Journal of Quantitative Criminology, Vol. 2, No. 4, 1986

Reassessing the Reliability and Validity of Self-Report Delinquency Measures

David Huizinga[1] and Delbert S. Elliott[1,2]

Several issues related to the reliability and validity of self-report delinquency measures are raised and discussed. These include problems associated with the use of internal consistency as the measure of reliability, the level of reliability or precision required for different types of analyses, problems with the content validity of self-report measures, problems of overreporting and underreporting, problems with the use of official records as a validity check on self-reports, and the lack of any good criterion as a major obstacle in assessing the empirical validity of self-report measures. In the light of these problems, some cautions about the use of self-report measures are made.

KEY WORDS: reliability; validity; self-report; delinquency.

1. INTRODUCTION

1.1. General Objectives

Few issues are as critical to the study of crime and delinquency as the question of the reliability and validity of our measures of this phenomenon. Much of the earlier debate on this issue centered on the relative merits and disadvantages of self-report measures as compared to official record measures, and for a number of years now criminologists have been polarized with respect to these two approaches to measuring crime. This resulted in part because there was limited information available on the reliability and validity of self-report measures and in part because these measures appeared to generate different basic findings regarding the volume and distribution of crime in the population and a different partitioning of subjects into criminal and noncriminal subgroups. These measure-related differences quickly became linked to ideological differences and theoretical preferences.

[1]Institute of Behavioral Science, University of Colorado, Boulder, Colorado 80309.
[2]To whom correspondence should be addressed at Campus Box 483, IBS, University of Colorado, Boulder, Colorado 80309.

0748-4518/86/1200-0293$05.00/0

The concern over the measurement of crime has now taken a slightly different direction. Currently many crime and delinquency researchers consider self-report measures to have acceptable levels of reliability and validity, i.e., the reliability and validity of these measures compare favorably to those of other standard measures employed routinely by social scientists (Hindelang *et al.*, 1981). It is also clear that official record measures of crime have *not* been replaced by self-report measures, and there is no sign that they are likely to be replaced in the near-future. Further, there is recent evidence that at least some of the earlier observed discrepancies in findings between self-report and official record measures were the result of differences in measure content and form, i.e., comparisons involving different offense sets and/or prevalence with incidence measures (Reiss, 1975; Hindelang *et al.*, 1979, 1981; Elliott and Ageton, 1980; Elliott and Huizinga, 1983). As a result, self-reported offender measures, self-reported victimization measures, and official record measures now tend to be viewed as alternative measures of crime which compliment one another, each having some strengths or advantages which the others lack and some limitations which are better addressed by the others. Each is considered a reasonably reliable and valid measure of crime which is more appropriate for certain research purposes than others (Garofalo and Hindelang, 1977).

The accumulated research on the reliability and validity of self-report delinquency measures has consistently supported the conclusion that these measures have acceptable levels of reliability and validity as judged by conventional social-science standards (e.g., Hindelang *et al.*, 1981; Sampson, 1985; Wyner, 1981; Hardt and Petersen-Hardt, 1977; Huizinga and Elliott, 1983). Still, the question of the reliability and validity of self-report measures continues to be a major issue. There are several reasons for this. First, the approach to validation has relied heavily (but not exclusively) upon official record measures of crime as the validation criterion. While correlation with alternative measures is a standard form of measure validation, since the validity of neither arrest nor self-report measures is beyond question, it leaves the issue of the true validity of these two measures unanswered. Second, there are conceptual, methodological, or interpretation problems with much of the earlier validation work. Third, a number of important validity issues have simply not been addressed. For example, the major emphasis has been on deliberate falsification and recall problems as sources of underreporting; relatively little attention has been given to sources of error leading to overreporting. Fourth, there is some evidence that while self-report measures are reliable and valid in general, they are differentially valid within certain subpopulations. For example, Hindelang *et al.* (1981) found that self-report measures have a lower reliability and validity for blacks and delinquents than for whites and nondelinquents. There are

grounds for questioning this finding (see Elliott, 1982), but if it were sustained by further research, it would seriously limit the appropriateness of self-report measures for certain research purposes. Finally, while these measures may meet minimum standards, it cannot be said that estimates of reliability and validity are uniformly high; there is an obvious need to work toward the further improvement of self-report measures of crime and delinquency.

In the following sections, several issues related to the reliability and validity of self-report delinquency measures are raised. Discussions of these issues include prior research findings and incorporate new information from the National Youth Survey.[3] In light of the problems described, some cautions about the use of self-report measures are made.

2. RELIABILITY

2.1. Definition of Reliability

The reliability of a measuring instrument is commonly defined as the level of precision of the instrument. In this context, the level of precision refers to the extent to which the measuring instrument would produce identical scores if it were used to make multiple measures of the same object or, equivalently, the amount of measurement error, when each measurement is considered as the sum of true score and error components.

This definition of reliability (along with assumptions of linearity and independence) leads to the use of the correlation between two repeated measurements as an estimate of the reliability of the measuring instrument (test–retest reliability). If only a single test is available, an estimate of reliability may be obtained by dividing the items included in the instrument or scale into two equivalent parts and obtaining the correlation between the scores of the two parts or by examining the internal consistency of the instrument or scale. This internal consistency estimate of reliability is commonly obtained from the Kuder–Richardson (1937) formula 20 or Cronbach's (1951) coefficient alpha. It should be carefully noted that if the items contained in a scale are dependent on more than one dimension, trait,

[3]The National Youth Survey (NYS) is a projected longitudinal study of delinquent behavior, alcohol and drug use, and problem-related substance use in the American youth population. To date, six waves of data have been collected on this national youth panel ($N = 1725$), covering the period from 1976 to 1983. The NYS employed a probability sample of households in the continental United States based upon a self-weighting, multistage, cluster sampling design. Annual involvement in delinquent behavior and substance use was selfreported by members of the youth panel in confidential, personal (face-to-face) interviews. In 1980 a search of police records was completed for each respondent in each location where the respondent lived between 1976 and 1978.

or attribute, then the internal consistency estimate of reliability may be poor and underestimate the actual reliability.

2.2. Some Methodological Issues

2.2.1. Reliability Is a Relative Measure

The reliability coefficient, as formulated above, is expressed as a fraction of the variance of the original scores. Letting r_{xx} denote reliability, then $r_{xx} = \sigma_t^2/\sigma_x^2 = (\sigma_x^2 - \sigma_e^2)/\sigma_x^2$, where σ_t^2 is the variance of the true scores, σ_x^2 is the variance of observed scores, and σ_e^2 is the variance of the measurement errors. Thus the degree of precision indicated by the reliability coefficient is appraised relative to the size of the total variance of the observed measures. The same level of absolute precision will provide a higher coefficient of reliability if the variance is large than if the variance is small. For example, suppose that in half of a sequence of measurements the magnitude of the errors made is +1 and in the other half it is −1, so that $\sigma_e^2 = 1$. If the variance of the measurements is relatively small, say, for instance, $\sigma_x^2 = 2$, then the reliability of the measurements is 0.5. If, on the other hand, the variance is large, say, for instance, $\sigma_x^2 = 10$, then the reliability of the measurements is 0.9. Thus the same level of absolute error (± 1) leads to different levels of reliability, depending on the magnitude of the variance of the measurements. For this reason, the use of the standard error of measurement or other indicators of absolute accuracy, in conjunction with the reliability coefficient, has been suggested (Helmstadter, 1970; Thorndike, 1951; Nunnally, 1972; APA, 1974).

One reason for recognizing that reliability is measured relative to the total variance is that any procedure that results in a restriction of the range of the measured variable may lower the reliability coefficient, even when the absolute level of accuracy remains unchanged. For example, an examination of the reliabilities of lower-variance subscales of a larger, more inclusive measure may indicate lower reliabilities for the subscales. Similarly, the reliabilities for certain subgroups of the sample from which measures were taken may have lower reliabilities. Thus, for self-reported criminal behavior, measures of only serious or violent criminal acts may have lower reliabilities than a more inclusive measure and the reliabilities for various demographic groups may be different, simply because of differing variances of the subscales or of differing variances between subgroups, and not because of differing levels of absolute accuracy.

2.2.2. What Index of Reliability to Use for Delinquency Measures?

As noted above, there are several different methods for estimating the reliability of a test or measure. For the purpose of determining the reliability

of self-reported indices of crime or delinquency (SRD indices), the use of a test–retest estimate of reliability seems more appropriate. There is, in general, no a priori reason to assume that an individual engaging in a particular delinquent behavior is likely to engage in other delinquent behaviors (especially if the described behaviors are relatively specific) or to engage in various delinquent behaviors at the same frequency. As a result, a SRD index is not likely to be undimensional, nor will the items be homogeneous (cf., Hindelang *et al.*, 1981). Thus, as noted in an early conference on self-reported delinquency measures (Hardt and Bodine, 1965), the use of split-half and other internal consistency measures of reliability is inappropriate. Similarly, creating two equivalent forms of the same SRD index that vary in any meaningful way may be impossible. It should also be noted that if interest focuses on the full range of delinquent behavior, the use of internal consistency measures of reliability to prune from the scale those items that have a low correlation with scale totals and a low correlation with other items, with the intent of increasing the reliability of the scale, is unwarranted. Not only may the actual or true reliability be unaffected by the removal of such items, but restricting the index to a more homogeneous set of items may eliminate from consideration important aspects of an individual's involvement in delinquent behavior.

In addition, many of the difficulties inherent in the test–retest method may also be minimized for SRD measures. Memory effects can be reduced by using moderate test–retest intervals and changes in respondents between test and retest are not likely to produce major changes in counts of behaviors. Because delinquency may not be a homogeneous domain, the question of variability due to the sampling of SRD items becomes inappropriate; different samples of items would be expected a priori to produce different scores. Finally, careful control of test and retest administration can and must be made. Clearly any substantial change in the administration of the test and retest may result in changes in self-reported delinquency.

