Our proposition was that it would work well to define crime seriousness. The logic is that things that are more valued, e.g., human life, are more protected and more punished when taken or violated. Therefore, crimes form a hierarchy of difficulty based on estimated risk. Less risky crimes are done more often and more risky crimes less often. Therefore, crime seriousness is determined by the value of the object and subsequently the risk in taking it. For example, if you asked thieves (stealing being frequent) why they refrain from killing people (murder being more rare), we think they would say that it is because murder is a much more serious crime (though they may use other words that connote seriousness, e.g., capital crime, life in prison, etc.).

We recognize that this relationship is not simple. For example, Rossi, Simpson, and Miller (1985) pointed out the complexity of crime seriousness is so fine grained as to make

endorsement probability seem “improbable” in assessing seriousness. For example, how can you take “cold-bloodedness” into account? We also noted earlier that there are clearly cultural, educational, gender, age, etc. differences that may challenge unidimensionality. While the Rasch model may not address all such issues, our analysis illustrates the capability of examining some such differences using differential item functioning (DIF) analysis.

Differential Item Functioning

While it is clear that females commit fewer crimes than males and that the crimes that females tend to commit are less violent and less serious, the issue in this study was not which gender commits more crimes or more serious crimes, but rather how we can measure someone’s criminality without bias. Bias may occur when there is differential item functioning (DIF). DIF refers to the observation that persons in different demographic groups score differently on an item even though they are at the same level on the underlying trait. DIF can produce “bias” if not corrected. Regarding such differential crime (item) seriousness per group, the study of sociodemographic variables such as age, gender, and race, using linear interval measurement as provided by the Rasch IRT model (Embretson and Reise 2000), could be informative of the sizes of group differences as measured on the Rasch ruler. These estimates could be used to adjust measures where they were perceived to be unfair, e.g., male standards unfairly applied to crimes such as rape where females are overwhelmingly the victims, or prostitution and commercialized vice that are most prevalent in females. Therefore, unlike other methods, IRT/Rasch measurement provides the capability to estimate the DIF and adjust for it so that measures may be more equally and fairly weighted across demographic groups using endorsement probabilities for persons and items rather than subjective opinions.

Crimes and Differential Item Functioning

Gender DIF. Employing the Self-Reported Delinquency Scale, Piquero et al. (2002) reported that gender DIF tended to follow gender role expectations, with males more likely to endorse theft, carrying a concealed weapon, hitting other students, sexual assault, and breaking and entering. Females were more likely to endorse running away from home, hitting a parent, and being "loud, rude or unruly in a public place." These findings were consistent with Kwan et al. (2000) discussed above.

To give an idea of raw prevalence by gender, U.S. population statistics indicated that, in 2005, females were most highly represented in the following offenses: prostitution and commercialized vice (74% females), runaways (58%), embezzlement (44%), and larceny-theft (42%), none of which are violent crimes (OJJDP Statistical Briefing Book 2007). Females were least represented in: forcible rape (only 2% females), gambling (2%), robbery (9%), and murder/manslaughter (10%), three of which are violent crimes (OJJDP Statistical Briefing Book, 2007).

Age DIF. Again, Kwan et al. (2000) found what they interpreted as substantial differences in crime seriousness among the age groups 18-30 and 46+ in Hong Kong. For example, 37% of the older group thought that "possession of arms" was even more serious than "rape," while only 5% of the younger group thought so. Many of the authors' interpretations of their findings had to do with local history that was experienced by the older group but not by the younger group. Piquero et al. (2002) found among 11 to 17 year olds that DIF concerned the younger, i.e., 11 and 12 year olds being more likely to endorse violent crimes, e.g., gangs and strong arming, and sexual offenses while the older adolescents were more likely to endorse crimes involving property.

Race/Ethnicity DIF. Institutional anomie theory would expect violence to be higher in areas facing greater poverty and change (Kim and Pridemore 2005; Messner and Rosenfeld 1997). It posits that lower-class African-American youth, especially males, are at most risk of selecting violent and criminal responses (Tatum 2000). Piquero et al. (2002) found racial/ethnic differences among whites, blacks, and Hispanics where stealing something <$5, hitting a parent, disorderly conduct, selling hard drugs, prostitution and strong arming teachers had large significant DIF coefficients ($p<.001$).

Summary. In summary, we might expect greater differences by gender and age than by race. In their study of the effects of race, gender, age, and social status on crime serious, Rossi, Simpson, and Miller (1985) concluded: "Perhaps the most outstanding feature of these findings concerning social characteristics of offenders was how slight were their effects. The largest and most consistent effect was that of gender" (p. 77).

Methods

Data Source

The data were comprised of 7,435 cases from the 77 studies involving persons being screened for substance abuse in three dozen locations around the United States that used the GAIN (described below). Over two thirds of these studies were conducted by independent investigators. They were funded by a wide range of organizations (e.g., the Center for Substance Abuse Treatment, National Institute on Alcohol Abuse and Alcoholism, National Institute on Drug Abuse, Robert Wood Johnson Foundation and Interventions Foundation) and conducted in a variety of institutional settings in screening for potential substance abuse treatment, including across adolescent and adult levels of care, student assistance programs, criminal and juvenile justice agencies, mental health agencies, and child protective service and family service

agencies. All data were collected as part of general clinical practice or specific research studies under their respective voluntary consent procedures and were subsequently de-identified.

Research studies were conducted under the supervision of Chestnut's Institutional Review Boards with general consents under federal guidelines (42 CFR Part 2) that explicitly allow record abstraction for program evaluation and development as long as the data are de-identified and kept confidential. Data obtained since the implementation of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) were covered by formal data sharing agreements between Chestnut Health Systems and each of the participating agencies. All interviews were conducted by interviewers with three to four days of training followed by rigorous field-based certification procedures. Full details about the GAIN in general and a working paper on the CVS specifically may be obtained at the following: http://www.chestnut.org/LI/gain/

Background Characteristics of the Sample

The data for this analysis came from 7,435 respondents who completed the CVS. As shown in Table 1, the sample was predominantly under 18 years of age (73%) and male (67%). Almost half were Caucasian (45%), a quarter were African American (26%), and the remainder Hispanic or mixed race. Of the top five primary drugs reported, marijuana was reported by 36% of the sample. The drug least often reported was opioids at 5%. Other drugs reported included amphetamines (8%), cocaine (10%), and alcohol (17%). Twenty-three percent of the sample reported other drugs. Valid measures were obtained on 7,424 persons, 99.9%.

Description of the Measure

The Crime and Violence Scale (CVS) is a count of increasingly violent strategies used for resolving interpersonal conflict in the past year and the types of drug-related, property, and

interpersonal crimes the respondent has committed. It includes serious crimes such as homicide and rape. It is based on the Conflict Tactic Scale introduced in the Family Violence Survey (Strauss 1990) and lay versions of the Federal Bureau of Investigation's (1993) uniform crime report categories introduced in the 1995 National Household Survey on Drug Abuse (Office of Applied Studies 1996) and predicts future crime and violence (White et al. 2003; White 2005).

