The Self-Report Method for Measuring Delinquency and Crime

by Terence P. Thornberry and Marvin D. Krohn

ABSTRACT

The self-report technique is one of three major ways of measuring involvement in delinquent and criminal behavior. The basic approach of the self-report method is to ask individuals if they have engaged in delinquent or criminal behavior, and if so, how often they have done so. In this chapter, we review the origins of the self-report method in the 1950s, the growth and refinement of this measurement technique since then, and its role in criminological research, especially longitudinal research on the etiology of delinquent and criminal behavior. Particular attention is paid to assessing the reliability and validity of self-reported measures of delinquency. We also discuss specialized data collection methods, such as random response techniques and audio assisted computer-based interviewing, that have the potential to increase the accuracy of responses. Overall, we conclude that the psychometric quality of the self-report method has increased considerably since its inception in the 1950s. Although there is much room for continued improvement, self-report data appear acceptably valid and reliable for most research purposes.

Terence P. Thornberry, Ph.D., is a Professor with, and former Dean of, the School of Criminal Justice, University at Albany, State University of New York. Marvin D. Krohn, Ph.D., is a Professor in the Department of Sociology, University of Albany, State University of New York.

The development and widespread use of the self-report method of collecting data on delinquent and criminal behavior is one of the most important innovations in criminological research in the 20th century. Currently, this method of data collection is used extensively both within the United States and abroad (Klein 1989). Because of its common use, we often lose sight of the major impact that self-report studies have had on the research concerning the distribution and patterns of crime and delinquency, the etiology of juvenile delinquency, and the juvenile justice system, including the police and courts.

Reliance on official sources introduces layers of potential bias between the actual behavior and the data.

Thorsten Sellin made the simple but critically important observation that "the value of a crime rate for index purposes decreases as the distance from the crime itself in terms of procedure increases" (1931, 337). Thus, prison data are less useful than court or police data as a measure of actual delinquent or criminal behavior because they are generated not only by the behavior of the perpetrators of offenses, but also by the behavior of police and court officials. Moreover, the reactions of the juvenile and criminal justice systems often rely on information from victims or witnesses of crime. A substantial amount of crime is not reported, but even many crimes reported or brought to the attention of law enforcement agents are not officially recorded. Thus, reliance on official sources, such as the Uniform Crime Reports (UCR) or the National Prison Statistics, introduces layers of potential bias between the actual behavior and the data. Yet, throughout the first half of this century, our understanding of the behavior of criminals and those who reacted to crime was based almost entirely on official data.

Although researchers were aware of many of these limitations, the dilemma they faced was how to obtain information closer to the source of criminal and delinquent behavior. Observing the behavior taking place would be one method, but given the illegal nature of the behavior and the potential consequences if caught, participants in crime and delinquency are reluctant to have their behavior observed. Even when observational studies were conducted, for example, in studies of gangs (e.g., Thrasher 1927), researchers could only observe a very small portion of crime that took place. Hence, although these studies generated theoretical ideas about why and how crimes took place, they had limited utility in describing the distribution and patterns of criminal behavior.

If one could not observe the behavior taking place, self-reports of delinquent and criminal behavior would be the nearest data source to the actual behavior. There was great skepticism about whether respondents would agree to tell researchers about their participation in illegal behaviors. However, early studies

(Porterfield 1943; Wallerstein and Wylie 1947) found that not only were respondents willing to self-report their delinquency and criminal behavior, they did so in surprising numbers.

The self-report methodology has become much more sophisticated in design, making it more reliable and valid and extending its applicability to a myriad of issues.

Since those early studies, the self-report methodology has become much more sophisticated in design, making it more reliable and valid and extending its applicability to a myriad of issues. These developments include the use of inventories with a wide array of delinquency items incorporating serious offenses; the use of open-ended frequency response sets instead of a relatively small number of categories; and the use of followup questions to eliminate trivial, and perhaps not criminal, acts. Much work has been done on improving the reliability and validity of self-reports, including specialized techniques to enhance the quality of self-report data. The use of self-report surveys within the context of longitudinal designs has given rise to other concerns that are not as problematic in cross-sectional research, such as construct continuity and testing or panel effects.

These developments have made self-report studies an integral part of the way crime and delinquency is studied. In this chapter, we review the history of the self-report methodology, assess the psychometric properties of self-report instruments, discuss the innovative ways in which the technique has been improved, examine the particular problems in using the technique within longitudinal designs, and suggest some future directions for the application of self-reports.

Historical Perspective

Early studies on delinquency and crime in America relied on official sources of data, such as police, court, and prison records. With these data, criminologists mapped the geography of crime (Park, Burgess, and McKenzie 1928; Shaw and McKay 1942) and, to the extent possible, identified the sociodemographic characteristics of delinquents and criminals. The data indicated that crime was disproportionately located in disadvantaged areas of the city and that those convicted of crime were more likely to be of lower class status and to be minority group members.

Although relying on official sources of data to make such generalizations, many scholars recognized that these data were not ideal for the task (Merton 1938; Sutherland 1939) because they did not tap "hidden delinquency" that

constituted the "dark figure of crime" (Gibbons 1979). An early study by Robison (1936) found that estimates of the number of delinquents doubled when they included those referred to unofficial agencies rather than sent through the Children's Court. Moreover, she reported that social status characteristics, including race and religion, seemed to be related to where children were referred. Robison concluded that "court figures alone are not only insufficient, but also misleading" (p. 76). Similar conclusions were reached by Murphy, Shirley, and Witmer (1946), after analyzing caseworker records of boys brought to the juvenile court. They found that less than 1.5 percent of law violations in the caseworker reports had resulted in official complaints.

Gibbons (1979) credits Edwin Sutherland for providing the impetus for self-report studies. Sutherland's (1949) landmark work on white-collar crime provided what Gibbons (p. 81) characterizes as the first important challenge to the prevailing wisdom that individuals from favored social backgrounds were unlikely to break the law. The apparent discrepancy between reports relying on official data about "street crimes" and Sutherland's observations about crime among the upper classes led criminologists to seek alternative means of measuring crime.

Austin Porterfield (1943, 1946) provided the first published results from a self-report survey on crime. Porterfield analyzed the juvenile court records of 2,049 delinquents from the Fort Worth, Texas, area and identified 55 offenses for which they had been adjudicated delinquent. He then surveyed 200 men and 137 women from three colleges in northern Texas to determine if and how frequently they had committed any of the 55 offenses. He found that every one of the college students had committed at least one of these offenses. The offenses committed by the college students were as serious as those committed by the adjudicated delinquents (although not as frequent), yet few of the college students had come into contact with legal authorities.

Inspired by Porterfield's findings, Wallerstein and Wylie (1947) sampled a group of 1,698 adult men and women and examined self-reports of their delinquent behavior committed before the age of 16. They mailed questionnaires containing 49 offenses to their sample. Almost all reported committing at least one delinquent act, and 64 percent of the men and 29 percent of the women had committed at least 1 of the 14 felonies included on their checklist.

The Porterfield and the Wallerstein and Wylie studies are methodologically unsophisticated. Evaluated on criteria used today, they are problematic in terms of sample representivity, selection of delinquency items, failure to examine the reliability and validity of these items, and reliance on descriptive analysis to examine poorly stated hypotheses. They are still landmark studies in the history

of self-report methodology, however, because they not only alerted criminologists to the existence of extensive hidden delinquency, but demonstrated a methodology for measuring such behavior.

Although the contributions of Porterfield and Wallerstein and Wylie are significant developments in the self-report methodology, the work of James Short and F. Ivan Nye (1957, 1958) "revolutionized ideas about the feasibility of using survey procedures with a hitherto taboo topic" and changed the thinking about delinquent behavior itself (Hindelang, Hirschi, and Weis, 1981, 23). What distinguishes Short and Nye's research from previous self-report methods is their attention to methodological issues—such as scale construction, reliability and validity, and sampling—and their explicit focus on the substantive relationship between social class and delinquent behavior.

Short and Nye collected self-report data from high school students in three Western communities varying in population from 10,000 to 40,000; from three Midwestern communities varying across rural, rural-urban fringe, and suburban areas; and from a training school for delinquents in a Western State. A 21-item list of criminal and antisocial behaviors was used to measure delinquency although most of their analyses employed a scale composed of a subset of only 7 items. Focusing on the relationship between delinquent behavior and the socioeconomic status (SES) of the adolescents' parents, Nye, Short, and Olson (1958) found that, among the different SES groups, relatively few differences in delinquent behavior were statistically significant.

Short and Nye's work stimulated much interest in both the use of the self-report methodology and the substantive issue concerning the relationship between some measure of social status (socioeconomic status, ethnicity, race) and delinquent behavior. The failure to find a relationship between social status and delinquency challenged prevailing theories built on the assumption that an inverse relationship did in fact exist, and suggested that the juvenile justice system might be using extralegal factors in making decisions concerning juveniles who misbehave. A number of studies in the late 1950s and early 1960s used self-reports to examine the relationship between social status and delinquent behavior (Akers 1964; Clark and Wenninger 1962; Dentler and Monroe 1961; Empey and Erickson 1966; Erickson and Empey 1963; Gold 1966; Slocum and Stone 1963; Vaz 1966; Voss 1966). These studies advanced the use of the self-report method by applying it to different, more ethnically diverse populations (Gold 1966; Clark and Wenninger 1962; Voss 1966), attending to issues concerning validity and reliability (Gold 1966; Clark and Tifft 1966; Dentler and Monroe 1961), and constructing measures of delinquency that specifically addressed issues regarding offense seriousness and frequency (Gold 1966). These studies found that although most juveniles engaged in some delinquency,

relatively few committed serious delinquency repeatedly. For example, Gold (1966) found that 88 percent of his sample committed one or more delinquent acts, but only 6 percent of the boys and none of the girls committed armed robbery. With few exceptions, these studies supported Short and Nye's general conclusion that if there were any statistically significant relationship between measures of social status and self-reported delinquent behavior, it was weak and did not mirror the findings of studies using official data sources.

During the 1960s, researchers began to recognize the true potential of the self-report methodology. By including questions about other aspects of adolescent life with a delinquency scale in the same questionnaire, researchers could explore etiological issues. Theoretically interesting issues concerning the family (Nye and Olson 1958; Dentler and Monroe 1961; Voss 1964; Stanfield 1966; Gold 1970), peers (Short 1957; Voss 1964; Reiss and Rhodes 1964; Matthews 1968; Erickson and Empey 1963; Gold 1970), and school (Reiss and Rhodes 1963; Elliott 1966; Kelly 1974; Polk 1969; Gold 1970) emerged as the central focus of self-report studies. The potential of the self-report methodology in examining etiological theories of delinquency was perhaps best displayed in Travis Hirschi's (1969) *Causes of Delinquency.*

The use of self-report studies to examine theoretical issues continued throughout the 1970s. In addition to several partial replications of Hirschi's arguments (Conger 1976; Hepburn 1976; Hindelang 1973; Jensen and Eve 1976), other theoretical perspectives such as social learning theory (Akers et al. 1979), self-concept theory (Jensen 1973; Kaplan 1972), strain theory (Elliott and Voss 1974; Johnson 1979), and deterrence theory (Waldo and Chiricos 1972; Silberman 1976; Jensen, Erickson, and Gibbs 1978; Anderson, Chiricos, and Waldo 1977) were evaluated using data from self-report surveys.

Another development during this period of time was the introduction of national surveys of delinquency and drug use. Williams and Gold (1972) conducted the first nationwide survey, with a probability sample of 847 boys and girls who were from 13 to16 years of age. Among the issues examined was the relationship between social status characteristics and delinquent behavior, for which they found little support.

One of the larger undertakings on a national level is the National Youth Survey (NYS), conducted by Elliott and colleagues (Elliott, Huizinga, and Ageton 1985). NYS began in 1976 surveying a national probability sample of 1,725 youths ages 11 through 17. The survey design corrected a number of methodological deficiencies of prior self-report studies and has been greatly instrumental in improving measurement of self-reported delinquent behavior. NYS is also noteworthy because it is a panel design, having followed the original respondents into their thirties.

Monitoring the Future (Johnston, O'Malley, and Bachman 1996) is a national survey on drug use that has been conducted since 1975. It began as an inschool survey of a nationally representative sample of high school seniors and has since expanded to include 8th and 10th grade students. It also conducts follow-up surveys by mail on representative subsamples of respondents from the previous 12th grade sample. Its findings have been the primary source of information on the trends in drug use among youths in this country.

Despite the expanding applications of this methodology, questions remained about just what self-report instruments measure. The discrepancy in findings regarding the social status-delinquency relationship, based on self-report versus official (and victim) data, continued to perplex scholars. Self-reports have come under increasing criticism on a number of counts, including sample selection and the selection of delinquency items. Gwynn Nettler (1978, 98) stated that "an evaluation of these unofficial ways of counting crime does not fulfill the promise that they would provide a better enumeration of offensive activity." Gibbons (1979, 84) was even more critical in his summary evaluation, stating: "The burst of energy devoted to self-report studies of delinquency has apparently been exhausted. This work constituted a criminological fad that has waned, probably because such studies have not fulfilled their early promise."