Given the advantages of the test–retest measure of reliability for SRD indices, it should be noted that the majority of researchers who have examined this issue (cited later in this section) have used a test–retest procedure. Commonly, the product-moment correlation between test and retest scores has been used as the reliability coefficient.

2.2.3. Errors Correlated with Scores

The basic derivation of reliability as a ratio of true-to-observed score variance and the use of the correlation between test and restest scores as an estimate of this ratio are predicated on the assumption that the errors have a mean of zero, are uncorrelated with each other, and are uncorrelated

with the true scores, both within and across the two tests. For SRD indices the assumptions that errors have a zero mean and are uncorrelated may be reasonable, but the assumption that the errors are uncorrelated with true scores is not, especially for indices that obtain a frequency or count of behaviors. It might be anticipated that if the number of times an individual has performed a behavior is zero or very small, then an accurate report would be obtained. However, if the number of times is large, then only a rough approximation could be given, since it would be difficult to remember and count each occurrence of the behavior. In fact, there may be some unknown function between frequency and magnitude of error. As a result, errors and true scores are likely to be correlated. An empirical example of the correlation between "true" scores and error scores using self-reported arrest data is provided by Wyner (1981), and related findings are given later in this paper. "Ever-variety" scale scores (see Hindelang *et al.*, 1981) are likely to face the same difficulty, since assuming that items have an equally likely chance of being under- or overreported, the magnitude of error is dependent simply on the number of items included in the scale.

The effect of the true and error score correlations is that the product-moment correlation overestimates the true reliability and the overestimate is proportional to the error-true score correlation. Thus for SRD indices the reported levels of reliability probably indicate higher levels of accuracy than is actually the case. Also, the standard error of measurement is not constant across the range of SRD scores and the computed error of measurement probably indicates a level of error that is too large for small SRD scores and is too small for large SRD scores. This is a general problem for measures of reliability of self-reported crime and delinquency, and there does not appear to be a simple, general correction or adjustment. Conceivably, the reliability of SRD should be obtained at several different scale levels and a comparison made between the reliabilities of the different levels.

Although the statistical treatment of correlated errors is beyond the scope of this paper, it is important to recognize that such errors provide serious problems for regression and correlation analyses and thus for linear models and path analysis (Chai, 1971; Chai and Frankfurter, 1974; Cochran, 1963) and for variance or interval estimation (Cochran, 1963). Thus it is an issue to which researchers employing self-reported crime or delinquency measures should pay greater attention. Initial considerations are under way (Wyner, 1981; Bridges, 1978) and the use of latent or unobservable variable models such as LISREL (Joreskog, 1970; Joreskog and Sorbom, 1978) has been considered and used with some preliminary success (Sampson, 1985). However, given the multivariate normal distribution assumptions underlying the current formulation of these models and the very highly skewed self-reported measures of crime, the latent-variable models may not provide a

panacea for the problems of correlated errors. Although there is some evidence about the robustness of maximum-likelihood estimation of latent-variable models in the face of nonnormality (Fuller and Hemmerle, 1966; Olson, 1979; Hubba and Harlow, 1983), the general adequacy of the method for use with delinquency data that commonly exhibit extreme skewness with only nonnegative scores and a mode and median (appropriately defined) of 0 is simply unknown. As a result, the degree of confidence that can be placed in the parameter estimates obtained and conclusions drawn from these estimates becomes questionable. Clearly both statistical and methodological techniques (e.g., interviewing, scoring, and scale construction techniques) for reducing and adjusting for correlated errors in self-reported measures of crime and delinquency are needed.

2.2.4. Reliability for Items or for Scales

Although reliability is most often thought of in terms of scales, the identical notions of true and error score variance are equally applicable to individual items. For SRD indices, where the items in scales may not be measuring an underlying uniform domain, examination of item reliabilities may be important. Conceivably, the correlation between the test and the retest responses to each item contained in a scale could be zero or even negative and yet the reliability or correlation of scale totals could be 1.0. Thus for studies employing SRD measures it would be useful to know the item reliabilities to ensure that scale reliabilities reflect consistency of reporting at the item level, and since many such studies report item-level findings, knowledge of item reliabilities would be useful directly.

2.2.5. Meaning and Adequacy of Different Levels of Reliability

Often when reliabilities of SRD are reported it is noted that the scales have an adequate reliability because the reliabilities are of the same magnitude as the reliabilities reported for various attitude and behavioral scales (Hindelang *et al.*, 1981; Wyner, 1981). This correspondence does not, however, indicate what is a good reliability or whether the reliability is in fact adequate for the purpose intended. A reliability of 0.7 translates into a standard measurement error of only slightly more than half of a standard deviation and a reliability of 0.99 indicates a standard error of one-tenth of a standard deviation.

For SRD indices with a small standard deviation a reliability of 0.7 may seem adequate. For example, for frequency indices with a standard deviation of 1 or 2 behaviors, a 0.7 reliability indicates that the standard error is around 1 behavior, or less. For the highly skewed large variance SRD scales, however, a reliability of 0.7 may indicate that the standard measurement

error exceeds the mean of the scale. Clearly this might challenge the usefulness or adequacy of that scale. Even a reliability of 0.99, implying a standard error of 0.10 standard deviation, may not be adequate if the scale variance is very large.

It should also be noted that certain levels of reliability may be adequate for some purposes but not for others. Kelley (1927) is often quoted in this regard. He noted that if differences of 0.26 standard deviation are considered important, then to make a discrimination at this level with chances of 5 to 1 being correct, a test must have a reliability of 0.50 for locating the mean and 0.90 for determining the difference between the means of two successive measurements. Similarly for individual data, to evaluate the level of an individual a reliability of 0.94 is needed and to evaluate differences in individual performance a reliability of 0.98 is required. While these minimal levels of reliability are based on particular assumptions that probably are not often obtained in practice, they do illustrate the point that the level of reliability needed is dependent on the purpose or use of the scale. They also illustrate that if some notion of the level of discrimination needed in a particular application can be expressed as a proportion of the true standard deviation and an approximate sample size is known, then an approximate level of reliability needed for the application can be determined. It is not simply a matter of judgment.

For the large-variance SRD scales, the level of reliability needed for different purposes may be of critical importance. Determining the level of delinquent involvement for group data may be reasonable, but if reliabilities are not large, determining or predicting individual levels of delinquency may be problematic.

2.2.6. Reported Levels of Reliability in Prior Research

In a brief and nonexhaustive review of the reliabilities reported in earlier delinquency studies, it became apparent that although only a few studies had formally examined the reliability of the SRD indices employed, those that had were reasonably consistent in reporting relatively high reliabilities for the total samples. Test-retest reliabilities in the 0.85–0.99 range were reported by several studies employing various scoring schemes and numbers of items and using test-retest intervals of from less than 1 hr to over 2 months (Kulik *et al.*, 1968; Belson, 1968; Hindelang *et al.*, 1981; Braukman *et al.*, 1979; Patterson and Loeber, 1982; Solnick *et al.*, 1981; Clark and Tift, 1966; Broder and Zimmerman, 1978). Other studies that approximate a test-retest reliability, although having different periods of delinquency reporting in test and retest, also indicate moderate correlations (Farrington, 1973; Bachman *et al.*, 1978; Blakely *et al.*, 1980). In addition,

some studies reported on changes in item responses to simple yes/no questions or card sorts, and again the level of agreement in test-retest situations was reasonably high, with 88 to 96% of the responses remaining unchanged (Belson, 1968; Dentler and Monroe, 1961; Paternoster, 1978; Broder and Zimmerman, 1978). In general, it appears that the reliability of SRD indices is quite high and would be considered adequate by the prevailing standards for attitude and other social-psychological measures.

There are, however, some findings which are not as positive. In a more comprehensive study of the reliability of SRD measures, Hindelang *et al.* (1981) examined reliabilities of different scoring procedures within different sex, race, and police-court record groups. All but one group had test-retest reliabilities in the 0.84–0.97 range. For black males with a police record, however, the reliabilities varied from 0.62 to 0.81, depending on scoring procedure. Patterson and Loeber (1982) report on the reliabilities of various subscales of a larger general measure of SRD and note that a scale consisting of only nonserious items had a reliability of 0.69. Thus there is some indication that the high reliabilities for total samples and total scales may not carry over to certain subgroups or subscales. It should also be noted that there is an indication that when a variety measure (i.e., a count of the number of different offenses committed) is used, the reported reliabilities are slightly higher than when a frequency measure (i.e., the number of all reported offenses) is used (Belson, 1968; Hindelang *et al.*, 1981).

In the National Youth Survey (NYS) test-retest reliabilities were obtained for a sample of respondents. The total set of respondents participating in the fifth-wave survey was stratified by race (white, black) and four levels of delinquent involvement. Within each of the eight strata, approximately 20 individuals were randomly selected to be included in the test-retest study. A total of 177 retest interviews was completed. All retest respondents were reinterviewed approximately 4 weeks after their initial interview. (The distribution of test-retest intervals is bell shaped, with a range of 21–35 days. The mean, median, and mode, however, all fall on the 28- to 29-day interval.) The retest interview was conducted in the same manner as the initial interview and in most cases involved identical interview situations, i.e., the same interview setting and interviewer. Complete details of the test-retest study are given by Huizinga and Elliott (1983).