The CVS consists of four conceptually distinct subscales with a total of 31 dichotomous items (Table 2). Its subscales are the: 12 item General Conflict Tactic Scale (GCTS), 7-item Property Crime Scale (PCS), 7-item Interpersonal Crime Scale (ICS), and the 5-item Drug Crime Scale (DCS). The item stem for the GCTS reads: *"During the past 12 months, have you done the following things?"* Response format is Yes/No (coded: no=0, yes=1). The item stem for the other scales reads: "During the past 12 months, how many times have you..." While the response set is in "times", it is dichotomized to as 0 for none and 1 for one or more times for this scale. This analysis focused on the CVS taken as a single 31-item measure of the construct of crime and violence (a.k.a., criminality) and is a measure of the breadth of the types of violence and crime a person has engaged in during the past year.

Psychometric Quality

Content validity. As noted earlier, the CVS content was derived from the Conflict Tactic Scale and the FBI Uniform Crime Report categories. To examine how this compared to the content of another commonly used measure representing a similar construct using self-report, a table was composed consisting of the 31 CVS items along with the 24 items of the Self-Reported Delinquency Scale or SRDS (Elliott, Ageton, and Huizenga 1985; Huizinga & Elliott 1986). The SRDS was designed to include items that were representative of the full range of acts for which juveniles could be arrested and consisted of 24 items with a nine point Likert-type

response scale ranging in frequency from 1=never, 2=once or twice a year and so on to 9= 2-3 times a week. While we recognize that delinquency is somewhat different from criminality, the literature has noted that self-report measures have lacked items concerning the more serious crimes (Thornberry and Krohn 2003) and, as noted above, they focused on personality indicators, arrest statistics, etc. for adults instead.

There was no way for us to equate or compare the CVS and the SRDS items mathematically in terms of their calibrations or measures because we had no sample that used both measures. Instead, we performed a content analysis of both scales to get some idea of how well the measures covered the spectrum of crimes for which adolescents and adults can be arrested. Though the measures were not directly comparable psychometrically since there were no common persons, the qualitative comparison may be enlightening in gaining perspective on their narrative content. We calculated the CVS measures for adolescents (<18) alone, as well as the measures for the full sample including adults. The adolescent calibrations were included to enable examination of age comparability to the Piquero et al. (2002) study which only included adolescents under 18(n= 1,719) using data from the first wave of the National Youth Survey, a longitudinal study of delinquent behavior among American youth (Huizinga & Elliott 1986). As a concurrent validity indicator, we included cost of crime data on 14 of the CVS indicators (McCollister, French, and Fang, in review).

In Rasch analysis the item hierarchy that is created by the item difficulty estimates provides an indication of construct validity (Smith 2001). The items should form a ladder of low seriousness symptoms on the bottom to high seriousness symptoms on the top. The SRDS and cost hierarchies were referenced to examine how well the Rasch-generated CVS hierarchy conformed**,** i.e., testing the validity of using frequency as a criterion for seriousness. Finally, we

used this table to provide an item by item annotated summary of the results of all analyses conducted in this study.

As described by Embretson and Reise (2001) and others, the Rasch model (Rasch 1960) is the only IRT model that provides linear, interval measurement (Bond and Fox 2007; Wright and Stone 1979). This informed our choice of the Rasch model since the intervals involved in scaling crimes and persons were a major consideration in depicting crime seriousness.

The CVS was analyzed with a Rasch dichotomous model (Rasch 1960; Wright and Stone 1979) with *Winsteps* statistical software (Linacre 2007). The dichotomous model estimates the probability that a respondent will choose a particular response category for an item as:

$$\ln \frac{Pni}{Pni+1} = B_n - D_i,$$

where ln is the natural logarithm, P_{ni} is the probability of respondent n endorsing item i, P_{ni} +1 is the probability of respondent n not endorsing item i, B_n is the person measure of respondent n, D_i is the difficulty of item i. These endorsement probabilities are concatenated over all person responses to all items to place persons and items on the Rasch ruler.

Unidimensionality. The Rasch model requires unidimensionality; that is, all items must be indicators of a single latent variable, in this case, criminality. While other small dimensions may be present, they must not be substantial enough to distort the measure of the construct of interest. The criteria for unidimensionality that we used were as follows. First, we required a high percentage of variance explained by the principal measurement dimension, e.g., > 40%. Second, we required a low percentage of residual variance explained by subsequent factors, e.g., < 15% explained by the first factor of residuals**.** As a comparison Reckase (1979) used 20% as substantial variance, so that our criteria were more conservative, i.e., requiring more variance explained by the measurement dimension and less of the rival factor. To further ensure

unidimensionality, there must be good fit (described below) of items to the model, i.e., less than 1.33 on both infit and outfit statistics for each item (Linacre 1998; Smith 2002).

Person reliability. Person reliability, also known as internal consistency, is typically estimated with Cronbach's alpha. We obtained alpha, as well as Rasch person reliability where our criterion for acceptable quality was .80. The estimate is based on the same concept as Cronbach's alpha. That is, it is the fraction of observed response variance that is reproducible:

$$R_P = \frac{SA_P^2}{SD_P^2},$$

The denominator represents total person variability (SD_P^2). The numerator represents the reproducible part of this variability, i.e., the amount of variance that can be reproduced by the Rasch model. The amount of variance that is reproducible with the Rasch model is called the adjusted person variability. The adjusted person variability is obtained by subtracting error variance from total variance ($SD_P^2 - SE_P^2 = SA_P^2$). This reproducible part then is divided by the total person variability to obtain a reliability estimate for persons (R_P) with values ranging between 0 and 1 (Wright and Masters 1982).

Person reliability is more conservative than alpha since it either estimates extreme scores as having high error or deletes the extreme scores from the estimate. Alpha is typically higher since it estimates extreme scores as perfectly measured. This can be particularly misleading when there are large subgroups with 0 on all measures (i.e., a floor effect) as seen in most measures of criminality.

Fit statistics as validity criteria. Rasch analysis provides fit statistics to test assumptions of fundamental measurement (Wright and Stone 1979). "Fitting the model" simply means meeting basic assumptions of measurement, e.g., high scorers should endorse or get right almost

all of the easy items. Understanding poor fit can lead to dropping items, improving them, or (when differences are real population differences), taking subgroup norms into account through calibration.

The fit of the data to the model is evaluated by fit statistics that are calculated for both persons and items. The Rasch model provides two indicators of misfit: infit and outfit. Calculation of Rasch fit statistics begins with the response residuals (y_{ni}) which estimate how far the actual response (x_{ni}) deviates from Rasch model expectations (E_{ni}).

$$y_{ni} = x_{ni} - E_{ni}$$

To standardize the residuals, we divide them by the item standard deviation, e.g., the formula for a z-score is the person's score on the item minus the item mean divided by the item SD which has a mean of 0 and standard deviation of 1 (For complete explanation of this abbreviated discussion with examples, please see Wright and Stone, 1979; Wright and Masters, 1982). If we sum the squared standardized residuals and divide by N we get the outfit statistics, i.e., the mean square (MNSQ) standardized response residuals with an expected value of 1.0. Outfit statistics are sensitive to unexpected responses farther from the person's measure:

$$outfitMNSQ = \frac{\sum Z_{ni}^2}{N}$$ where Z = the standardized response residual.

For the infit statistic, the standardized response residuals are weighted by the individual variance (W_{ni}) to lessen the impact of unexpected responses far from the person's measure:

$$infitMNSQ = \frac{\sum Z_{ni}^2 W_{ni}}{\sum W_{ni}}$$

The infit is sensitive to unexpected or too much random behavior affecting responses to items near the person ability level or item difficulty level and the outfit is outlier sensitive. Mean square or outfit statistics are defined such that the model-specified uniform value of randomness is 1.0 (Wright and Stone 1979). Person fit indicates the extent to which the person's performance is consistent with the way the items are used by the other respondents.). Items with high infit mean squares show a confused or random pattern that is more serious than outfit and reflects that these items are poor indicators of the construct (Bond and Fox 2007). The items with high outfit mean squares are items with more unexpected responses than are consistent with the model at the tails (i.e., endorsing high severity items but not low severity items). Using Wilson's (2005) criteria for both infit and outfit, an item in this study was regarded as problematical if its misfit was higher than 1.33 mean square.