Two studies were particularly instrumental in pointing to the flaws in self-report measures. Hindelang, Hirschi, and Weis (1979) illustrated the problems encountered when comparing results from studies using self-reports with those using official data. They employed a third source of data on crime—victimization data—and compared characteristics of offenders from the three data sources. They concluded that there is more similarity in those characteristics when comparing victimization data with UCR data than between self-report data and the other two sources. They argued that self-report instruments do not include many of the more serious crimes for which people are arrested, which are included in victimization surveys. Thus, self-reports tap a different domain of behaviors than either of the other two sources, and discrepancies in observed relationships when using self-reports should not be surprising. The differential domain of crime tapped by early self-report measures could also explain the discrepancy in findings regarding the association between social status and delinquency.

Elliott and Ageton (1980) also explored the methodological shortcomings of self-reports. They observed that a relatively small number of youths commit a disproportionate number of serious offenses. However, most early self-report instruments truncate the response categories for the frequency of offenses and do not include serious offenses in the inventory at all. In addition, many of the samples did not include enough high-rate offenders to clearly distinguish them

The development of instruments to better measure serious and very frequent offenses and the suggestion to acquire data from high-risk samples coincided with a substantive change in the 1980s in the focus of much criminological work on the etiology of offenders. The identification of a relatively small group of offenders who commit a disproportionate amount of crime and delinquency led to a call to focus research efforts on the "chronic" or "career" criminals.

from other delinquents. By allowing respondents to report the number of delinquent acts they committed rather than specifying an upper limit (e.g., 10 or more), and by focusing on high-rate offenders, Elliott and Ageton found relationships between engaging in serious delinquent behavior and measures of social status that are more consistent with results from studies using official data.

The Hindelang, Hirschi, and Weis (1979) and the Elliott and Ageton (1980) studies both suggested designing self-report studies so that they would acquire sufficient data from those high-rate, serious offenders most likely to come to the attention of authorities. They also suggested a number of changes in the way in which we measure self-report data to reflect the fact that some offenders contribute disproportionately to the rate of serious and violent delinquent acts.

The development of instruments to better measure serious and very frequent offenses and the suggestion to acquire data from high-risk samples coincided with a substantive change in the 1980s in the focus of much criminological work on the etiology of offenders. The identification of a relatively small group of offenders who commit a disproportionate amount of crime and delinquency led to a call to focus research efforts on the "chronic" or "career" criminals (Wolfgang, Figlio, and Sellin 1972; Blumstein et al. 1986). Blumstein and his colleagues' observation that we need to study the career of criminals—including early precursors of delinquency, maintenance through the adolescent years, and later consequences during the adult years—was particularly important in recognizing the need for examining the lifecourse development of high-risk offenders with self-report methodology.

The self-report methodology continues to advance, both in terms of its application to new substantive areas and the improvement of its design. Gibbons' (1979) suggestion that self-reports were just a fad whose use was likely to disappear is clearly wrong. Rather, with improvements in question design, administration

technique, reliability and validity, and sample selection, this technique is being used in the most innovative research on crime and delinquency. The sections that follow describe the key methodological developments that have made such applications possible.

Development of the Self-Report Method

Self-report measures of delinquent behavior have advanced remarkably in the 30-odd years since their introduction (see Thornberry 1989, 347–350). The prototypical "early" self-reported delinquency scale was developed by Short and Nye (1957; Nye and Short 1957). The inventory included 21 items, but most analyses were limited to 9, and in many cases 7, items that formed a Guttman scale of delinquency. The scale items refer to trivial forms of delinquent behavior—for example, there is no item measuring violent behavior, and the most serious theft item concerns stealing things worth less than $2. Moreover, subjects were only afforded a four-category response set ("no," "once or twice," "several times," and "often"), and the reference period for the instrument ("since you began grade school") was both long and somewhat varied for these high school respondents.

Since its introduction by Short and Nye, considerable attention has been paid to the development and improvement of the psychometric properties of the self-report method. The most sophisticated and influential work was done by Elliott and his colleagues (Elliott and Ageton 1980; Elliott, Huizinga, and Ageton 1985; Huizinga and Elliott 1986) and by Hindelang, Hirschi, and Weis (1979, 1981). From their work, a set of characteristics for acceptable (i.e., reasonably valid and reliable) self-report scales has emerged. Four of the most salient characteristics are the inclusion of a wide array of delinquency items, serious offenses, frequency response sets, and followup questions.

Inclusion of a wide array of delinquency items

The domain of delinquency and crime covers a wide range of behaviors, from truancy and running away from home to aggravated assault and homicide. If the general domain of delinquent and criminal behavior is to be represented in a self-report scale, it is necessary for the scale to cover that same wide array of human activity. Simply asking about a handful of these behaviors does not accurately represent the theoretical construct of crime. In addition, empirical evidence suggests that crime does not have a clear unidimensional structure that would facilitate the sampling of a few items from a theoretically large pool to represent adequately the entire domain.

These considerations suggest that an adequate self-report scale for delinquency will be relatively lengthy. A large number of individual items are required to represent the entire domain of delinquent behavior, to represent each of its subdomains, and to ensure that each subdomain—e.g., violence, drug use—is itself adequately represented.

It is essential that a general self-reported delinquency scale tap serious as well as less serious behaviors. Failure to do so misrepresents the domain of delinquency and contaminates comparisons with other data sources.

Inclusion of serious offenses

Early self-report scales tended to ignore serious criminal and delinquent events and concentrated almost exclusively on minor forms of delinquency. As a result, only certain subdomains of delinquency, such as petty theft and status offenses, were measured, even though theoretical interest and conclusory statements focused on juvenile delinquency broadly construed.

It is essential that a general self-reported delinquency scale tap serious as well as less serious behaviors. Failure to do so misrepresents the domain of delinquency and contaminates comparisons with other data sources. In addition, it misrepresents the dependent variable of many delinquency theories that set out to explain serious, repetitive delinquency (e.g., Elliott, Huizinga, and Ageton 1985; Thornberry 1987).

Inclusion of frequency response sets

Many self-report studies rely on response sets with a relatively small number of categories, which tend to censor high-frequency responses. For example, Short and Nye (1957) used a four-point response, with the most extreme category being "often." As a result, a respondent who committed a theft 5 times would be treated the same as a respondent who committed the act 50 times. Aggregated over many items, the use of limited response sets has the consequence of lumping together occasional and high-rate delinquents rather than discriminating between these behaviorally different groups.

When frequency responses are used, a number of specific indicators can be constructed from the basic inventory. The three most common are prevalence, incidence, and variety. Prevalence refers to the proportion or percentage of *people* who report involvement in delinquency—the percentage of the sample who answer "yes." Incidence (also called frequency) refers to the number of delinquent *acts* reported—the total number of times the person reports committing different acts. Variety refers to the number of different *types* of delinquency

reported by the person. For example, if the index has six items (offense types), the variety score can vary from 0 to 6. Each of these basic measures can also be created for different time periods. One of the most common is a "lifetime" or "ever" measure; for example, an "ever-prevalence" measure of marijuana use would indicate the percentage of the people who had ever used marijuana. Most measures are time-limited; for example, referring to offenses committed during the past year or past 6 months.

Inclusion of followup questions

Self-report questions seem to have an inherent tendency to elicit reports of trivial acts that are very unlikely to elicit official reactions, or even acts that are not violations of the law. This occurs more frequently with less serious offenses but also affects responses to serious offenses. For example, respondents have listed such pranks as hiding a classmate's books in the respondent's locker between classes as "theft," or roughhousing between siblings as "serious assault."

Some effort must be made to adjust or censor the data to remove these events if the delinquency of respondents is to be reflected properly and if the rank order of respondents with respect to delinquency is to be portrayed properly. Two strategies are generally available. First, one can ask a series of followup questions designed to elicit more information about the event, such as the value of property stolen, the extent of injury to the victim, and the like. Second, one can use an open-ended question asking the respondent to describe the event, and then probe to obtain information necessary to classify the act. Both strategies have been used with some success.

Summary

Recent examinations of the self-report method have identified a number of shortcomings in earlier scales and suggested ways of improving the technique's psychometric properties. The more salient suggestions include the following:

- Self-report scales should include a wide range of delinquent acts so that the general domain of delinquency, as well as its various subdomains, is adequately represented.
- The scale should include serious as well as minor acts.
- A frequency scale should be used to record responses so that high-rate offenders can be isolated from low-rate offenders.
- Extremely trivial, nonactionable acts that are reported should be identified and eliminated from the data.

These procedures improve our ability to identify delinquents and discriminate among different types of delinquents. Thus, they are likely to improve the validity, and to some extent the reliability, of self-report scales. These are clearly desirable qualities.

To gain these desirable qualities, however, requires a considerable expansion of the self-report schedule. This can be illustrated by describing the major components of the index currently being used in the Rochester Youth Development Study (Smith and Thornberry 1995), as well as in the other two projects of the Program of Research on the Causes and Correlates on Delinquency (see Browning et al. 1999). The inventory includes 32 items tapping general delinquency and 12 tapping drug use, for a total of 44 items. For each of these items, the subjects are asked if they ever committed the act, and if so, if they had committed the act in the past 6 months. For the most serious of each type of delinquency reported in the past 6 months, subjects are asked to describe the event by responding to the question: "Could you tell me what you did?" If that open-ended question does not elicit the information needed to describe the event adequately, a series of probe questions, which vary from 2 to 14 probes depending on the offense, are asked.

Although most of these specific questions are skipped for most subjects, since delinquency remains a rare event, this approach to measuring self-reported delinquency is a far cry from the initial method of using a few categories to respond to a small number of trivial delinquencies, with no followup items. In the remaining pages, we evaluate this approach for measuring delinquent and criminal behavior.

Reliability and Validity

For any measure to be scientifically worthwhile, it must possess both reliability and validity. Reliability is the extent to which a measuring procedure yields the same result on repeated trials. For example, if a bathroom scale were reliable, it would yield the same reading of your weight if you got on and off that scale 10 times in a row. If it were unreliable, the reading of your weight would vary somewhat, even though your true weight would not change in the space of time it would take you to get on and off the scale 10 times.

No measure is absolutely, perfectly reliable. Repeated use of a measuring instrument will always produce some variation from one application to another. That variation can be very slight to quite large. So the central question in assessing the reliability of self-reported delinquency measures is not whether the measure is reliable but *how* reliable it is; reliability is always a matter of degree.

Validity is a much more abstract notion than is reliability. The best definition of validity is something like as follows: A measure is valid to the extent to which it measures the concept you set out to measure, and nothing else. Whereas reliability focuses on a particular property of the measure—namely, its stability over repeated uses—validity concerns the crucial relationship between the theoretical concept you are attempting to measure and what you actually measure.

For example, let us say we are interested in measuring an individual's actual involvement in criminal behavior over the past year. During that time period, there are some people who never commit a crime, while others do. Our measure would be completely valid if it accurately identified as criminals all of the people who did commit crimes and also accurately identified as noncriminals all of the people who did not commit crimes. That is, the measure would accurately reflect our theoretical concept—involvement in crime during the past year. As with reliability, the assessment of validity is not an either/or proposition. There are no perfectly valid measures, but some are more valid than others.

Even though both validity and reliability are always a matter of degree, we often see statements that assert that a particular measure "is valid" or that another measure "is unreliable." That is a shorthand way of saying that the first measure possesses sufficient validity for the analytic purpose at hand, but the second measure does not.

The relationship between validity and reliability is asymmetrical. A particular measure can be highly reliable but have little or no validity. For example, a bathroom scale could be very consistent but the calibration could be off by 50 pounds. In this case, we have very reliable measures, but every one of them would be wrong—too low or too high by 50 pounds.

In contrast, if a measure is valid, it is also reliable. Since a valid measure is one that accurately measures what it sets out to measure, by definition it must be consistent, yielding the same estimates time after time.

All scientifically adequate measures must possess high levels of both validity and reliability. We now turn to an assessment of whether self-reported measures of delinquency are psychometrically acceptable.

Assessing reliability

There are two classic ways of assessing the reliability of social science measures: "test-retest" reliability and internal consistency. Huizinga and Elliott (1986) make a convincing case that the test-retest approach is fundamentally more appropriate for assessing self-reported measures of delinquency.

Internal consistency

Internal consistency simply means that multiple items measuring the same underlying concept should be highly intercorrelated. This can be illustrated by returning to our example of weight and bathroom scales. If you weighed yourself on 15 different bathroom scales, you would get slightly different readings on each, but the answers should be highly correlated; that is, your weight should be the same. If one scale differed substantially from the others, you would likely throw it out as being inaccurate. The same approach is used in assessing attitudes and opinions. Many questions tapping the same concept are asked, and the expectation is that the answers on these items will be highly intercorrelated. For example, in assessing attachment to parents, one could ask an adolescent to respond to statements such as "I think my mother is really terrific" and "I have a great deal of respect for my mother." It is reasonable to expect that an adolescent strongly attached to his or her mother would respond positively to both statements, and an adolescent who is very alienated from his or her mother would respond negatively. That is, across these and similar items, responses would be highly correlated.

