4

Juvenile Delinquency Treatment: A Meta-Analytic Inquiry into the Variability of Effects

Mark W. Lipsey

One need not look beyond the daily newspaper to establish that crime is a matter of considerable concern in our society. Far less obvious is what should be done about it. As with almost any important matter, this is one on which opinions can differ sharply, not only in the political arena but among social science researchers and criminological experts as well.

Among the many approaches to crime prevention that have been advocated are punishment (deterrence), amelioration of social conditions that produce crime, target hardening, community prevention (e.g., neighborhood watches), and a host of other such notions. Of particular interest here are the options articulated for dealing with an actual or potential perpetrator once he or she (usually he) is identified. The two major schools of thought have been labeled "just desserts" (i.e., punishment proportionate to the offense) and "rehabilitation" (i.e., treatment aimed at reforming the miscreant and preventing future criminal behavior) (Cullen and Gilbert 1982). While these approaches are not necessarily mutually exclusive in practice, they differ so greatly in philosophy that they do not easily coexist within a given criminal justice program.

One domain within which rehabilitation has particular appeal is that of juvenile crime. While juvenile crime is often serious, and unquestionably represents a large proportion of the total criminal activity in a community, the nature of adolescence is generally seen as justifying special handling, a concept institutionalized in the separate juvenile

courts in which minors are tried. The most important feature of adolescence, of course, is that it is a formative period marked by behavior that will not necessarily be continued into adulthood. A rehabilitative strategy that shapes the delinquent offender toward more prosocial behavior during this formative stage, therefore, is particularly attractive. Also, since youth have a potentially long adulthood before them, the payoff in reduced criminality over a lifetime resulting from effective preventive intervention at an early age can be substantial.

The potential benefits of rehabilitation for juvenile delinquents can be attained, of course, only if effective intervention is applied. Unfortunately, the question of whether delinquency intervention in general or various specific varieties of intervention are in fact effective has not been convincingly resolved despite decades of research by behavioral scientists. It was the goal of the study described in this chapter to make the most comprehensive and probing attempt to date to synthesize and interpret the large body of research on the effects of preventive or rehabilitative treatment for delinquency. Moreover, given the history of controversy and uncertainty about the results of this body of research, it was deemed important to also attempt to determine why it has proven so difficult to interpret.

This chapter provides an overview of a large meta-analytic survey of the delinquency treatment research conducted over the last four decades. Before turning to a description of the methods and results of this investigation, a brief look at the history of previous attempts to review the delinquency treatment literature will provide some useful context.

Previous Research Reviews

Best known among the reviews of the criminal rehabilitation literature is Lipton, Martinson, and Wilks's broad survey of research on correctional treatments (1975). They made a detailed examination of 231 separate studies involving interventions for both juveniles and adults. Martinson's widely quoted conclusion was that "with few isolated exceptions, the rehabilitative efforts that have been reported so far have had no appreciable effect on recidivism" (1974, p. 25). Greenberg, who updated the Lipton et al. review, echoed that pessimism: "The blanket assertion that 'nothing works' is an exaggeration, but not by very much" (1977, p. 141).

Reviews that have focused exclusively on delinquency treatment have reached similar conclusions. Romig (1978), for example, attempted to

identify the characteristics of successful treatment of delinquents, cataloging the available studies with a level of detail rivaling that of Lipton et al. In each category of treatment, however, he found relatively few convincing positive results to report. More critical stances on delinquency treatment research were taken by Wright and Dixon (1977) and Lundman, McFarlane, and Scarpitti (1976). They too reported finding little evidence of significant effects on juvenile crime.

The predominantly negative reviews of rehabilitation that dominated the 1970s were not without challenge. Palmer (1975, 1983), for example, argued that they overlooked many positive instances of success in their haste to generalize and gave little attention to the issues of fit between the type of juvenile and the type of treatment. In a similar vein, Gendreau and Ross (1979) offered "bibliotherapy" to discouraged professionals in the form of a summary of correctional treatments that had produced positive evaluation results. Even more pointed remarks came from Gottfredson (1979), who satirically itemized the "treatment destruction techniques" that critics could use to discredit any promising treatment concept.

The tone for the 1980s was set by the reports of the National Academy of Science's (NAS) Panel on Research on Rehabilitative Techniques (Sechrest, White, and Brown 1979; Martin, Sechrest, and Redner 1981). The first report combed through the available treatment evaluation research, and reviews of that research, and reluctantly concluded that there was indeed little evidence of successful treatment for either juveniles or adults: "Although a generous reviewer of the literature might discover some glimmers of hope, those glimmers are so few, so scattered, and so inconsistent that they do not serve as a basis for any recommendation other than continued research" (Sechrest, White, and Brown 1979, p. 3). *But* the NAS panel emphasized the possibility that the problem was the nature of the evidence rather than the failure of the concept. In particular, they identified a variety of factors essential to credible evaluation research that they found lacking in this literature—well-controlled designs, sensitive measures, strong and well-implemented treatments, and the like.

For purposes of making a broad assessment of delinquency treatment effects, the most important development since the National Academy of Science's report has been the rise of meta-analysis as a technique for aggregating the continuously growing research literature. Unfortunately, the meta-analyses to date have been limited efforts that, in many ways, have raised more questions than they have answered.

The most extensive of these meta-analyses was conducted by Garrett (1984, 1985), who focused on adjudicated delinquents placed in resi-

dential facilities, either community or institutional. She examined 111 studies of the effects of treatment programs in such facilities on a wide range of outcome variables. On the other end of the spectrum, Kaufman (1985) restricted his meta-analysis to "prevention" treatment of preadjudicated at-risk juveniles. He looked only at delinquency outcome measures used in randomized research designs using a sample of 20 studies. The meta-analysis by Gottschalk et al. (1987) fell somewhere between these other two efforts. Their sample of 90 studies included only treatment of adjudicated delinquents, but was not restricted to treatment in residential facilities; indeed, only about half the sample was residential.

All three of these meta-analyses found a positive grand mean effect size for the better-designed studies, averaged over studies and outcome measures. The magnitude of this overall effect ranged from around one-fourth to one-third of a standard deviation superiority for the mean treatment group outcome compared with the mean control group outcome. These three studies differed, however, in their assessment of the statistical significance of this effect (Garrett did not test; Kaufman reported significance; Gottschalk et al. reported nonsignificance) and on most other topics examined. In particular, results were inconsistent with regard to the relative efficacy of different treatment modalities, the role of amount or frequency of treatment, and the relationship of research design to study outcome.

More recently, Whitehead and Lab (1989) conducted a meta-analysis of 50 delinquency treatment studies published in journals since 1975 that used control groups and dichotomous recidivism measures. They adopted the Phi coefficient for an effect size index and chose, apparently arbitrarily, a value of .20 as the minimum necessary to consider an effect worthwhile (equivalent to an effect size of about .41 in standard deviation units). A simple count, using no statistical analysis, showed a minority of studies yielding effects as large as .20 and Whitehead and Lab concluded that, therefore, treatment was not effective. From their summary table, the mean Phi coefficient can be computed as .12, equivalent to a difference of about .25 standard deviation units between treatment and control (Cohen 1988). Despite the authors' disparaging conclusions, therefore, this meta-analysis also yielded a positive mean effect of about the same order of magnitude as the previous efforts.

Andrews et al. (1990) responded to the negative conclusions of the Whitehead and Lab meta-analysis with their own reanalysis, augmented by additional studies. They distinguished between appropriate correctional services, defined as those delivered to high-risk cases us-

ing modes of treatments matched with client learning styles, and various categories of inappropriate services. They found a mean Phi coefficient of .30 for appropriate services (equivalent to .63 standard deviation units), which was significantly larger than the mean values for inappropriate services. The grand mean over all the studies in their analysis was an effect size of .21 standard deviation units, quite comparable to those found in virtually all the previous meta-analyses.

While meta-analytic reviews of delinquency treatment are reaching somewhat more favorable conclusions than earlier conventional reviews, the efforts of this sort to date have been quite circumscribed. Moreover, the different approaches and restrictions adopted by the various meta-analysts make it difficult to compare their results or find interpretable patterns across them.

The meta-analysis reported here was designed to improve on previous reviews of delinquency treatment research, both conventional and meta-analytic, in the following ways: (1) broadening the coverage of the literature by making an exhaustive search for relevant studies, both published and unpublished; (2) coding sufficient detail from each eligible study to support a probing analysis of the correlates of measured treatment effects, including those stemming from the research methods used as well as from the nature and circumstances of treatment; (3) applying state of the art statistical analysis for meta-analytic data to properly assess the magnitude of the effects found in these studies and the sources of variability in those effects.

This chapter is a preliminary report of the major results from that meta-analytic investigation. It focuses primarily upon the variability in the delinquency treatment effects found in the research literature and identification of the major sources of that variability.

Methods

Eligibility Criteria

Research reports were defined as eligible for inclusion in this meta-analysis according to a set of detailed criteria specified prior to the search for relevant studies and periodically revised to incorporate new distinctions required by ambiguous instances that had to be resolved during the search. In abbreviated form, these criteria were as follows:

1. There had to be some intervention or treatment, broadly defined, that had as its aim (explicitly or implicitly) the reduction, preven-

tion, treatment, remediation, and so forth, of delinquency or antisocial behavior problems similar to delinquency. Delinquency was defined as behavior chargeable under applicable laws whether or not apprehension occurs or charges are brought; antisocial behavior was defined as actions that are threatening, disruptive, or damaging to property, to other persons, or to self. The large category of treatments targeted solely on substance abuse and no other component of antisocial behavior, however, was excluded.