Differential Item Functioning (DIF) for Age, Gender, and Race

As Bond and Fox (2007) noted, the Rasch model requires that relative item estimates, i.e., item difficulty estimates, remain invariant across subgroups of persons, e.g., females and males. A DIF contrast is simply the estimate of the difference between groups among group members who are at the same level on the construct. It allows us to examine whether items have significantly different meanings for different groups. Bond and Fox suggest that items that show DIF should be investigated to determine what may be inferred about the underlying construct and what that implies about the subsamples of persons detected. In other words, it is an important validity criterion concerning the appropriateness of items and their calibrations. A significant DIF contrast was based on $\geq$ .9 logit difference for all comparisons which is approximately half a standard deviation (SD = 1.87) for the items. Half a standard deviation is a common criterion for clinical significance (Conrad et al. 2007; Norman et al. 2003).

For a complete treatment of Rasch analysis, we recommend Bond and Fox (2007) which includes a glossary of Rasch measurement terminology and Conrad and Smith (2004) for a brief summary with useful references. Terminology may also be accessed online via Rasch Measurement Transactions located at http://www.rasch.org/rmt/.

Results

Psychometric Quality

Content validity. The items of both the CVS and the SRDS are displayed in Table 3 in descending order of seriousness with their respective Rasch measures, annotated evaluation comments, and cost where available. The CVS includes crimes of higher seriousness such as *24.ForcedSex,* 25.*Homicide,* and *26.Arson* (item numbers keyed to Table 2) that are not in the SRDS. As a result the CVS covers a greater range. Most of the SRDS items are in the mid-range. In general, the CVS item hierarchy seems appropriate except for *24.ForcedSex, 29.TradedSex, 15.Forgery/BadChecks*, *1.DiscussedItCalmlySettledIt*, and *2.LeftRoomRatherThanArgue*. *24.ForcedSex* is higher than *25.Homicide,* which is counter-intuitive and counter to the cost data where the cost of rape was $245,032 and the cost for homicide was over 30 times higher at $8,635,611. The calibration for *29.TradedSex* seems too high since it is higher in seriousness than violent crimes such as *20.ArmedTheft, 26.Arson,* and *12.UseGunOrKnifeOnSomeone*. *15.Forgery/BadChecks* is also much higher than supported by the cost data at $5,133 since this is lower than most of the costed items below it. Two other items, in the middle range, may be unsupported by the costs. Specifically, *31.IllegalGambling* is only costed at $8 so it is too high, and *22.HurtOtherNeedMedicalAttn* is similar to aggravated assault at $127,573 so it may be too low. The items *1.DiscussedItCalmlySettledIt* and *2.LeftRoomRatherThanArgue* are not crimes.

For the SRDS, Sexual Assault was the sixth highest item in the hierarchy. Shouldn't it be higher than those above it such as *Stole >$50* and *Prostitution*? Aggravated assault was also too low since it is the 16th highest out of 24 items. Additionally, *GangFights* was the 3rd lowest. Regarding the intent of the SRDS to measure crimes for which adolescents can be arrested, is *Sexual Intercourse* such a crime? Is *Runaway*? The items with dollar values keep decreasing in worth as time goes by so that their meaning is changing as well. How would this affect pretest/posttest assessments? Should the SRDS be adjusted for inflation? In general, it appears that a number of the SRDS items are somewhat dated and should be revised and recalibrated using current data.

Dimensionality. The variance explained by the CVS measure was substantial (Reckase 1979) at 45%, and the percentage of residual variance explained by the first factor of residuals was 11% (below our 15% criterion). Both of these results were supportive of the interpretation that the CVS was unidimensional, a requirement of the Rasch model.

Item fit. Two CVS items had problematic fit statistics (Table 2 which also contains complete items keyed by number), *1.DiscussCalmlySettleIt* with 1.52 infit and 3.01 outfit and *2.LeftRatherThanArgue* with 1.32 infit and 3.36 outfit. While the latter did not quite reach the 1.33 criterion for poor infit, it was close enough to cause concern. Of course, these are the two CVS items that are not crimes. Therefore, we concluded that while they may be useful in the questionnaire to set up other items (i.e., to create a positive response bias before asking questions about stigmatizing or illegal behavior), they should be dropped from the CVS score and analysis. *29.TradedSex* had the highest outfit (outfit mnsq=4.18). *15.ForgeryBadChecks* was the next most misfitting item, i.e., outfit, followed by *24.ForcedSex*, *27.DUI*, and *16.Theft(store)*. These last five items had good infit. We interpreted this to mean that they were good items but that

there was a subgroup of people that endorsed them without endorsing more common crimes and who may be worth studying further.

Person reliability. The Cronbach's alpha (which includes extreme scores) of CVS was high (.91). The Rasch person reliability (which excludes extreme scores) of CVS was good (.81). These estimates were calculated after deleting the two misfitting items.

Differential Item Functioning (DIF) for Gender, Age, Race

The figures below present easily interpretable graphs of the relationships of the various groups on the CVS items which were arranged in ascending measure order, i.e., interpreted as increasing seriousness. The data that formed these graphs are provided in Conrad et al. (2009), available at: www.chestnut.org/li/gain that contains the information to compute differences between groups on each item. In the following section, we provide some interpretation with the results, to avoid both a meaningless list and unnecessary repetition.

Gender DIF. The items, with DIF contrast in parentheses, *5.ActuallyThrewThingAtOne (.91), 7.SlapAnotherPerson (1.29), 31.IllegalGambling (-1.41), 20.ArmedTheft(money) (-.96),* and *29.TradedSex (2.19)* were significantly different for males and females where a negative sign indicated the item was easier (more common) for males to endorse, and positive was easier for females to endorse. The "Gender DIF" figure (Figure 1) shows that *24.ForcedSex* was easier for females to endorse than it was for males, but this is counterintuitive. It should not be easier for females to endorse this item since it is intended to indicate the crime of rape. As we noted earlier, this is historically one of the least frequent crimes of females. Therefore, the DIF value that shows endorsing *24.ForcedSex* as easier for females to endorse than it is for males seemed counterintuitive. The actual item reads: "Made someone have sex with you by force when they did not want to have sex?" This may be an item that males refuse to endorse out of fear of

punishment or the desire to give a socially appropriate response. In fact, 24 people endorsed Forced Sex and 15 were males. Nine females endorsed it, which was about the same percentage, both less than 1% of the respective gender. In this item, is it possible that some females interpreted the word "made" as "persuaded" or "seduced?" Would men interpret this item the same way? In any case, this item, the most misfitting crime, appears to be ineffective in that men seem to be endorsing it too little and women, perhaps, too much. The item should be revised and the revision thoroughly tested in qualitative interviews.