This expectation is much less reasonable for behavioral inventories such as self-report measures of delinquency, however. Current self-report measures typically include 30 or 40 items measuring a wide array of delinquent acts. Just because someone reports being truant is no reason to expect he or she would be involved in theft or vandalism. Similarly, if someone reports being involved in assaultive behavior, there is no reason to assume that he or she has been involved in drug sales or loitering. Indeed, given the relative rarity of involvement in delinquent acts, it is very likely that most people will respond negatively to most items and affirmatively to only a few. This is especially the case if we are asking about the past year or the past 6 months. Because of this, there is no strong underlying expectation that the responses will be highly intercorrelated. Therefore, an internal consistency approach to assessing reliability is not particularly appropriate.

Test-retest reliability

Thus, we will focus on the test-retest method of assessing reliability. This approach is quite straightforward. A sample of respondents is administered a self-reported delinquency inventory (the test); then, after a short interval, the same inventory is readministered (the retest). In doing this, the same questions and the same reference period should be used at both times.

It is also important to pay attention to the time lag between the test and the retest. If it is too short, answers to the retest likely will be a function of memory;

respondents are likely to remember what they said the first time and simply repeat it. If so, estimates of reliability would be inflated. On the other hand, if the time period between the test and the retest is too long, responses to the retest would probably be less accurate than those to the test simply because of memory decay. In this case, the reliability of the scale would be underestimated. There is no hard and fast rule for assessing the appropriateness of this lag, but the optimal time lag appears to be in the range of 1 to 4 weeks.

The simplest way of deriving a reliability coefficient for the test-retest method is to correlate the first and second sets of responses. The correlations should be reasonably high, preferably in the range of 0.70 or greater.

A number of studies have assessed the test-retest reliability of self-reported delinquency measures. In general, the results of these studies indicate that these measures are acceptably reliable. The reliability coefficients vary somewhat, depending on the number and types of delinquent acts included in the index and the scoring procedures used (e.g., simple frequencies or ever-variety scores). But scores well above 0.80 are common. In summarizing previous literature in this area, Huizinga and Elliott (1986, 300) stated:

> 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 hour to over 2 months (Kulik et al., 1968; Belson, 1968; Hindelang et al., 1981; [Braukmann] et al., 1979; Patterson and Loeber, 1982; [Skolnick] et al., 1981; Clark and [Tifft], 1966; Broder and Zimmerman, 1978).

Perhaps the most comprehensive assessment of the psychometric properties of the self-report method was conducted by Hindelang, Hirschi, and Weis (1981). Their self-report inventory was quite extensive, consisting of 69 items divided into the following major subindexes: official contacts, serious crimes, delinquency, drugs, and school and family offenses. To see whether the method of administration matters, some subjects were interviewed and others responded on a questionnaire. For both types of administration, some subjects responded anonymously and others were asked to provide their names.

To maximize variation in the level of delinquency, the study sample was selected from three different populations in Seattle, Washington. The first consisted of students without an official record of delinquency attending Seattle schools. The second consisted of adolescents with a police record but no court record, and the third group consisted of adolescents with a juvenile court record. Within these three major strata, subjects were further stratified by gender, race, and, among the whites, socioeconomic status.

Several self-reported measures of delinquency were created. The major ones include an ever-variety score (the number of delinquent acts the respondents report ever having committed), a last year variety score (the same type of measure for the past year), and a last year frequency score (the total number of times respondents report committing each of the delinquent acts).

As indicated earlier, internal consistency methods can be used to assess the reliability of self-reported responses. The classic way of doing so is with Cronbach's alpha. Although mindful of the limitations of internal consistency approaches, Hindelang, Hirschi, and Weis (1981) report alpha coefficients for a variety of demographic subgroups and for the ever-variety, last year variety, and last year frequency scores. The coefficients range from 0.76 to 0.93. Most of the coefficients are above 0.8, and 8 of the 18 coefficients are above 0.9.

Hindelang, Hirschi, and Weis (1981) also estimated test-retest reliabilities for these three self-report measures for each of the demographic subgroups. Unfortunately, only 45 minutes elapsed between the test and the retest, so it is quite possible that the retest responses are strongly influenced by memory effects. Nevertheless, they report substantial degrees of reliability for the self-report measures. Indeed, most of the test-retest correlations are above 0.9.

Thus, whether an internal consistency or test-retest approach is used, the Seattle data indicate a substantial degree of reliability for a basic self-reported delinquency measure. Hindelang, Hirschi, and Weis point out that reliability scores of this magnitude are higher than those typically associated with many attitudinal measures and conclude that "the overall implication is that in many of the relations examined by researchers, the delinquency dimension is more reliably measured than are many of the attitudinal dimensions studied in the research" (1981, 82).

The other major assessment of the psychometric properties of the self-report method was conducted by Huizinga and Elliott, using data taken from the well-known National Youth Survey. NYS began in 1976 with a nationally representative sample of 1,725 American youths between the ages of 11 and 17. At the fifth interview, 177 respondents were randomly selected and reinterviewed approximately 4 weeks after their initial assessment. Based on these data, Huizinga and Elliott (1986) estimated test-retest reliability scores for the general delinquency index and for several subindexes. They also estimated reliability coefficients for frequency scores and for variety scores.

The general delinquency index appears to have an acceptable level of reliability. The test-retest correlations are 0.75 for the frequency score and 0.84 for the variety score. For the various subindexes—ranging from public disorder

offenses to the much more serious index offenses—the reliabilities vary from a low of 0.52 (for the frequency measure of felony theft) to a high of 0.93 (for the frequency measure of illegal services). In total, Huizinga and Elliott (1986) report 22 estimates of test-retest reliability—across indexes and across frequency and variety scores—and the mean reliability coefficient is 0.74.

Another way of assessing the level of test-retest reliability is by estimating the percentage of the sample who changed their frequency responses by two or less. If the measure is highly reliable, one would expect few such changes. For most subindexes, there appears to be acceptable precision and reliability based on this measure. For example, for index offenses, 97 percent of the respondents changed their answers by two delinquent acts or less. Huizinga and Elliott (1986, 303) summarize these results as follows:

> Scales representing more serious, less frequently occurring offenses (index offenses, felony assault, felony theft, robbery), have the highest precision, with 96 to 100 percent agreement, followed by the less serious offenses (minor assault, minor theft, property damage), with 80 to 95 percent agreement. The public disorder and status scales have lower reliabilities (in the 40 to 70 percent agreement range), followed finally by the general SRD [self-reported delinquency] scale, which, being a composite of the other scales, not surprisingly has the lowest test-retest agreement.

Huizinga and Elliott also report little evidence of differential reliability across various subgroups. They found no consistent differences across sex, race, class, place of residence, or delinquency level in terms of test-retest reliabilities (see also Huizinga and Elliott 1983).

Summary

Overall, these studies suggest that the self-report method possesses acceptable reliability for most analytic purposes. Test-retest correlations are often 0.80 or higher, and self-reported delinquency responses are no less reliable than other social science measures. That is particularly impressive considering the sensitive nature of the topic: unreported criminal activity. Although this assessment is generally positive, it does not mean that there are no reliability problems for self-reported responses. Some subindexes have low reliabilities, and more research is needed to identify which indexes are most reliable across different samples and which are least reliable. Despite these concerns, it appears that self-reports of delinquent acts are fairly stable over time. As Hindelang, Hirschi, and Weis conclude: “If self-report measurement is flawed, it is not here, but in the area of validity” (1981, 84).

Assessing validity

Recall that validity refers to the accuracy of a measure. A measure is valid to the extent to which it accurately measures the concept that you set out to measure. There are several ways to assess validity. We will concentrate on three: content validity, construct validity, and criterion validity.

Content validity

Content validity is a subjective or logical assessment of the extent to which the measure adequately reflects the full domain, or the full content of the concept being measured. For example, if one were interested in assessing arithmetic ability among grade school children and had a test that only included questions on addition, the test would lack content validity. That is, by not containing questions to assess subtraction, multiplication, and division, the test would not measure the full domain, or the full content, of the concept of arithmetic ability.

Note that our assessment implied that we have a clear definition of what is contained in the concept of arithmetic ability. Only by knowing that arithmetic includes these four basic functions can we draw the conclusion that a test that measures only one of them is inadequate in terms of its content validity. As in all assessments of validity, content validity requires a clear theoretical definition of the concept.

To argue that a measure has content validity, we must meet the following three criteria. First, we must define the domain of the concept clearly and fully. Second, we must create questions or items to cover the whole range of the concept under investigation. And third, we must sample items or questions from that range so that the ones that appear on the test are representative of the underlying concept.

In our case, we are interested in measuring involvement in delinquency and crime. A reasonable definition of delinquency and crime is the commission of behaviors that violate criminal law and that place the individual at some risk of arrest if such a behavior were known to the police. Can we make a logical case that self-report measures of delinquency are valid in this respect?

As noted before, the earlier self-report inventories contained relatively few items to measure the full range of delinquent behavior. For example, the Short and Nye (1957) inventory only contains 21 items and most of their analysis was conducted with a 7-item index. Similarly, Hirschi's self-report measure (1969) is based on only 6 items. More importantly, the items included in these scales are clearly biased toward the minor or trivial end of the continuum. For example, Hirschi's inventory includes only one item measuring a violent crime:

"Not counting fights you may have had with a brother or sister, have you ever beaten up on anyone or hurt anyone on purpose?"

More recent self-report measures appear much better in this regard. For example, the Hindelang, Hirschi, and Weis (1981) index includes 69 items that range from status offenses, such as skipping class, to violent crimes, like serious assault and armed robbery. The NYS index (Elliott, Huizinga, and Ageton 1985) has 47 items designed to measure all but 1 (homicide) of the 8 UCR Part I offenses and 60 percent of the 21 Part II offenses, as well as offenses that juveniles are likely to commit. The self-report inventory used by the three projects of the Program of Research on the Causes and Correlates of Delinquency has 32 items measuring delinquent behavior and 12 measuring substance use. These more recent measures, although not perfect, tap into a much broader range of delinquent and criminal behavior. As a result, they appear to have reasonable content validity.

Construct validity

Construct validity refers to the extent to which the measure being validated is related in theoretically expected ways to other concepts or constructs. In our case, the key question is: Are measures of delinquency based on the self-report method correlated in expected ways with other variables?

In general, self-report measures of delinquency and crime, especially the more recent longer inventories, appear to have a high degree of construct validity. They are generally related in theoretically expected ways to basic demographic characteristics and to a host of theoretical variables drawn from various domains such as individual attributes, family structure and processes, school performance, peer relationships, and neighborhood characteristics. Hindelang, Hirschi, and Weis offer one of the clearer assessments of construct validity (1981, 127ff). They correlate a number of etiological variables with different self-report measures, collected under different conditions (e.g., interviews or questionnaires). With a few nonsystematic exceptions, the correlations are in the expected direction and of the expected magnitude.

Overall, construct validity may offer the strongest evidence for the validity of self-reported measures of delinquency and crime. Indeed, if one examines the general literature on delinquent and criminal behavior, virtually all theoretically expected relationships are actually observed for self-report measures of delinquency and crime. It is unfortunate that this approach is not used to assess validity more formally and systematically.

Criterion validity

Criterion validity "refers to the relationship between test scores and some known external criterion that adequately indicates the quantity being measured" (Huizinga and Elliott 1986, 308). There is a fundamental difficulty in assessing the criterion validity of self-reported measures of delinquency and crime and, for that matter, all measures of delinquency and crime. Namely, there is no "gold standard" against which to judge the self-report measure. That is, there is no fully accurate assessment to use as a benchmark. In contrast, to test the validity of self-reports of weight, one could ask people to self-report their weight and then weigh them on a scale, an external criterion. Given the secretiveness of criminal behavior, however, there is nothing comparable to a scale in the world of crime. As a result, the best approach is to compare different flawed measures of criminal involvement to find similar responses. The similarity of results from different measurement strategies heightens the probability that the various measures are tapping into the underlying concept of interest. Although not ideal, this is the best possible approach in this area of inquiry.

There are several ways of assessing criterion validity. One of the simplest is called "known group validity." In this approach, one compares scores for groups of people who are likely to differ in terms of their underlying involvement in delinquency. For example, one would expect the delinquency scores of seminarians to be lower than the delinquency scores of street gang members.

Over the years, a variety of group comparisons have been made to assess the validity of self-report measures. They include comparisons between individuals with and without official arrest records, between individuals convicted and not convicted of criminal offenses, and between institutionalized adolescents and high school students. In all cases, these types of comparisons indicate that the group officially involved with the juvenile justice system self-reported substantially more delinquents act than the other group. (See, for example, the work by Hirschi, 1969; Hardt and Petersen-Hardt 1977; Erickson and Empey 1963; Farrington 1973; Hindelang, Hirschi, and Weis 1981; Short and Nye 1957; Voss 1963; Kulik, Stein, and Sarbin 1968).

