Chapter 2

STUDY METHODOLOGY

Introduction

Chapter 1 discussed the three goals of this study.

- **Sentencing preferences**: Do people with different ethnic backgrounds or from different parts of the county, for instance, favor different sentences?

- **Characteristics of crimes and criminals that influence preferred sentences**: How much does it matter if bystanders are injured in the course of a street crime? For white collar crimes, how important is the amount of money illegally taken?

- **Comparison of the federal guidelines and public sentencing views**: The sentencing guidelines make important distinctions between trafficking in crack cocaine and trafficking in other controlled substances. Does the American public make those same distinctions?

To carry out credible research bearing on these issues requires that sound measurement procedures be designed and implemented, that a representative sample of Americans be obtained, and that the data collected be properly analyzed. Sound measurement procedures are essential. If the measurement procedures are inadequate, either because they fail to capture the phenomena of interest or because they are dominated by noise, then it matters little what the sample is or how the data are analyzed. If the sample does not represent the American population, one cannot draw any useful conclusions about what the American people think about federal sentencing. And if the data are improperly analyzed, there is nothing appropriate to generalize about.

This chapter considers each of these methodological concerns: measurement, sampling, and data analysis.

Instrument Design

As described in Chapter 1, conventional research seeking to elicit the public's views on criminal sanctions typically employs questionnaires with batteries of brief questions. For example, conventional questions are often worded as follows: "Do you favor the death penalty for people who are convicted of murder?"; "What is the appropriate sentence for possession of marijuana?"; "Do you favor truth in sentencing laws that require prisoners to serve their full terms?"; or, "Do you favor life imprisonment without the possibility of parole for offenders convicted for a third time of violent crimes?"

Although these kinds of questions may be of interest to politicians, administrators and policy makers who want to understand public responses to current rhetorical clichés, such questions are not appropriate for this study. As prescribed in the guidelines, federal sentences are explicitly determined by a number of factors, including specified characteristics both of the crime and the offender. In the guidelines it matters if a firearm is used, and it matters whether the offender has had prior prison terms;

sentences are adjusted accordingly. Consequently, asking a question about the appropriate sentence for bank robbery without providing additional information that the guidelines use cannot inform federal sentencing policy. The guideline sentence for bank robbery in fact depends on whether a weapon is used and on the offender's prior record. Accordingly, the appropriate answer to a questionnaire item such as "What do you think is the right sentence for bank robbery?" is "It depends!" In short, simple "one liner" questions cannot be responsive to the mandate to study the correspondence or lack of correspondence between public opinion on sentencing for federal crimes and the sentences prescribed in the guidelines.

Equally important, "one liner" questions about sentencing will fail to capture the views of thoughtful respondents who appreciate that, to a significant degree, sentences must respond to the nature of the crime and the background of the offender. Individuals who might support long prison terms for individuals convicted of smuggling illegal immigrants into the United States for profit, might support short prison terms for individuals who illegally smuggle family members into the United States. If the motive for the smuggling is not specified, thoughtful respondents may refuse to answer, or may make assumptions about motives with sentences ranging accordingly. Both interpretations could undermine the goals of our study.

Finally, if one wants to understand the rationale used by the American public when deciding on appropriate federal sentences, one must examine the impact of factors that are taken into account. Does it matter for the sentence given if the offender has a family to support? How important is it that the offender was the mastermind behind a savings and loan fraud compared to simply participating in a fraud designed by others? How much does it matter if the actions of environment polluters destroy wildlife habitats or if their actions had no discernible effects on such habitats?

Beyond simply describing *how* Americans sentence federal offenders, such information is necessary to inform one or more theories of sentencing. For example, if the American public follows the logic of deterrence theories in sentencing white collar offenders, that should be manifest by their giving prison terms to white collar criminals that increase proportionately with the amount of money illegally taken.