2. The majority of the subjects to whom the treatment was applied had to be juveniles, defined as persons age 21 or younger. To exclude childhood behavior problems without legal implications, however, studies involving juveniles below the age of 12 were not included unless the antisocial behaviors treated were clearly of a type chargeable as delinquent offenses.
3. There had to be measured outcome variables with quantitative results reported that included at least one delinquency measure. Additionally, there had to be some comparison that contrasted one or more designated treatments with one or more designated control conditions on those outcome variables.
4. The treatment versus control comparison groups used in the research had to be based on random assignment of subjects to conditions or, if assignment was nonrandom, there had to be both premeasures and postmeasures on the outcome variable; some evidence of matching prior to treatment; or a range of measures of such characteristics as prior delinquency history, sex, and age which allowed some assessment of the similarity of the treatment and control groups prior to treatment. Pre-test-post-test studies with no control group and post-test-only comparisons between nonrandom groups with no information about group equivalence were not eligible.
5. To maintain some homogeneity in cultural context and social meaning of delinquency, studies had to be set in the United States or a substantially similar English-speaking country (e.g., Canada, Britain, Australia) and reported in English. The juvenile subjects in the study, however, were not required to be English-speaking or "Anglo"; for example, studies of Latino delinquents set in the United States qualified.
6. To restrict the studies to the relatively modern era with regard to criminal justice practices and conceptions of delinquency, studies were eligible only if the date of reporting or publication was 1950 or later, that is, post–World War II.

Identification and Retrieval of Eligible Research Reports

Bibliographic citations for potentially eligible research reports were obtained primarily from three sources. One initial source was the bibliographies of previous literature reviews and meta-analyses—for example, those cited earlier in this chapter. The major source was a comprehensive search in the bibliographic databases of the Dialog system. For this purpose an extensive set of keywords was developed around alternative expressions of the concepts "research," "delinquency," and "treatment." Keyword searches for studies with title, abstract, or index terms representing conjunctions of these three concepts were then conducted in all the Dialog databases that were judged potentially relevant. Appendix 4.A lists the databases examined. In a few instances where keyword search was judged less than optimal, it was supplemented by manual searches through the bibliographic volumes themselves.

A third source of bibliographic information was citations within the reports that were identified by the above procedures and subsequently retrieved and screened for eligibility.These and other incidental sources produced more than 8,000 citations for which available information indicated potential relevance to the meta-analysis. The bibliography is quite complete through 1986 and is currently being updated to include more recent material. It should be noted that the procedures for generating this bibliography included no restrictions according to type of report or nature of publication. Thus books, technical reports, conference papers, theses, and dissertations, as well as published journal articles, were included.

As much as possible of the material identified in this bibliography was located at university libraries in the southern California area, through interlibrary loan or through bulk purchases of microfiche from relevant services. Reports that could not be located through these channels (mostly technical reports and conference papers) were pursued by writing directly to authors whenever addresses could be found either in the original citation or in membership directories for such organizations as the American Psychological Association, American Society of Criminology, and American Sociological Association. At the time of this writing, search and retrieval activities are still continuing but the preponderance of identified material has been either located and screened for eligibility or declared unretrievable after persistent effort.

Coding of the Studies

Each eligible report was coded by a doctoral student in psychology who had been trained in the task through study of a detailed coding manual and supervised practice coding. The coding scheme consisted of 154 items that a coder completed on the basis of the text of the selected report. Additionally, certain of these items were repeated if a study had multiple outcome measures or breakdowns of the results for subsets of subjects or different times of measurement. Table 4.3 (presented later) lists the major variables coded. A brief description of the major categories of information extracted by the coding follows.

EFFECT SIZE. The major treatment group versus control group comparison was selected for each study and all quantitative outcome variables contrasting those two groups were identified. These variables were then divided into those that indexed delinquent behavior and those that represented other behavior or characteristics: for example, school grades, self-esteem. For each such outcome variable, a coding was made of the direction of the effect—that is, whether it favored the treatment group, control group, or neither. Studies without direction of effect information for at least one outcome variable were dropped from further consideration.

Where sufficient quantitative information was reported, an effect size estimate was then computed for each outcome variable. The effect size index used for this purpose was Cohen's *d* (Cohen 1988), defined generally as the difference between the treatment group mean score and the control group mean score divided by the pooled standard deviations of those scores. The resulting effect size was given a positive value if treatment group performance was superior or "better" than control group performance and a negative value if control group performance was superior. Effect size was thus represented as the number of standard deviation units by which the treatment group outperformed the control group on the identified outcome variable. This is the basic formulation developed and elaborated for meta-analysis by Glass and Hedges (Glass, McGaw, and Smith 1981; Hedges 1981; Hedges and Olkin 1985).

When means and standard deviations were not reported for an outcome variable, which occurred with unfortunate frequency, the effect size was estimated, if possible, from whatever statistical information was reported—*p, t, or F* values, contingency tables, and the like. A common, but not universal, form for reporting delinquency outcome

in the studies was recidivism rate: the proportion of subjects in each experimental group who were rearrested, reconvicted, or whatever subsequent to treatment. Such proportion and percentage data were converted to effect size estimates using the arcsine transformation described in Cohen (1988).

Effect size estimates were computed using the statistics available for the comparison of treatment and control groups on each outcome variable without attempting to adjust for any lack of comparability between the groups at the time of measurement. These effect sizes, of course, may be biased upward or downward by such factors as nonrandom designs that yield initial nonequivalence between the experimental groups and attrition from either or both groups after assignment to experimental conditions. The approach that was taken to this problem was to code separately the information that was available in each study regarding these matters, that is, nonequivalence and attrition. That information was subsequently used in the analysis to determine the nature of its relationship with effect size and, where relationships were found, to partial them out statistically before considering what effects might be attributable to treatment. Similarly, the details of the various measures—for example, source, type, and period covered for delinquency measures—were coded so that differences in effect size that were related to the metric used could also be statistically controlled in later analyses.

In addition to the overall "aggregate" comparison between treatment and control groups on each outcome variable, many of the studies also reported results for subgroups: for example, males versus females. When possible, effect size estimates were also computed for these "breakdown" groups separately: for example, effect size for males and effect size for females. Moreover, both "aggregate" and "breakdown" comparisons sometimes had follow-up measures, that is, outcome variables measured at more than one time subsequent to treatment. When possible, effect size estimates were separately computed for each follow-up comparison.

The full coding on effect size, therefore, represented delinquency and nondelinquency outcome variables for the aggregate treatment versus control comparison, any breakdowns of that comparison, and any follow-up comparisons of either the aggregate or the breakdown comparison. This chapter reports only on delinquency outcome for the aggregate comparison at the first point of measurement subsequent to treatment. It is these data that give the broadest overview of the effects of delinquency treatment.

METHOD VARIABLES. To enable inquiry into the relation between the methodology used in a study and the effects found in that study, a wide range of information was coded about study design, measures, samples, attrition, and the like. Particular attention was given to the extent of initial equivalence between treatment and control groups prior to application of the treatment.

STUDY CONTEXT. When available, information was coded about the year and form of publication of each study, country in which it was conducted, source of funding, and characteristics of the researcher—for instance, institutional affiliation and discipline.

NATURE OF TREATMENT. An important set of issues, of course, has to do with the characteristics of the treatments used in the various studies and their relationships to the study outcome. Accordingly, information was coded regarding the treatment type, setting, sponsorship, duration, intensity, and a wide range of other such features.

NATURE OF SUBJECTS. Study outcome may also vary with the nature of the juvenile subjects treated. The coding scheme recorded, where available, information about the demographic characteristics of the juveniles (e.g., race, sex, age), prior delinquency history, and other such matters.

Results

The analyses presented here focus on the distribution of measured delinquency effects in the studies coded for this meta-analysis; that is, on the effect size indices for the treatment versus control group comparisons on delinquency outcome variables. Results on other (nondelinquency) outcomes will be reported in later papers. At the time of this writing 443 studies were coded and available for analysis and the results that follow are based on that set.

Effect Size Distribution for Delinquency Outcomes

Many of the studies had multiple delinquency outcome variables. Creating a distribution of effects for all of them would have overrepresented studies that reported more variables and underrepresented those that reported fewer variables. This distortion was judged undesirable because of both the statistical dependencies created by using multiple

effects from a single study and the potential misrepresentation of the pattern of outcomes across studies. Instead, an analysis of all delinquency measures was first done to identify the types of variables that were most commonly used in this literature. Then, when multiple delinquency outcomes were available in a study, a single one was selected for analysis according to criteria designed to identify that representing the category most widely used in other studies. This selection was done blindly with regard to the effect size on the various candidate measures. Since the most common delinquency outcome measure was some variation on the concept of recidivism—rearrest, reconviction, and so on—this process acted to favor measures of that sort. The measures selected in this manner will be referred to as "primary" delinquency measures. (Table 4.3, which will be discussed in more detail later, provides descriptive information about these measures, especially in items 32 through 37.)

After selection of the primary delinquency measure for each study, the effect size for that measure was weighted by the coefficient developed by Hedges (1981) to correct for bias in estimation. Effect sizes based on small samples tend to run larger than the population values that they estimate and must be reduced proportionately. The specific weighting coefficient used for all effect sizes in this study was $1 - (3/(4n_t + 4n_c - 9))$ where n_t is the sample size for the treatment group and n_c is the sample size for the control group (Hedges 1981; Hedges and Olkin 1985). Where it is necessary to be specific, this formulation of the effect size will be referred to as the "n-adjusted effect size."

DIRECTION AND SIZE OF EFFECTS. We are now in a position to examine the effect size distribution for evidence regarding the efficacy of delinquency treatment. The treatment effect information that covers the largest number of studies is the coding of the simple direction of the difference between the treatment and control groups. A difference favoring treatment indicates that the treatment group had a better outcome (less delinquency) than the control group; a difference favoring the control group indicates that the treatment group had a worse outcome; a difference favoring neither group indicates that they were exactly equal or that the original study reported no significant difference without providing actual values. Table 4.1 shows the breakdown of direction of effect for the primary delinquency measures, one from each of the 443 studies in the present analysis.

If, in the general case, treatment is not effective in reducing delinquency we would still expect some positive and negative differences due to sampling error, but would expect them in equal proportions. If

Table 4.1 Direction of Treatment versus Control Group Differences on Primary Delinquency Outcome Measure for All Studies

	N	%
Favors Treatment	285	64.3
Favors Control	131	29.6
Favors Neither	27	6.1
Total	443	

Binomial test (by z approximation) that population proportions are .50/.50: $z = 7.32$ $p<.001$ (hypothesis rejected).

we take the relatively few cases in which neither treatment nor control group is favored and divide them evenly between the other two categories, the proportions can be tested by the normal distribution approximation to the binomial to determine if they depart from the expectation of a 50-50 split (Siegel 1956). As indicated in Table 4.1, the large skew toward differences that favor the treatment group is statistically significant.