The more serious crimes *20.ArmedTheft(money)*and *23.ArmedTheft(other)*, and one less serious crime, *31.IllegalGambling*, were substantially easier for males to endorse. For females, it was substantially easier to endorse less serious offenses such as *Slapping someone*, *5.ActuallyThrewThingAtOne*, and *29.TradedSex*. The complete item for *29.TradedSex* is: "*Traded sex for food, drugs, or money?"* Because *29.TradedSex* was harder for males to endorse and the sample was heavily weighted with males constituting two-thirds, *29.TradedSex* attained a much higher calibration than it should have, i.e., the third highest overall but second highest for males alone. In other words, since trading sex is very rare for males and males made up two-thirds of the sample, *29.TradedSex* was one of the rarest crimes overall even though it was not that rare when we look at females alone, i.e., the 10th highest for females.

Age DIF. In Figure 2, the items *13.PropertyDamage (-.95), 15.ForgeryBadChecks (1.58), 21.FightingHitting (-.91), 24.ForcedSex (2.01), 26.Arson (-1.31), 29.TradedSex (2.52)* were significantly different for adolescents and adults where a negative sign indicated the item was easier for adolescents to endorse, and positive was easier for adults to endorse. *29.TradedSex, 15.ForgeryBadChecks*, and *24.ForcedSex* were > 0.9 logit, i.e., easier for adults to endorse. This finding mirrors the gender DIF findings because adult females tended to endorse

29.TradedSex, 15.ForgeryBadChecks, and *24.ForcedSex*. Therefore, adults and females were driving the DIF on these items. Additionally, one could make the argument that child prostitution is more serious than adult prostitution so that child prostitution would have a higher seriousness calibration than adult prostitution. Or one could argue that child prostitution is not a crime that the child is committing. Rather, with child prostitution, the child is the victim rather than the criminal. If this were the case, the question arises as to whether the item should be dropped from a child's measure.

Race DIF. African Americans tended to endorse *24.ForcedSex* more than the other groups, but the small numbers involved and the other problems noted above do not allow meaningful interpretation of this finding. We concluded that we could not determine any meaningful differential item functioning by race, and we dropped race from further analysis. Full details can be found in Conrad et al. (2009), available at: www.chestnut.org/li/gain.

Interaction of Gender and Age DIF

Figure 3 and Table 4 display the interaction of gender and age DIF. Figure 3 makes it clear that the major differences in item responses are not simply by gender or by age but by the interaction of these two demographics. Additionally, the significant DIF is usually between adult females and adolescent males. Examining Figure 3 from left to right, there is no significant DIF, i.e., >0.9, for the first five items. Then the significant contrasts for adult females vs. adolescent males, where the items are easier for adolescent males, occurs for *21.FightingHitting, 10.BeatUpSomeone, 13.PropertyDamage, 22.HurtOtherNeedMedAttn, 31.IllegalGambling, 18.BreakAndEnter,* and *26.Arson*. The significant contrasts that are easier for adult females compared to adolescent males are *7.SlapAnotherPerson, 15.ForgeryBadChecks, 29.TradedSex,*

and *24.ForcedSex*. Because of the small number of responses and the suspicion of refusal to answer or misinterpretation by males, we will disregard *24.ForcedSex*.

We can also see in Figure 3 how the item calibrations were driven by adolescent males because they made up the largest proportion in the sample, 3,900 which was 53% in Table 4. Specifically, the dash symbol which represents adolescent males in Figure 3 rises from left to right almost monotonically which indicates that it has the greatest influence on the calibrations. In contrast, the adult female symbol, asterisk, varies the most from the adolescent males as well as from the other groups. As one would expect, adult females tended to vary most similarly with adolescent females. Additionally, we see in Table 4 that the person reliability was lowest for adult females, and the difference was greatest for adolescent males vs. adult females. The CVS had higher person reliability for adolescents (.84) than it did for adults (.74).

Discussion

An important issue that was raised here was that of the construct validity of the hierarchy overall and for males and females split into adult and adolescent groups. When we examined Figure 3, we saw that adult females usually had the highest or lowest item calibration. The most extreme differences were between adolescent males and adult females. What this means is that adolescent males and adult females had very different hierarchies for crime. More to the point is that, when a sample is predominantly composed of adolescent males as ours was in this study, the measures of adult females will tend to be biased upward. This is because women's measures will tend to be composed of less violent acts and less serious crimes such as *7.SlapAnotherPerson, 15.ForgeryBadChecks,* and *29.TradedSex.* On the other hand, adolescent boys will be endorsing more serious and violent crimes such as *22.HurtOtherNeedMedAttn, 20.ArmedTheft(money),* and *23.ArmedTheft(other)).*

A typical way to handle this is to simply say that the former three are female-biased items and drop them. If we drop these "biased" items, everyone's measures should drop, but we would expect that females' measures would drop more because females tended to endorse these items more. This is, in fact, what happens. When we dropped these three items, 917 (38%) of the females' measures decreased, and 918 (18%) of the males' measures decreased. Using a .5 decrease as an arbitrary cutoff for the purpose of comparison, 40 (1.64%) females had decreases of greater than .5 while only 18 (0.36%) males had decreases of greater than .5. By dropping items we observed that, on average, females' criminality dropped more than males. This shows that, indeed, gender does make a difference on certain items in calculating criminality. The problem with dropping items is that the differences in the patterns of crime are real, not bias due to bad items. In other words, the items work well within groups, but not when the groups are pooled.

Another way to handle this problem without dropping the items is to anchor all of the items at their common calibration except for the three biased items (see Conrad et al. 2007 for details of this procedure). These three are allowed to "float." Anchoring ensures that both males and females will be measured on the same scale, thereby enabling comparison. However, the three items that we judged as biased will be unanchored, so that their calibrations may be estimated separately for males and females. To do this, we must have separate runs for males and females. When we did this analysis, all of the females' measures *decreased* by an average of .078 logit while all of the males' measures *increased* by an average of .035 logit. While this is only about a tenth of a logit difference (.078 + .035 = .113) between men and women on average, depending on where the seriousness cutoff was set, some women would drop below that cutoff while some men would go above. Another way of saying this is that the rankings would change

for some men and women whereby removing the bias against women would tend to lower their criminality scores relative to the men while men's scores would go up relative to the women.

To recap, in either of the above scenarios, dropping items or calibrating them separately, women's criminality measures were adjusted downward relative to men's measures. The choice of biased items was made based on fit statistics and the item logic, i.e., no reason to assume that violent crimes should be more serious for women, so these were not dropped nor calibrated separately. The items that were judged to be biased were either dropped or calibrated separately which enabled an "objective" adjustment. In the separate calibration, the fact that most of the items were anchored created a common ruler for both men and women so that their measures could be compared even though three items were calibrated separately.

Of course, one can object to the assumptions we made. However, our point is not to insist on these assumptions but to demonstrate how to adjust measures when the judgment is made that certain items are biased or work differently by gender or other subgroup. Another potential objection is that the item hierarchies are simply different for men and women so that entirely different rulers should be used depending on gender or that separate rulers are needed based on gender and age. The problem with this argument is that most items do not appear to be biased either psychometrically or logically. Additionally, the advantage of having common items is that they may be anchored to create a common ruler so that men and women may be compared to each other in terms of criminality. If we had separate rulers, we could not use them to compare the criminality of men vs. women. As a result, there would still be potential for bias depending on how the rulers were used.

Less reliable for adults. However, an observation that would support building separate rulers or adding appropriate items to the current CVS is the fact that the CVS was found to be

less reliable for adults than it was for adolescents, and that it is the least reliable for female adults at .72 vs. .85 for male adolescents. This is a substantial difference which means that female adult measures have much more error in them. Of course, more error in the measures increases the likelihood of error in clinical decision-making. This may suggest the need for more qualitative cognitive work with adult women to see if the items and reliability can be improved.