In short, designing an appropriate measurement instrument requires that plausible variations in different kinds of crime be represented. Factorial survey methods are one excellent way to allow for that variation in a form that can be delivered to a large number of subjects. A factorial survey is based on combining sample surveys with one or more factorial experiments, each of which can deal with relatively rich crime descriptions that reflect more fully the complexity of criminal actions.

More specifically, the factorial survey approach involves presenting to respondents short descriptions or vignettes of complex phenomena about which judgments are to be made. The vignettes are constructed by systematically varying critical elements in the phenomenon. When respondents provide evaluations of each vignette, the contrasts between those evaluations provide empirical clues as to the weighting given by respondents to each vignette feature. For example, if vignettes describing crimes in which guns are used receive sentences that are on the average two years longer than vignettes describing crimes in which a gun is not used, one can estimate the average weight given by respondents to the use of a gun.

Actual crimes are composed of a number of features; whether a gun is used is but one possible feature. Consequently, factorial survey methods rely on vignettes that also are composed of a number of features. These features are called dimensions. Whether a gun is used could be one dimension. Whether

a bystander is injured could be another dimension. How much money is taken could be yet another dimension.

Each dimension is composed of two or more categories, known as "levels".[1] Each level for a given dimension represents how the vignette varies along that dimension. To repeat the gun example, a dimension called "weapon use" might have two levels: 1) used a gun, 2) did not use a gun. The dimension called "amount of money taken" might have five levels: 1) $100, 2) $500, 3) $1,000, 4) $5,000, and 5) $10,000.

The vignette dimensions are assembled just as in a factorial experiment: hence, the name "factorial survey." The dimensions are fully crossed with one another: That is, all possible combinations of levels from each of the dimensions used are present in the set of vignettes used. This guarantees that the design is balanced and that the factors are independent of one another. This independence enormously simplifies the data analyses permitting uncontaminated estimates of the weights given by respondents to various dimensions and levels.

Crossing just the weapon dimension and the amount of money dimension leads to ten possible combinations of weapon use and amount of money taken. By adding more dimensions, fully crossed, one can build up not just a large number of combinations of levels, but rather more complicated crime descriptions. For example, one might describe a bank robbery in which the offender threatened a teller with a gun, but did not injure the teller or bystanders in any manner, and got away with $10,000. In the getaway attempt, the car jumped a sidewalk and killed a pedestrian.

One might wish to add characteristics of the offender to bank robbery vignette. One dimension might be whether the offender had any prior felony convictions, and another might be whether the offender had a family to support; both might affect the sentence respondents give.

In principle, one can keep adding dimensions, allowing for increasingly complex vignettes. However, at some point the amount of information exceeds what respondents can digest. Respondents will then either refuse to cooperate, or give up trying to seriously evaluate each of the vignettes. It is critical, therefore, to exclude all but the most important dimensions. Experience suggests that respondents can usually handle three or four dimensions with ease, and often can handle many more. And even three or four dimensions can produce very complex and realistic descriptions in some research settings. But a lot depends on who the respondents are and on the complexity of the material included. As a result, pretesting is a vital part of vignette design.

There is a second limitation. Even with a relatively few fully crossed dimensions, the number of combinations may become too large for any respondent to evaluate. Depending on the complexity of each vignette, that upper bound will likely be somewhere between twenty and fifty vignettes each. In practice, this means that each respondent is presented with a sample from the population of all possible vignettes. Typically, a simple random sample of vignettes is presented to each respondent. This turns the factorial design into a fractional factorial design and guarantees that, on the average, there will be no association between the characteristics of the respondent and the characteristics of the vignettes. Just as in the fully crossed design, random sampling greatly simplifies later data analyses.

---

[1] The term "level" is used in the experimental design literature. It is not to be confused with "offense level" as used in the guidelines.

Choosing which dimensions should be included in a factorial survey usually involves prior knowledge of the topic, past research, and whatever social science theory may be relevant. In this study, however, the relevant dimensions are given in the guidelines. Although we did not incorporate all of the dimensions that are in the guidelines, we did include most of the important ones, *i.e.*, those that count heavily in calculating the guideline sentences.