To get information about the magnitude (not just the direction) of the differences between treatment and control groups in these studies, we must turn to the computed effect size values. Since not all studies in the database provided sufficient information for computation of effect size, a smaller set is available for this analysis ($n = 397$). Direction of effect breakdowns for this subset of studies (not shown) were virtually identical with those shown in Table 4.1 for all studies.

Figure 4.1 presents the distribution of n-adjusted effect size values on the primary delinquency outcome measures and summary statistics for the distribution. The mean and median effect size values are positive, though numerically modest, showing that on average these studies found lower delinquency for treatment groups than control groups. The unweighted mean of .172 is the value most directly comparable to the results of the other delinquency meta-analyses reviewed in the introduction to this chapter. Recall that those values were in the range of .20 to .33 but represented more highly selected sets of studies.

It should be noted that the effect size values represented in Figure 4.1 give equal representation to each study irrespective of its sample size (other than the slight correction already applied for biased estimation in small samples). Thus one effect size is contributed to this distribution by a study of $n = 10$ and, similarly, only one is contributed by a study of $n = 1{,}000$. Since effect size estimates based on large

Figure 4.1 Distribution of Unweighted *n*-Adjusted Effect Sizes for Primary Delinquency Measures

Count	ES	
2	−1.20	xx
2	−1.10	xx
0	−1.00	
1	−.90	x
1	−.80	x
4	−.70	xxxx
6	−.60	xxxxxx
7	−.50	xxxxxxx
9	−.40	xxxxxxxxx
11	−.30	xxxxxxxxxxx
26	−.20	xxxxxxxxxxxxxxxxxxxxxxxxxx
33	−.10	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
49	.00	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
20	.00	xxxxxxxxxxxxxxxxxxx
48	.10	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
33	.20	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
38	.30	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
18	.40	xxxxxxxxxxxxxxxxx
25	.50	xxxxxxxxxxxxxxxxxxxxxxxx
13	.60	xxxxxxxxxxxxx
14	.70	xxxxxxxxxxxxxx
12	.80	xxxxxxxxxxxx
4	.90	xxxx
8	1.00	xxxxxxxx
1	1.10	x
2	1.20	xx
1	1.30	x
3	1.40	xxx
0	1.50	
3	1.60	xxx
1	1.70	x
2	1.80	xx

Summary Statistics:	
Number of cases	397
Unweighted mean	.172
Median	.100
Standard deviation	.438
Variance	.192

samples are statistically more reliable than those based on small samples, some accommodation of the sample size differences must be made in order to test the statistical significance of the means of the effect size distribution.

Hedges and Olkin (1985) have shown that the optimal procedure is to weight each effect size inversely by its variance (which reflects sam-

Table 4.2 Statistical Tests for Effect Size Means and Homogeneity

A. *n*-Adjusted Effect Sizes for All Studies		
Inverse-variance weighted ES mean	.103	(n = 397)
.99 confidence interval for mean	.083 to .123	
Inverse-variance weighted ES variance	.089	
Homogeneity test statistic	H = 1319.00	df = 237
Chi-square .01 critical value	273.78	
B. *n*-Adjusted Effect Sizes for Studies with Random Assignment		
Inverse-variance weighted ES mean	.110	(n = 294)
.99 confidence interval for mean	.086 to .134	
Inverse-variance weighted ES variance	.080	
Homogeneity test statistic	H = 904.14	df = 293
Chi-square .01 critical value	351.46	
C. *n*-Adjusted Effect Sizes for Studies with Random Assignment and No Appreciable Attrition from Experimental Groups		
Inverse-variance weighted ES mean	.140	(n = 78)
.99 confidence interval for mean	.094 to .186	
Inverse-variance weighted ES variance	.090	
Homogeneity test statistic	H = 281.08	df = 77
Chi-square .01 critical value	107.98	

ple size). This was done in the present data with the restriction that treatment and control n were separately Windsorized at 300 to prevent a few very large studies from dominating the results.

With inverse-variance weighted effect sizes, a confidence interval can be determined and the statistical significance of the mean effect size for the distribution in Figure 4.1 can be assessed. Table 4.2 (part A) summarizes the statistical information for this procedure: the inverse-variance weighted mean, the confidence interval, and some homogeneity statistics that will be discussed later.

The inverse-variance weighted effect size mean shown in Table 4.2 (part A) is positive and very similar to the median value for the unweighted effect sizes shown in Figure 4.1. The .99 confidence interval does not include zero; thus the positive mean effect displayed here is statistically significant.

One might wonder, however, if the positive value of the mean effect size is simply a reflection of biased results in the studies represented. In particular, since many of these studies did not use randomly assigned control groups, the positive mean effect size may indicate only initial nonequivalence between the treatment and comparison group reappearing as a pseudo-treatment effect in the outcome measures. If, for example, the juveniles selected for treatment were less delinquency

prone, on average, than those selected for comparison groups, post-treatment outcome measurement would be expected to show differences favoring treatment.

Table 4.2 (parts B and C), reports the results of two tests of this possibility. First, the mean effect size and confidence interval were determined for that subset of studies which reported random assignment to experimental conditions (n = 294). Second, and more probing, the mean effect size and confidence interval were computed for the subset of studies which reported both random assignment and no (or trivial) attrition from the experimental groups between assignment and outcome measurement (n = 78). This latter case excludes studies that used random assignment but, prior to outcome measurement, may have lost that initial equivalence because of differential attrition from the treatment and control groups.

As Table 4.2 shows, the mean effect size for each of these selected subsets of studies is positive and, indeed, very similar to the mean for all the studies together. The modest differences are in the positive direction; that is, the better controlled studies yielded slightly larger mean effect sizes than the general mix of studies. Moreover, the confidence intervals indicate that the mean effect sizes for the selected subsets of studies are statistically significant. It does not appear that the overall positive effect size mean can be attributed to bias resulting from inclusion of results from poorly controlled studies.

The answer to the general question "Does treatment reduce delinquency?" therefore appears to be "Yes, on average there is a positive effect." But, while positive and statistically significant, the mean effect sizes found here appear relatively modest. If we take the inverse-variance weighted mean for the distribution of effects on the primary delinquency measures as the standard, treated juveniles showed about .10 standard deviation units less delinquency subsequent to treatment than did the control juveniles. At first impression, this sounds quite trivial.

This figure is more meaningful if we translate it into something more directly relevant than standard deviation units. Since the modal measure represented in these data is a rearrest recidivism rate, one alternative is to express the mean effect in those terms. If we assume that control groups without treatment recidivate at the rate of 50 percent, which is about the mean value for those studies that used simple dichotomous recidivism measures (n = 208), we can convert the treatment-control difference from standard deviation units to percentages using the arcsine transformation from Cohen (1988). This procedure shows that .10 standard deviation units is equivalent to a decrease of 5

percentage points from a 50 percent baseline. In other words, the mean treatment effect of .10 standard deviation is equivalent to a reduction in average recidivism from 50 to 45 percent. This formulation of the effect is much more interpretable and, while it still shows that the result is modest, does reveal that it is not trivial. A reduction in recidivism of 5 percentage points from a baseline of 50 percent amounts to a 10 percent decrease in recidivism (5/50). While a 10 percent average drop may not be spectacular, it cannot be said to be obviously negligible.

Moreover, the true effect represented in these studies is almost certainly larger than these figures indicate. Effect sizes are attenuated by the unreliability of the study outcome measures upon which they are calculated. With few exceptions, the measures in the present collection of studies represent some aspect of officially recorded delinquency—arrests, probation violations, reconvictions, and the like. It is well known that officially recorded contacts represent a small proportion of the total number of delinquent behaviors in which a juvenile engages (Williams and Gold 1972). As a result, it is largely a matter of chance whether a particular delinquent act eventuates in an officially recorded contact with an agent of law enforcement or the juvenile justice system. This large chance component makes such delinquency measures very unreliable. Lipsey (1982, 1983) estimated their reliability to be around .20–.30.

We can correct the inverse-variance weighted mean effect size of .10 for the attenuation that would result from the low reliability of the delinquency outcome measures at issue simply by dividing by the square root of the reliability (Hedges 1981). If we assume reliability of .25, the resulting deattenuated mean effect size is .20; that is, it doubles. Translating this into simple dichotomous recidivism terms, we find that it is equivalent to a decrease in a treatment group of 10 percentage points from a control baseline of 50 percent recidivism. Or, in overall percentage terms, it is equivalent to a 20 percent decrease in recidivism (10/50). Without the masking effect of highly unreliable delinquency measures, therefore, the overall treatment effect found in this meta-analysis could be quite large enough to have practical significance.

Despite this relatively positive finding with regard to delinquency treatment, some care must be taken in the conclusions drawn at this point. What we have shown is that the average treatment versus control difference in these studies favors the treatment group. The extent to which that difference reflects the efficacy of the treatments employed, rather than some other feature of these studies—for example, some methodological characteristic—is still in some doubt (though Ta-

ble 4.2 seems to rule out one of the more obvious possibilities). The grand mean effect size averaged over so many diverse studies is rather like a main effect in a complex analysis of variance design: It generalizes over all the other factors and interactions that may be influencing the outcome to make a crude overall comparison. Before interpreting that main effect, we need to determine whether it is equally representative of the results of all the types of studies in the database. This, in turn, requires testing of the homogeneity of the effect size distribution.

HOMOGENEITY TESTS. If the values in the effect size distribution are tightly clustered around the mean—for example, varying no more than would be expected by sampling error—that mean is a reasonable representation of the outcome of each and all of the studies. If the variation is great, however, the mean may not represent any distinct group of studies and may be quite misleading. Of particular concern in the present context is the possibility that methodologically low-quality studies would spuriously yield larger effect sizes than higher-quality studies, thus biasing the distribution upward and overstating the magnitude of the actual effects of treatment.

Hedges (1982) has developed a test of the homogeneity of effect sizes that is useful in this regard. It requires computation of a term, H, that can be tested with the chi-square distribution. Table 4.2 reports the summary statistics for the homogeneity tests on the distributions of effect sizes for all studies and for the selected subsets of studies. All three distributions show significant heterogeneity. Indeed, for the full set of studies the H statistic, which is a sum of squares term, shows more than three times as much heterogeneity as would be expected on the basis of sampling error alone.

The task to which we now turn is attempting to identify the sources of the variability in the effect size distributions. In particular, we want to try to determine the extent to which variation in study methodology contributes to effect size variation in contrast to differences among studies on the substantive factors of treatment and subject type.