Limitations of the Study. It was beyond the scope of this study to conduct further construct validation, e.g., concurrent or predictive using theoretically appropriate variables. However, the CVS has been shown to have good predictive validity in studies by White, et al. (2003) and White (2005).

Conclusions

The CVS is useful as a measure of the construct of crime and violence, but we found areas for improvement. Two items may be dropped from scoring, *1.DiscussCalmlySettleIt* and *2.LeftRatherThanArgue* since they are not crimes and had high misfit estimates as a result. The person reliability of the CVS is especially strong for adolescent males at .85 and adolescent females at .82, but it is relatively weaker for adult males at .76 and adult females at .72. The observation of different hierarchies for gender and age and their interaction is important theoretically since it helps us to understand that one size does not fit all when it comes to calibrating the seriousness of crimes for males vs. females and for youth vs. adults and for the interaction of gender and age. This finding supports the need for development of better items and concomitant measures for adults, especially for adult females.

References

Anderson, P. R., and Newman D. J. 1998. *Introduction to criminal justice.* Boston, MA: McGraw-Hill.

Black, D. 1979. Common sense in the sociology of law. *American Sociological Review 44*: 18-27.

Bond, T.G. and Fox, C.M. 2007. *Applying the Rasch Model: Fundamental measurement in the human sciences*, 2nd Ed. Mahwah, NJ: Erlbaum Associates.

Casement, M.R., St. George, D.M., Tallent, K.A., and Bonnett, D.M. 1994. *OAD-related violence prevention tools for planning in your community.* Rockville, MD: Center for Substance Abuse Prevention (CSAP) Training System, New and Emerging Issues Project.

Cohen, M. A. 1988. Some new evidence on the seriousness of crime. *Criminology 26*: 343–353.

Conrad, K. J., Conrad, K. M., Dennis, M. L., Riley, B.B., Chan, Y. F., and Funk, R. 2009. *Validation of the Crime and Violence Scale (CVS) to the Rasch measurement model, GAIN methods report 1.1*. Chicago, IL: Chestnut Health Systems. Available at: www.chestnut.org/li/gain.

Conrad, K.J., Dennis, M.L., Bezruczko, N., Funk, R., and Riley, B. 2007. Substance use disorder symptoms: Evidence of differential item functioning by age. *Journal of Applied Measurement 8*: 373-387.

Conrad, K.J. and Smith, E.V. 2004. Applications of Rasch analysis in health care. *Medical Care* 42 (Suppl I).

Dennis, M. L., Titus, J.C. White, M. et al. 2003. *Global Appraisal of Individual Needs (GAIN): Administration guide for the GAIN and related measures.* Version 5. Bloomington, IL:

Chestnut Health Systems. Available at: www.chestnut.org/li/gain. Accessed January 30, 2009.

Elliott, D.S., Ageton, S.S., and Huizinga, D. 1985. *Explaining delinquency and drug use*. Beverly Hills, CA: Sage.

Embretson, S.E. and Reise, S.P. 2000. *Item response theory for psychologists.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Federal Bureau of Investigations. 1993. *1993 Uniform Crime Reports.* Washington DC: Available at: http://www.fbi.gov/ucr. Accessed May 23, 2006.

Gottfredson, M.R. and Hirshi, T. 1990. *A general theory of crime*. Stanford, CA: Stanford University Press.

Hare, R. D. 2003. *Manual for the Revised Psychopathy Checklist (2nd ed.).* Toronto, ON, Canada: Multi-Health Systems.

Hindelang, M.J., Hirshi, T., and Weis, J.G. 1981. Correlates of delinquency: The illusion of discrepancy between self-report and official measures. *American Sociological Review 44*: 995-1014.

Huizinga, D., and Elliott, D.S. 1986. Reassessing the reliability and validity of self report delinquency measures. *Journal of Quantitative Criminology 2*:293-327.

Jessor, R., Donovan, J.E., and Costa, F.M. 1991. *Beyond adolescence: Problem behavior and young adult development.* Cambridge, England: Cambridge University Press.

Kim, S. and Pridemore, W.A. 2005. Poverty, socioeconomic change, institutional anomie, and homicide. *Social Science Quarterly 86*(s1): 1377–1398.

Kwan, Y.K., Chiu, L.L., Ip, W.C., and Kwan, P. 2002. Perceived crime seriousness consensus and disparity. *Journal of Criminal Justice 30*: 623-632.

Kwan, Y.K., Ip, W.C., and Kwan, P. 2000. A crime index with Thurstone's scaling of crime severity. *Journal of Criminal Justice 28*: 237-244.

Linacre, J.M. 1998. Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement 2*: 266-283.

Linacre, J.M. 2007 *Winsteps Rasch Measurement* (Version 3.63.0). www.winsteps.com. Author.

McCollister, K.E., French, M.T. and H. Fang. In review (revise and resubmit). The cost of crime to society: New crime-specific estimates for policy and program evaluation. *Drug and Alcohol Dependence*.

Messner, S.F., and Rosenfeld, R. 1997. *Crime and the American Dream.* 2nd ed. Belmont, CA: Wadsworth.

Norman, G.R., Sloan, J.A., and Wyrwich, K.W. 2003. Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care 41*: 582-592.

Office of Applied Studies, Substance Abuse and Mental Health Services Administration; 1996. *Preliminary estimates from the 1995 National Household Survey on Drug Abuse.* Rockville, MD: Author.

OJJDP Statistical Briefing Book. 2006. *Juvenile Offenders and Victims: 2006 National Report*, Chapter 3. Washington, D.C.: Office of Juvenile Justice and Delinquency Prevention . Data Source: Federal Bureau of Investigation. Supplementary Homicide Reports for the years 1980–2002 [machine-readable data files]. Washington, D.C.: FBI. Available online: http://ojjdp.ncjrs.gov/ojstatbb/offenders/qa03105.asp?qaDate=2002. Released on March 27, 2006. Adapted from Snyder, H. and Sickmund, M.

OJJDP Statistical Briefing Book. 2007. Adapted from Snyder, H. Juvenile Arrests 2005. [Forthcoming]. Washington, D.C.: Office of Juvenile Justice and Delinquency Prevention. Available online: http://ojjdp.ncjrs.gov/ojstatbb/crime/qa05101.asp?qaDate=2005. Released on March 19, 2007.

Pepper, J. and Petrie, C. 2003. Overview. In Pepper, J. and Petrie, C., eds. *Measurement problems in criminal justice research: Workshop summary*. Washington, DC: National Academies Press.

Piquero, A.R., MacIntosh, R., and Hickman, M. 2002. The validity of a self-reported delinquency scale: comparisons across age, gender, race, and place of residence. *Sociological Methods & Research 30*: 492-529.

Piquero, A.R., Brame, R., and Moffitt, T.E. 2005. Extending the study of continuity and change: Gender differences in the linkage between adolescent and adult offending. *Journal of Quantitative Criminology 21*: 219-243.

Ramchand, R., MacDonald, J.M., Haviland, A., and Morral, A.R. 2009. A developmental approach for measuring the severity of crimes. *Journal of Quantitative Criminology 25*: 129-153.

Rasch, G. 1960. Probabilistic *models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogiske Institut. (Republished Chicago: The University of Chicago Press: 1980).

Reckase, M. 1979. Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics 4*: 207-230.