Some illustrative findings from this study are contained in Table I. Test-retest correlations are provided for both frequency-scored and variety-scored scales. In addition, a measure of absolute accuracy is provided that gives the percentage of respondents who changed their frequency responses by two or less. Assuming that the mean of the test and retest scores closely approximates the expected value of the observed scores, then given the classic assumptions for the derivation of the reliability coefficient, the latter

Table I. Reliability Indices of the National Youth Survey Self-Reported Delinquency Scales

Scale	Test-retest reliability indices: 1980, Total sample ($N = 177$)				
	Test scale mean	Correlation frequency score	Correlation variety score	Percentage of sample with test-retest difference of 2 or less	Standard error of measurement
General delinquency	45.1	0.750	0.844	37.8	56.0
Index offenses	0.6	0.651	0.869	97.0	1.7
Felony assault	0.3	0.673	0.762	98.3	0.6
Minor assault	1.5	0.585	0.565	84.6	2.3
Robbery	0.1	0.837	0.747	99.4	0.4
Felony theft	0.4	0.523	0.884	98.9	2.0
Minor theft	0.8	0.802	0.759	96.6	1.7
Property damage	1.2	0.880	0.578	91.5	2.1
Illegal services	1.7	0.930	0.792	93.8	3.5
Public disorder	12.0	0.915	0.691	66.9	14.1
Status	33.6	0.489	0.798	50.0	69.1

	Anglos ($N = 100$)			Blacks ($N = 76$)		
	Correlation frequency score	Correlation variety score	Percentage of sample with test-retest difference of 2 or less	Correlation frequency score	Correlation variety score	Percentage of sample with test-retest difference of 2 or less
General delinquency	0.818	0.768	36.4	0.651	0.926	39.7
Index offenses	0.503	0.708	97.0	0.976	0.919	98.7
Felony assault	0.653	0.695	98.0	0.828	0.820	98.7
Minor assault	0.551	0.636	91.0	0.665	0.491	76.0
Robbery	0.551	0.636	91.0	0.665	0.804	76.0
Felony theft	0.805	0.882	99.0	0.859	0.884	98.7
Minor theft	0.789	0.579	95.0	0.889	0.911	98.7
Property damage	0.892	0.478	91.0	0.735	0.771	92.2
Illegal services	0.931	0.770	92.0	0.810	0.856	96.1
Public disorder	0.918	0.670	58.6	0.419	0.664	77.6
Status	0.865	0.788	54.3	0.279	0.816	45.5

measure indicates the proportion of respondents whose response errors are less than or equal to one delinquent act. Data for the total sample and for whites and blacks are provided.

Several interesting findings are illustrated in Table I. First, correlation-type measures of reliability indicate relative, not absolute, levels of precision. For example, within the total sample the highly skewed, low-variance, frequency-scored UCR index scale has a reliability of 0.65, although 97% of the respondents provided test-retest differences of two or less. For the larger-variance general delinquency scale the reliability is 0.75, yet only 38% of the sample have differences of less than two for this scale. Second, in terms of reliability, neither frequency scores nor variety scores outperform the other. One scoring procedure may be more reliable for a given scale but less reliable for another. Third, the adequacy of the SRD scales, as measured by their reliability, varies by scale, by subgroup, and by scoring method. For example, the reliability of the variety-scored property damage scale is low for Anglos only. On the other hand, the frequency-scored status offense scale is low only for blacks. There is no evidence here that blacks have systematically lower reliabilities than whites for any of the measures of reliability. These findings are illustrative of the findings from the more comprehensive NYS reliability study (Huizinga and Elliott, 1983). There were no consistent differences across sex, race, class, place of residence, or delinquency level, in the sense that no one group had consistently lower or higher reliabilities across a majority of scales. Similar findings held for various bivariate subgroup classifications (i.e., sex by race, sex by class, etc.). Fourth, in terms of absolute precision, there is a general ordering of the SRD scales. Scales representing more serious, less frequently occurring offenses (index offenses, felony assault, felony theft, robbery) have the highest precision, with 96 to 100% agreement, followed by the less serious offenses (minor assault, minor theft, property damage), with 80–95% agreement. The public disorder and status scales have lower reliabilities (in the 40 to 70% agreement range), followed finally by the general SRD scale, which, being a composite of the other scales, not surprisingly has the lowest test-retest agreement.

In general, the reliabilities of the individual items included in the NYS delinquency measure are over 0.5, with the majority of reliabilities ranging from 0.65 to 1.00. Although there are some items with low reliabilities, for the most part the reliabilities at the item level are in the same range as the reliabilities for scales. Thus, the reliabilities of the scales do not result from a fortuitous combination of item scores, but reflect the reliabilities of the underlying items.

Additional information about the absolute reliability of responses is indicated by the standard error of measurement which is presented in Table

I. As noted earlier, for the highly skewed delinquency scales, the standard error of measurement is probably not constant across the range of the scores and indicates a level of error that is too large for all scores and too small for large scores. This results from individuals engaging in none or only a few acts having a better chance of reliably recalling and reporting each such act, while those engaged in many acts can give only a rough approximation. Even with these difficulties, the estimated measurement errors give some notion of the size of "average errors." These errors are large relative to the mean values and, in most instances, exceed the mean values. While this most likely results from large errors accompanying large scores and thus does *not* indicate that all scores are "more error than accurate measurement," it does suggest that correlational measures of reliability may not adequately reflect the absolute precision of SRD indices.

2.2.7. Some Empirical Evidence of the Correlation Between Scores and Magnitude of Errors

Some notion of differences in the magnitude of the errors made by less frequent and more frequent offenders is indicated by the proportions of these offender types who change their responses by more than two behaviors. Earlier it was noted that more frequent offenders might be anticipated to have larger errors in reporting. Defining a low-delinquency group as having five or fewer reported offenses and a high-delinquency group as having six or more reported offenses, approximately 60% of the low-delinquency group had test–retest differences on the general SRD measures that were two or less, and only about 20% of the high-delinquency group were this precise. While the exact magnitude of error is not indicated by these data, they clearly suggest that errors made by high-frequency offenders are likely, on the average, to be larger than those made by less frequent offenders. While the proportion of individuals within these two delinquent groups with difference scores of less than two varies by particular scales, the low-delinquency group always has the largest such proportion, as illustrated in Table II.

Also included in Table II is the correlation between the original test scores and the absolute value of the test–retest difference scores. Assuming, for a given individual, that the mean of the two test scores approximates the true score, then the test–retest difference provides an indicator of the magnitude of the response error (the magnitude is approximately one-half the difference score). These correlations clearly indicate, across various SRD scales, that as the number of reported delinquent acts increases, so does the magnitude of the response errors.

Considering the test–retest difference as a simple linear function of the original test scores, the *b* coefficient and constant are also given in Table

Table II. Percentage of Low- and High-Delinquency Groups who Have a Test-Retest Difference of Two or Less and Correlation and Regression Coefficients of First Test Score with Test-Retest Differences

Scale	Percentage with test-retest difference of 2 or less		Correlation and regression coefficients of first test score with absolute value of test-retest differences		
	Low delinquency ($N=82$)	High delinquency ($N=95$)	r	b_1	b_0
General delinquency	59.5	19.4	0.73	0.46	3.93
Index offenses	100.0	95.8	0.92	0.70	−0.02
Felony assault	100.0	96.8	0.81	0.96	−0.03
Minor assault	98.8	72.3	0.73	0.71	0.21
Robbery	100.0	98.9	0.97	0.65	0.01
Felony theft	100.0	97.9	0.99	0.987	−0.06
Minor theft	100.0	93.7	0.92	0.57	0.06
Property damage	97.6	86.3	0.99	0.74	0.12
Illegal services	97.6	90.5	0.83	0.35	0.36
Public disorder	85.0	51.6	0.35	0.16	4.96
Status	71.8	33.3	0.90	0.76	−0.20

II. In most cases the regression line passes close to the origin, indicating that those individuals reporting zero or one behavior make very small errors. The exceptions to this are the general delinquency and public disorder scales. Given earlier assumptions, the percentage error of a response is approximately one-half the slope of the regression line, so the size of the b_1 coefficients suggests that rather sizable errors would be anticipated when responses are large.

2.2.8. Percentage of Persons who Change Their Response from Positive to Never or from Never to Positive

Although not directly involved in the usual examination of reliability for delinquency measures, it is of interest to examine a particular kind of change in SRD scores from test to retest. While small changes in reported delinquency would be expected, it might be anticipated that individuals will accurately remember and report whether they ever engaged in particular behaviors during the last year. Thus, it would be expected that never (or 0) responses on the original test would remain never on the retest and, similarly, that positive responses would remain positive. The percentage of the test-retest sample who "changed their minds" about whether they had engaged in the offenses contained in each scale is given in Table III. Separate percentages are given for the total sample and for whites and blacks.

Table III. Percentage of Sample who Changed Their Response from Never to Positive or from Positive to Never on Test and Retest

Scale	Total sample	Whites	Blacks
General delinquency	9.88	13.13	5.48
Index offenses	8.52	8.00	9.21
Felony assault	7.39	7.00	7.89
Minor assault	19.43	14.00	26.67
Robbery	2.27	1,00	3.95
Felony theft	3.41	3.00	3.95
Minor theft	10.17	15.00	3.90
Property damage	14.12	18.00	9.00
Illegal services	5.08	6.00	3.90
Public disorder	14.29	14.14	14.47
Status	10.00	10.87	9.09

Examination of Table III indicates that in many instances a substantial proportion of the sample changed from positive to never or from never to positive. Examination of the direction of change (data not presented) indicates that in most cases a slightly larger proportion of the changes is from positive to never. Changes in the more serious offense specific scales (felony assault, robbery, and felony theft) are comparatively smaller than changes in the less serious offense specific scales. A more comprehensive examination of these changes indicated that although there were differences between various sex, race, and social-class subgroups on some scales, there did not appear to be any consistent differences such that one group had a consistently greater or lower proportion of changes across a majority of scales.