Analyzing Effect Size Variability

If some of the variability in effect sizes is systematically related to differences in the studies from which they originate, we should be able to find a pattern of correlations between the relevant study characteristics and effect size. Our ability to investigate such correlations is limited by the availability of variables representing study characteristics in the meta-analysis which, in turn, is limited by what authors report when

they write up their studies. Table 4.3 lists the major study characteristics that were coded in the present meta-analysis, reports the frequency breakdown on each for the 443 studies in the present database, and indicates the proportion of studies for which information on the item was unavailable.

For purposes of analyzing effect size variability, study characteristics were grouped into 11 clusters (all but "Outcome" on Table 4.3). These clusters, in turn, represent three larger categories—study context, method, and treatment. The clusters are listed descriptively below with a shorthand label for each. They are sequenced from the more fundamental and general methodological issues to the more study-specific issues of treatment and study context. The items included in each are marked with an asterisk or double asterisk in Table 4.3.

Method
- Experimental groups, sample size, sampling (Samples)
- Initial equivalence of experimental groups (Equivalence)
- Attrition from experimental groups (Attrition)
- Characteristics of the control condition (Control)
- Characteristics of the delinquency outcome measures (Measures)
- Information about the effect size computation (ES Info)

Treatment
- Characteristics of subjects/clients treated (Subjects)
- Amount or intensity of treatment (Dosage)
- Characteristics of the condition (Treatment)
- Treatment philosophy and context (Tx Philos)

Study Context
- Country, publication year, author's discipline, etc. (Context)

A straightforward approach to analyzing the variability of a single dependent variable (effect size in this case) as a function of various independent variables (such as those in the above clusters) is multiple regression. To employ this technique, however, a number of procedural and conceptual issues must be faced.

One problem, noted earlier, is the uneven sample sizes upon which the effect sizes are based. A study with a large sample should be given more weight in the analysis than one with a small sample since it represents information about the response of more people and yields more reliable results. Hedges and Olkin (1985) have shown that the same inverse-variance weights that were used earlier to compute effect size means, confidence intervals, and homogeneity statistics can be applied

Table 4.3 Descriptive Data for Major Variables Coded

	N	%		N	%
STUDY CONTEXT					
1. Country of Study			Conference paper	7	1.6
United States	407	91.9	Missing	0	0.0
Canada	12	2.7	6. Year of Publication		
Britain	15	3.4	1950–1959	5	1.1
Other	7	1.6	1960–1969	58	13.1
Missing	2	0.5	1970–1979	207	46.7
2. Author's Discipline			1980–1987	166	37.5
Psychology	135	30.5	Missing	7	1.6
Criminal justice	68	15.3			
Sociology	43	9.7	METHOD		
Education	43	9.7	*Experimental Groups, Sample Size, Sampling (Samples)*		
Social work	24	5.4			
Psychiatry/medicine	12	2.7			
Political science	8	1.8			
Other	7	1.6			
Missing	103	23.3			
3. Author's Affiliation					
Academic	246	55.5			
Government agency	53	12.0			
Program agency	88	19.9			
Research firm	30	6.8			
Other	2	0.5			
Missing	24	5.4			
4. Source of Research Funding					
Agency/organization	127	28.7			
Federal	126	28.4			
State/local government	56	12.6			
Funded, unknown source	15	3.4			
No funding indicated	117	26.4			
Missing	2	0.5			
5. Type of Publication					
Journal/book chapter	168	37.9			
Technical report	192	43.3			
Dissertation/thesis	44	9.9			
Book	32	7.2			
7. Number of Treatment Groups in Design**					
One	364	82.2			
Two	52	11.7			
Three	18	4.1			
More	6	1.4			
Missing	3	0.7			
8. Number of Control Groups in Design					
One	341	77.0			
Two	79	17.8			
Three	15	3.4			
More	2	0.4			
Missing	6	1.4			
9. Post-Test Total Sample Size**					
1–25	38	8.6			
26–50	59	13.3			
51–75	47	10.6			
76–100	39	8.8			
101–150	65	14.7			
151–200	46	10.4			
201–300	47	10.6			
301–500	30	6.8			
501–800	21	4.7			
801+	27	6.1			
Missing	24	5.4			

Table 4.3 *(Continued)*

	N	%
10. Method Quality: Representativeness of Sampling**		
Low	140	31.6
Moderate	164	37.0
High	138	31.2
Missing	1	0.2
11. Method Quality: Statistical Power**		
Low	227	51.2
Moderate	113	25.5
High	103	23.3
Missing	0	0.0
Initial Equivalence of Experimental Groups (Equivalence)		
12. Unit on Which Assignment to Experimental Groups Based		
Individual	409	72.3
Intact group	20	4.5
Program area	10	2.3
Missing	4	0.9
13. Procedure for Assignment to Groups**		
Random after matching	61	13.8
Random, no matching	134	30.2
Regression discontinuity	4	0.9
Wait list control	12	2.7
Nonrandom, matched on pretest	14	3.2
Nonrandom, matched on individual features	37	8.4
Nonrandom, matched on demographics	71	16.0
Matched groupwise	22	5.0
Random with serious degradation	27	6.1
Individual selection (e.g., by need)	32	7.2
Convenience comparison group	28	6.3
Missing	1	0.2
14. Confidence/Explicitness of Assignment Procedure*		
Very low	1	0.2
Low	13	2.9
Moderate	36	8.1
High	111	25.1
Very high	279	63.0
Missing	3	0.7
15. Method Quality: Treatment/Control Group Comparability*		
Low	88	19.9
Moderate	202	45.6
High	153	34.5
Missing	0	0.0
16. Rating: Overall Similarity of Treatment and Control*		
Very similar 1	12	2.7
2	114	25.7
3	141	31.8
4	75	16.9
5	58	13.1
6	27	6.1
Very different 7	3	0.7
Missing	13	2.9
17. Confidence/Explicitness of Group Similarity*		
Very low	2	0.5
Low	15	3.4
Moderate	190	42.9

	N	%
High	188	42.4
Very high	37	8.4
Missing	11	2.5
18. Researcher's Comparison of Treatment/Control Equivalence*		
No comparisons made	95	21.4
No statistically significant differences	96	21.7
Significant differences unimportant	27	6.1
Significant differences uncertain	51	11.4
Significant differences important	26	5.9
Descriptive differences unimportant	71	16.0
Descriptive differences uncertain	46	10.4
Descriptive differences important	20	4.5
Missing	11	2.5
19. Direction of Treatment/Control Pre-Test Difference*		
Favors treatment	64	14.4
Favors control	53	12.0
Favors neither	7	1.6
Missing	319	72.0
20. Direction of Treatment/Control Sex Difference**		
Favors treatment	53	12.0
Favors control	55	12.4
Favors neither	99	22.3
Missing	236	53.3
21. Direction of Treatment/Control Age Difference**		
Favors treatment	94	21.2
Favors control	71	16.0
Favors neither	66	14.9
Missing	212	47.9
22. Direction of Treatment/Control Ethnicity Difference*		
Favors treatment	64	14.4
Favors control	67	15.1
Favors neither	50	11.3
Missing	262	59.1
23. Direction of Treatment/Control Delinquency History Difference*		
Favors treatment	66	14.9
Favors control	59	13.3
Favors neither	32	7.2
Missing	286	64.6
24. Direction of Treatment/Control Delinquency Typology Difference**		
Favors treatment	24	5.4
Favors control	22	5.0
Favors neither	23	5.2
Missing	374	84.4
Attrition from Experimental Groups (Attrition)		
25. Treatment Group *N* Change from Pre- to Post-Test**		
Gain	9	2.0
Loss	108	24.4
No difference	241	54.4
Missing	85	19.2
26. Control Group *N* Change from Pre- to Post-Test**		
Gain	11	2.5
Loss	96	21.7
No difference	246	55.5

Table 4.3 *(Continued)*

	N	%		N	%
Missing	90	20.3	31. Number of Delinquency Outcome Measures Not Codable*		
27. Method Quality: Attrition Problems**			None	330	74.5
Low	136	30.7	One	43	9.7
Moderate	183	41.3	Two	24	5.4
High	115	26.0	Three	12	2.7
Missing	9	2.1	Four	8	1.8
Characteristics of the Control Condition (Control)			Five	6	1.4
28. Type of Control Condition**			More	11	2.4
No treatment	57	12.9	Missing	9	2.0
Wait list	17	3.8	32. Weeks After Treatment Begins When Primary Measure Taken**		
Minimal contact	32	7.2	1–13	79	17.8
Treatment as usual	307	69.3	14–26	114	25.7
Placebo	18	4.1	27–52	111	25.1
Other	6	1.4	53–112	61	13.8
Missing	6	1.4	113+	32	7.2
29. Confidence/Explicitness of Control Condition*			Missing	46	10.4
Very low	0	0.0	33. Period Covered in Primary Delinquency Measurement, Weeks**		
Low	4	0.9	1–13	60	13.5
Moderate	45	10.2	14–26	131	29.6
High	129	29.1	27–52	130	29.3
Very high	255	57.6	53–112	52	11.7
Missing	10	2.3	113+	30	6.8
Characteristics of the Delinquency Outcome Measures (Measures)			Missing	40	9.0
30. Number of Delinquency Outcome Measures Codable**			34. Type of Delinquency Represented in Primary Measure*		
One	164	37.0	Antisocial behavior	24	5.4
Two	86	19.4	Unofficial delinquency	19	4.3
Three	65	14.7	School disciplinary	12	2.7
Four	38	8.6	Arrests/police contact	195	44.0
Five	27	6.1	Probation contact	35	7.9
More	47	10.6			
Missing	16	3.6			