Although comparisons across known groups are helpful, they offer a very minimal test of criterion validity. The real issue is not whether groups differ but whether *individuals* have similar scores on the self-report measure and on other measures of criminal behavior. As mentioned previously, the basic problem in this area is there is no perfect benchmark against which to judge the self-report measures. Thus, a variety of external criteria have been used (see the discussion in Hindelang, Hirschi, and Weis 1981, 97–101). The two most common approaches are to compare self-reported delinquency scores with official arrest records and with self-reports of official arrest records.

The premise behind these comparisons is quite simple. If the measures are valid, they should produce similar scores for both the prevalence and frequency of delinquent and criminal involvement. That is, if the self-report measure identifies certain individuals as essentially nondelinquent, we should not expect to find them in official records. In contrast, if the self-report measures identify individuals as highly delinquent, we should expect both to find them in official records and to have extensive criminal histories. If this is the case, the two measures would be positively correlated and the correlation would suggest that the measures have some degree of validity. As with reliability assessment, the most sophisticated examinations of this topic have been conducted by Hindelang, Hirschi, and Weis (1981) and Huizinga and Elliott (1986).

We can begin by examining the correlation between self-reported official contacts and official measures of delinquency as presented by Hindelang, Hirschi, and Weis (1981). In this case, the correlations are quite high, ranging from 0.70 to 0.83. Correlations of this magnitude are reasonably large for this type of data.[1] Adolescents seem quite willing to self-report their involvement with the juvenile justice system.

The generally high level of concordance between self-reports of being arrested or having a police contact and having an official record has been observed in other studies as well. For example, Hardt and Petersen-Hardt (1977) found that 78 percent of the juveniles with police records self-report that they have been arrested. Similar results are reported by Hathaway, Monachesi, and Young (1960) and, for status offenses, by Rojek (1983). When convictions are examined, even higher concordance rates are reported by Blackmore (1974) and Farrington (1977).

The most important comparison presented by Hindelang, Hirschi, and Weis (1981) is between self-reported delinquent behavior and official measures of delinquency. It is important because these are *independent* measures of an individual's involvement in delinquent behavior. One is based on self-reports and one is based on official police records. Hindelang, Hirschi, and Weis present correlations using a number of different techniques for scoring the self-report measures. However, we will focus on the average correlation across these different measures and on the correlation based on the ever-variety scores, as presented in their figure 2 (1981, 113).

Overall, these correlations are reasonably high, somewhere around 0.60 for all subjects. The most important data, though, are presented for race-by-gender groups. For white and African-American females and for white males, the correlations range from 0.58 to 0.65 when the ever-variety score is used. For correlations averaged across the different self-report measures, the magnitudes

range from 0.50 to 0.60. For African-American males, however, the correlation is at best moderate. For the ever-variety self-reported delinquency score, the correlation is 0.35, and the average across the other self-reported measures is 0.30.

Putting this together leads to a somewhat mixed assessment of the validity of self-report measures based on the Seattle data. On the one hand, the overall validity of self-report data seems to be in the moderate to strong range. For the link between self-reported delinquent behavior and official measures of delinquency (the only link based on independent sources of data), the overall correlations are smaller but still acceptable. On the other hand, if we look at the issue of differential validity, there appears to be a substantial difference between African-American males and other respondents. Official measures of delinquency and self-report measures of delinquency are not correlated very highly for African-American male adolescents. It is hard to determine whether this is a problem with the self-report measures, the official measures, or both. We will return to a discussion of this issue after additional data are presented.

The majority of individuals who have been arrested self-report their delinquent behavior, and the majority of offenses they commit are also reported.

Huizinga and Elliott (1986), using data from NYS, also examine the correspondence between self-reports of delinquent behavior and official criminal histories. They recognize that there can be considerable slippage between these two sources of data, even when the same event is recorded in both datasets. For example, an adolescent can self-report a gang fight, but it may be recorded in the arrest file as disturbing the peace; an arrest for armed robbery can be self-categorized as a mugging or theft by the individual. Because of this, Huizinga and Elliott provide two levels of matching. In one, there is "a very tight match of the self-report behavior to the arrest behavior," and in the second, there is a broad match "in which any self-reported offense that could conceivably have resulted in the recorded arrest was allowed" (1986, 317). The analysis provides information on both the percentages of youths who provide tight and broad matches to their arrest records and the percentage of arrests that are matched by self-reported behavior.

As expected, there are substantial differences in results, depending on whether tight or broad matches are used. For the tight matches, almost half of the respondents (48 percent) concealed or forgot at least some of their offensive behavior, and about a third (32 percent) of all of the offenses were not reported. When the broad matches are used, however, the percentage of respondents concealing or forgetting some of their offenses drops to 36 percent, and the percent of arrest

offenses not self-reported drops to 22 percent. Although the rates of underreporting are substantial, it should also be noted that the majority of individuals who have been arrested self-report their delinquent behavior, and the majority of offenses they commit are also reported.

The reporting rates for gender, race, and social class groupings are quite comparable to the overall rates, with one exception. As with the Seattle data, African-American males substantially underreport their involvement in delinquency.

The most recent major study assessing the criterion validity of self-reported measures was conducted by Farrington and colleagues (1996), using data from the middle and oldest cohorts of the Pittsburgh Youth Study. The Pittsburgh study, one of three projects in the Program of Research on the Causes and Correlates of Delinquency, uses the same self-reported delinquency index as described earlier for the Rochester Youth Development Study. In this analysis, Farrington and colleagues classified each of the boys in the Pittsburgh study into one of four categories based on their self-reports: no delinquency, minor delinquency only, moderate delinquency only, and, finally, serious delinquency. They then used juvenile court petitions as an external criterion to assess the validity of the self-reported responses. Both concurrent and predictive validity were assessed; the former used court petitions prior to the first self-report assessment, and the latter used court petitions after the first self-report assessment.

Overall, this analysis suggests that there is a substantial degree of criterion validity for the self-report inventory used in the Program of Research on the Causes and Correlates of Delinquency. Respondents in the most serious category based on their self-report responses are significantly more likely to have juvenile court petitions, both concurrently and predictively. For example, the odds ratio of having a court petition for delinquency is about 3.0 for the respondents in the most serious category versus the other three.

Farrington and colleagues (1996) also present information on the issue of differential validity. Their results indicate that African-American males are no more or less likely to self-report delinquent behavior than are white males. With few exceptions, the odds ratios comparing self- reported measures and official court petitions are significant for both African-Americans and whites. In some cases, the odds ratios are higher for whites, and in other cases, they are higher for African-Americans.

These researchers also compared the extent to which boys with official court petitions self-reported being apprehended by the police. Overall, about two-thirds of the boys with court petitions answered in the affirmative. Moreover, there was no evidence of differential validity. Indeed, the African-American

respondents were more likely to admit being apprehended by the police than were the white respondents. Farrington and colleagues conclude that "concurrent validity for admitting offenses was higher for Caucasians but concurrent validity for admitting arrests were higher for African-Americans. There were no consistent ethnic differences in predictive validity" (1996, 509).

Other studies have also examined the concordance between self-reports of delinquent behavior and official records. For example, Elliott and Voss (1974) examined this issue in a high school sample drawn from southern California. Overall, they found that 83 percent of the arrest offenses were self-reported by the respondents, but the rate varied by offense type. In general, more serious offenses were more likely to be underreported than were minor offenses. Based on a school sample from Honolulu, Voss (1963) found that 95 percent of arrest offenses were reported in the self-report inventories.

Rather than relying on police records as the external criterion, Gold (1970) relied on reports by friends and classmates. He found that, of the respondents whose friends had said they engaged in delinquent acts, 72 percent self-reported delinquencies, 17 percent concealed their delinquent acts, and, in 11 percent of cases, the outcome was uncertain.

The previous studies have all focused on types of delinquent or criminal behavior that have no true external criterion for evaluating validity. There is an external criterion for one class of criminal behavior; namely, substance use. Physiological data—for example, from saliva or urine—can be used to independently assess recent use of various substances. The physiological data can then be compared with self-reports of substance use to assess the validity of the self-report instruments. A few examples of this approach can be offered.

We begin with a study of a minor form of deviant behavior, adolescent tobacco use. Akers and colleagues (1983) examined tobacco use among a sample of junior and senior high school students in Muscatine, Iowa. The respondents provided saliva samples, which were used to detect nicotine use by the level of salivary thiocyanate. The students also self-reported whether they smoked and how often they smoked. The self-report data had very low levels of either underreporting or overreporting tobacco use. Overall, Akers and colleagues estimated that 95 to 96 percent of the self-reported responses were accurate and valid.

The Arrestee Drug Abuse Monitoring (ADAM) program, formerly the Drug Use Forecasting (DUF) program, sponsored by the National Institute of Justice, is an ongoing assessment of the extensiveness of drug use for samples of arrestees in cities throughout the country. Individuals who have been arrested and brought to central booking stations are interviewed and asked to provide

urine specimens. Both the urine samples and the interviews are provided voluntarily; there is an 80-percent cooperation rate for urine samples and a 90-percent cooperation rate for interviews. The urine specimens are tested for 10 different drugs: cocaine, opium, marijuana, PCP, methadone, benzodiazepines, methaqualone, propoxyphene, barbiturates, and amphetamines. The arrestees also are interviewed, and some interviews include a self-reported drug use inventory. Assuming that urine samples provide a reasonably accurate estimate of actual drug use, they can be used to validate self-reported information.

DUF compares 1988 urinalysis test results for male arrestees with self-reported drug use (U.S. Department of Justice 1990, 12); the results vary considerably by type of drug. There generally is a fairly high concordance for marijuana use. For example, in New York City, 28 percent of the arrestees self-report marijuana use, and 30 percent test positive for marijuana use. Similarly, in Philadelphia, 28 percent self-report marijuana use, and 32 percent test positive. The worst comparison in this particular examination of DUF data came from Houston, where 15 percent of arrestees self-report marijuana use and 43 percent test positive.

For more serious drugs, however, underreporting is much more common. For cocaine, for example, 47 percent of New York City arrestees self-reported use, while 74 percent tested positive. Similar numbers were generated in Philadelphia, where 41 percent self-reported cocaine use, but 72 percent tested positive. Similar levels of underreporting have been observed in other cities for other hard drugs, such as heroin.

The data collected in DUF differ considerably from those collected in typical self-report surveys. The sample is limited to people just arrested, who then are asked to provide self-incriminating evidence to a research team while in a central booking station. How this setting affects the results is not entirely clear. On the one hand, individuals are likely to be reluctant to provide additional self-incriminating evidence after having just been arrested. On the other hand, if one has just been arrested for a serious crime like robbery, auto theft, or burglary, admitting to recent drug use may not be considered a big deal. In any event, one has to be cautious in using these data to generalize to the validity of typical self-report inventories.

Summary

We have examined three different approaches to assessing the validity of self-reported measures of delinquency and crime: content validity, construct validity, and criterion validity. Several conclusions appear warranted, especially for the more recent self-report inventories.

On the one hand, the self-report method for measuring this rather sensitive topic—undetected criminal behavior—appears to be reasonably valid. The content validity of the recent inventories is acceptable, the construct validity is quite high, and the criterion validity appears to be in the moderate-to-strong range. Putting this all together, one could conclude that for most analytic purposes, self-reported measures are acceptably accurate and valid.

On the other hand, despite this general conclusion, there are still several substantial issues concerning the validity of self-report measures. First, the validity of the earlier self-report scales, and the results based on them, are at best questionable. Second, based on the results of the tests of criterion validity, there appears to be a substantial degree of either concealing or forgetting past criminal behavior. Although the majority of respondents report their offenses and the majority of all offenses are reported, there is still considerable underreporting.

Third, there is an unresolved issue of differential validity. Compared with other race-gender groups, the responses provided by African-American males appear to have lower levels of validity. Specifically, Hindelang, Hirschi, and Weis (1981) and Huizinga and Elliott (1986) report that African-American males self-report fewer of the offenses found in their official criminal histories. More recently, however, Farrington and colleagues (1996), using data from the Pittsburgh Youth Study, find no evidence of differential validity. It seems that the level of difference in the validity of self-reports for African-American males versus other groups has yet to be determined. If it is less, the processes that bring it about are frankly not understood. This is perhaps the most important methodological issue concerning the self-report method and should be a high priority for future research efforts.

Fourth, based on studies of self-reported substance use, there is some evidence that validity may be less for more serious types of offenses. In the substance use studies, the concordance between the self-report and physiological measures was strongest for adolescent tobacco use, and then for marijuana use; it was weakest for hard drugs, such as cocaine and heroin. A similar pattern is also seen for several studies of self-reported delinquency and crime (e.g., Elliott and Voss 1974; Huizinga and Elliott 1986).