The design of the study and the feasibility of the factorial survey approach as a data collection modality was tested in a large pre-test designed by the authors and fielded by the Social Science Research Institute of University of California at Los Angeles. Using a quota sample, 200 respondents were chosen to fill out a booklet containing 50 vignettes. The analysis of the pre-test data led to the reduction in the size of the respondent vignette sample size and changes in the wording of alternatives. Analysis of the resulting pre-test data indicated that the factorial survey approach was not only feasible but promised to provide fruitful findings.[2]

The dimensions used in the this study are shown in Appendix A. The major dimension is "Crime Type," which consists of twenty major crime categories, ranging from drug trafficking to kidnapping. These were selected from among the major federal crimes considered in the guidelines. The Crime Types chosen represent either a major portion of the case load in federal courts (such as drug trafficking) or are of special interest (such as environmental violations.) Within each of the major crime categories, there are a number of levels, each constituting an instance of that crime category. For example, drug trafficking is represented by 20 different levels, each indicating the illegal drug involved and the dollar value of the drugs trafficked.

For each crime category, there are typically one or more additional dimensions, often specific to that crime category. For example, the drug trafficking vignettes include dimensions describing the role of the felon in drug trafficking and the use of weapons in the crime. The fraud crime category includes a dimension for the amount of money lost by victims.

In addition, each vignette also contains a short description of the convicted felon consisting of a level from each of the following dimensions: 1)previous felony record, 2) family status, 3) employment status, 4) gender. Thus, respondents are asked to take into account not only the specific offense, but also the background of the offender, before passing sentence. An offender's previous record plays an important role in the guidelines, modifying sentences for all crimes. In contrast the other three offender attributes are not explicitly recognized in the guidelines as grounds for modifying sentences. These three offender characteristics were included because it was thought that the public might consider them relevant to sentencing.

In principle, it is desirable to allow levels to be picked without taking into account the other levels in a vignette. In practice, it is often necessary to restrict certain combinations in order to avoid empirically impossible combinations. Although such exclusions can complicate the analysis, they avoid presenting respondents with vignettes which are bizarre or make little sense. For example, it is not possible to have a police officer with a prior felony record, so previous record is not used with crimes in which a police officer is the perpetrator. A few such restrictions were built into the vignettes as the documentation in Appendix A indicates.

---

[2] Pre-test findings are reported in Rossi, Peter H. and Berk, Richard A. *Fair and Just Punishments: American Views on Sentencing in the Federal Courts. A Report on a Pre-Test for a National Survey.* (1993)

As can be seen in the vignette examples in Chapter 3 (Figure 3.1) the task given to respondents was to mark a desired sentence for each vignette. Respondents were asked to choose between 1) probation, 2) a prison term, and 3) a death sentence. For prison term, respondents were also asked to provide a sentence length in months for sentences up to a year and in years for sentences over one year. Although death is not recognized in the guidelines as a recommended sentence, it was included as an option because the pre-test results indicated that a small number of respondents wanted to have the opportunity to register death as a choice.

When the vignettes were assembled, we were able to specify the probability with which each Crime Type and the probabilities with which each dimension and each level within dimensions would appear in the vignettes. To provide the most information about the most common crimes, the selection probabilities were specified so that the more common crimes, dimensions, and levels appeared more frequently among our vignettes. For example, the probability that a vignette would be a drug trafficking crime was about 0.21 while the probability that the vignette would be an immigration crime was about 0.052; most of the crime probabilities were around 0.05. In a similar fashion, the probability was fixed at 0.80 that the offender was a man and at 0.20 that the offender was a woman.