The magnitude of the percentages of individuals who change their mind about whether or not they have engaged in various kinds of delinquent behavior clearly suggests a moderate level of error in many of the SRD indices. Although for group analyses the positive-to-never and never-to-positive changes may "cancel" much of the error, for individual data the "error" is rather large. As noted above, this is especially true for the less serious or minor scales, where over a quarter of some subgroups changed their minds about whether they had ever (in the last year) engaged in certain minor delinquent behaviors. Thus, the lack of response consistency to the question of ever committing particular offenses suggests that at least the minor SRD indices may not be very reliable.

As an overview of the NYS findings, it appears that the levels of reliability are somewhat lower than those reported in the studies cited earlier. Also, there is variation in the reliabilities of various scales; not all

scales are equally reliable for all subgroups. Scales involving more serious behaviors have a higher absolute reliability, but given their small variances, their test–retest correlations often do not reflect this precision. In addition, there is evidence that the magnitude of response errors is positively correlated with the number of delinquent acts reported, that the standard error of measurement is relatively large for most SRD scales, and that a sizable proportion of individuals changed their minds in the period between test and retest about whether they have ever engaged in particular delinquent acts.

2.3. Summary

In the preceding it has been noted that because delinquent behavior is most likely not a homogenous domain, the use of test–retest correlations as measures of the reliability of SRD indices is more appropriate. The vast majority of studies examining the reliability of SRD indices has followed this prescription and generally has found the reliabilities to lie in the eighties and nineties. While this level of reliability is often said to be adequate in the light of prevailing standards for attitude measurement, there are some major difficulties inherent in the reliabilities of SRD indices.

First, measurement error is most likely correlated with delinquency scores so that correlations become inaccurate representations of reliability. Second, the levels of reliability (as measured by correlations) are low for certain groups and/or certain SRD scales. There is less variation by scale and group with respect to the absolute measures. Third, although the reliability of SRD scales may be considered adequate for some purposes, e.g., determining group norms, the levels reported may be questionable for determining individual differences and, especially, for determining change scores.

Clearly, the reliability of SRD scales is an issue that requires further examination and it would be inappropriate to assume on the basis of current evidence that the reliability of SRD indices is adequate for all subgroups or for all purposes. While the current evidence is promising and the reliabilities reported compare favorably with those of other social-psychological measures, further effort in determining and improving the reliability of SRD measures is necessary and some care should be taken in the use of these scales in future delinquency research. Investigators should not rely on the findings of others or assume that the reliability of SRD scales has been proven adequate but should continue to examine the reliability of the scales they employ. Some suggestions for the construction of SRD instruments that may influence their reliability are given by Huizinga and Elliott (1984).

3. VALIDITY

3.1. Definition

The validity of a psychological or behavioral test is commonly defined as the evidence that the test measures what it was intended to measure or that it represents what it appears to represent. Thus to determine the validity of indices of delinquent or criminal behavior, it becomes important to delineate carefully what is being measured or represented. The term delinquent has been used in various ways, e.g., to describe persons or groups, to describe illegal behaviors, and as a synonym for deviant, with the result that the meaning of the term delinquent is often ambiguous. However, what is being measured by a delinquency index for most current researchers is the commission of behaviors that are violations of criminal statutes or such violations that are actually acted upon by formal law-enforcement agencies. This definition is important not only because it is a necessary prerequisite to determining if a measure is valid but also because it indicates what ostensibly is being measured is a count of specific behaviors. Underlying the delinquency measures are, although perhaps unknowable, absolute true scores of delinquent behavior. Thus delinquency is not an abstract construct and a variety of empirical indicators can play a more prominent role in the determination of the validity of a given measure of delinquent behavior.

Given a relatively precise definition of what is being measured, three major approaches to the demonstration of validity are often described. Content validity refers to the subjective evaluation that the test items seem plausible and relevant and that the universe of behavior being measured is adequately sampled by the test items. Empirical or criterion validity refers to the relationship between test scores and some known external criterion that accurately indicates the quantity being measured. Construct validity involves the use of theoretical hypotheses about the relationship of test scores to other theoretical variables and the empirical justification of those hypotheses.

In general, based on the first or second of these indicators of validity, almost all researchers in crime and delinquency that have investigated the validity of their self-reported measures of delinquent behavior conclude that these measures are reasonably valid or are valid in the sense that they compare favorably with the validity of other measures employed in the social sciences (cf. Hindelang *et al.*, 1981, pp. 114, 213). However, it should be carefully noted that most such researchers, including the authors of this article, have a vested interest in producing a positive evaluation of the validity of either official data or self-reports of delinquency (or both), since a negative evaluation would challenge years of individual research effort. The conclusions concerning validity are not made by disinterested parties.

In the validity literature, only two articles provide strong cautionary notes. Gould (1969) suggests that given the problems inherent in both arrest and self-report data, there may be no measure of delinquent behavior in which criminologists can place a high degree of trust, and Bridges (1978) concludes from a more technical examination that biases and correlated errors may seriously distort our measures of crime and delinquency.

Construct validity has seldom, if ever, been used in delinquency research. The problem of simultaneously examining both tests of theory and validity issues within the same study generally precludes examination of construct validity. However, many variables theoretically linked to delinquency have been shown to be correlated with self-reported delinquency measures, and even when the correlations are not those specified by a given theory, the researchers have concluded that the theories are misspecified and not that the self-report measures are invalid. Thus, in a very loose sense, there is some indication of the construct validity of SRD measures.

In the following sections a brief review of findings relative to the content and empirical validity of self-reported measures of delinquency is given. A more detailed review of some of the studies cited is given by Hindelang *et al.* (1981).

3.2. Content Validity

3.2.1. Face Validity

Face validity refers to the evaluation of what the items included in an index appear to measure. Many of the indices of self-reported delinquency that have been used include items that do not involve violations of criminal statutes or involve such trivial infractions that they would rarely result in official action even if observed or discovered. Although many of the items included in some SRD indices are about criminal violations, others are not, and the summative scales or indices constructed from the total set of items thus do not appear to have a uniform or consistent face validity. More recently this problem has been recognized and at least partially corrected by the elimination of items that involve only trivial or noncriminal infractions. However, many of the SRD indices in use include such items and thus may fail the test of validity [a notable exception is the set of items employed by Hindelang *et al.* (1981)].

A related problem concerns the nature of responses to items which, on the surface, appear to be about serious-offense behavior. Questioning respondents about offenses they have reported reveals that some responses are about trivial events that do not match the severity of the offense described. This source of error results in inflated estimates of involvement

in delinquent behavior, i.e., it constitutes a form of overreporting. In one national study it was estimated that 22% of the responses were too trivial to be charged or considered delinquent (Gold and Reimer, 1975), and in the NYS it was estimated that 36% of the responses either were inappropriate (i.e., did not involve behaviors which belonged in the category of behaviors identified by the question) or were too trivial to be charged (Huizinga and Elliott, 1983). The amount of misclassification (inappropriate responses) by respondents is quite low (4%) in the NYS. A disproportionate number of classification errors involved theft items (e.g., reporting a bicycle theft in response to an auto-theft question or the theft of a stereo valued at $150 in response to a theft of $5–50 question). There were no consistent differences in the rate of inappropriate responses by sex, race, class, age, or place of residence.

The vast majority of overreporting (trivial responses) in the NYS involved items concerning minor assault. However, the remaining items, especially felony assault (including robbery) and property damage items, also had a sizable proportion of responses that were considered trivial. There was no evidence, however, of a differential distribution of trivial responses by sex, race, social class, or place of residence (urban, suburban, rural). Exactly why the interview situation, instruction sets, or wording of items causes some respondents to report trivial events to serious items is not clear, but some combination of those factors illicited reports of trivial events. This is problematic because unless detailed information about each reported event is obtained, thereby greatly expanding the length of the SRD questions, knowledge about the triviality of reported offenses cannot be obtained. Perhaps investigative efforts directed at interview situations and item wordings that eliminate trivial responses without altering responses about serious offenses would reveal the factors underlying the trivial-response problem and appropriate alterations to SRD instruments could be made. Since there were no sex, race, class, or age differentials in the reporting of trivial events, this overreporting problem may not be a serious one for estimating the social correlates of criminal behavior. But it poses a serious problem for comparisons of self-reported offense rates with NCS or UCR rates and potential problems for etiological studies.

The face validity of SRD scales made up from individual items is also of concern. Frequently scales include both serious and nonserious or trivial items, with the result that scale scores are dominated by the more frequently occurring trivial offenses. The titles of such scales, however, are commonly based on the more serious items, and thus the scales masquerade as measures of serious-offense behavior, while the scores more accurately reflect only trivial behaviors. In addition, some summative scales include overlapping items (e.g., theft at school and theft under $5), which may result in double

counts of the same behavior in scale totals. The face validity of such scales is certainly questionable.

3.2.2. Sampling or Logical Validity

Sampling validity refers to the question of whether the items included in a scale form an adequate and representative sample of the domain of behavior being investigated. Often it appears that measures of self-reported delinquency have not conformed to this notion of validity. The samples of behaviors commonly do not cover the full range of delinquent behavior, most especially underrepresenting serious delinquent behavior. In addition, some measures have excluded serious behaviors that were originally part of the SRD index in order to improve the internal consistency of the measure. (As noted in Section 2, the latter seems an inadvisable procedure.) Over time, since the introduction of SRD indices by Nye and Short (1957), measures of self-reported delinquency have been expanded to include both a wider range of delinquent behaviors and a larger number of more serious offenses (see Hindelang *et al.*, 1981; Elliott and Ageton, 1980) so that currently some SRD measures appear to have a much greater sampling validity. However, even in these more extended measures, it is doubtful that all behaviors that result in arrest are adequately covered. Rarely is the development of a SRD index described by presenting frequency counts of behaviors that have in fact resulted in an arrest and an indication of how the SRD items tap the variety of the more frequently occurring offense behaviors that result in arrest. Without such information it becomes a matter of opinion whether a particular SRD index has a satisfactory sampling validity. In the development of the NYS measure, every offense listed in the UCR which accounted for more than 1% of juvenile arrests was included in the scale (Elliott and Ageton, 1980). While current indices are much improved, it remains, at least in part, an empirical question whether they have an adequate sampling validity, a question that should be investigated.