What then are the psychometric properties of self-reported measures of delinquency and crime? With respect to reliability, this approach to measuring involvement in delinquency and crime appears to be acceptable. Most estimates of reliability are quite high, and there is no evidence of differential reliability. With respect to validity, the conclusion is a little murkier; we find a considerable amount of underreporting and a potential problem of differential reporting for African-American males. Nevertheless, content and construct validity

appear to be quite high, and criterion validity would be in the moderate to strong range overall. Perhaps the conclusion Hindelang, Hirschi, and Weis reached in 1981 (p. 114) is still the most reasonable:

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

Specialized Response Techniques

Because of the sensitive nature of asking people to report undetected criminal behavior, there has always been concern about how best to ask these questions to maximize accurate responses. Some early self-report researchers favored self-administered questionnaires, and others favored personal, face-to-face interviews. Similarly, some argued that anonymous responses were inherently better than nonanonymous responses. In their Seattle study, Hindelang, Hirschi, and Weis (1981) directly tested these concerns by randomly assigning respondents to one of four conditions: nonanonymous questionnaire, anonymous questionnaire, nonanonymous interview, and anonymous interview. Their results indicate that there is no strong method effect in producing self-report responses, and that no one approach is consistently better than the others. Similar results are reported by Krohn, Waldo, and Chiricos (1974). Some research, especially in the alcohol and drug use area, has found a methods effect. For example, Aquilino (1994) finds that admission of alcohol and drug use was lowest in telephone interviews, somewhat higher in face-to-face interviews, and highest in self-administered questionnaires (see also Aquilino and LoSciuto 1990; Turner, Lessler, and Devore 1992). Although evident, the effect size typically is not very great.

Although basic method effects do not appear to be very strong, there is still concern that in all of these approaches to the collection of survey data, respondents will feel vulnerable about reporting sensitive information. Because of that, a variety of more specialized techniques have been developed to protect respondents' confidentiality, hopefully increasing the level of reporting.

Randomized response technique

The randomized response technique assumes that the basic problem with the validity of self-reporting responses is that respondents are trying to conceal

sensitive information; that is, they are unwilling to report undetected criminal behavior if others, including the researchers, might link the behavior to them. Randomized response techniques allow respondents to conceal what they really did, while at the same time providing useful data to the researchers. There are a variety of ways of accomplishing this. We can illustrate how the basic process works with a simple example of measuring the prevalence of marijuana use.

Imagine an interview setting in which there is a screen between the interviewer and respondent so that the interviewer cannot see what the respondent is doing. The interviewer asks the sensitive question: "Have you ever smoked marijuana?" The interviewer gives the following special instructions: "Before answering, please flip a coin. If the coin lands on heads, please answer "yes" regardless of whether or not you smoked marijuana. If the coin lands on tails, please tell me the truth." Thus, the interviewer cannot know whether a "yes" response is truthful or is produced by the coin landing on heads. In this way, the respondent can admit to sensitive behavior but other people, including the interviewer, cannot know whether the admission is truthful.

From the resulting data, though, we can estimate the prevalence of marijuana use. Say we receive 70 "yes" responses from a sample of 100 respondents. Fifty of those would be produced by the coin landing on heads and can simply be ignored. Of the remaining 50 respondents, however, 20 said "yes" because they have smoked marijuana, so the prevalence of marijuana use is 20 out of 50, or 40 percent.

This technique is not limited to "yes" or "no" questions or to flipping coins. Any random process can be used as long as we know the probability distribution of bogus versus truthful responses. From these data, we can estimate prevalence, variety, and frequency scores and means and variances, and we can correlate the information with other variables, just as we do with regular self-report data.

Weis and Van Alstyne (1979) tested a randomized response procedure in the Seattle study. Based on their data, they concluded that the randomized response approach is no more efficient in eliciting positive responses to sensitive items than are traditional methods of data collection. This finding is consistent with the overall conclusion in the Hindelang et al. (1981) Seattle study that the method of administration does not significantly affect the validity of self-report responses.

The other major assessment of the randomized response technique was conducted by Tracy and Fox (1981). They sampled people who had been arrested in Philadelphia and sent interviewers to their homes. The interviewers did not

know that the sample consisted only of people with official arrest records. Respondents were asked if they had been arrested and, if so, how many times. Since this information was already known from the arrest records, the validity of the self-reported responses could be assessed. (This is much like the "reverse record check" technique used in victimization surveys; see Cantor and Lynch in this volume.) The Tracy and Fox study employed two methods of data collection, a randomized response procedure and a regular self-report interview.

The results indicate that the randomized response approach does make a difference. For all respondents, there was about 10 percent less error in the randomized response technique. For respondents who had been arrested only once, the randomized response approach actually increased the level of error. But for recidivists, the randomized response technique reduced the level of error by about 74 percent.

The randomized response technique also generated random errors (errors not correlated with other important variables). The regular self-reported interview, however, generated systematic error or bias. In this approach, underreporting was higher for females, African-American females, respondents with high need for approval, lower income respondents, and those with a larger number of arrests.

Overall, it is not clear to what extent a randomized response approach generates more complete and accurate reporting. The two major studies of this topic produce different results: Weis and Van Alstyne (1979) report no effect, and Tracy and Fox (1981) report sizable and positive effects. It should be noted, however, that Tracy and Fox's results only generalize to self-reports of being arrested, and may or may not apply to self-reports of undetected delinquent behavior. The value of the randomized response approach requires additional research, which should be conducted within the context of audio and computer-assisted interviewing, the topic to which we now turn.

Computer-assisted interviewing

Advances in both computer hardware and software have made the introduction of computers in the actual data collection process not only a possibility but, according to Tourangeau and Smith (1996, 276), "perhaps the most commonly used method of face-to-face data collection today." The use of computers in the data collection process began in the 1970s with computer-assisted telephone surveys (Saris 1991). This technique is used by the National Crime Survey and described in Cantor and Lynch in this volume. The technology was soon adapted to the personal interview setting, either with the interviewer administering the schedule, the Computer-Assisted Personal Interview (CAPI), or with the respondent self-administering the schedule by reading the questions on the

computer screen and entering his or her responses, the Computer-Assisted Self-Administered Interview (CASI). It is also possible to have an audio version in which the questions are recorded and the respondent listens to them, rather than having them read by the interviewer or having the respondent read them. This is called an Audio Computer-Assisted Self-Administered Interview (ACASI).

The use of computerized tools is one of two trends that have transformed survey research in the United States; the other trend is the collection of increasingly sensitive information concerning illegal and embarrassing behaviors.

Tourangeau and Smith (1996) suggest that the use of computerized tools is one of two trends that have transformed survey research in the United States; the other trend is the collection of increasingly sensitive information concerning illegal and embarrassing behaviors. One reason for the use of computer-assisted data collection that is particularly relevant for this chapter, is its potential for collecting sensitive information in a manner that increases the confidentiality of responses. By not having the interviewer read the questions or be involved in the recording of answers, the respondent does not have to reveal potentially embarrassing behavior directly to another person. In addition, the responses cannot be overheard by other people (e.g., family members or teachers) who might be nearby. Of course, the same advantage could be acquired by administering a paper-and-pencil self-administered questionnaire. However, computer-assisted techniques have other potential advantages.

A key advantage of computer-assisted administration of interview schedules over questionnaires is that they allow for the incorporation of complex branching patterns (Saris 1991; Beebe et al. 1998; Wright, Aquilino, and Supple 1998; Tourangeau and Smith 1996). For example, many delinquency checklists include a series of followup questions if the respondent answers affirmatively to having committed a particular type of delinquent behavior within a specified period of time. The branching of these followup items can be quite complex; respondents who are asked to read and follow the skip patterns can easily miss important items. Computer software can program the skip patterns and increase the probability that the respondent will answer all appropriate questions. An added advantage of computer-assisted presentation is that the respondent does not see the implications of answering in the affirmative to questions with multiple followups. Respondents may be reluctant to indicate that they have committed a delinquent act if they realize that an affirmative answer will trigger a series of followup questions (Thornberry 1989).

Computer software can also identify inconsistent and incomplete responses. Thus, if a respondent indicates that he has never been arrested but also indicates that he has spent time in a juvenile correctional facility, the program can identify this inconsistency and either prompt the respondent to clarify the issue or prompt the interviewer to ask for clarification. Computer-assisted administration can also decrease incomplete responses and reduce the number of "out of range" responses (Wright, Aquilino, and Supple 1998).

An audiotape on which questions are read to the respondent (ACASI) has two additional advantages. First, it circumvents the potential problem of illiteracy; the respondent does not have to read the questions. Second, in situations where other people might be nearby, the questions and responses are not heard by anyone but the respondent. Hence, the respondent can be more assured that answers to sensitive questions will remain private.

Although computer-assisted administration of sensitive questions provides some obvious advantages in terms of efficiency of presentation and data collection, the key question concerns the difference in the responses elicited when such technology is used. Tourangeau and Smith (1996) reviewed 18 studies that have compared different modes of data collection. The types of behavior examined include health problems (e.g., gastrointestinal problems), sexual practices, abortion, and alcohol and drug use. Tourangeau and Smith indicate that self-administered techniques generally elicit higher rates of problematic behaviors than those administered by an interviewer. Moreover, computer-assisted self-administered interviews elicit higher rates than either self-administered questionnaires or paper-and-pencil interviews administered by an interviewer. Also, ACASI (audio computer-assisted self-administered interviews) elicit higher rates than CASI.

In their own research, Tourangeau and Smith (1996) compared different modes of administration with a sample of adults ages 18 to 45. Respondents were asked questions regarding their sexual behavior as well as their use of alcohol and drugs. Data were collected using CAPI, CASI, and ACASI. With CAPI, the questions appeared on the computer screen and were read by the interviewer, who then entered the responses. With CASI, the respondent entered the responses. With ACASI, the questions appeared on the screen while a digitized recording was provided to the respondent via earphones. They found that ACASI and CASI elicited higher rates of drug use and sexual behavior than CAPI. For example, respondents who were administered CAPI reported a lifetime prevalence rate for drug use of 44.8 percent, compared with 58 percent under CASI and 66.3 percent under ACASI. The same trend was evident for other measures of drug use and for sexual activity, although in some cases the differences were not statistically significant. Tourangeau and Smith conclude that by allowing

respondents to interact directly with the computer, respondents are convinced of the "legitimacy and scientific value of the study" (p. 301). Other studies comparing administration modes have found that the level of reporting may be contingent on characteristics of respondents and the setting. For example, Wright, Aquilino, and Supple (1998) found that adolescents reported higher levels of alcohol and drug use in the computer mode than in the paper-and-pencil mode. However, mode effects were not evident for young adult respondents.

Estimates of prevalence rates of illegal and embarrassing behavior appear to be higher when computer-assisted techniques, particularly those involving self-administration, are used. The higher prevalence rates need to be externally validated. The added benefits of providing for schedule complexity and consistency in responses make these techniques attractive, and it is clear that they will continue to be used with increasing frequency.

Self-Report Measures Across the Lifecourse

Criminological research has increasingly come to rely on longitudinal panel designs using self-report measures of antisocial behavior to understand the dynamics of offending careers.

One of the most significant developments in criminology over the past 15 years has been the emergence of a "lifecourse" or developmental focus (Farrington, Ohlin, and Wilson 1986; Thornberry and Krohn forthcoming; Thornberry 1997; Jessor 1998; Weitekamp 1989). Theoretical work has expanded from a narrow focus on the adolescent years to encompass the entire criminal career of individuals. This can extend from precursors of delinquency manifested in early childhood (Moffitt 1997; Tremblay et al. 1998), through the high-delinquency years of middle and late adolescence, on into adulthood when most, but not all, offenders decrease their participation in illegal behavior (Moffitt 1997; Thornberry and Krohn forthcoming; Sampson and Laub 1990; Loeber et al. 1998). Research on "criminal careers" (Blumstein et al. 1986) has documented the importance of examining such issues as the age of onset (Krohn, Thornberry, and Rivera forthcoming) and the duration of criminal activity (Wolfgang, Thornberry, and Figlio 1987). In addition, a growing body of research has demonstrated that antisocial behavior is fairly stable from childhood to adulthood (Farrington 1989a; Huesmann et al. 1984; Olweus 1979; Moffitt 1993). Much of this work has relied primarily on the use of official data. However, criminological research has increasingly come to rely on longitudinal panel designs using self-report measures of antisocial behavior to understand the dynamics of offending careers. The use of self-report

techniques in longitudinal studies over the lifecourse introduces a number of interesting measurement issues.

Construct continuity

Although many underlying theoretical constructs, such as involvement in illegal behaviors, remain constant over time, their behavioral manifestations can change as subjects age. Failure to adapt measures to account for these changes inevitably leads to age-inappropriate measures with reduced validity and reliability. To avoid this, measures need to adapt to the respondent's developmental stage to reflect accurately the theoretical constructs of interest (Campbell 1990; Patterson 1993; Le Blanc 1989; Weitekamp 1989). In some cases, this may mean defining the concept at a level to accommodate the changing contexts in which people act at different ages. In other cases, it may mean recognizing that different behaviors at different ages imply consistency in behavioral style (Campbell 1990, 7).