Rossi,P.H., and Henry, J.P. 1980. Seriousness: A measure for all purposes? In Klein, M.W., and Teilmann, K.S. (eds.), *Handbook of Criminal Justice Evaluation*, Sage, Beverly Hills, CA.

Rossi, P., Simpson, J.E., and Miller, J.L. 1985. Beyond crime seriousness: Fitting the punishment to the crime. *Journal of Quantitative Criminology 1*:59-90.

Rossi, P., Waite, E., Bose, C., and Berk, R. 1974. Seriousness of crime: normative structure and individual differences. *American Sociological Review 39*: 224-237.

Sellin, T., and Wolfgang, M.E. 1964. *The measurement of delinquency*. Wiley, New York.

Smith, E.V. 2001. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement 2*, 281-311.

Smith, E.V. 2002. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement 3*: 205-231.

Stone, M.H. 2000. Thurstone's crime scale re-visited. *Popular Measurement Spring*: 53-54.

Strauss, M.A. 1990. Conflict tactic scale. In: Strauss, M.A., Gelles, R.J., eds. *Physical violence in American Families: Risk Factors and Adaptations to Violence in 8,145 Families*. Durham, NH: University of New Hampshire.

Tatum, B. L. 2000. Toward a Neocolonial model of adolescent crime and violence. *Journal of Contemporary Criminal Justice 16*: 157-170.

Thornberry, T.P. and Krohn, M.D. 2003. Comparison of self-report and official data for measuring crime, 43-94. In Pepper, J. and Petrie, C., eds. *Measurement problems in criminal justice research: Workshop summary*. Washington, DC: National Academies Press.

Thurstone, L.L. 1927. The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology 21*: 384-400.

Walters, G.D., White, T.W., and Denney, D. 1991. The Lifestyle Criminality Screening Form. *Criminal Justice and Behavior 18*: 406-418.

Walters, G.D. 2002. The Psychological Inventory of Criminal Thinking Styles: A review and meta-analysis. *Assessment 9*: 278-291.

White, M. 2005. Predicting violence in juvenile offenders: The interaction of individual, social, and environmental influences. *Offender Substance Abuse Report*: 83–90.

White, M.K., Funk, R., and White, W. 2003. Predicting violent behavior in adolescent cannabis users: The GAIN-CVI. *Offender Substance Abuse Report 3*:67–69.

Wilkins, L. T. 1980. World crime: To measure or not to measure? In G. R. Newman (Ed.), *Crime and deviance: A comparative perspective* (pp. 17–41). London: Sage Publications.

Wilson, M. 2005. *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum Associates.

Wright, B.D., and Masters, G.N. 1982. Rating scale analysis. Chicago: University of Chicago, MESA Press.

Wright, B.D., and Stone, M.H. 1979. *Best test design.* Chicago: University of Chicago, MESA Press.