3.3. Empirical or Criterion Validity

In examining the empirical validity of SRD measures, various means of determining the relationship between SRD and some external criterion have been employed. These include known groups—in which the differences in SRD between groups presumed to have differences in delinquent behavior are demonstrated; correlational—in which the relationship of SRD scales with a criterion variable is examined; and official record checks—in which a check is made to determine if an individual with an officially recorded

offense reports a behavior matching the offense behavior. The evidence relating to each of these forms of empirical validity is briefly reviewed.

3.3.1. Known Group Validity

Differences in SRD between various groups expected to have different levels of delinquency have been examined by several studies. Differences between those with self-reported police contact and those with no self-reported contact have been investigated by Elmhorn (1965) and, indirectly, by Christie *et al.* (1965). Differences between groups defined by various official records and classifications have also been examined in several studies including differences among those with no arrests, one arrest, and two or more arrests (Hirschi, 1969; Hardt and Peterson-Hardt, 1977); those with different numbers of arrests and a group of incarcerated youth (Erickson and Empey, 1963); groups with different numbers of arrests and a group that had gone to court (Marsden and Meade, 1981); those convicted vs those not convicted (Farrington, 1973); those who had gone to court vs those who had not been to court (Hindelang *et al.*, 1981); and those institutionalized in a training school vs those not institutionalized (Short and Nye, 1957, 1958; Voss, 1963; Kulik *et al.*, 1968).

In all cases involving official records or self-report of official contact, the groups that would be anticipated to have higher delinquent involvement (those with greater official involvement) had substantially and usually statistically significant higher mean SRD scores. Although few formally examined the ability of the SRD measures actually to discriminate between groups, most studies would appear to allow some moderately accurate classification into the known groups. In terms of this rather minimal check in validity, self-report measures of delinquency are clearly indicated as being valid.

Differences between the mean SRD scores of groups defined by different levels of variables related to delinquent behavior have also been investigated. These variables include teacher reports or expectations of delinquent behavior (Elliott and Voss, 1974; Hackler and Lautt, 1969), the delinquency of friends (Hardt and Petersen-Hardt, 1977), the SO scale of the CPI (Stein *et al.*, 1970), the perceived personal risk of punishment for delinquent acts (Jensen *et al.*, 1978), and levels of obedience, being a class bully, and other indicative variables (Dentler and Monroe, 1961). As with official records, again all groups anticipated to have greater delinquent involvement have higher mean SRD scores. Thus, those who teachers nominate, who have a greater number of delinquent friends, who have lower socialization scores, who have a low perceived risk of punishment, who are less obedient, or who are class bullys, as groups, have higher SRD scores. As a result, in

terms of the differences between the groups defined by these other variables, the SRD indices would appear to be valid.

3.3.2. Correlational Validity

Stronger evidence for the validity of a measure is provided by its correlation with a criterion related to the behavior being measured. A number of factors important to SRD measures affect the magnitude of the measure-criterion correlation, however. Among these are the relationship between the criterion and the underlying behavior and the reliabilities and nonconstant biases of the measure and the criterion. Because most criterion measures used in the examination of the validity of SRD measures are not particularly accurate indicators of the volume of delinquent behavior, correlations between SRD and criterion variables are not expected to be high. Also, since the correlations are affected by the reliabilities of both measures, and the reliabilities of at least the SRD indices are known to be only moderate, the correlations would not be expected to be high.

The correlational validity of SRD measures has been examined using official data, other self-reported indicators of delinquent involvement, reports on respondents behavior by others, and other variables presumed to be related to SRD as criterion measures. The correlations among SRD and arrests or official contacts are generally low, varying from essentially zero for both property and person offenses (Rojek, 1983), to 0.16 (Gould, 1969; Bridges, 1978), to 0.27–0.32 depending on the scoring method or particular scales (Hirschi, 1969; Huizinga and Elliott, 1983), to 0.34–0.56 depending on the scale, scoring method, and sex (Hindelang *et al.*, 1981). The relationship between SRD and self-reported official contacts is much higher, with correlations ranging in the 0.60s for various scales (Hindelang *et al.*, 1981).

The level of these relationships between SRD and official contacts or self-reported official contacts raises a number of issues that are beyond the scope of this paper. Clearly if official data are an accurate reflection of individual involvement in delinquent behavior, then SRD measures do not appear to be very valid. It is more likely, however, that the frequency of delinquent behavior is not tied very tightly to arrests or contacts, and other problems with the accuracy of official data coupled with problems of reliability result in the low reported correlations.

Indirectly related to the validity of SRD are the correlations between self-reported arrests or contacts and official records. These correlations vary from 0.66 (Wyner, 1981) to 0.51–0.80, depending on the sex of the respondent (Hindelang *et al.*, 1981). Additionally, the correlation between self-reported court appearances and officially recorded court appearances is in the 0.80s

(Hindelang *et al.*, 1981). As the level of official processing becomes more advanced, perhaps reflecting more accurate records and more salient events for respondents, the correlation between self-reports and official records appears to increase. [For a similar finding among prison inmates, see Petersilia (1978).] These correlations are thus consistent with the possibility of measurement problems in both official and self-report data at the offense behavior level.

Overall, the relationships between SRD and official measures do not provide very strong evidence for the validity of SRD measures. However, it was not anticipated that the correlations would be large, and in general, the correlations are positive, indicating that the relationship between self-reported and official indicators is essentially as expected.

The correlation between SRD indices and the delinquency of friends provides another indicator of the validity of the SRD indices. Although the use of this correlation as a check on validity requires the assumption that the level of an individual's delinquency is related to the delinquency of friends, and this, perhaps, raises some theoretical issues, there is a consistency to this correlation that reflects on the validity issue. Many studies have examined the relationship between SRD and the delinquency of friends, and they generally find at least a moderate correlation between these two variables (e.g., Hackler and Lautt, 1969; Gold, 1970; Elliott and Voss, 1974; Hindelang *et al.*, 1981; Elliott *et al.*, 1985). While most of the reported product-moment correlations are in the 0.30s, using ordinal categories and gamma as in an index of association, Hindelang *et al.* (1981) found associations ranging from 0.41 to 0.88, depending on SRD scale, sex, and race. They also found moderate associations for different sex and race groups from a reanalysis of other data sets. The consistency of these findings suggests that, to the extent that an individual's delinquency is indicated by the delinquency of friends, some moderate level of validity is being demonstrated for various SRD indices.

3.3.3. Record Checks

One of the most frequently used methods for investigating the validity of SRD measures has been an examination of whether offenses or official actions reported by others will be admitted on a self-report index. These examinations have included whether individuals will self-report the behaviors evidenced by peer reports of their offense behavior and whether they will self-report acts reported or known to the police. While only indirectly related to SRD indices, examinations have also been made of whether individuals will self-report known arrests, court appearances, and convictions.

While on the surface these record checks appear as a strong check on the validity of SRD measures, there are some inherent difficulties in their use (Elliott, 1982). First, obtaining a match between officially recorded and self-reported behavior may be difficult. Official agencies have a wide discretion in the classification of a particular offense behavior, and the classification used may distort the actual event. In addition, a police officer may not know the full details of an event and may charge an individual with a less serious offense than the offense actually occurring. On the other hand, the perception of an event by the offender may be quite different from that of the police officer or some witness or victim filing a complaint, and youthful offenders may not know whether they have been formally arrested or not. The second major difficulty with official record checks as indicators of validity stems from the fact that they can examine only the responses for those individuals and offenses that have come to the attention of official agencies. The absence of a police record cannot be equated with no involvement in delinquent behavior. Whether findings about particular offenses from this select group which has penetrated the justice system can be generalized to larger samples is unknown. Conceivably, the validity of these particular SRD responses for this select group may be either lower or higher than the validity of the total set of responses of this group or of the responses in general population samples. As a result, official record checks can provide only a partial indicator of the validity of SRD indices. In some respects, this comparison is a better validity check on officially recorded events, since there should be a self-report match for every event recorded, whereas it might reasonably be expected that only 1 of every 100 or so self-reported offenses will be matched with an official record, assuming no error in either measure.

One "record check" that did not depend on official records involved the use of peer reports of the delinquent behavior of their friends (Gold, 1970). Examining whether the offenses reported by peers were also reported on an SRD inventory, Gold found that 72% of the youth confessed all offenses, 17% were concealers, and 11% were uncertain.

Record checks that examine whether offenses known to the police are reported on SRD indices have shown that a high proportion of such offenses is in fact admitted. In a school sample in Honolulu, Voss (1963) found that 95% of the official offenses were admitted on an SRD index. Elliott and Voss (1974) found that 83% of the official offenses for a school sample in California were self-reported on a SRD index and that the actual rate of reporting varied by offense from 57 to 100%, Using a more lenient criterion that required that some self-reported offenses be at least as serious as an official offense, 96% of the official responses were reported on the SRD measure, although serious offenses were less well reported (81%) than

nonserious offenses (98%). Hindelang *et al.* (1981), employing a community sample of youth from Seattle, found that the self-reporting rate of official offenses varies by race, with whites admitting 90% of their official offenses and blacks 65% of their official offences. While there is, thus, some evidence of differential validity by offense type and by race, it appears that a high percentage of offenses known to police is reported on SRD indices.