	N	%
Court contact	80	18.1
Parole contact	25	5.6
Institutional disciplinary	15	3.4
Institutionalization	28	6.3
Catchment area indicator	4	0.9
Missing	6	1.4
35. Range of Offenses Covered in Primary Measure*		
All offenses	385	86.9
Status offenses only	10	2.3
Other restricted	37	8.4
Missing	11	2.5
36. Type of Scaling of Primary Delinquency Measure*		
Dichotomous recidivism	247	55.8
Summed dichotomy	9	2.0
Frequency or rate	141	31.8
Severity index	11	2.5
Event timing	6	1.4
Rating of amount	8	1.8
Other	8	1.8
Missing	13	2.9
37. Source of Data for Primary Delinquency Measure**		
Self-report, juvenile	33	7.4
Therapist, teacher, etc.	14	3.2
School records	14	3.2
Police records	127	28.7
Probation records	41	9.3
Court records	114	25.7
Institutional records	52	11.7
Other records	5	1.1
Missing	43	9.7
38. Validity of Primary Delinquency Measure Demonstrated?		
Yes	16	3.6
No	427	96.4
39. Reliability of Primary Delinquency Measure Demonstrated?		
Yes	22	5.0
No	421	95.0
40. Sensitivity of Primary Delinquency Measure Demonstrated?		
Yes	1	0.2
No	442	99.8
41. Rating: Overlap of Measure with Content of Treatment**		
Very low 1	137	30.9
2	82	18.5
3	58	13.1
Moderate 4	61	13.8
5	39	8.8
6	25	5.6
Very high 7	37	8.4
Missing	4	0.9
42. Rating: Potential for Social Desirability Bias*		
Very low 1	311	70.2
2	62	14.0
3	20	4.5
Moderate 4	9	2.0
5	14	3.2
6	9	2.0
Very high 7	14	3.2
Missing	4	0.9
43. Confidence/Explicitness re Overlap and Social Desirability*		

Table 4.3 *(Continued)*

	N	%		N	%
Very low	2	0.5	Missing	7	1.6
Low	6	1.4	47. Confidence/Explicitness re Delinquency Risk*		
Moderate	46	10.4			
High	218	49.2	Very low	1	0.2
Very high	164	37.0	Low	4	0.9
Missing	7	1.6	Moderate	59	13.3
44. Method Quality: Psychometric Properties of Primary Measure**			High	160	36.1
			Very high	213	48.1
			Missing	6	1.4
Low	286	64.6	48. Proportion of Juveniles with Prior Offense History*		
Moderate	106	23.9			
High	51	11.5	None	16	3.6
Missing	0	0.0	Some	62	14.0
45. Method Quality: Blinding in Collection of Outcome Data*			Most	68	15.3
			All	206	46.5
			Some, can't estimate	50	11.3
Low	287	64.8	Missing	41	9.3
Moderate	90	20.3	49. Predominant Type of Prior Offenses		
High	55	12.4			
Missing	11	2.5			
			No priors	18	4.1
TREATMENT			Mixed	149	33.6
Characteristics of Subjects/ Clients Treated (Subjects)			Person crimes	6	1.4
			Property crimes	91	20.5
			Status offenses	39	8.8
46. Level of Delinquency Risk/Involvement**			Other	11	2.3
			Missing	129	29.1
Nondelinquent, normal	3	0.7	50. Aggressive History of Juveniles*		
Nondelinquent, symptomatic	26	5.9	No	116	26.2
Predelinquents	64	14.4	Yes, some juveniles	91	20.5
Delinquents	155	35.0	Yes, most juveniles	7	1.6
Institutionalized, nonjuvenile justice	7	1.6	Yes, all juveniles	7	1.6
			Some, can't estimate	69	15.6
Institutionalized, juvenile justice	87	19.6	Missing	153	34.5
			51. Sex of Juveniles*		
Mixed, low end	37	8.4	No males	10	2.3
Mixed, high end	33	7.4	Some males	26	5.9
Mixed, full range	24	5.4	Mostly males	188	42.4

		N	%
All males		154	34.8
Some, can't estimate		18	4.1
Missing		47	10.6
52. Average Age of Juveniles at Time of Treatment**			
6–11		8	1.8
12		7	1.6
13		38	8.6
14		92	20.8
15		100	22.6
16		83	18.7
17		22	5.0
18		15	3.4
19		19	4.3
20–21		6	1.4
Missing		53	12.0
53. Predominant Ethnicity of Juveniles**			
Anglo		143	32.3
Black		52	11.7
Hispanic		8	1.8
Other minority		2	0.5
Mixed, none >60%		70	15.8
Mixed, can't estimate		32	7.2
Missing		136	30.7
54. Rating: Overall Heterogeneity of Treated Juveniles**			
Very homogeneous	1	2	0.5
	2	98	22.1
	3	142	32.1
Moderately heterogeneous	4	82	18.5
	5	67	15.1
	6	23	5.2
Very heterogeneous	7	4	0.9
Missing		25	5.6
55. Confidence/Explicitness re Information on Heterogeneity			
Very low		2	0.5
Low		25	5.6
Moderate		209	47.2
High		179	40.4
Very high		2	0.5
Missing		26	5.9
56. Source of Clients for Treatment**			
Voluntary, family		14	3.2
Non–criminal justice agency		33	7.4
Criminal justice agency, voluntary		142	32.1
Criminal justice agency, mandatory		201	45.4
Multiple sources		14	3.2
Researcher solicits		30	6.8
Missing		9	2.1
Amount or Intensity of Treatment (Dosage)			
57. Duration, Weeks from First to Last Treatment Event**			
1–6		69	15.6
7–13		60	13.5
14–26		108	24.4
27–39		51	11.5
40–52		52	11.7
53–78		9	2.0
79–112		18	4.1
113+		10	2.3
Missing		66	14.9
58. Frequency of Treatment Contact**			
Continuous		71	16.0
Daily		55	12.4
2–4 per week		48	10.8
1–2 per week		151	34.1

Table 4.3 *(Continued)*

		N	%
Less than weekly		45	10.2
Missing		73	16.5
59. Mean Hours Contact per Week*			
Less than 1		45	10.2
1–2		108	24.4
3–5		44	9.9
6–10		30	6.8
11–20		12	2.7
21–30		9	2.0
31–50		10	2.3
51–100		4	0.9
Continuous		70	15.8
Missing		111	25.1
60. Mean Total Number of Hours of Contact**			
1–10		65	14.7
11–20		32	7.2
21–40		42	9.5
41–100		40	9.0
101–200		37	8.4
201–1,000		35	7.9
1,000+		8	1.8
Continuous		71	16.0
Missing		113	25.5
61. Confidence/Explicitness of Information on Treatment Amount*			
Very low		7	1.6
Low		46	10.4
Moderate		111	25.1
High		127	28.7
Very high		95	21.4
Missing		57	12.8
62. Rating: Amount of Meaningful Contact**			
Trivial	1	15	3.4
	2	59	13.3
	3	82	18.5
Moderate	4	71	16.0
	5	90	20.3
	6	70	15.8
Substantial	7	21	4.7
Missing		35	7.9
63. Rating: Intensity of Treatment Event**			
Weak	1	11	2.5
	2	54	12.2
	3	108	24.4
Moderate	4	119	26.9
	5	75	16.9
	6	27	6.1
Strong	7	9	2.0
Missing		40	9.0
64. Confidence/Explicitness re Ratings of Amount/Intensity*			
Very low		9	2.0
Low		34	7.7
Moderate		191	43.1
High		162	36.6
Very high		15	3.4
Missing		32	7.3
65. Evidence of Degradation in Treatment Delivery**			
Yes		132	29.8
Possible		68	15.3
No		195	44.0
Missing		48	10.8
66. Method Quality: Integrity of Treatment Implementation*			
Low		194	43.8
Moderate		158	35.7
High		87	19.6
Missing		2	0.5
Characteristics of the Treatment Condition (Treatment)			
67. Role of Researcher in Treatment**			

	N	%
Delivered treatment	28	6.3
Planned, controlled	162	36.6
Influential, no direct role	57	12.9
Independent of treatment	157	35.4
Missing	39	8.8
68. Treatment Modality; Therapy Type**		
Juvenile Justice Interventions		
Probation, regular	2	0.5
Probation, added counseling	36	8.1
Probation, restitution	12	2.7
Probation, other enhancement	37	8.4
Parole, regular	2	0.5
Parole, enhanced	15	3.4
Institutionalization, regular	4	0.9
Institutionalization, added counseling	43	9.7
Institutionalization, community residential	13	2.9
Institutionalization, other enhancement	33	7.4
Deterrence, shock contact	11	2.5
All other juvenile justice interventions	7	1.6
Non–Juvenile Justice Interventions		
Residential, camp	21	4.7
School, added counseling	17	3.8
School, other enhancement	26	5.9
Individual counseling	23	5.2
Group/family counseling	33	7.4
Other counseling	14	3.2
Behavioral therapy	24	5.4
Skill/employment training	36	8.1
Service broker, multimodal	29	6.5
All other	5	1.1
Missing	0	0.0
69. Confidence/Explicitness re Treatment Modality*		
Very low	0	0.0
Low	0	0.0
Moderate	47	10.6
High	140	31.6
Very high	251	56.7
Missing	5	1.1
70. What the Treatment Attempts to Change		
Broadband delinquency	238	53.7
Status offenses	21	4.7
Other specific offenses	16	3.6
School performance	22	5.0
Psychological attribute	52	11.7
Social attribute	52	11.7
Skill level	27	6.1
Other	10	2.3
Missing	5	1.1
71. Who Administers the Treatment**		
Criminal justice personnel	112	25.3
School personnel	19	4.3
Public mental health personnel	44	9.9
Private mental health personnel	77	17.4

Table 4.3 *(Continued)*

	N	%
Non mental health counselors	44	9.9
Laypersons	85	19.2
Researcher	14	3.2
Other	16	3.6
Missing	32	7.2
72. Format of Treatment Sessions**		
Juvenile alone	22	5.0
Juvenile and provider	123	27.8
Juvenile group	180	40.6
Juvenile with family	47	10.6
Mixed	44	9.9
Other	10	2.3
Missing	17	3.8
73. Treatment Site a Public Facility**		
Yes, criminal justice	138	31.2
Yes, non–criminal justice	86	19.4
No, private	132	29.8
Mixed	31	7.0
Other	17	3.8
Missing	39	8.8
74. Treatment Site a Residential/Institutional Setting**		
Yes	123	27.8
No	302	68.2
Mixed	7	1.6
Missing	11	2.5
75. Formal Setting*		
Yes	311	70.2
No	65	14.7
Mixed	36	8.1
Missing	31	6.8
Treatment Philosophy and Context (Tx Philos)		
76. Treatment Level of Theoretical Development**		
Black box label	60	13.5
Action strategy	134	30.2
Conceptual rationale	140	31.6
Hypothesis testing	40	9.0
Integrated theory	68	15.3
Missing	1	0.2
77. Treatment Etiological Orientation**		
Individual	163	36.8
Individual, mixed	106	23.9
Sociological, micro	88	19.9
Sociological, macro	32	7.2
Labeling	23	5.2
Sociological, mixed	24	5.4
Missing	7	1.6
78. Program Age*		
New (<2 years)	277	62.5
Established	155	35.0
Defunct	5	1.1
Missing	6	1.4
79. Program Sponsorship**		
Researcher, one cohort	112	25.3
Researcher, multiple cohorts	34	7.7
Independent private	41	9.3
Public, non–criminal justice	85	19.2
Public, criminal justice	165	37.2
Missing	6	1.4
80. How Fully Treatment Is Described*		
Detailed	62	14.0
General	166	37.5
Descriptive label	188	42.4
No description	22	5.0
Missing	5	1.1