In previous sections, our discussion has focused on the problems with how self-reported delinquent behavior has been defined and measured when sampling adolescents. When applying the self-report technique to both younger children and adults, these definitional issues are magnified. We recognize that different items may be needed to measure the same underlying construct to maintain the age-appropriateness of the measure. Therefore, the construct continuity of the different measures of delinquency or antisocial behavior becomes of paramount importance.

Self-report measures for children

Although antisocial behavior is quite stable, it has been likened to a chimera (Patterson 1993), with manifestations that change and accumulate with age. At very young ages (2 to 5 years), behavioral characteristics such as impulsivity, noncompliance, disobedience, and aggression are seen as early analogs of delinquent behavior. At these young ages, self-report instruments are not practical because of the age of the respondents. Rather, researchers have measured these key indicators either through parental reports or through observational ratings. Many studies of youngsters at these ages have used Achenbach's (1992) Child Behavior Checklist (CBCL), a parent-completed inventory with versions for children as young as 2 to assess "externalizing" problem behaviors.[2] Studies using either CBCL, some other parental or teacher report of problem behaviors, or observational ratings have demonstrated that there is a relationship between these early manifestations of problem behavior and antisocial behavior in school-age children (Belsky,

Woodworth, and Crnic 1996; Campbell 1987; Richman, Stevenson, and Graham 1982; Shaw and Bell 1993).

Starting at school age, the range of antisocial behaviors expands to include stubbornness, lying, bullying, and other externalizing problems (Loeber et al. 1993). School-age children, even those as young as first grade, begin to exhibit delinquent behaviors. However, self-report instruments of delinquent behavior have rarely been administered to preteen-age children (Loeber et al. 1989). There are some studies that have administered self-report instruments to youngsters as young as 10 or 11 years of age, slightly modifying the standard delinquency items (Elliott, Huizinga, and Ageton 1985).

Loeber and colleagues (1989) provide one of the few attempts to not only gather self-report information from children younger than the age of 10, but also examine the reliability of those reports. They surveyed a sample of 849 first grade and 868 fourth grade boys using a 33-item self-reported antisocial behavior scale. This is a younger age version of the self-reported delinquency index used by the three projects of the Program of Research on the Causes and Correlates of Delinquency. Items that were age appropriate were selected, and some behaviors were placed in several different contexts to make the content less abstract for the younger children. A special effort was made to ensure that the child understood the question by preceding each behavior with a series of questions to ascertain whether the respondents knew the meaning of the behavior. If the child did not understand the question, the interviewer gave an example and then asked the child to do the same. If the child still did not understand the question, the item was skipped.

The parents and teachers of these children were also surveyed, using a combination of the appropriate CBCL and delinquency items. To examine the validity of the child self-reported antisocial behavior scale, comparable items contained in the parent and teacher CBCL were compared with the self-report items.

Loeber et al. (1989) report that the majority of boys understood most of the items. First grade boys did have problems understanding the items regarding marijuana use and sniffing glue, and fourth grade boys had difficulty understanding the question regarding sniffing glue.

A substantial minority of the first grade boys reported damaging property (26 percent) and stealing (26 percent), while over half of the fourth grade boys reported vandalizing (51.2 percent) and stealing (53.1 percent). An even higher percentage of both first and fourth graders reported a violent offense (66.3 percent and 91.2 percent, respectively), but these items included hitting siblings and other students.

Loeber and colleagues (1989) recognized the difficulty of assessing the accuracy of self-reported delinquent behavior among elementary school children, who are unlikely to have court or police records. As an initial step, Loeber and colleagues compared the children's self-reports with parental reports about similar behaviors. They found surprisingly high concordance between children's and parents' reports about the ever-prevalence of delinquent behavior. This is especially true for behaviors that are likely to come to the attention of parents, such as aggressive behaviors and school suspension. Concordance was higher for first graders than it was for fourth graders, which Loeber and colleagues suggest would be expected, since parents are more likely to know about misbehavior at younger ages.

These findings are encouraging and suggest that self-report instruments, if administered with concern for the respondents' age, can be used for very young children. Loeber and colleagues (1989) suggest that another measure of the utility of these measures will be their predictive validity. If self-reports of delinquent behavior in the first or fourth grades predict later delinquency, there is further reason to be confident in this methodology's applicability for elementary school samples.

Self-report measures for adults

The interest in assessing antisocial behavior across the lifespan has also led to an increasing number of longitudinal surveys that have followed respondents from their adolescent years into early adulthood (e.g., Elliott 1994; Huizinga et al. 1998; Loeber et al. 1998; Farrington 1989b; Le Blanc 1989; Hawkins, Catalano, and Miller 1992; Krohn, Lizotte, and Perez 1997). The concern in constructing self-report instruments for adults is to include items that take into account the different contexts in which crime occurs at these ages (e.g., work instead of school), the opportunities for different types of offenses (e.g., domestic violence, fraud), the inappropriateness or inapplicability of offenses that appear on adolescent self-report instruments (e.g., status offenses), and the potential for very serious criminal behaviors, at least among small subset of chronic violent offenders.

Weitekamp (1989) has criticized self-report studies not only for being predominantly concerned with the adolescent years but, when covering the adult years, for also using the same items as for juveniles. He argues that even such studies as NYS (Elliott 1994) do not include many items that are more serious, and therefore appropriate for adults, than the items included in the original Short and Nye study (1957). Weitekamp asserts that we need to use different instruments during different life stages. Doing so, however, raises questions about

construct continuity similar to those discussed in constructing self-report inventories with very young children. If the researcher wants to document the change in the propensity to engage in antisocial behavior throughout the lifecourse, he or she must assume that the different items used to measure antisocial behavior at different ages do indeed measure the same underlying construct. Le Blanc (1989) suggests that a strategy of including different but overlapping items on instruments covering different ages across the lifespan is the best compromise.

The use of self-report studies in longitudinal research has generated a number of issues regarding the definition and measurement of antisocial behavior. If the researcher wants to examine the development of antisocial behavior across the lifespan, a definition and measurement of delinquent behavior limited to the standard used in research on adolescents will not suffice. Expanding that definition to encompass behaviors that take into account antisocial acts by very young children and more serious offenses by adults that may take place in different social contexts requires a well-considered definition of the construct that these different behaviors represent. Indeed, ultimately the resolution of this issue relies on a strong theoretical foundation that provides a clear definition of antisocial behavior. The utility of such a definition and the measurements that derive from it will be assessed in examining the correlations across different stages in the lifespan.

Panel or testing effects

Developments in self-report methods have improved the quality of data collected and have expanded the data's applicability to the study of antisocial behavior throughout the lifecourse. Although these advances are significant, they have increased the potential for the data to be contaminated by testing or panel effects (Thornberry 1989).

Testing effects are any alterations of the respondent's response to an item or scale that are caused by the prior administration of the same item or scale (Thornberry 1989, 351). With the use of self-reports in longitudinal research, respondents are administered the same or similar items across waves of data collection. Improvements in self-report instruments have led to the inclusion of a longer list of items to tap more serious offenses, and often, a number of followup questions are asked. The more acts a respondent admits to, the longer the overall interview will take. The concern is that this approach will make respondents increasingly unwilling to admit to delinquent acts because those responses will lengthen the interview. This effect would probably be unequally distributed because respondents with the most extensive involvement in delinquency would lose the most time by answering affirmatively to the delinquency items. Over successive administrations of the self-report instrument, respondents

would learn that positive responses lengthen their interview time, and the amount of underestimation of delinquency rates would increase.

It is also possible that the simple fact that a respondent is reinterviewed may create a generalized fatigue, decreasing the respondent's willingness to respond to self-report items. Research using the National Crime Survey of victimization found that the reduction in reporting was due more to the number of prior interviews than to the number of victimizations reported in prior interviews (Lehnen and Reiss 1978).

Three studies have examined testing effects in the use of self-report studies; all are based on data from NYS (Elliott, Huizinga, and Ageton 1985). They were conducted by Thornberry (1989), Menard and Elliott (1993), and Lauritsen (1998). NYS surveyed a nationally representative sample of 1,725 11- to 17-year-old youths in 1976. They reinterviewed the same subjects annually through 1981. These data allow researchers to examine the age-specific prevalence rates by the number of times a respondent was interviewed. For example, some respondents were 14 at the time of their first interview, some were 14 at their second interview (the original 13-year-old cohort), and some were 14 at their third interview (the original 12-year-old cohort). Because of this, a 14-year-old prevalence rate can be calculated from data collected when respondents were interviewed for only the first time, from data collected when they were interviewed a second time, and so on. If a testing or panel effect plays a role in response rates, the more frequently respondents are interviewed, the lower the age-specific rates should be. Thus the 14-year-old rate from a second interview would be lower than the 14-year-old rate based on a first interview.

Thornberry analyzed these rates for 17 NYS self-report items representing the major domains of delinquency and the most frequently occurring items. The overall trend seemed to suggest that either a panel or testing effect was occurring. For all offenses except marijuana use, comparisons between adjacent waves indicated that the age-specific prevalence rates decreased more often than they increased. For example, comparing the rate of gang fights from wave to wave, Thornberry found that for 67 percent of the comparisons, there was a decrease in the age-specific prevalence rates, whereas there was an increase in only 20 percent of the comparisons, and there was no change in 13 percent. The magnitude of the changes were, in many cases, substantial. For example, for stealing something worth $5 to $50, the rate drops by 50 percent for 15-year-olds from wave 1 to wave 4 (Thornberry 1989, 361).

NYS did not introduce the detailed followup questions to the delinquency items until the fourth wave of data collection. The data analyzed by Thornberry (1989) show the decline in reporting occurred across all waves. Hence, it appears

that the panel design itself, rather than the design of the specific questions, had the effect of decreasing prevalence rates. Thornberry suggests that panel and testing effects could be a serious threat to the use of self-reports in longitudinal research and calls for a more thorough investigation of this issue. The observed decline in the age-specific rates could be due to an underlying secular drop in offending during these years. Cross-sectional trend data from the Monitoring the Future (MTF) study, which cannot be influenced by a testing effect, do not indicate any such secular decline (see Thornberry 1989).

Menard and Elliott (1993) reexamined this issue using both the NYS and MTF data. They rightfully point out that comparisons across these studies need to be undertaken cautiously because of differences in samples, design features, item wording, and similar concerns. Menard and Elliott's analysis also shows that at the item level, declining trends are more evident in the NYS than in the MTF data (1993, 439). They go on to show that most of these year-to-year changes are not statistically significant, however. They then use a modified Cox-Stuart trend test to examine short-term trends in delinquency and drug use. Overall, the trends for 81 percent of the NYS offenses are not statistically significant, and about half of the MTF trends are. But, an examination of the trends for the 16 items included in their table 2 indicates that there are more declining trends in the NYS data, 9 of 16 for the 1976–80 comparisons and 7 of 16 for the 1976–83 comparisons, than there are for the MTF data, 3 of 16 in both cases. Menard and Elliott focus on the statistically significant effects that indicate fewer declining trends in NYS than is evident when one focuses on all trends, regardless of the magnitude of the change.

More recently, Lauritsen (1998) examined this topic using data from the first five waves of NYS. Specifically, she used hierarchical linear models (HLM) to estimate growth curve models for general delinquency and for serious delinquency. HLM models make fuller use of the data and include tests for statistical significance. She limited her analysis to four of the seven cohorts in NYS, those who were ages 11, 13, 15, and 17 at wave 1.

For those who were age 13, 15, or 17 at the start of NYS, involvement in both general delinquency and serious delinquency decreased significantly over the next 4 years. For the 11-year-old cohort, the rate of change was also negative but not statistically significant. This downward trajectory in the rate of delinquent behavior for all age cohorts is not consistent with theoretical expectations or what is generally known about the age-crime curve. Also, as Lauritsen points out, it is not consistent with other data on secular trends for the same time period (see also Thornberry 1989; Osgood et al. 1989).

Finally, Lauritsen examined whether this testing effect is due to the introduction of detailed followup questions at wave 4 of the NYS or whether it appeared to be produced by general panel fatigue. Her analysis of individual growth trajectories indicates that the decline is observed across all waves. Thus, she concludes, as Thornberry did, that the reduced reporting is unlikely to have been produced by a change in survey administration, namely, by the addition of followup questions.

Overall, Lauritsen offers two explanations for the observed testing effects. One concerns generalized panel fatigue, suggesting that as respondents are asked the same inventory at repeated surveys, they become less willing to respond affirmatively to screening questions. The second explanation concerns a maturation effect in which the content validity of the self-report questions would vary with age. For example, how respondents interpret a question on simple assault, and the type of behavior they consider relevant for responding to the question, may be quite different for 11- and 17-year-olds. Of course, both of these processes may operate.