A total of 101,040 vignettes were generated for the study.[3] These vignettes were incorporated in 2,526 independently generated booklets.[4] Each booklet contained 40 vignettes. Respondents marked sentences they wanted for each vignette in the booklet. These booklets, in turn, were the source of the sentencing information we analyze in the chapters to follow. In addition, two vignettes were presented to the respondent before the booklet was given to acquaint the respondent with the nature of the task and to provide interviewers with an opportunity to explain the task in great detail. These two "practice" vignettes were identical for all respondents and represented respectively a serious crime and a minor crime.[5]

To better understand the rationale behind the sentences given by respondents, we also designed a brief questionnaire to obtain some biographical information on respondents and some measures of attitudes that might be related to sentencing preferences. The questionnaire was administered orally by interviewers who recorded answers on the instrument. The respondents were asked about past involvement with the criminal justice system, personal experiences with crimes, some general ideological tendencies. In addition, the usual assortment of background questions were asked concerning socio-economic status and demographic characteristics. A copy of the questionnaire can be found in Appendix B.

Sampling and Data Collection

---

[3] The vignettes were generated by a computer program that randomly assembled vignettes by picking levels from each of the dimensions needed for the crime randomly chosen for each vignette, and then printed out the resulting vignette.

[4] More booklets were produced than were intended to be used to ensure that each interviewer had on hand a sufficient supply. Because each booklet was independently generated it did not matter whether, say, the 1225th booklet or the 12th booklet was used.

[5] The texts of the two "practice" vignettes are shown in Appendix C.

Data collection for this survey took place over several months starting in January 1994. Under contract with the U. S. Sentencing Commission, the Response Analysis Corporation designed the sample for the study and used its field staff to conduct the needed interviews, using improved versions of the survey instruments used in the pre-test.

The sample design for the study called for 1,500 face-to-face interviews with a sample representative of adults (18 years old or older) living within the continental United States with sufficient English language reading ability to handle the sentencing task. In designing the sampling strategy, three design issues had to be resolved: First, there was a need to balance the lower costs of cluster sampling against the loss in precision that necessarily follows from choosing households in clusters (compared to probability sampling without clustering). The overall plan was to employ a multi-stage cluster sample using 1) primary sampling units — usually counties or MSAs, 2) census tracts or enumeration districts, 3) blocks, 4) households, and 5) one respondent within each household. At each level, probability sampling was used.

Given budget constraints, a key decision was how many secondary units to select. The best compromise was to sample 100 specific locations. Fewer sites would have reduced costs, but produced less precise population estimates. More sites would have produced more precise population estimates, but at a higher price.[6]

Second, respondents had to be selected at random within households in a manner that did not increase non-response. The Kish method is perhaps the best from a technical point of view, but places a high burden on respondents. The "most-recent-birthday" approach does not seem to have a rigorous theoretical justification, but places a very light burden on respondents. The Trodahl-Carter procedure was selected as a compromise between the two. This procedure attempts to assure that only eligible persons in the sampled households have an equal probability of being chosen as respondents.

Third, the booklet of vignettes was to be self-administered. Consequently, literacy in English was essential. (All of the booklets were in English.)[7] Individuals who could not speak or understand English were relatively easy to identify. Individuals with limited English literacy skills were a challenge to identify. Moreover, the process of inquiring or "testing" risked alienating respondents and increasing the refusal rate. The solution employed presented each respondent with two vignettes that were the same for each respondent. These were used to gauge the literacy of respondents and provide practice in evaluating the crime vignettes. Respondents who could not read the two vignettes readily acknowledged it. For respondents who could apparently manage the material, interviewers reviewed the sentences given for the two practice vignettes and asked about any apparent anomalies. In particular, one of the crimes was quite

---

[6] In their final report to the Commission, the contractor, Response Analysis Corporation, reported that 100 specific locations were selected from among the 1,600 secondary sampling units in the 100 primary sampling units in the contractor's area probability sampling frame. Response Analysis Corporation maintains a national multistage area probability sampling frame consisting of 100 primary sampling units. These primary sampling units usually consist of single counties or groups of counties.