		N	%
OUTCOME			
Descriptive Outcome			
81. Tone of Report			
Positive		315	71.1
Neutral		102	23.0
Negative		22	5.0
Missing		4	0.9
82. Author's Interpretation of Study Result			
Success		226	51.0
Mixed		123	27.8
Failure		60	13.5
No conclusion		20	4.5
Missing		14	3.2
Statistical Outcome, Primary Delinquency Measure			
83. Direction of Treatment/Control Difference at Post-Test			
Favors treatment		285	64.3
Favors control		131	29.6
Favors neither		18	4.1
Missing		9	2.0
84. Statistical Significance of Post-Test Difference			
Significant		97	21.9
Not significant		177	40.0
Missing		169	38.1
85. Unadjusted Post-Test Effect Size			
−2.00 to −1.00		4	0.9
−0.99 to −0.50		14	3.2
−0.49 to −0.25		25	5.6
−0.26 to −0.01		79	17.8
0.00		21	4.7
+0.01 to +0.25		111	25.1
+0.26 to +0.50		72	16.3
+0.51 to +1.00		54	12.2
+1.00 to +2.00		17	3.8
Missing		46	10.4
Statistical Information re Effect Sizes/Outcomes (ES Info)			
86. Confidence/Explicitness of Information for Post-Test Effect Size**			
Highly estimated	1	4	0.9
Moderate estimation	2	5	1.1
Some estimation	3	12	2.7
Slight estimation	4	33	7.4
No estimation	5	334	75.4
Missing		55	12.4
87. Type of Post-Test Means Reported**			
Arithmetic		167	37.7
Median		2	0.5
Proportion		242	54.6
Other		9	2.0
Missing		23	5.2
88. Type of Post-Test Variances Reported			
Standard deviation		125	28.2
Variance		1	0.2
Standard error		4	0.9
Proportion		215	48.5
Other		5	1.1
Missing		93	21.0
89. Type of Statistical Test Researcher Used for Post Difference			
No report		103	23.3
t, F, z		107	24.2
Chi-square		93	21.0
Nonparametric		16	3.6
ANCOVA		15	3.4

Table 4.3 *(Continued)*

	N	%		N	%
Blocked	2	0.5	Low	105	23.7
Other	3	0.7	Moderate	206	46.5
Missing	104	23.5	High	130	29.3
90. Method Quality:			Missing	2	0.5
Controls for Subject Heterogeneity			92. Confidence/Explicitness for Overall Method Quality Ratings**		
Low	209	47.2			
Moderate	151	34.1			
High	83	18.7	Very low	2	0.5
Missing	0	0.0	Low	8	1.8
91. Method quality:			Moderate	58	13.1
Appropriateness of Statistical Analysis*			High	303	68.4
			Very high	71	16.0
			Missing	1	0.2

*Variables included in initial cluster definitions for multiple regression analyses.
**Variables included in pared-down clusters for hierarchical multiple regression analysis.

to this situation. The approach used here to analyze the variation in effect sizes, therefore, is a weighted multiple regression in which the contribution of each case (study) to the analysis is weighted by the inverse variance of the effect size.

Additionally, not all the potential predictor variables for these analyses were in the form of graduated or continuous measures appropriate for correlational analysis; many were categorical. Categorical variables with more than two categories were recoded into a rank order sequence that reflected the natural progression of the categories if there was one. If there was not, conceptually similar and small n categories were aggregated and each case was dummy coded, 1 or 0, to index membership in each of the resulting categories.

Another issue that arises in this analysis is how to handle the missing values, since a fair number are sprinkled throughout the items in the predictor clusters. A two-step procedure was used to resolve this matter. First, a missing value indicator for each potential predictor variable was created in dichotomous form: 1 if a value was present and 0 if it was missing. These dichotomies were then correlated with the effect size dependent variable to determine if there was any relationship between missing data in a set of studies and the effect sizes found in those studies. Then, in the regression analyses, the means of nonmiss-

ing values on a predictor were substituted for each missing value in order to keep the number of cases up. If the proportion of missing values was under 10 percent, no further adjustments were made. If the proportion of missing values was greater than 10 percent, however, the correlation for the missing value dichotomy was examined. If nonsignificant, no further adjustments were made; if significant, the missing value dichotomy was itself entered into the regression equation along with the original variable from which it was derived. With this procedure, all cases could be used in the analysis, but any information about effect size carried by the fact that information on an item was missing for some studies was retained. Although a number of these "missing value" codes were involved in the preliminary regression analysis, none proved sufficiently strong as predictor variables to be retained in the final regression model.

Finally, some attention must be paid to the multicollinearity of the predictor variables and variable clusters, that is, the correlations and confoundings among the predictors themselves. To the extent that there are appreciable correlations, especially among the clusters of variables that are the primary focus in the present analysis, decisions must be made about where to allocate the confounded variance and in what sequence the clusters should be entered as predictors into the analysis.

One cluster—study context—proved to have no predictive power beyond that available in the other clusters and was dropped from further consideration. Once the specifics of the method and treatment used in a study were accounted for, items such as discipline of the author and year of publication that constituted the study context cluster added nothing else. This is not surprising since we would not expect the author's training and other such matters to have influence on study results except by way of the character of the specific treatments and methods employed in the study.

CLUSTERS OF PREDICTOR VARIABLES. We first examine the relationship of each individual cluster of variables to effect size. This is done by constructing a single weighted multiple regression for each cluster in which only the variables from that cluster are entered as predictors. The question to be answered here is whether any of these clusters show notable correlations with effect size and thus potentially explain some of its variance. More particularly, we would like to know if the variability in effect sizes primarily reflects differences in methodology used in the various studies or if it primarily reflects differences in the treatments and treatment circumstances under investigation. If the former,

Table 4.4 Multiple Correlation of Predictor Clusters with Effect Size (Diagonals) and with Each Other (Off-Diagonals)

Method										
Samples	.20*									
Equivalence	.08	.28*								
Attrition	.11	.10	.22*							
Control	.01	.16*	−.14*	.08						
Measures	.04	.27*	.09	.16*	.28*					
ES Info	.06	.02	−.07	.05	.15*	.10				
Treatment										
Subjects	.11	.04	.02	.12*	.08	.12	.19			
Dosage	.03	.07	.05	−.01	.09	.04	.07	.24*		
Treatment	.12	.11	.16*	.02	.09	.19*	.11	.09	.40*	
Tx Philos	−.01	.09	.06	.07	.11	.17*	.04	.01	.18*	.20*
	Samp	Equi	Attr	Cont	Meas	ESIn	Subj	Dosa	Trea	TxPh
	Method						Treatment			

*$p < .05$

we have methodological bias that must be accounted for; if the latter, we have interesting differences in the effectiveness of treatment that bear further investigation.

The diagonals of the matrix in Table 4.4 report the multiple correlations between each cluster of independent variables and effect size. All the clusters having to do with treatment produced relatively large multiple correlations, as did some of the method clusters. In particular, clusters having to do with the sampling, equivalence between experimental groups, attrition, and characteristics of the delinquency outcome measures used were moderately correlated with effect size.

We must, however, consider the possibility of confoundings among the variables represented in the clusters. Dosage, for example, might correlate with effect size because studies that used high dosage also happen to frequently use a design that biases effect sizes upward. The off-diagonal correlations in Table 4.4 show the relationships among the clusters. They are obtained by using the regression equation for each cluster to compute a predicted effect size for each case and then correlating those predicted values. Some of those correlations are quite low, showing little relationship between the variables in one cluster and those in another, but others are large enough to raise a question about the independence of the relationship of the respective clusters with effect size. In particular, there are four statistically significant correlations

showing confoundings between a method cluster and a treatment cluster. What appear to be relationships between the nature of the treatment and the resulting effect may therefore only reflect confounded method artifacts.

The next step in the analysis was to use hierarchical multiple regression with these clusters to examine their conjoint relationship with effect size. To conserve degrees of freedom, variables were dropped from each cluster if neither the zero-order correlation with effect size nor the beta coefficient in the multiple regression equation for the cluster reached .10, so long as this did not make the cluster size smaller than three variables or omit a variable of unusual conceptual interest (e.g., whether subjects were randomly assigned). The variables remaining in these pared-down clusters are marked with a double asterisk in Table 4.3.

HIERARCHICAL MULTIPLE REGRESSION. The pared-down clusters were stepped into the hierarchical weighted multiple regression in the order indicated on Table 4.4 and the listing above. Entering all the method clusters before any of the treatment clusters made it possible to examine the independent contribution of treatment characteristics beyond those explainable by methodological characteristics with which they were confounded. Within the method category the sequence allowed investigation of the successive influence of the nature of the samples, initial group equivalence, subsequent attrition, the nature of the control group, particulars of outcome measurement, and particulars of the effect size information. This sequence was chosen to reflect the approximate temporal sequence of the major methodological steps in mounting experimental research. That is, samples are drawn before assignment to groups, assignment precedes attrition, and so forth. Thus where there are confoundings between clusters, the contested variance in the effect size distribution is assigned to the methodological step that comes earliest in the sequence.