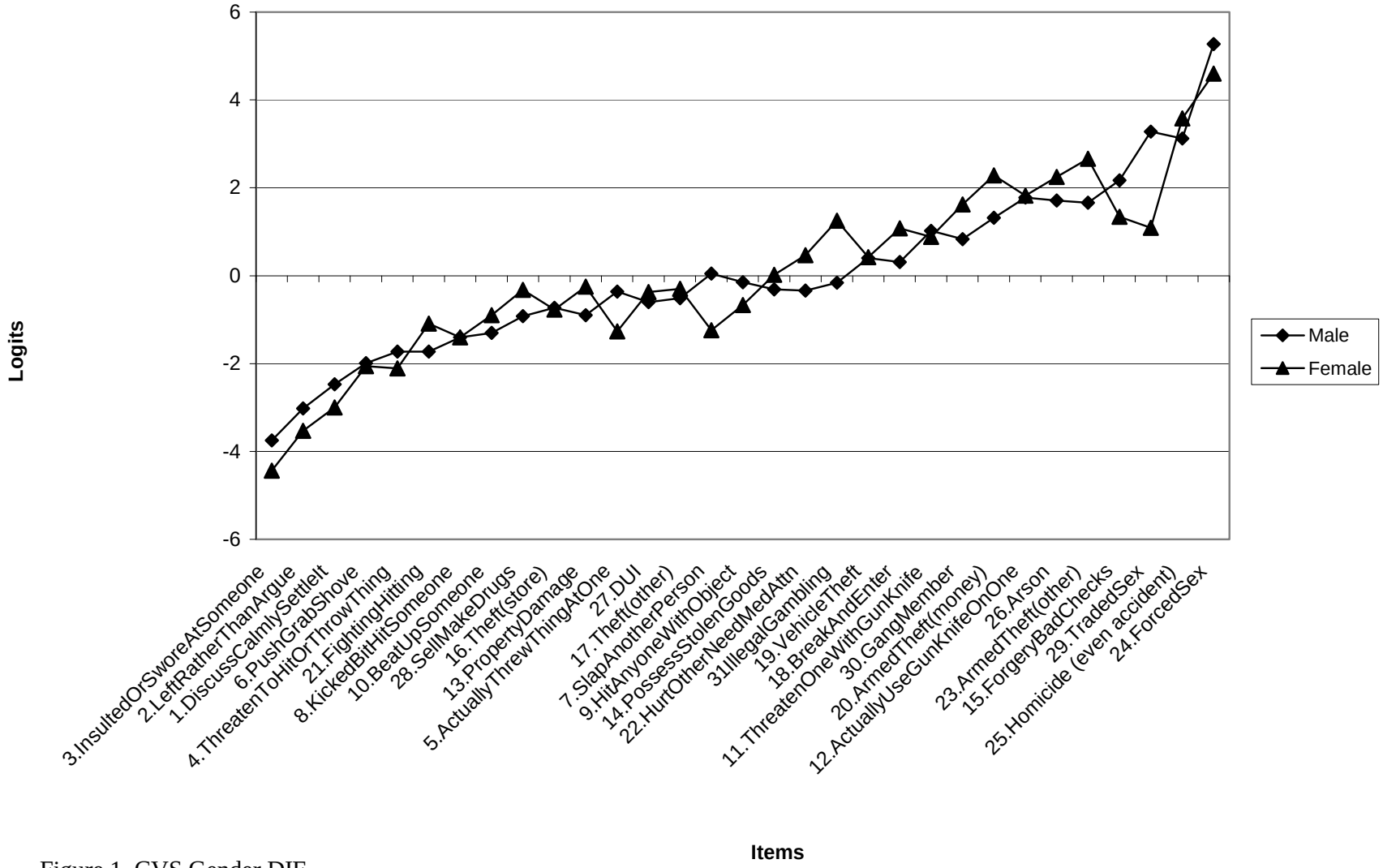

Figure 1. CVS Gender DIF

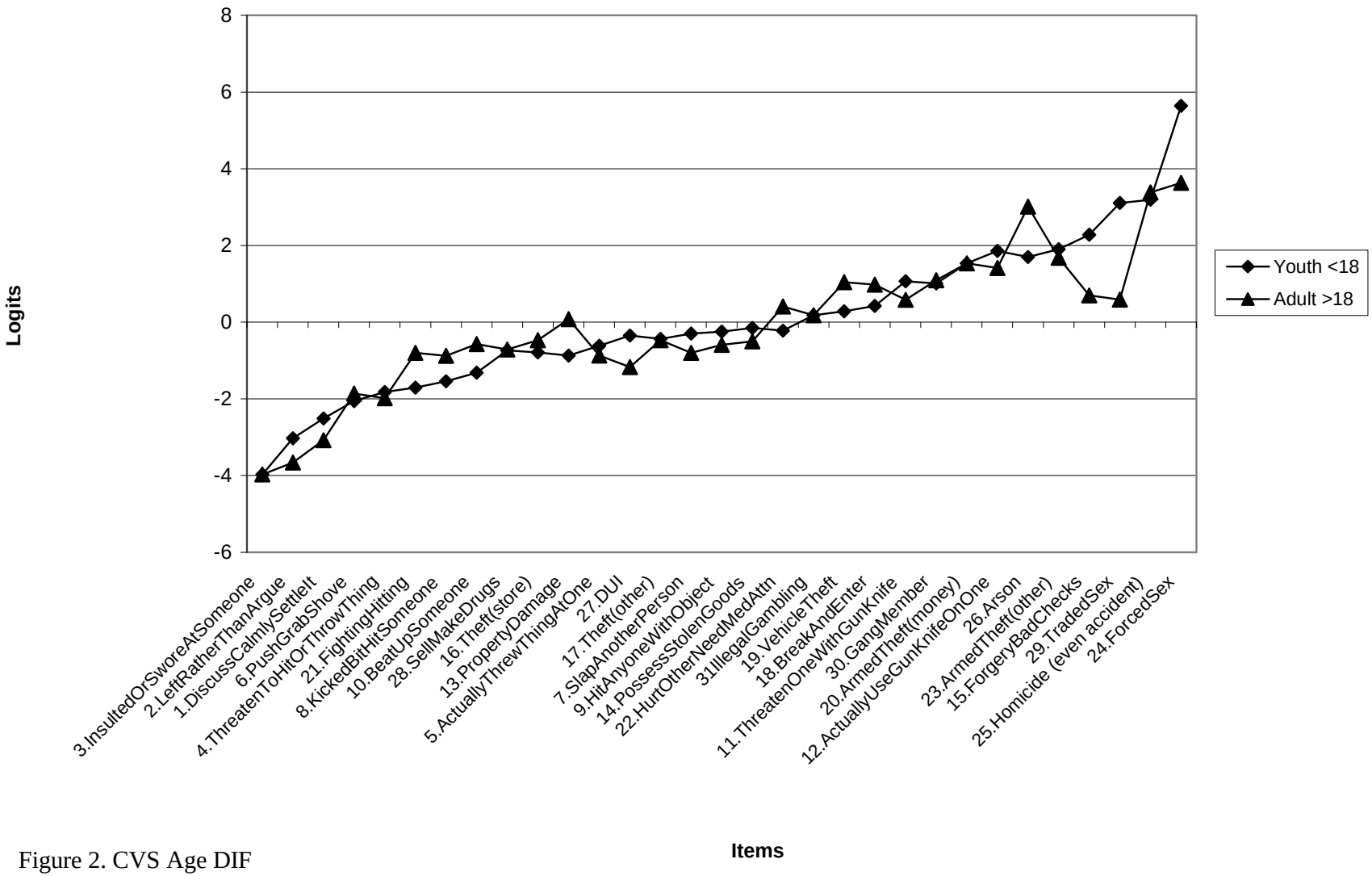

Figure 2. CVS Age DIF

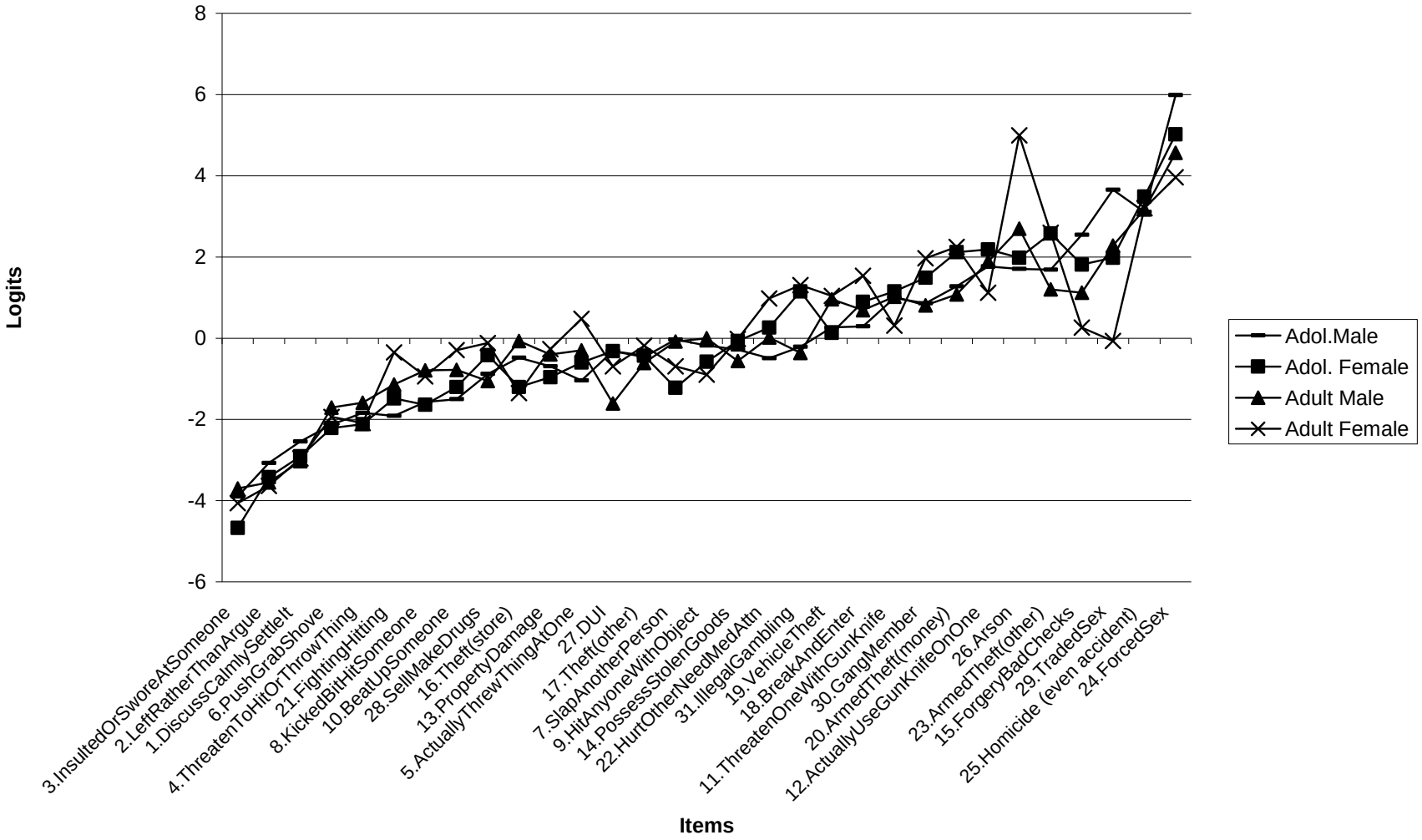

Figure 3. Interaction of Age and Gender DIF

Table 1. Demographic Characteristics of the Sample (N=7435[a])

	Percent	**Number**
Age		
< 18 years	72.5	5388
>18 years	27.5	2047
Gender		
Male	67.1	4992
Female	32.7	2437
Race		
African American	25.7	1913
Caucasian	45.2	3360
Hispanic	10.8	806
Mixed/other	17.7	1314
Drug (primary, most severe)		
Alcohol	20.6	1527
Amphetamines	11.0	820
Marijuana	49.1	3654
Cocaine	10.9	808
Opiates	5.3	393
Other drug	2.9	214