While the above record checks have examined offense *behavior*, it is also useful to determine how many *individuals* are concealing their delinquent behavior. Conceivably, only a few individuals may account for the majority of unreported official offenses. Gibson *et al.* (1970), in a sample of British schoolboys, found that 83% admitted all official convictions on a SRD inventory, 9% made at least partial admissions, and only 8% made no relevant admissions. Thus only 8% deliberately concealed or failed to recall their convictions.

Record checks of self-reported arrests or police contacts have also been made. Hardt and Petersen-Hardt (1977) indicate that 78% of individuals with police records acknowledge this fact on self-report instruments and Hathaway *et al.* (1960) found that 80% of such individuals report their involvement on the MMPI. Rojek (1983) found that 75% of status offenders reported all of their arrests. Hirschi (1969) found somewhat lower rates of accurate reporting of being picked up by the police, with only 60% of those with official records admitting this event. The admission rates of official convictions also appear to be quite high. Blackmore (1974) reports on two separate studies in which 92 and 97% of court-convicted delinquents admitted at least one such conviction and Farrington (1977) reports a 94% figure for a similar record check.

Record checks of self-reported official contacts also provide some indication of the amount of overreporting on self-report measures. Although there is some question whether self-reported official actions that cannot be verified result from inaccuracies in the official record (see Chaiken and Chaiken, 1982) or from exaggeration on the part of respondents, high levels of overreporting would seem suspicious. Estimates of the number of individuals who report official contact when there is no official record vary from 10 to 30% (Hardt and Petersen-Hardt, 1977; Hathaway *et al.*, 1960; Hirschi, 1969). In addition, Rojek (1983) indicates that among those with an arrest, 14% overreport their number of arrests. There is thus some indication of potential exaggeration on the part of respondents to self-report questionnaires.

In the National Youth Survey a record search was performed in which the police records of the towns and cities in which each respondent lived and in all jurisdictions in a 10-mi radius of each respondent's home were searched. This wider search proved important, as nearly 50% of all arrests

for respondents were found outside of the immediate jurisdiction in which the respondent lived. In order to determine if reported arrests were matched by self-reported behaviors, two levels of matching were created. In the first a very tight match of the self-reported behavior to the arrest behavior was required. In the second a more broad match was allowed, in which any self-reported offense that could conceivably have resulted in the recorded arrest was allowed. Complete details of the NYS record search and definitions of matching offenses are given by Huizinga and Elliott (1983). The percentages of youth providing tight and broad matches to their arrest records and the percentage of arrests that are matched by SRD behaviors are given in Table IV, for the total sample and by sex, race, and social-class groups. The race groups are provided only for males since the number of arrests among black females was insufficient for analysis.

As indicated in Table IV, the percentage of individuals that acknowledge all of their arrests and the number of arrest offenses admitted on the SRD items are substantially smaller than those reported in earlier studies. Assuming that tight and broad matches provide estimates of minimum and maximum values and assuming that the arrest records are accurate, then somewhere between 36 and 48% of the respondents with an arrest record were concealing or forgetting at least some of their offense behavior and somewhere between 22 and 32% of the arrest offenses were not reported on the SRD inventory. While there is some variation by sex and class, the concealment or forgetting rate among black males is extreme. This issue is considered in the next section.

Requiring individuals to provide matching SRD responses to all arrests is a rather stringent requirement that does not take into account the likelihood that at least some arrests may not have matching SRD responses, for reasons noted above. Also, it is difficult to understand why a respondent would provide matching SRD responses for a majority of arrests but deliberately conceal other offenses. As a result, the number of individuals that have matching SRD responses for more than half of their arrests is also given in Table IV, for both tight- and broad-match criteria. As expected, the proportion of arrested youth who have matching responses on more than half of their arrests is substantially higher than the proportion with matches on all arrests, across all subgroups. Even on this more relaxed criterion, however, only 78% of arrestees provided broad matches to over half of their arrests, so that more than 20% still would appear to be concealing or forgetting at least some of their offense behavior.

Given the findings from the various official record checks, what conclusions seem warranted? First, it appears that the majority of arrested individuals will self-report officially known offenses. The assertion that most such individuals will deliberately hide their delinquent behavior on survey

Table IV. National Youth Survey, 1976–1978: Number and Percentage of Youth who Have Matching Self-Reported Delinquency (SRD) Responses to Known Arrests and Number of Arrests with Matching Self-Reported Deinquency Responses

	Number of youth									
	Participating in police record search (*N*)	With at least 1 arrest (*N*)	With narrow matches on all arrests		With broad matches on all arrests		With narrow matches on more than 50% of arrests		With broad matches on more than 50% of arrests	
			N	%	*N*	%	*N*	%	*N*	%
Total sample	1452	126	65	52	80	64	87	69	98	78
Sex: Male	769	99	48	49	61	62	68	69	77	78
Female	683	27	17	63	19	70	19	70	21	78
Race (males only): Anglo	610	70	40	57	48	69	53	76	57	81
Black	115	18	2	11	5	28	7	39	11	61
Other	44	11	6	55	8	73	8	73	9	82
Social class: Middle	351	16	10	63	11	69	12	75	13	81
Working	432	33	18	55	23	70	25	76	27	82
Lower	597	63	33	53	39	62	42	67	48	76

	Number of arrests				
	Total number of arrests (*N*)	With a narrow match		With a broad match	
		N	%	*N*	%
Total sample	276	188	68	216	78
Sex: Male	230	155	67	181	79
Female	46	33	72	36	76
Race (males only): Anglo	139	106	76	116	84
Black	57	23	40	35	61
Other	34	26	76	30	88
Social class: Middle	20	14	70	15	75
Working	89	74	83	79	88
Lower	135	82	61	99	73

instruments does not appear to be true. Second, whether self-reported measures of delinquency are seen as valid is an issue for debate. Clearly, on the basis of the official record checks, SRD measures are not perfectly valid and the degree to which the measures appear to be valid depends on whether one "sees the cup as being mostly full or partially empty." Using the NYS data, which present perhaps the lowest record-check validity estimates for juvenile studies, and assuming that the official records are accurate and that the findings from the arrested sample can be generalized to the total sample, it then appears that at least 20% of the respondents may be concealing or forgetting some part of their delinquent behavior and that, overall, approximately 20% of the delinquent behavior among respondents is not being reported on the SRD measure. If the necessary assumptions are correct, clearly this is a substantial error, and even allowing some leeway for inaccurate official records, the findings suggest a sizable level of *underreporting* on the part of youthful respondents. Third, because of potential errors in official records, the magnitude of *overreporting* in self-report instruments is difficult to determine. However, if the errors in official records are not too large, the official record checks also give some indication of *overreporting* on the part of respondents. Further, the earlier discussion of the rate of reporting trivial events suggests substantial levels of overreporting (i.e., 22–32% of all reported offenses). Overall, the magnitude of overreporting appears to be at least as great as that of underreporting. While these two sources of error tend to offset one another on a global measure of delinquency, this may not be the case on more specific scales or for particular subgroups (e.g., blacks).

3.4. Differential Validity of SRD

In the preceding review there has been some indication that the level of validity of SRD measures may differ in different subgroups. In this section the question of differential validity is examined more completely. It should be noted that most of the evidence concerning differential validity comes from record checks and is thus limited to samples of arrestees and arrest behaviors. Whether these findings can be generalized to total samples or to all offense behaviors of arrestees requires some questionable assumptions.

To some extent the findings concerning differential validity are mixed. Gold (1970) found, in matching peer reports of offense behavior and self-reported behavior, that although there were some sex differences on individual items, there were no differences between sex by race groups, at least differences that were statistically significant (at even the 0.20 level). These findings involved relatively small sample sizes, however. Similarly, in a study involving a sample of status offenders and using log-linear models,

Rojek (1983) found that neither sex nor race nor their interaction had an effect on the concordance of self-reported arrest and official arrest. There was an effect by delinquency level, however, with those reporting two or more arrests having less accurate reporting rates. Hardt and Peterson-Hardt (1977) report that in their official record check there was no indication of a racial difference but some indication of differences by social class, with lower-class youth forgetting or concealing more of their official contacts.

In contrast, data from Hirschi (1969) indicate a substantial race difference in the level of underreporting of being picked up by the police, with 16% of white boys and 36% of black boys underreporting their official contacts. Employing data from a more comprehensive study, Hindelang *et al.* (1981) found substantial differences by race, especially among males. Overall, white males reported 90% and black males 67% of their official offenses. The underreporting by black males was even more pronounced among serious offenses, with a 20% underreporting rate for whites and a 57% underreporting rate for blacks. The difference between white and black females is in the same direction but smaller for all offenses, 15 vs 27%, and for serious offenses, 50 vs 59%. Although there are sex differences on specific items, the overall rate of underreporting is similar for both sexes (18–19%). There were also no consistent differences by social class.

Differences in underreporting by seriousness of offense are also clearly indicated across all sex by race groups, although the number of official contacts for serious offenses for females is relatively low, so the accuracy of these rates may be questionable. The rates of underreporting for serious vs nonserious offenses are 20 vs 5% for white males, 57 vs 18% for black males, 50 vs 13% for white females, and 59 vs 21% for black females, yielding an overall 35% serious and 11% nonserious underreporting rate. Data from Elliott and Voss (1974) also indicate a difference by crime type, with a 19% underreporting rate for serious crime, as opposed to 2% for nonserious crime. In contrast to the underreporting of serious offenses found in the juvenile studies, it should be noted that in a sample of adult prison inmates, Petersillia (1978) reports that arrests for more serious offenses were more accurately self-reported than arrests for less serious offenses. Conceivably the reasons for underreporting may be different for juveniles than for adults.