Within the treatment category, any number of reasonable sequences might be adopted. The sequence that was chosen (subjects, dosage, treatment, treatment philosophy) was designed to be conservative about attributing effects to specific treatment modalities if they could be accounted for by more general factors. Stepping the subject cluster into the analysis as the first of the treatment clusters, for example, ensured that no effects would be attributed to dosage and treatment modality if they could alternatively be accounted for by differences among types of subjects in their responsiveness to treatment. Similarly, entering dosage before treatment modality ensured that no effect would be attributed to specific treatment types that might only be a general func-

tion of the amount or intensity of treatment delivered, irrespective of type. Treatment philosophy, on the other hand, is a general factor (philosophy, nature of setting, etc.), but it was stepped in last on the presumption that these matters should have only indirect influence on treatment outcomes. The only interesting aspect of treatment philosophy, in other words, is what influence it might have that cannot be explained by the specifics of the subjects, dosage, and treatment type.

In addition to examining the relative influence of the different variable clusters themselves, it is interesting to consider the possibility of interactions among the clusters. The regression weights from preliminary analysis were used to construct factors combining the individual variables within each cluster into a single composite variable. The cross-products of these factors could then be entered as additional predictor variables in the regression analysis to examine the influence of cluster level interactions. Since the total number of cross-product terms for 11 clusters is quite large, testing of interactions was limited to two-way interactions (e.g., dosage by treatment modality) and, further, to those cross-products that seemed most promising in preliminary analysis.

A cluster of cross-product terms representing interactions among the method clusters was entered in the analysis after the last method cluster but before the first treatment cluster. Similarly, a cluster of cross-products representing interactions between method clusters and treatment clusters was entered after the last treatment cluster. Finally, a cluster representing interactions among treatment clusters was entered after everything else.

Table 4.5 reports the summary results for this stepwise regression procedure and indicates the variance accounted for by each cluster as it is added to the regression equation. The method clusters and method interactions altogether have a multiple correlation of .50 with effect size, accounting for 25 percent of the variability in effect size. Of the method clusters, all but the one representing the nature of control groups (Control) and the one encoding effect size information (ES Info) made statistically significant contributions to predicting effect size.

Even more interesting, perhaps, is the strong relationship of the treatment clusters to effect size above and beyond what could be attributed to the method variables. Addition of these clusters and their interactions increased the multiple correlation from .50 to .68 and accounted for an additional 22 percent of the variance in effect size. All of the treatment clusters made statistically significant contributions to predicting effect size except those dealing with subject characteristics (Subjects). Most of the contribution of treatment variables came from the cluster having to do with the treatment modality (Treatment). The

Table 4.5 Summary Table for Stepwise Hierarchical Inverse-Variance Weighted Multiple Regression Using All Clusters to Predict Effect Size on the Primary Delinquency Measure

Step	Variable Cluster	Cumulative Multiple R	Cumulative R-Square	R-Square Change	Change as Proportion of Total R-Square
	Method			.25	.53
1	Samples	.20	.04	.04*	.09
2	Equivalence	.31	.10	.06*	.12
3	Attrition	.36	.13	.03*	.07
4	Control	.40	.16	.03	.06
5	Measures	.44	.20	.04*	.08
6	ES Info	.46	.21	.01	.03
7	Meth x Meth	.50	.25	.04*	.09
	Treatment			.22	.47
8	Subjects	.51	.26	.01	.02
9	Dosage	.53	.29	.03*	.07
10	Treatment	.63	.40	.11*	.24
11	Tx Philos	.65	.42	.02*	.04
12	Tx x Meth	.68	.46	.04*	.09
13	Tx x Tx	.68	.47	.01	.02

*$p < .05$

effect size found in a delinquency treatment study thus depends substantially upon the methodological characteristics of the study, but it is also importantly influenced by the nature and circumstances of the treatment under study, as indeed we would expect.

Overall, therefore, the clusters of predictor variables included in this analysis accounted for nearly 50 percent of the variability in the effect size distribution. Of that, the largest share (53 percent) was associated with methodological variables, but the independent contribution of treatment variables was also considerable.

At this point, we can ask how well the multiple regression model performed in accounting for the total variability in effect size among the studies. As shown in Table 4.2, the variance of the distribution of *n*-adjusted effect sizes was calculated to be .089, a value more than three times as great as expected from sampling error alone. The variance of the residuals from the multiple regression was .047, or 53 percent of the total (consistent with an $R^2 = .47$). Testing those residuals for homogeneity yielded $H = 798.61$ ($df = 311$), to be compared with

an alpha = .01 critical chi-square value of 371.17. While substantially reduced, significant heterogeneity still remained in the effect size distribution after fitting the multiple regression model.

Despite its statistical significance, however, it seems unlikely that the variation in the effect sizes not accounted for by the model was meaningful or important. The variance of the residuals, .047, includes a portion of approximately .024 (27 percent of total variance) attributable to sampling error (computed using techniques from Hedges 1984). Additionally, it almost certainly includes measurement error in the effect size values themselves, many of which were estimated from limited statistical information available in the study reports and subject, further, to whatever errors that coders may have made in computations with the information. A recoding of 25 studies (approximately every fifteenth) yielded a correlation of about .90 between effect size estimates for different coders, but this does not reflect the error inherent in estimating effect size from incomplete statistical information as was sometimes done. If the overall measurement error in effect size is as high as 20 percent of the nonsampling error variance (i.e., reliability coefficient = .80), then another 15 percent of the total variance must be measurement error (.20 (.089 − .024)/.089). With 47 percent of the variance accounted for by the multiple regression model, 27 percent by sampling error, and 15 percent by measurement error in the effect size estimates, only about 11 percent is left unaccounted for. Little of the variability remaining after fitting the multiple regression model, therefore, is likely to be meaningful despite its statistical significance.

CLUSTER-LEVEL RELATIONSHIPS WITH EFFECT SIZE. Detailed discussion and interpretation of the weightings of the individual predictor variables in each cluster that resulted from the multiple regression exceeds the scope of this chapter. Moreover, some refinement of the coding and categorization beyond the present preliminary form is doubtless necessary before such detailed scrutiny will be fully rewarding. It is possible, however, to give a general characterization of the relationship between each major cluster of variables (excluding interactions) and the distribution of effect sizes. A summary of those relationships is presented in Table 4.6.

METHOD. The method cluster that accounted for the largest proportion of variance in effect size was that dealing with the pre-treatment equivalence of the treatment and control groups used in the study (Equivalence). Not surprisingly, the greater the magnitude and number of differences between the treatment and control groups prior to treat-

Table 4.6 General Nature of the Multiple Regression Results for Each Major Variable Cluster

Cluster	R^2 Change	
Method		
Samples	.04	Larger studies with larger sample sizes were associated with smaller effect sizes.
Equivalence	.06	Specific dimensions of initial nonequivalence between treatment and control groups (e.g., sex, delinquency type) were associated with larger or smaller effect sizes. Overall method of subject assignment (e.g., random vs. nonrandom), however, was not associated with effect size.
Attrition	.03	Greater attrition from either treatment or control group was associated with smaller effect sizes.
Control	.03	Control groups receiving some contact, e.g., "treatment as usual" in the juvenile justice system, were associated with smaller effect sizes than "no treatment" controls except for probation treatment as usual.
Measures	.04	Large number of delinquency outcome measures, long spans of time covered in those measures, and weak reliability and validity were associated with smaller effect sizes.
ES Info	.01	Less explicit reporting of statistical results was associated with larger effect sizes as was more explicit reporting of general methodological procedures.
Treatment		
Subjects	.01	Juveniles with more indication of delinquency (higher "risk") were associated with larger effect sizes.
Dosage	.03	Longer duration treatment and that judged to provide larger amounts of meaningful contact were associated with larger effect sizes.
Treatment	.11	(1) Treatment provided by the researcher or situations where the researcher was influential in the treatment setting were associated with larger effect sizes.

Table 4.6 *(Continued)*

Cluster	R^2 Change	
		(2) Treatment in public facilities, custodial institutions, and the juvenile justice system were associated with smaller effect sizes.
		(3) Behavioral, skill-oriented, and multimodal treatment was associated with larger effect sizes than other treatment approaches.
Tx Philos	.02	Treatment judged to have a more sociological and less psychological orientation was associated with larger effect sizes.

ment, the greater were the delinquency differences subsequent to treatment. More surprising was the finding that the nature of subject assignment to groups (random versus nonrandom), often viewed as synonymous with design quality, had little relationship to effect size. What mattered far more was the presence or absence of specific areas of nonequivalence—for example, sex differences—whether they occurred in a randomized design or not.

Loss of equivalence between treatment and control groups can also occur after a study begins via attrition. While the Attrition cluster played a smaller role in effect size than initial nonequivalence, it was appreciable nonetheless. Curiously, attrition from both the treatment and control groups appears to suppress effect sizes. This is the result that would occur if more amenable juveniles tended to drop out of treatment groups and/or more delinquent juveniles tended to drop out of control groups.

Other important design issues were sample size (Samples) and the type of control group selected (Control). Studies with larger samples tended to have smaller effect sizes. On first blush, this may appear to be a reflection of the upward bias known to occur in estimation of effect sizes from small samples (Hedges 1981; Hedges and Olkin 1985). Statistical adjustments were applied to the effect size values in order to control that bias, however. More likely, there is a general size of study effect here—small studies may be done more carefully, have more consistently delivered treatments, and the like. It is notable in this regard that studies having more outcome variables and more experimental groups also showed smaller effect sizes.

Whatever the study size, control groups that received some attention—for example, "treatment as usual" in a juvenile justice setting—showed less contrast with treatment groups (smaller effect size) than those control groups that received no treatment at all. Since the treatments studied in juvenile justice contexts are often augmentations to services that can already be extensive (e.g., custodial care), this is not surprising. The one exception, "treatment as usual" for probation services, is consistent with this pattern since probation contact is usually quite minimal.

The remaining method variable cluster of consequence was that dealing with the nature of the delinquency outcome measurement (Measures). Although collectively these variables were correlated with effect size, no readily interpretable pattern was evident. Other than number of delinquency measures, which was probably part of the study-size effect discussed above, the strongest relationship was a tendency for delinquency measures covering a longer period of time post-treatment to be associated with smaller effect sizes.

TREATMENT. Of primary interest in Table 4.6 are those clusters that show an important influence of the type and circumstances of treatment upon delinquency. Since all the method clusters were stepped into the regression analysis prior to any of these clusters, we can have some confidence that any relationships that emerge represent characteristics of effective treatment rather than confoundings with influential method variables.