[a] Numbers may not add up to 100% due to missing values

Table 2. Crime and Violence Scale: Subscales, Items and Fit Statistics

Crime and Violence Scale: Subscales and Items	**Rasch Item Calibrations and Fit Statistics**		
Items	Measure	Infit	Outfit
General Conflict Tactic Scale			
1. Discussed it calmly and settled the disagreement?	-2.65	1.52	3.01
2. Left the room or area rather than argue?	-3.19	1.32	3.36
3. Insulted, swore, or cursed at someone?	-3.97	0.95	1.19
4. Threatened to hit or throw something at another person?	-1.86	0.9	0.88
5. Actually threw something at someone?	-0.67	0.97	0.84
6. Pushed, grabbed, or shoved someone?	-2.01	0.83	0.75
7. Slapped another person?	-0.4	1.09	0.95
8. Kicked, bit, or hit someone?	-1.4	0.87	0.8
9. Hit or tried to hit anyone with something (an object)?	-0.32	0.91	0.74
10. Beat up someone?	-1.17	0.87	0.76
11. Threatened anyone with knife or gun?	0.98	0.91	0.74
12. Actually used a knife or gun on someone?	1.79	0.93	0.87
Property Crime Scale			
13. Purposely damaged or destroyed property that did not belong to you?	-0.71	0.91	1
14. Bought, received, possessed, or stolen goods?	-0.21	0.94	0.94
15. Passed bad checks, forged, or altered a prescription, or took money from an employee?	1.91	1.11	2.17
16. Taken something from a store without paying for it?	-0.73	1.07	1.38
17. Other than from a store, taken money or property that didn't belong to you?	-0.45	0.97	1.02
18. Broken into a house or building to steal something or just to look around?	0.5	0.98	1.17
19. Taken a car that didn't belong to you?	0.4	0.99	1.23
Interpersonal Crime Scale			
20. Used a weapon, force, or strong-arm methods to get money or things from a person?	1.53	0.84	0.59
21. Hit someone or gotten into a physical fight?	-1.52	0.79	0.76

22. Hurt someone badly enough they needed bandages or a doctor?	-0.12	0.85	0.74
23. Used a knife or gun or some other thing, like a club, to get something from a person?	1.87	0.82	0.52
24. Made someone have sex with you by force when they did not want to have sex?	5.07	1.04	1.44
25. Been involved in the death or murder of another person (including accidents)?	3.22	0.98	0.92
26. Intentionally set a building, car, or other property on fire?	1.84	0.98	0.86
Drug Crime Scale			
27. Driven a vehicle while under the influence of alcohol or illegal drugs?			
28. Sold, distributed, or helped to make illegal drugs?	-0.52	1.11	1.55
29. Traded sex for food, drugs, or money?	-0.74	0.95	1.14
30. Been a member of a gang?	2.34	1.19	4.18
31. Gambled illegally?	1.02	1.01	1.14
	0.19	1.03	1.18

Table 3. Content Validity of CVS and the SRDS with Items Side-by-Side in Descending Rasch Calibration

CVS Meas	CVS Items	CVS Validity Eval & Cost	SRDS Meas	SRDS Items	SRDS Validity Evaluation
n=5,317, <18 (full)	*Alpha=.90, person r = .81*	*with alternative term for crime*	*n=1,719, <18*	*Alpha=.76*	
5.66 (5.07)	Forced sex	Costs $245,032 for rape; gender DIF; candidate for revision			
3.18 (3.22)	Homicide (even accident)	Costs $8,635,811 (not highest?)			This supports the criticism
3.10 (2.34)	Prostitution	Too high due to gender DIF			of delinquency scales that
2.26 (1.91)	Forgery/badchecks	$5,133 high rank given this cost			they lack the more
1.87 (1.86)	Armed theft (other)				serious crimes.
1.83 (1.78)	Use gun or knife on someone	$127,573 aggravated assault			
1.66 (1.84)	Arson	$16,139	1.75	Stole motor vehicle	
			1.75	Sold hard drugs	
1.49 (1.53)	Armed theft (money)	$45,334 robbery	1.61	Strong-armed teachers	
1.03 (0.98)	Threaten with gun or knife				
.96 (1.01)	Gang member		0.96	Prostitution	
		CVS has a greater range	0.96	Stole >$50 worth	$ worth decreases in time
		of seriousness of crimes.	0.94	Sexual assault	Should be highest
.37 (0.51)	Break and enter		0.64	Panhandled	Higher than assault?
.23 (0.4)	Vehicle theft	$10,806	0.52	Strong-armed others	
		CVS could use 1 or 2 items	0.51	Strong-armed students	Most of the SRDS items
		in the mid-range, but there	0.35	Broke into building, car	are located in the central
		are no big gaps, i.e., >.5.	0.03	Sold marijuana	range of seriousness.
.13 (0.18)	Illegal gambling	$8	0.22	Joyriding	Therefore, it lacks the
			-0.06	Stole $5-$50 worth	ability to measure persons
			-0.11	Hit parent	at the high end and has
-0.20 (-0.21)	Possess stolen goods		-0.01	Runaway	some large gaps at the
-0.28 (-0.12)	Hurt other=need medical attn	$127,573 same as agg. assault?	-0.09	Aggravated assault	low end.
-0.31 (-0.32)	Hit anyone with something		-0.3	Carry hidden weapon	Piquero et al. found that
-0.35 (-0.4)	Slap another person		-0.38	Hit teacher	the 9 pt. rating scale did
-0.41 (-0.53)	DUI	$29			not function well so this
-0.50 (-0.44)	Theft (other)	$7,100 burglary			does not help.

CVS Meas	CVS Items	CVS Validity Eval & Cost	SRDS Meas	SRDS Items	SRDS Validity Evaluation
-0.68 (-0.67)	Threw something at someone		-0.61	Bought stolen goods	
-0.81 (-0.74)	Sell/make drugs				
-0.85 (-0.73)	Theft (store)	$3,966 includes all types			
-0.93 (-0.71)	Property damage	$5,099 vandalism	-0.86	Gang fights	Too low?
-1.38 (-1.17)	Beat someone up	$19,000 simple assault	-1.29	Sexual intercourse	Can one be arrested?
-1.60 (-1.4)	Kicked, bit or hit someone		-1.38	Stole <$5 worth	
-1.77 (-1.52)	Fighting/hitting				
-1.89 (-1.86)	Threaten to hit or throw s'thing				
-2.12 (-2.01)	Push grab or shove someone		-2.24	Disorderly conduct	
-2.57 (-2.65)	Discuss it calmly and settle it	High infit and outfit. Not crime.	-2.92	Hit students	
-3.09 (-3.19)	Left room rather than argue	High infit and outfit. Not crime.			
-4.03 (-3.97)	Insulted or swore at someone				

Table 4. Descriptive statistics and psychometric properties of the CVS when calibrated separately by age and gender

	Males		Females	
Statistics	Adolescent	Adult	Adolescent	Adult
N of cases	3900	1092	1478	954
Mean	-1.44	-2.39	-1.68	-2.58
SD	1.61	1.55	1.58	1.41
Min.	-4.66	-4.87	-5.15	-5.22
Max.	4.93	2.15	5.55	2.36
Person reliability	0.85	0.76	0.83	0.72