Findings from the National Youth Survey also indicate some level of differential validity. As indicated in Table IV, differences between males and females are found for the number of youth who provide narrow matches to all arrests, but this difference diminishes as the less stringent criteria of broad matches and matches to a majority of arrests are used. Differences between the sexes in the underreporting of arrest behaviors are relatively small. Similarly differences by social class are usually small, and the ordering

of classes by level of underreporting varies by the criteria used, although in all cases the greatest amount of underreporting is found in the lower class. Due to sample sizes for youth with arrest records, differences by race could be examined only among males. Striking differences in the levels of underreporting can be seen, with black males having substantially higher levels of underreporting than whites or others, across all criteria, for both persons and behaviors. Only 11% of blacks provided matching SRD responses to all of their arrests, although this figure is raised to 61% when matching responses to a majority of their arrests. Although sample sizes precluded a comparison of race by seriousness of offense, differences in reporting rates for index and nonindex offenses suggest some differential validity by type of offense. Using the narrow-match criteria, 77% of the nonindex arrests had matching SRD responses, while only 35% of the index arrests were so reported. Using the broad-match criteria, this difference is much smaller but in the same direction, 80 vs 72%.

Assuming that the findings can be generalized from arrested to general samples, several conclusions appear warranted. While it appears that there are some sex differences on particular items, overall levels of underreporting do not vary by sex. Findings concerning social class are mixed, but generally there are few substantial or consistent class differences. The two largest studies with comprehensive arrest and SRD data clearly provide evidence of differentials by race. Most extreme is the underreporting by black males and, in one study, evidence of underreporting by black females as well. In addition, there is some indication that rates of underreporting are greater for the more serious offenses. While it is possible that the magnitudes of the differentials encountered are in part dependent on police practices and errors and biases in official data, they nevertheless provide a cautionary note about the interpretation of results from SRD studies, especially results concerning race. A description of some of the factors that may influence the size of the race differential and analyses of problems arising from this differential are given by Hindelang *et al.* (1981).

Assuming that there is a potential for blacks, and black males in particular, to have a larger underreporting rate on record checks of official data, a major issue arises as to why this is the case. There are a number of possibilities including lying or deliberate falsification, forgetting and lower salience of events, difficulty with coding behavioral events, difficulty with "paper-and-pencil" tests, acquiescence and social desirability, and inaccurate or invalid arrest data. There is, however, relatively little evidence concerning this issue. Evidence that a larger proportion of blacks may deliberately falsify responses is given by Bachman *et al.* (1984). They found that blacks had larger amounts of missing data, which they interpreted as a greater lack of trust and deliberate omission. They also found that a larger

proportion of blacks than whites indicated that they would be unwilling to report marijuana or heroin use if they were in fact using these substances. On the other hand, when faced with a psychological stress evaluator, and the investigators claim that this device could identify subjects who were lying, a larger proportion of white than black males changed their responses from no to yes (Hindelang *et al.*, 1981), thus suggesting that whites are more prone to deliberate falsification than blacks. Also, it is possible that police are more likely to stop and officially record contacts with blacks, and thus blacks may be more likely to be picked up and falsely or incorrectly charged with an offense (Elliott, 1982).

Some evidence that the differential in underreporting does not result from acquiescence or social desirability is provided by Hindelang *et al.* (1981). These authors also found that the race differential was smaller in face-to-face interviews than in self-administered questionnaires, thus suggesting that difficulty in reading and paper-and-pencil tests may result in some level of underreporting.

Given the potential problems in obtaining matches between arrest and SRD offenses (as noted above, the offense as seen by a police officer and a juvenile may be quite different), when less restrictive categories of offense are employed for matching purposes, the race differential is reduced in both the NYS and the study by Hindelang *et al.* (1981). However, the differential is not eliminated in either study.

Finally, the sampling issue in reverse record checks should again be noted. If arrested youth have a lower ability to respond accurately on tests and questionnaires, then they form a select sample of all respondents, and if blacks have an even greater problem in this area [some indication of this is given by Hindelang *et al.* (1981) and Chaiken and Chaiken (1982)], then they form a select sample not only of blacks but of black offenders. Conceivably, the failure to report recorded offenses is characteristic of a small number of black offenders and generalization to all black offenders or the black population is erroneous.

In light of the above, unequivocal answers about why there is a white–black differential in the underreporting of officially recorded offenses are unknown, as is the exact magnitude of the differential. It is our judgment that the strength of the evidence suggests that while various factors may reduce the level of the differential, some difference in the reporting of known arrest offenses remains, and a research effort directed at understanding the reasons underlying this differential would be profitable.

3.5. Summary

As an overview of the validity of self-reported offender measures, the consideration of the content validity of these measures indicated some

potential problems. Examination of the face validity of these measures suggested that they often included trivial items that either were not law violations or were such trivial infractions that in only very specialized circumstances would they result in official action. There is also evidence that items involving seriously delinquent behaviors lead to reports of trivial behaviors and thus to overreporting on some items. The sampling validity of the items contained in self-reported measures is also of concern. The construction of these measures needs to ensure that the full range of delinquent behavior is included. Often, serious offenses have not been adequately represented in prior measures.

In examining the empirical validity of the self-reported offender measures, the examination of known group validity consistently indicated substantial and often large differences on self-report measures between groups presumed to have a low or high involvement in delinquent behavior. The correlational validity of these measures, as indicated by their correlation with other criterion variables presumed to indicate levels of delinquent behavior, was generally quite small. However, none of the criterion variables that have been used are very good indicators of the level of individual delinquent behavior, and as a result, the low correlations would be anticipated. The lack of any good criterion variables provides a major obstacle to the examination of the validity of self-reported offender measures. Without such variables, no truly adequate test of validity can be made. Finally, official record checks indicate that some, and usually the majority of, "officially known" individuals will report the majority of their known offenses, including their serious offenses. However, these record checks also indicate sizable levels of underreporting, especially among blacks, and in general the rate of underreporting was larger for more serious offenses.

4. CONCLUSION

It has become customary, as Hindelang *et al.* (1981) note, for researchers employing self-reported offender data to preface their work with a brief review of research on the reliability and validity of these measures and to reach the general conclusion that these measures are reasonably reliable and valid or that at least the reliability and validity of these measures compare favorably with those of other social-science measures. However, the discussion of the reliability and validity of self-reported offender data presented above suggests that the quality of these measures cannot be taken for granted, nor are the reliabilities and validities sufficiently high that these measures can be used without question. Although at times the psychometric properties of SRD compare favorably with those of other social-science measures, there are instances where they clearly do not meet this criteria.

Particularly problematic are the lower validities among black respondents. In addition, these are measures of countable behaviors, not loosely defined attitudes, matching the levels of reliability and validity of other social-science variables does not mean that the SRD measures are particularly good or that they would meet the standards commonly required in other academic fields.

We believe that self-report measures are among the most promising of our measures of criminal behavior and are, perhaps, the only measures capable of meeting the needs of both descriptive and etiological research efforts. As a result, while research projects employing SRD measures are likely to be continuing, attempts to improve this methodology should be undertaken. Toward this end, bounding and other recall aids should be routinely employed with self-report delinquency measures, as should in-depth follow-up questions. Additional attention must also be given to careful item selection, wording, and scale construction. While suggestions such as these for the potential improvement of self-report delinquency measures can be made, it is clear that further research is needed to improve the reliability and validity of these measures and to understand the conditions and circumstances associated with both over- and underreporting. Such research is necessary if the full potential of self-report offender measures is to be realized.

ACKNOWLEDGMENTS

The National Youth Survey, discussed in this paper, was supported by a series of grants from the Center for Studies of Antisocial and Violent Behavior, NIMH (MH27552), for the period June 1975 through May 1986. Supplemental funding for the second and third years of the study was received from the National Institute for Juvenile Justice and Delinquency Prevention, Office of Juvenile Justice and Delinquency Prevention, U.S. Department of Justice (78-JN-AX-0003). The points of view or opinions expressed in this paper are those of the authors and do not necessarily represent the official position or policies of the Department of Health and Human Services or the Department of Justice.

REFERENCES

APA, Inc. (1974). *Standards for Educational and Psychological Tests*, American Psychological Association, Washington, D.C.

Bachman, J. G., O'Malley, P. M., and Johnston, J. (1978). *Adolescence to Adulthood: Change and Stability in the Lives of Young Men. Youth in Transition, Vol. VI*, Institute of Social Research, University of Michigan, Ann Arbor.

Bachman, J. G., Johnston, L. D., and O'Malley, P. M. (1984). *Questionnaire Responses from the Nations High School Seniors, 1982*, Institute of Social Research, University of Michigan, Ann Arbor.

Belson, W. A. (1968). The extent of stealing by London boys and some of its origins. *Adv. Sci.* 25: 171–184.

Blackmore, J. (1974). The relationship between self reported delinquency and official convictions amongst adolescent boys. *Br. J. Criminol.* 14: 172–176.

Blakely, C. H., Kushler, M. A., Parisian, J. A., and Davidson, W. S. (1980). Self-reported delinquency on an evaluation measure. *Crim. Just. Behav.* 7: 369–386.

Braukman, C. J., Kirigin, K. A., and Wolf, M. M. (1979). Social learning and social control perspectives in group home delinquency treatment research. Paper presented at the American Society of Criminology Meetings, Philadelphia.

Bridges, G. (1978). Errors in the measurement of crime: An application of Joreskogs method of analysis of general covariance structures. In Wellford, C. (ed.), *Quantitative Studies in Criminology*, Sage, Beverly Hills, pp. 9–29.

Broder, P. K., and Zimmerman, J. (1978). *Establishing the Reliability of Self-Reported Delinquency Data*, National Center for State Courts, Williamsburg, Va.

Chai, J. (1971). Correlated measurement errors and the least squares estimator of the regression coefficient. *J. Am. Stat. Assoc.* 66: 478–483.