The cluster of variables representing subject characteristics (Subjects) was stepped into the analysis first among the treatment clusters to test the possibility that certain juveniles were especially responsive to treatment, whatever its nature. While there was a slight tendency for studies of juveniles with higher risk levels—that is, greater involvement with delinquency—to show larger effect sizes, the overall influence of this cluster was small and statistically nonsignificant. The prospect of such a relationship, however, deserves further scrutiny in later analysis. Targeting high-risk juveniles was one of the criteria for "clinically relevant" treatment in the Andrews et al. (1990) meta-analysis cited in the introduction to this chapter.

In similar spirit, the cluster of variables dealing with the amount or intensity of treatment (Dosage) was entered into the analysis next. This permitted consideration of the possibility that the size of the treatment dose was more important than the specific nature of the treatment administered. As the National Academy of Science's review of correc-

tional treatment observed, weak and incompletely delivered treatments cannot be expected to have meaningful effects (Sechrest, White, and Brown 1979).

The regression analysis did show a modest positive relationship between effect size and the duration, frequency, and amount of treatment. The relationship seems to be weakened, however, by an unexpected confounding. Some of the treatment dosage variables are such that they are naturally higher for juveniles in institutional care—for example, frequency of contact. As a category, treatment in institutional context seems to be associated with smaller effect sizes. This results in a somewhat curvilinear relationship in which effect size increases with amount of treatment up to amounts associated with institutional care (i.e., "continuous" frequency of contact) and then declines. Subsequent analysis of this relationship will require more refined breakdowns among treatment categories than those used in the present analysis.

By far the strongest relationship with effect size was found for the cluster of variables representing treatment modality and the nature of the treatment provider (Treatment). These relationships showed three different facets. First, treatments that were delivered by the researcher, or in which the researcher had a considerable influence, showed larger effect sizes. A cynical interpretation of this pattern might suggest that these larger effects stemmed from some interest on the part of the researcher in making the treatment look good. It is at least equally plausible, however, that treatment delivered or administered by the researcher for research purposes was better implemented and monitored than the typical practices of service agencies. If such is the case, the "researcher involvement" variable becomes a more general indicator for treatments mounted with enthusiasm and careful administrative control—a circumstance that may well lead to larger effects.

The second facet of the variables in the treatment cluster was an association between smaller effect sizes and treatments provided in public facilities, or within the juvenile justice system or custodial institutions. Since these findings overlap considerably with the pattern of findings for specific treatment modality, we turn to them now.

The most influential variables in the Treatment cluster were those that dummy-coded various specific treatment types separately for juvenile justice and non–juvenile justice sponsors. A rather consistent pattern emerged which is most easily seen by looking at the mean effect size for each category of treatment. Since we want to examine treatment effects unconfounded by method effects, the mean effect sizes

for each treatment category were computed from the multiple regression residuals after all method clusters were removed (adding back the grand mean, of course). To make these mean effect sizes more interpretable, each was also translated into the equivalent reduction it represented in a dichotomous recidivism rate when compared with a hypothetical control group with 50 percent recidivism. Table 4.7 reports the results.

Treatment modality is often described rather crudely in the source studies upon which this analysis relies, often by no more than a label or phrase. It is correspondingly difficult to code into a meta-analysis in any definitive way. It would be a mistake, therefore, to focus on any particular category in Table 4.7 and draw a general conclusion about the efficacy of treatments offered under such various conventional labels as "restitution" or "counseling." This would also contribute to the unfortunate tendency in delinquency treatment to advocate a "magic bullet," a specific treatment concept alleged to be a superior approach to delinquency. Moreover, the categories in Table 4.7 include instances of varying efficacy ranging above and below the category mean and they overlap considerably for those many treatments with multiple elements.

A more appropriate approach to interpreting Table 4.7 is to examine the broader patterns in the ranking of more and less effective treatment modalities. Viewed this way, there is striking consistency in both the juvenile justice and non–juvenile justice treatments. In both cases, the more structured and focused treatments (e.g., behavioral, skill-oriented) and multimodal treatments seem to be more effective than the less structured and focused approaches (e.g., counseling). It will be the task of subsequent analysis of these data to better tease apart the various treatment parameters that account for this ranking.

It is noteworthy that the best of the treatment types, both inside and outside the juvenile justice system, show delinquency effects of meaningful practical magnitude, in the range of 10–20 percentage points reduction in recidivism. Since these are reductions from a presumed 50 percent control group baseline, they represent decreases of 20–40 percent (i.e., 10/50 to 20/50). It is also interesting that the treatment types that show this large order of effects are, with few exceptions, those defined as most "clinically relevant" in the Andrews et al. review (1990).

Finally, it should be noted that a number of treatment approaches were associated with mean effect sizes of virtually zero. This family of treatments simply may not work, as many critics have charged. Further, a couple of treatment categories appear to produce negative ef-

Table 4.7 Residualized Effect Size Estimates After Removal of Method Variance for Different Treatment Modalities

Treatment Modality	Effect Size	Equivalent Recidivism Change from 50% Control
Juvenile Justice		
Employment (4)	.37	−.18
Multimodal (12)	.25	−.12
Behavioral (8)	.25	−.12
Institutional, other (9)	.20	−.10
Skill-oriented (15)	.20	−.10
Community residential (12)	.16	−.08
Any other juvenile justice (5)	.14	−.07
Probation/parole, release (16)	.11	−.05
Probation/parole, reduce caseload (11)	.08	−.04
Probation/parole, restitution (13)	.08	−.04
Individual counseling (20)	.08	−.04
Group counseling (39)	.07	−.03
Probation/parole, other enhancement (7)	.07	−.03
Family counseling (6)	.02	−.01
Vocational (9)	−.18	+.09
Deterrence (9)	−.24	+.12
Non–Juvenile Justice		
Skill-oriented (17)	.32	−.16
Multimodal/broker (29)	.21	−.10
Behavioral (31)	.20	−.10
Group counseling (17)	.18	−.09
Casework (7)	.16	−.08
Family counseling (29)	.10	−.05
Advocacy (4)	.10	−.05
Other counseling (5)	.06	−.03
School class/tutor (14)	.00	−.00
Individual counseling (24)	−.01	+.00
Any other non–juvenile justice (3)	−.01	+.00
Employment/vocational (22)	−.02	+.01

Note: The number of studies in each category is reported in parentheses.

fects—most notably, deterrence treatments. This category includes shock incarceration and the "scared straight" program model that received considerable publicity a few years ago.

Whatever patterns one discerns in these results, they do indicate

that the specifics of what is done in delinquency treatment are important. No generalized placebo or Hawthorne effect is likely to be able to account for the differential outcomes of different approaches.

The final cluster of treatment-related variables to be entered in the regression represented those that indicated something about the treatment philosophy: its etiological orientation, level of theory development, and related matters. This cluster was only weakly related to effect size. It appears that there is little in the reported treatment philosophy, above and beyond the characteristics of its subjects, dosage, and treatment type, that influences the size of effects.

Conclusions

What is presented here is only the most general analysis of the measured effects from delinquency treatments studies. While it was demonstrated that the grand mean of those effects is positive, indicating at least modest overall treatment effects, the primary focus of this phase of the investigation has been upon the variability of effects. This variability was shown to be far in excess of what would be expected simply on the basis of sampling error. It follows that there must be some circumstances in which studies yield large effects and others in which they yield small effects. The challenge is to discover the nature of those circumstances.

If research results are shaped primarily by the methods chosen, we should know which aspects of the methods are most important and investigate the bias they introduce. If, on the other hand, some substantial portion of the variability in measured delinquency effects stems from the nature of the treatments applied and the characteristics of the juvenile recipients of those treatments, then it behooves us to discover which treatment circumstances produce the largest effects and put that information to practical use.

The analyses presented in this chapter are less concerned with the details of these issues than with charting the overall domain. The results indicate that both method and treatment influence the effects of delinquency treatment studies. Although method variables collectively seem to play a somewhat greater role, the largest single category of influences is the nature of the treatment itself. Subsequent work using this database will focus on closer specification of the details of method and treatment, and their interaction, that are most important in shaping study outcome.

The pattern of the general results presented here throws some light

on the checkered history of research reviews in delinquency treatment. The grand mean effect size is perilously close to zero. While not so close as to justify the "nothing works" rhetoric of the 1970s, convincing positive effects would be difficult to discern in any sample from this literature. This would be especially true if the sample was of modest size and if the review primarily used "box score" techniques that keyed on the statistical significance of individual study findings. The sample sizes used in this literature (median around 60 in each experimental group) do not yield sufficient statistical power for an individual study to find statistical significance for effect sizes in the range of .10–.20 standard deviation units.

Moreover, the wide variability in effects found in this literature means that different reviews that sampled different portions of it could come, quite honestly, to rather different conclusions. On the high end of the distribution are studies that show impressively large effects, as Gendreau and Ross (1979), Palmer (1975), Andrews et al. (1990), and others have asserted. On the low end of the distribution, and even in the middle, a considerable number of studies can be produced that show insignificant and even apparently negative effects, as Martinson (1974), Whitehead and Lab (1989), and others have insisted. If the heterogeneity of the distribution of effects in delinquency treatment research is as large as an elephant, perhaps it is no wonder that each reviewer, grasping here a tail of the distribution and there a hump, describes the beast so differently.

Appendix 4.A Bibliographic Databases Used in Search

Arts and Humanities Citation Index
Books in Print
British Books in Print
British Education Index
Child Abuse and Neglect
Criminal Justice Periodical Index
CRISP: National Institute of Mental Health
Dissertation Abstracts Online
ERIC (Educational Resources Information Center)
Family Resources
Federal Research in Progress
Library of Congress Books
Medline
Mental Health Abstracts
National Criminal Justice Reference Service
National Technical Information Service
PAIS International (Public Affairs Information Service)
Psychological Abstracts
Social Science Citation Index
Sociological Abstracts
SSIE Current Research (Smithsonian Science Information Exchange)
U.S. Government Printing Office Publications
U.S. Political Science Documents

Notes: The research reported in this paper was funded by the National Institute of Mental Health, Antisocial and Violent Behavior Branch (MH39958 and MH42694), and the Russell Sage Foundation.

There were 443 studies involved in the analysis presented in this paper. The full bibliography of studies can be obtained from the author at the Psychology Department, Claremont Graduate School, Claremont, CA 91711.