[7] Because each vignette booklet was virtually a unique collection, translation costs were prohibitive. However, persons excluded because of language difficulties amount only to less than 5 percent. See Table 2.1.

serious (bank robbery using a gun) and the other was much less serious (possession of a small amount of marijuana for person use). If a longer sentence was given for marijuana possession than for bank robbery, the interviewer asked for a brief explanation. Any reasonably cogent explanation was accepted, and the respondent was encouraged to proceed. If the explanation was totally unreasonable, the respondent was asked whether he/she would be able to read and comprehend the booklet with forty crime vignettes. If the respondent indicated an inability to undertake the task, the interview was terminated.

This method seemed to work well. Overall, designated respondents in 89 percent of the appropriate housing units (occupied non-group quarters) were eligible for the study.[8] Of the ineligible, 60 percent did not speak English and 40 percent were insufficiently literate. And only 2.5 percent of the total needed to justify their answers to the two vignettes.

Table 2.1 shows the disposition of all the 3,018 housing units chosen in the sample. Attempts were made by the interviewing staff to contact all the chosen units repeatedly. Some of the units (8.6%) were not usable because they were vacant or were “group quarters”. Others could not be contacted after many attempts (9.2%) and others (17.5%) refused to be interviewed sufficiently to determine their eligibility for the study. There were also households which were ineligible (7.2%). Counting those who completed the interview as a proportion of those who were eligible, the overall completion rate was 70.1percent, an acceptable response rate by current standards.

Although the resulting sample is a probability sample, the probability of selection was not equal for all respondents. Because this is a household sample, respondents living in multi-adult households had smaller probabilities of selection than those who lived in single-adult households. Properly to represent the total eligible population, it was necessary to devise a set of weights to compensate for those inequalities. A "base weight” was calculated reflecting the probability of selection of each respondent through the various stages of selection. In brief, the base weight for each respondent was computed as the product of the inverses of the probabilities of selection at each of the five stages. Even if the response rate had been 100 percent, the base weight would have been necessary; it follows from the design of the multi-stage sample.

In addition, a weight was computed for each respondent to correct for differential response rates among sub-groups of respondents. Respondents were each classified into one of 27 weighting classes, defined by crossing the nine census divisions with three size stratifications. In effect, this strategy assumes that the characteristics of non-respondents are the same within each of the 27 strata. That is, the strategy assumes that non-respondents in small towns in New England are like small town New Englanders. This assumption is certainly false if taken literally, but similar weighting strategies are routinely used in the analysis of many sample surveys.[9]

---

[8] Because only persons living in conventional housing units fell into the sample, persons living in other housing arrangements—for example, nursing homes and shelters for the homeless or battered women—are not represented in the sample.

[9] The alternative of using unweighted data is more likely to result in biases because without weights all non-respondents are assumed to be like the average respondent in the entire sample, thereby ignoring all regional and size-of-place differences.

Finally, post-stratification weights were constructed. These weights force the weighted marginals to mirror a set of target marginals for the relevant population from the latest available Current Population Survey. The target marginals were for: 1) age, 2) gender, 3) race — black or white, and 4) census division. For example, respondents under 30 years of age were a bit under-represented in the sample, compared to the figures from Current Population Survey. The post-stratification weight for respondents under 30 increased their representation accordingly. This assumes, of course, that the Current Population Survey is more representative than our sample.[10] Although the larger sample size for the Current Population Survey guarantees smaller sampling errors, that leaves unaddressed non-response and response errors. Thus, we have no way of knowing whether the post-stratification weights really improve matters. In any event, these post stratification weights were not used in the analyses.

Table 2.1: Sample Disposition for All Sampled Housing Units in the Study

Total Initial Sample	**3,018**	**100%**
Unusable Units	**262**	**8.6%**
Vacant Dwellings	218	7.2%
Group Quarters	44	1.4%
Usable Units	**2,786**	**91.4%**
Eligibility Unknown	814	26.7%
No Contact	282	9.2%
Initial Refusal	532	17.5%
Eligible	1,753	57.4%
Completed	1,737	57.4%
Not Completed	16	0.5%
Ineligible	219	7.2%
Language Barrier	131	4.3%
Not Literate	88	2.9
Age < 18	0	0.0%
Eligibility Rate	1,753/(1,753+219)	**89%**

[10]Given that response rates in the Current Population Survey are typically higher than 90 percent, that survey suffers less from selection biases.