The studies by Thornberry and by Lauritsen suggest that there is some degree of panel bias in self-report data collected in longitudinal panel studies. The analysis by Menard and Elliott indicates that this is still just a suggestion, as the necessary comparisons between panel studies and cross-sectional trend studies are severely hampered by lack of comparability in item wording, administration, and other methodological differences. Also, if there are testing effects, neither Thornberry nor Lauritsen is arguing that they are unique to NYS. It just so happens that the sequential cohort design of NYS makes it a good vehicle for examining this issue. The presumption, unfortunately, is that if testing effects interfere with the validity of the NYS data, they also interfere with the validity of other longitudinal data containing self-report information. This is obviously a serious matter, as etiological research has focused almost exclusively on longitudinal designs in the past 20 years. Additional research to identify the extensiveness of testing effects, their sources, and way of remedying them are certainly a high priority.

Conclusions

The self-report method for measuring crime and delinquency has developed substantially since its introduction a half century ago. It is now a fundamental method of scientifically measuring criminality and forms the bedrock of etiological studies. The challenges confronting this approach to measurement are daunting; after all, we are asking individuals to tell us about their own, undetected criminality. Despite this fundamental challenge, the technique seems to be successful and capable of producing valid and reliable data.

Early self-report scales had substantial weaknesses, containing few items and producing an assessment of only minor forms of offending. Gradually, as the underlying validity of the approach became evident, the scales expanded in terms of breadth, seriousness, and comprehensiveness. Contemporary measures typically cover a wide portion of the behavioral domain included under the construct of crime and delinquency. These scales are able to measure serious as well as minor forms of crime, with such major subdomains as violence, property crimes, and drug use; to measure different parameters of criminal careers such as prevalence, frequency, and seriousness; and to identify high-rate as well as low-rate offenders. This is substantial progress for a measurement approach that began with a half dozen items and a four-category response set.

The self-report approach to measuring crime has acceptable, albeit far from perfect, reliability and validity. Of these two basic psychometric properties, the evidence for reliability seems stronger. There are no fundamental challenges to the reliability of these data. Test-retest measures (and internal consistency measures) indicate that self-reported measures of delinquency are as reliable as, if not more reliable than, most social science measures.

Validity is much harder to assess, as there is no "gold standard" against which to judge the self-reports. Nevertheless, current scales seem to have acceptable levels of content and construct validity. The evidence for criterion validity is less clear cut. At an overall level, criterion validity seems to be in the moderate to strong range. Although there is certainly room for improvement, the validity appears acceptable for most analytic tasks. At a more specific level, however, there is a potentially serious problem with differential validity. Two of the major assessments of criterion validity, by Hindelang, Hirschi, and Weis (1981) and by Huizinga and Elliott (1986), found lower validity for African-American males. The more recent assessment by Farrington and colleagues (1996) did not. Additional research on this topic is imperative.

Although basic self-report surveys appear to be reliable and valid, researchers have experimented with a variety of data collection methods to improve the quality of reporting. Several of these attempts have produced ambiguous results; for example, there is no clear-cut benefit to mode of administration (interview versus questionnaire) or to the use of randomized response techniques. There is one approach that appears to hold great promise, however. Audio-assisted computerized interviews produce increased reporting of many sensitive topics, including delinquency and drug use. Greater use of this approach is warranted.

In the end, the available data indicate that the self-report method is an important and useful way to collect information about criminal behavior. The

skepticism of early critics like Nettler (1978) and Gibbons (1979) has not been realized. Nevertheless, the self-report technique can clearly be improved. The final issue addressed in this chapter is suggestions for future research.

Future directions

Much of our research on reliability and validity simply assesses these characteristics; there is far less research on improving their levels. For example, it is likely that both validity and reliability would be improved if we experimented with alternate items for measuring the same behavior and identified the strongest ones. Similarly, reliability and validity vary across subscales (e.g., Huizinga and Elliott 1986); improving subscales will not only help them but also the overall scale as they are aggregated.

Throughout this chapter, we discussed the issue of differential validity for African-American males. It is crucial to learn more about the magnitude of this bias and its source, if it exists. Future research should address this issue directly and attempt to identify techniques for eliminating it. These research efforts should not lose sight of the fact that the problem may be with the criterion variable (official records) and not the self-reports.

The self-report method was developed in and for cross-sectional studies. Using it in longitudinal studies, especially ones that cover major portions of the life-course, creates a new set of challenges. Maintaining the age-appropriateness of the items, while at the same time ensuring content validity, is a knotty problem that we have just begun to address. There is some evidence that repeated measures may create testing effects. More research is needed to measure the size of this effect and to identify methods to reduce its threat to the validity of self-report data in the longitudinal studies that are so crucial to etiological investigation.

One of the most promising developments in the self-report method is the advent of audio-assisted computerized interviews. This technique offers increased confidentiality to the respondent in an interview setting. Although somewhat expensive and complicated to design, the early studies indicate that it may be worth the effort.

Finally, we recommend that methodological studies be done in a crosscutting fashion so that several of these issues—reliability and validity, improved item selection, assessing panel bias, etc.—can be addressed simultaneously. It is particularly important to examine all of these methodological issues when data are collected using audio-assisted computerized interviewing. For example, studies that have found differential validity or testing effects have all used paper-and-pencil interviews. Whether these same problems are evident under the enhanced confidentiality of audio interviews is an open question. It is clearly a high-priority one as well.

There is no dearth of work that can be done to assess and improve the self-report method. If the progress over the past half century is any guide, we are optimistic that the necessary studies will be conducted and that they will improve this basic means of collecting data on criminal behavior.

Notes

1. This is particularly the case given the level of reliability of self-reported data (see the section "Assessing reliability"). By adding random error to the picture, poor reliability attenuates or reduces the size of the observed correlation coefficients.

2. CBCL also assesses internalizing problem behavior.

References

Achenbach, T.M. 1992. *Manual for the Child Behavior Checklist/2–3 and 1992 profile.* Burlington: University of Vermont.

Akers, Robert L. 1964. Socio-economic status and delinquent behavior: A retest. *Journal of Research in Crime and Delinquency* 1:38–46.

Akers, Robert L., Marvin D. Krohn, L. Lanza-Kaduce, and M. Radosevich. 1979. Social learning and deviant behavior: A specific test of a general theory. *American Sociological Review* 44:636–655.

Akers, Robert L., James Massey, William Clarke, and Ronald M. Lauer. 1983. Are self-reports of adolescent deviance valid? Biochemical measures, randomized response, and the bogus pipeline in smoking behavior. *Social Forces* 62 (September): 234–251.

Anderson, Linda S., Theodore G. Chiricos, and Gordon P. Waldo. 1977. Formal and informal sanctions: A comparison of deterrent effects. *Social Problems* 25:103–112.

Aquilino, William S. 1994. Interview mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly* 58:210–240.

Aquilino, William S., and Leonard LoSciuto. 1990. Effects of interview mode on self-reported drug use. *Public Opinion Quarterly* 54:362–395.

Beebe, Timothy J., Patricia A. Harrison, James A. McRae, Jr., Ronald E. Anderson, and Jayne A. Fulkerson. 1998. An evaluation of computer-assisted self-interviews in a school setting. *Public Opinion Quarterly* 62:623–632.

Belsky, J., S. Woodworth, and K. Crnic. 1996. Troubled family interaction during toddlerhood. *Development and Psychopathology* 8:477–495.

Belson, W.A. 1968. The extent of stealing by London boys and some of its origins. *Advancement of Science* 25:171–184.

Blackmore, J. 1974. The relationship between self-reported delinquency and official convictions amongst adolescent boys. *British Journal of Criminology* 14:172–176.

Blumstein, A., J. Cohen, J.A. Roth, and C.A. Visher, eds. 1986. *Criminal careers and "career criminals."* Washington, D.C.: National Academy Press.

Braukmann, C.J., K.A. Kirigin, and M.M. Wolf. 1979. Social learning and social control perspectives in group home delinquency treatment research. Paper presented at the 1979 Annual Meeting of the American Society of Criminology, November, Philadelphia.

Broder, P.K., and J. Zimmerman. 1978. *Establishing the reliability of self-reported delinquency data.* Williamsburg, Virginia: National Center for State Courts.

Browning, Katharine, David Huizinga, Rolf Loeber, and Terence P. Thornberry. 1999. *Causes and correlates of delinquency program.* Fact Sheet, FS 99100. Washington, D.C.: U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention.

Campbell, S.B. 1990. B*ehavioral problems in preschool children: Clinical and developmental issues.* New York: Guilford Publications.

———. 1987. Parent-referred problem three-year-olds: Developmental changes in symptoms. *Journal of Child Psychology & Psychiatry* 28:835–845.

Clark, J.P., and L.L. Tifft. 1966. Polygraph and interview validation of self-reported delinquent behavior. *American Sociological Review* 31:516–523.

Clark, J.P., and E.P. Wenninger. 1962. Socioeconomic class and area as correlates of illegal behavior among juveniles. *American Sociological Review* 28:826–834.

Conger, Rand. 1976. Social control and social learning models of delinquency: A synthesis. *Criminology* 14:17–40.

Dentler, R.A., and L.J. Monroe. 1961. Social correlates of early adolescent theft. *American Sociological Review* 26:733–743.

Elliott, Delbert S. 1994. Serious violent offenders: Onset, developmental course, and termination—The American Society of Criminology 1993 presidential address. *Criminology* 32 (1): 1–21.

——— . 1966. Delinquency, school attendance, and dropout. *Social Problems* 13:306–318.

Elliott, Delbert S., and S.S. Ageton. 1980. Reconciling race and class differences in self-reported and official estimates of delinquency. *American Sociological Review* 45 (1): 95–110.

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

Elliott, Delbert S., and H.L. Voss. 1974. *Delinquency and dropout.* Lexington, Massachusetts: D.C. Heath.

Empey, LaMar T., and Maynard Erickson. 1966. Hidden delinquency and social status. *Social Forces* 44 (June): 546–554.

Erickson, M., and L.T. Empey. 1963. Court records, undetected delinquency, and decision-making. *Journal of Criminal Law, Criminology, and Police Science* 54 (December): 456–469.

Farrington, David P. 1989a. Early predictors of adolescent aggression and adult violence. *Violence and Victims* 4:79–100.

———. 1989b. Self-reported and official offending from adolescence to adulthood. In *Cross-national research in self-reported crime and delinquency,* edited by Malcolm W. Klein. Los Angeles: Kluwer Academic Publishers.

———. 1977. The effects of public labelling. *British Journal of Criminology* 17:112–125.

———. 1973. Self-reports of deviant behavior: Predictive and stable? *Journal of Criminal Law and Criminology* 64:99–110.

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

Farrington, David P., Lloyd E. Ohlin, and James Q. Wilson. 1986. *Understanding and controlling crime: Toward a new research strategy.* New York: Springer-Verlag.

Gibbons, Don C. 1979. *The criminological enterprise: Theories and perspectives.* Upper Saddle River, New Jersey: Prentice Hall.

Gold, M. 1970. *Delinquent behavior in an American city.* Belmont, California: Brooks/Cole Publishing Company.

———. 1966. Undetected delinquent behavior. *Journal of Research in Crime and Delinquency* 3:27–46.

Hardt, R.H., and S. Petersen-Hardt. 1977. On determining the quality of the delinquency self-report method. *Journal of Research in Crime and Delinquency* 14:247–261.

Hathaway, R.S., E.D. Monachesi, and L.A. Young. 1960. Delinquency rates and personality. *Journal of Criminal Law, Criminology, and Police Science* 50:433–440.

Hawkins, J. David, Richard F. Catalano, and Janet Y. Miller. 1992. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. *Psychological Bulletin* 112 (1): 64–105.

Hepburn, J.R. 1976. Testing alternative models of delinquency causation. *Journal of Criminal Law and Criminology* 67:450–460.

Hindelang, M.J. 1973. Causes of delinquency: A partial replication and extension. *Social Problems* 20:471–487.

Hindelang, M.J., T. Hirschi, and J.G. Weis. 1981. *Measuring delinquency.* Beverly Hills: Sage Publications.

———. 1979. Correlates of delinquency: The illusion of discrepancy between self-report and official measures. *American Sociological Review* 44:995–1014.

Hirschi, T. 1969. *Causes of delinquency.* Berkeley: University of California Press.

Huesmann, L.R., L.D. Eron, M.M. Lefkowitz, and L.O. Walder. 1984. The stability of aggression over time and generations. *Developmental Psychology* 20:1120–1134.

Huizinga, D., and D.S. Elliott. 1986. Reassessing the reliability and validity of self-report delinquent measures. *Journal of Quantitative Criminology* 2 (4): 293–327.

———. 1983. *A preliminary examination of the reliability and validity of the National Youth Survey self-reported delinquency indices.* National Youth Survey Project Report 27. Boulder, Colorado: Behavioral Research Institute.