Chai, J., and Frankfurter, G. (1974). Errors in variables for the simple linear model: The effects of correlated errors of measurement on interval estimation and hypothesis testing. Proceedings of the Social Statistics Section, American Statistical Association.

Chaiken, J. M., and Chaiken, M. R. (1982). *Varieties of Criminal Behavior*, Rand Corp., Santa Monica.

Chambliss, W. J., and Nagasawa, R. H. (1969). On the validity of official statistics: A comparative study of white, black and Japanese high school boys. *J. Res. Crime Delinq.* 6: 71–77.

Christie, N., Andenaes, J., and Skirbekk, S. (1965). A study of self reported crime. In Christiansen, K. O. (ed.), *Scandinavian Studies in Criminology, Vol. 1*, Tavistock, London, pp. 86–116.

Clark, J. P., and Tift, L. L. (1966). Polygraph and interview validation of self reported delinquent behavior. *Am. Sociol. Rev.* 31: 516–523.

Cochran, W. C. (1963). *Sampling Techniques*, Wiley and Sons, New York.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.

Dentler, R. A., and Monroe, L. J. (1961). Social correlates of early adolescent theft. *Am. Sociol. Rev.* 26: 733–743.

Elliott, D. S. (1982). A review essay on "Measuring Delinquency" by M. J. Hindelang, T. Hirschi, and J. G. Weis. *Criminology* 20: 527–537.

Elliott, D. S., and Ageton, S. S. (1980). Reconciling race and class differences in self-reported and official estimates of delinquency. *Am. Sociol. Rev.* 45(1): 95–110.

Elliott, D. S., and Huizinga, D. (1983). Social class and delinquent behavior in a national youth panel: 1976–1980. *Criminology* 21: 149–177.

Elliott, D. S., and Voss, H. L. (1974). *Delinquency and Dropout*, D. C. Heath, Lexington, Mass.

Elliott, D. S., Huizinga, D., and Ageton, S. S. (1985). *Explaining Delinquency and Drug Use*, Sage, Beverly Hills.

Elmhorn, K. (1965). Study in self reported delinquency among school children in Stockholm, In Christiansen, K. O. (ed.), *Scandinavian Studies in Criminology, Vol. 1*, Tavistock, London, pp. 117–146.

Erickson, M., and Empey, L. T. (1963). Court records, undetected delinquency and decision-making. *J. Crim. Law Criminol. Police Sci.* 54: 456–469.

Farrington, D. P. (1973). Self-reports of deviant behavior: Predictive and stable? *J. Crim. Law Criminol.* 64: 99–110.

Farrington, D. P. (1977). The effects of public labelling. *Br. J. Criminol.* 17: 112–125.

Fuller, E. L., and Hammerle, W. J. (1966). Robustness of the maximum likelihood estimation procedure in factor analysis. *Psychometrika* 31: 255–266.

Garofalo, J., and Hindelang, M. J. (1977). *An Introduction to the National Crime Survey*, U.S. Government Printing Office, Washington, D.C.

Gibson, H. B., Morrison, S., and West, D. J. (1970). The confession of known offenses in response to a self-reported delinquency schedule. *Br. J. Criminol.* 10: 277–280.

Gold, M. (1970). *Delinquent Behavior in an American City*, Wadsworth, Belmont, Calif.

Gold, M., and Reimer, D. J. (1975). Changing patterns of delinquent behavior among Americans 13 through 16 years old: 1967–1972. *Crime Delinq. Lit.* 7: 483–517.

Gould, L. C. (1969). Who defines delinquency: A comparison of self-reported and officially reported incidences of delinquency for three racial groups. *Soc. Problems* 16: 325–336.

Hackler, J. C., and Lautt, M. (1969). Systematic bias in measuring self-reported delinquency. *Can. Rev. Sociol. Anthropol.* 6: 92–106.

Hardt, R. H., and Bodine, G. F. (1965). *Development of Self-Report Instruments in Delinquency Research*, Youth Development Center, Syracuse University, Syracuse, N.Y.

Hardt, R. H., and Peterson-Hardt, S. (1977). On determining the quality of the delinquency self-report method. *J. Res. Crime Delinq.* 14: 247–261.

Hathaway, R. S., Monachesi, E. D., and Young, L. A. (1960). Delinquency rates and personality. *J. Crim. Law Criminol. Police Sci.* 50: 433–440.

Helmstadter, G. C. (1970). *Research Concepts in Human Behavior*, Appleton-Century-Crofts, New York.

Hindelang, M. J., Hirschi, T., and Weis, J. G. (1979). Correlates of delinquency: The illusion of discrepancy between self-report and official measures. *Am. Sociol. Rev.* 44: 995–1014.

Hindelang, M. J., Hirschi, T., and Weis, J. G. (1981). *Measuring Delinquency*, Sage, Beverly Hills.

Hirschi, T. (1969). *Causes of Delinquency*, University of California Press, Berkeley.

Huba, G. J., and Harlow, L. L. (1983). Comparison of maximum likelihood generalized least squares, ordinary least squares and asymptotically distribution free parameter estimates in drug abuse latent variable causal models. *J. Drug. Educ.* 13: 387–404.

Huizinga, D., and Elliott, D. S. (1983). A preliminary examination of the reliability and validity of the national youth survey self-reported delinquency indices, National Youth Survey Project Report 27, Behavioral Research Institute, Boulder, Colo.

Huizinga, D. and Elliot, D. S. (1984). Self-reported measures of delinquency and crime: Methodological issues and comparative findings. National Youth Survey Project Report 30, Behavorial Research Institute, Boulder, Colo.

Jensen, G. F., Erickson, M. L., and Gibbs, J. P. (1978). Perceived risk of punishment and self-reported delinquency. *Soc. Forces* 57: 57–78.

Jessor, R., Graves, T. D., Hanson, R. C., and Jessor, S. L. (1968). *Society, Personality and Deviant Behavior: A Study of a Tri-Ethnic Community*, Holt, Rinehart and Winston, New York.

Joreskog, K. (1970). A general method for analysis of covariance structures. *Biometrika* 57: 239–251.

Joreskog, K. G., and Sorbom, D. (1978). *LISREL IV: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*, National Educational Resources, Chicago.

Kelley, T. L. (1927). *Interpretation of Educational Measurements*, World Book, New York.

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrica* 2: 151–160.

Kulik, J. A., Stein, K. B., and Sarbin, T. R. (1968). Disclosure of delinquent behavior under conditions of anonymity and non-anonymity. *J. Consult. Clin. Psychol.* 32: 506–509.

Lab, S. P., and Allen, R. B. (1981). Issues in the use of self-report and official crime data. Paper presented at the 1981 meetings of the American Society of Criminology, Washington, D.C.

Marsden, M. E., and Meade, A. C. (1981). The extent, distribution and timing of delinquency: Implications of alternative defining criteria and approaches to measurement. In Meade, A. C. (ed.) *Youth and Society: Studies of Adolescent Deviance*, Institute for Juvenile Research, Chicago.

Nunally, J. C. (1972). *Educational Measurement and Evaluation*, McGraw-Hill, New York.

Nye, F. I., and Short, Jr., J. F. (1957). Scaling delinquent behavior. *Am. Sociol. Rev.* 22: 326–331.

Olsson, V. (1979). On the robustness of factor analysis against crude classification of the observation. *Multivar. Behav. Res.* 14: 485–500.

Paternoster, R. (1978). *The Labelling Effects of Police Apprehension: Identity, Exclusion, and Secondary Deviance*, Unpublished Ph.D dissertation, Florida State University, Tallahassee.

Patterson, G. R., and Loeber, R. (1982). The understanding and prediction of delinquent child behavior. Research proposal to NIMH. Oregon Social Learning Center, Eugene.

Petersillia, J. (1978). Validity of criminality data derived from personal interviews. In Wellford, C. (ed.), *Quantitative Studies in Criminology*, Sage, Beverly Hills.

Reiss, A. J., Jr. (1975). Inappropriate theories and inadequate methods as policy plaques: self-reported delinquency and the law. In Demèrath, N. J., III, *et al.* (eds.), *Social Policy and Sociology*, Academic Press, New York.

Rojek, D. G. (1983). Social status and delinquency: Do self-reports and official reports match. In Waldo, G. P. (ed.), *Measurement Issues in Criminal Justice*, Sage, Beverly Hills.

Sampson, R. J. (1985). Sex differences in self-reported delinquency and official records: A multiple group structural modeling approach. *J. Quant. Criminol.* 1: 345–366.

Short, J. F., Jr., and Nye, F. I. (1957). Reported behavior as a criterion of deviant behavior. *Soc. Problems* 5: 207–213.

Short, J. F., Jr. and Nye, F. I. (1958). Extent of unrecorded juvenile delinquency: tentative conclusions. *J. Crim. Law and Criminol.* 49: 296–302.

Solnick, J. V., Braukmann, C. J., Bedlington, M. M., Kirigin, K. A., and Wolf, M. M. (1981). Parent-youth interaction and delinquency in group homes. *J. Abnorm. Child Psychol.* 9: 107–119.

Stein, K. B., Vadum, A. C., and Sarbin, T. R. (1970). Socialization and delinquency: A study of false negatives and false positives in prediction. *Psychol. Rec.* 20: 353–354.

Thorndike, R. L. (1951). Reliability. In Linquest, E. F. (ed.), *Educational Measurement*, American Council of Education, Washington, D.C.

Voss, H. L. (1963). Ethnic differentials in delinquency in Honolulu. *J. Crim. Law Criminol.* 54: 322–327.

Wyner, G. A. (1981). Response errors in self-reported number of arrests. In Bohrnstedt and Borgatta (eds.), *Social Measurement*, Sage, Beverly Hills.