Usability Rate	2,786/(2,786+262)	**91.4%**
Response Rate	1737/(1753+724)	**70.1%**

In short, weights were computed that, in principle, can reduce possible biases in our sample. We stress the "in principle," however. Only the weights to adjust for the probability of selection in the multistage sample are guaranteed to improve matters. The other weights rest on untestable assumptions that could conceivably make things worse. Fortunately, the results we report in later chapters change little whether weighted or not.

Data Analysis

There are three major complications that need to be considered in any analysis of the vignette crime data. Although these will be discussed in far more depth when they are faced in subsequent chapters, we discuss them initially now.

Respondents were allowed to designate one of three sentence types: probation, imprisonment, and death. In addition, a small but significant proportion[11] of respondents wrote in "life" in lieu of designating a number of years in prison. If probation and prison are represented as sentence lengths, they may be combined into a single variable (with probation coded as zero years in prison) and analyzed in a relatively routine fashion. Although the proper functional form remains an issue (*e.g.*, linear versus logarithmic), time in prison is a perfectly good equal interval variable. However, life imprisonment has no fixed length of imprisonment, which means that the sentence length equivalent of a life sentence does not exist. In addition, a death sentence has no apparent sentence length equivalent of any kind. Hence, the first data analysis problem was how properly to fit life and death sentences into a variable representing sentencing severity.

In the substantive chapters various options for analyzing the sentences given are considered at length. We chose to translate arbitrarily both life imprisonment and death into sentence length equivalents, but then to analyze the data in a fashion so that only their ordinal positions in the sentence length scale matters: much of the analysis focuses on medians rather than means. Using this strategy, the only assumptions are that life imprisonment is worse than the longest sentence in which sentence length is specified, and that death is worse than life in prison.

The second data analysis problem derives from possible non-independence among the sentences given by individual respondents. That is, one risks within-respondent correlated disturbances for almost any analysis that treats vignettes as the units of analysis. In effect, the vignette framework calls for repeated measures (*i.e.*, 40 randomly selected crime vignettes per respondent), and a hierarchical design results (*i.e.*, vignettes nested with respondents).

[11] About 2 percent of all vignettes were given "life" as prison terms. Because interviewers were instructed to discourage "life" as a sentence, this percentage is undoubtedly an underestimate of the popularity of this sentencing alternative.

It is well known that unaddressed correlated disturbances lead to optimistic estimates of the standard errors in which one risks underestimating the role of sampling variation. Accordingly, the effective sample is smaller than the nominal sample. In later chapters various analytic options are considered in detail. But suffice it to say that with a nominal sample size defined as the number of respondents times the number of vignettes per respondent, the loss of power that might be caused by correlated disturbances is far too little to make an important difference. Effects that are large enough to be of substantive interest have very large t-values that would be well over 2.0 if the effective sample size rather than the nominal sample size were used.

The final data analysis problem is whether to use the vignette as the unit of analysis or the respondent as the unit of analysis. The formal solution in principle is to employ a hierarchical model in which both units are used at once. Vignettes are the micro units and respondents are the macro units. However, our response variable does not meet the assumptions of conventional hierarchical models (*i.e.*, special cases of the general linear model). And as described in Chapter 5, our use of robust regression procedures to buffer the results against the arbitrary coding of life sentences and death sentences precludes hierarchical analyses. Unfortunately, hierarchical modeling has not been extended to the robust methods we employ.

Therefore, separate analyses are conducted for each of the two possible units of analysis. This approach is justifiable in part because, by design, each respondent receives his or her own simple random sample of vignettes. Consequently, characteristics of the vignettes are independent of characteristics of respondents, and the two kinds of units can be analyzed separately without bias.