Huizinga, David, Anne W. Weiher, Scott Menard, Rachele Espiritu, and Finn Esbensen. 1998. Some not-so-boring findings from the Denver Youth Survey. Paper presented at the 1998 Annual Meeting of the American Society of Criminology, November, Washington, D.C.

Jensen, G.F. 1973. Inner containment and delinquency. *Journal of Criminal Law and Criminology* 64:464–470.

Jensen, Gary F., Maynard L. Erickson, and Jack P. Gibbs. 1978. Perceived risk of punishment and self-reported delinquency. *Social Forces* 57:57–78.

Jensen, Gary F., and Raymond Eve. 1976. Sex differences in delinquency. *Criminology* 13:427–448.

Jessor, Richard. 1998. *New perspectives on adolescent risk behavior.* New York: Cambridge University Press.

Johnson, R.E. 1979. *Juvenile delinquency and its origins.* Cambridge: Cambridge University Press.

Johnston, Lloyd D., Patrick M. O'Malley, and Jerald G. Bachman. 1996. *National survey results on drug use from the Monitoring the Future study, 1975–1995.* Rockville, Maryland: U.S. Department of Health and Human Services.

Kaplan, H.B. 1972. Toward a general theory of psychosocial deviance: The case of aggressive behavior. *Social Science and Medicine* 6:593–617.

Kelly, D.H. 1974. Track position and delinquent involvement: A preliminary analysis. *Sociology and Social Research* 58:380–386.

Klein, Malcolm W., ed. 1989. *Cross-national research in self-reported crime and delinquency.* Los Angeles: Kluwer Academic Publishers.

Krohn, Marvin D., Alan J. Lizotte, and Cynthia M. Perez. 1997. The interrelationship between substance use and precocious transitions to adult statuses. *Journal of Health and Social Behavior* 38 (March): 87–103.

Krohn, Marvin D., Terence P. Thornberry, and Craig Rivera. Forthcoming. Later careers of very young offenders. In *Child delinquents: Development, interventions, and service needs,* edited by Rolf Loeber and David P. Farrington. Thousand Oaks, California: Sage Publications.

Krohn, Marvin D., Gordon P. Waldo, and Theodore G. Chiricos. 1974. Self-reported delinquency: A comparison of structured interviews and self-administered checklists. *Journal of Criminal Law and Criminology* 65 (4): 545–553.

Kulik, J.A., K.B. Stein, and T.R. Sarbin. 1968. Disclosure of delinquent behavior under conditions of anonymity and non-anonymity. *Journal of Consulting and Clinical Psychology* 32:506–509.

Lauritsen, Janet L. 1998. The age-crime debate: Assessing the limits of longitudinal self-report data. *Social Forces* 76 (4) (June): 1–29.

Le Blanc, Marc. 1989. Designing a self-report instrument for the study of the development of offending from childhood to adulthood: Issues and problems. In *Cross-national research in self-reported crime and delinquency,* edited by Malcolm W. Klein. Los Angeles: Kluwer Academic Publishers.

Lehnen, R.G., and A.J. Reiss. 1978. Response effects in the National Crime Survey. *Victimology: An International Journal* 3:110–124.

Loeber, Rolf, David P. Farrington, Magda Stouthamer-Loeber, Terrie E. Moffitt, and Avshalom Caspi. 1998. The development of male offending: Key findings from the first decade of the Pittsburgh Youth Study. *Studies on Crime and Crime Prevention* 7 (2): 1–31.

Loeber, Rolf, Magda Stouthamer-Loeber, Welmoet B. Van Kammen, and David P. Farrington. 1989. Development of a new measure of self-reported antisocial behavior for young children: Prevalence and reliability. In *Cross-national research in self-reported crime and delinquency,* edited by Malcolm W. Klein. Los Angeles: Kluwer Academic Publishers.

Loeber, R., P. Wung, K. Keenan, B. Giroux, and M. Stouthamer-Loeber. 1993. Developmental pathways in disruptive child behavior. *Development and Psychopathology* 5:101–133.

Matthews, Victor M. 1968. Differential identification: An empirical note. *Social Problems* 14:376–383.

Menard, Scott, and Delbert S. Elliott. 1993. Data set comparability and short-term trends in crime and delinquency. *Journal of Criminal Justice* 21 (5): 433–445.

Merton, R.K. 1938. Social structure and anomie. *American Sociological Review* 3:672–682.

Moffitt, Terrie E. 1997. Adolescence-limited and life-course-persistent offending: A complementary pair of developmental theories. In *Developmental theories of crime and delinquency,* edited by Terence P. Thornberry. Vol. 7 of *Advances in criminological theory.* New Brunswick, New Jersey: Transaction Publishers.

———. 1993. Life-course-persistent and adolescence-limited antisocial behavior: A developmental taxonomy. *Psychological Review* 100:674–701.

Murphy, F.J., M.M. Shirley, and H.L. Witmer. 1946. The incidence of hidden delinquency. *American Journal of Orthopsychiatry* 16 (October): 686–696.

Nettler, G. 1978. *Explaining crime.* New York: McGraw-Hill.

Nye, F.I., and James F. Short, Jr. 1957. Scaling delinquent behavior. *American Sociological Review* 22 (June): 326–331.

Nye, F. Ivan, James F. Short, Jr., and V. Olson. 1958. Socioeconomic status and delinquent behavior. *American Journal of Sociology* 63:381–389.

Olweus, D. 1979. Stability and aggressive reaction patterns in males: A review. *Psychological Bulletin* 86:852–875.

Osgood, D. Wayne, Patrick O'Malley, Jerald Bachman, and Lloyd Johnston. 1989. Time trends and age trends in arrests and self-reported illegal behavior. *Criminology* 27:389–418.

Park, Robert E.K., Ernest W. Burgess, and Roderick D. McKenzie. 1928. *The city.* Chicago: University of Chicago Press.

Patterson, G.R. 1993. Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology* 61:911–919.

Patterson, G.R., and R. Loeber. 1982. The understanding and prediction of delinquent child behavior. Research proposal to the U.S. Department of Health and Human Services, National Institute of Mental Health. Eugene: Oregon Social Learning Center.

Polk, K. 1969. Class strain and rebellion among adolescents. *Social Problems* 17:214–224.

Porterfield, Austin L. 1946. *Youth in trouble.* Fort Worth: Leo Potishman Foundation.

———. 1943. Delinquency and outcome in court and college. *American Journal of Sociology* 49 (November): 199–208.

Reiss, Albert J., Jr., and A.L. Rhodes. 1964. An empirical test of differential association theory. *Journal of Research in Crime and Delinquency* 1:5–18.

———. 1963. Status deprivation and delinquent behavior. *Sociological Quarterly* 4:135–149.

Richman, N., J. Stevenson, and P.J. Graham. 1982. *Preschool to school: A behavioural study*. London: Academic Press.

Robison, Sophia Moses. 1936. *Can delinquency be measured?* New York: Columbia University Press.

Rojek, D.G. 1983. Social status and delinquency: Do self-reports and official reports match? In *Measurement issues in criminal justice,* edited by Gordon P. Waldo. Beverly Hills: Sage Publications.

Sampson, R.J., and J.H. Laub. 1990. Crime and deviance over the life course: The salience of adult social bonds. *American Sociological Review* 55 (October): 609–627.

Saris, Willem E. 1991. *Computer-assisted interviewing.* Newbury Park, California: Sage Publications.

Sellin, Thorsten. 1931. The basis of a crime index. *Journal of Criminal Law and Criminology* 22:335–356.

Shaw, Clifford, and Henry D. McKay. 1942. *Juvenile delinquency and urban areas.* Chicago: University of Chicago Press.

Shaw, D.S., and R.Q. Bell. 1993. Developmental theories of parental contributors to antisocial behavior. *Journal of Abnormal Child Psychology* 21:35–49.

Short, James F. 1957. Differential association and delinquency. *Social Problems* 4:233–239.

Short, J.F., Jr., and F.I. Nye. 1958. Extent of unrecorded juvenile delinquency: Tentative conclusions. *Journal of Criminal Law and Criminology* 49:296–302.

———. 1957. Reported behavior as a criterion of deviant behavior. *Social Problems* 5:207–213.

Silberman, M. 1976. Toward a theory of criminal deterrence. *American Sociological Review* 41:442–461.

Skolnick, J.V., C.J. Braukmann, M.M. Bedlington, K.A. Kirigin, and M.M. Wolf. 1981. Parent-youth interaction and delinquency in group homes. *Journal of Abnormal Child Psychology* 9:107–119.

Slocum, W.L., and C.L. Stone. 1963. Family culture patterns and delinquent-type behavior. *Marriage and Family Living* 25:202–208.

Smith, Carolyn, and Terence P. Thornberry. 1995. The relationship between childhood maltreatment and adolescent involvement in delinquency. *Criminology* 33:451–481.

Stanfield, R. 1966. The interaction of family variables and gang variables in the aetiology of delinquency. *Social Problems* 13:411–417.

Sutherland, Edwin H. 1949. *White-collar crime.* New York: Holt, Rinehart and Winston.

———. 1940. White collar criminality. *American Sociological Review* 5:1–12.

———. 1939. Principles of criminology. 3d ed. Philadelphia: J.B. Lippincott.

Thornberry, Terence P. 1997. *Developmental theories of crime and delinquency.* New Brunswick, New Jersey: Transaction Publishers.

———. 1989. Panel effects and the use of self-reported measures of delinquency in longitudinal studies. In *Cross-national research in self-reported crime and delinquency,* edited by Malcolm W. Klein. Los Angeles: Kluwer Academic Publishers.

———. 1987. Toward an interactional theory of delinquency. *Criminology* 25:863–891.

Thornberry, Terence P., and Marvin D. Krohn. Forthcoming. The development of delinquency: An interactional perspective. In *Handbook of law and social science: Youth and justice,* edited by Susan O. White. New York: Plenum Press.

Thrasher, F. 1927. *The gang: A study of 1,313 gangs in Chicago.* Chicago: University of Chicago Press.

Tourangeau, Roger, and Tom W. Smith. 1996. Asking sensitive questions: The impact of data collection, mode, question format, and question context. *Public Opinion Quarterly* 60:275–304.

Tracy, Paul E., and James A. Fox. 1981. The validity of randomized response for sensitive measurements. *American Sociological Review* 46 (April): 187–200.

Tremblay, Richard E., Christa Japel, Daniel Perusse, Pierre McDuff, Michel Boivin, Mark Zoccolillo, and Jacques Montplaisir. 1999. The search for the age of "onset" of physical aggression: Rousseau and Bandura revisited. *Criminal Behaviour and Mental Health* 9:8–23.

Turner, Charles F., Judith T. Lessler, and James Devore. 1992. Effects of mode of administration and wording on reporting of drug use. In *Survey measurement of drug use: Methodological studies,* edited by Charles F. Turner, Judith T. Lessler, and Joseph C. Gfroerer. Washington, D.C.: U.S. Department of Health and Human Services.

U.S. Department of Justice, National Institute of Justice. 1990. *1988 Drug Use Forecasting annual report: Drugs and crime in America.* Research in Action, NCJ 122225. Washington, D.C.

Vaz, E.W. 1966. Self reported juvenile delinquency and social status. *Canadian Journal of Corrections* 8:20–27.

Voss, H.L. 1966. Socio-economic status and reported delinquent behavior. *Social Problems* 13:314–324.

———. 1964. Differential association and reported delinquent behavior: A replication. *Social Problems* 12:78–85.

———. 1963. Ethnic differentials in delinquency in Honolulu. *Journal of Criminal Law and Criminology* 54:322–327.

Waldo, Gordon P., and Theodore G. Chiricos. 1972. Perceived penal sanction and self-reported criminality: A neglected approach to deterrence research. *Social Problems* 19:522–540.

Wallerstein, J.S., and C.J. Wylie. 1947. Our law-abiding law-breakers. *Probation* 25:107–112.

Weis, J.G., and D.V. Van Alstyne. 1979. The measurement of delinquency by the randomized response method. Paper presented at the 1979 Annual Meeting of the American Society of Criminology, November, Philadelphia.

Weitekamp, Elmar. 1989. Some problems with the use of self-reports in longitudinal research. In *Cross-national research in self-reported crime and delinquency,* edited by Malcolm W. Klein. Los Angeles: Kluwer Academic Publishers.

Williams, Jay R., and Martin Gold. 1972. From delinquent behavior to official delinquency. *Social Problems* 20 (2): 209–229.

Wolfgang, Marvin E., Robert M. Figlio, and Thorsten Sellin. 1972. *Delinquency in a birth cohort.* Chicago: University of Chicago Press.

Wolfgang, Marvin E., Terence P. Thornberry, and Robert M. Figlio. 1987. *From boy to man, from delinquency to crime.* Chicago: University of Chicago Press.

Wright, Debra L., William S. Aquilino, and Andrew J. Supple. 1998. A comparison of computer-assisted and paper-and-pencil self-administered questionnaires in a survey on smoking, alcohol, and drug use. *Public Opinion Quarterly* 62:331–353.