'WHAT WORKS?' REVISITED AGAIN:

A META-ANALYSIS OF RANDOMIZED FIELD EXPERIMENTS IN INDIVIDUAL-LEVEL INTERVENTIONS

By Anthony J. Petrosino

A DISSERTATION SUBMITTED TO

THE GRADUATE SCHOOL-NEWARK, RUTGERS UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Written Under the Direction of
Dr. James O. Finckenauer
of the School of Criminal Justice and approved by:

Dr. James O. Finckenauer, Chairperson . . . .

Dr. David Weisburd, Member. . . . .

Dean Ronald V. Clarke, Member. . . . .

Dean Todd Clear, Outside Member ... .
(Florida State University)

Dr. Paul Lerman, Outside Member. . . . .
Rutgers, School of Social Work

Newark, New Jersey
October, 1997

ABSTRACT OF THE THESIS

'What Works?' Revisited Again: A Meta-Analysis of Randomized Field Experiments in Individual-Level Interventions

By Anthony J. Petrosino

Dissertation Chair: James O. Finckenauer

Meta-analysis, the statistical analysis of prior research reports, was used to assess the overall impact of individual-level interventions on subsequent offending. To be included, each study had to employ a randomized experimental design, be written in English, be available during 1950-1993, and include a quantifiable outcome measure of crime.

Interventions ranged from treatment programs (e.g., counseling) to deterrence strategies (e.g., arrest) to delinquency prevention programs (e.g., casework with 'at-risk' kids). Multiple search strategies retrieved over 300 eligible experiments; the alternation technique was used to select 150 for the analysis.

Each report was coded using a 196 item instrument. Intercoder reliability was assessed on a random subset of studies, with an overall rate of agreement of 80%. For those interventions which could be classified, most were rehabilitative in focus (N=115); deterrence-oriented (N=23) and delinquency prevention programs (N=9) were less frequent.

Cohen's effect size (*d),* which standardizes the experimental effect across studies, served as the common metric. Percentages, frequencies, test values and means were converted to *(d)* with the use of specialized meta-analytic software. A correction for small sample bias and a sample-size weighting technique developed by Hedges and Olkin (1985) were both applied to the data.

The findings indicated that the global mean effect size was highly unstable. The equal-weighted effect size (*d*) for rehabilitation programs was -.20, a finding congruent with prior 'what works?' meta-analyses. This effect was much larger than that for deterrence and prevention. When introducing sample-size weighting, the global mean for rehabilitation programs dropped to -.03, which was smaller than the deterrence effect (-.07). Neither type of outcome measure or internal validity rigor of the experiment altered these findings.

A homogeneity test, however, indicated considerable variability in effect size across rehabilitation experiments. Two independent variables were then introduced as moderating variables: sample size and type of subjects. Small sample rehabilitation studies (10-100 subjects total) reported much larger effects than bigger experiments (300+ subjects). Experiments with juvenile subjects (those under 18) were more effective than programs for adults. Discussion of substantive findings, recommendations for strengthening meta-analysis, and a future research agenda were provided.

Table of Contents

List of Tables and Figures

TABLES

FIGURES

Acknowledgements

It is difficult to complete a project as broad as this one without a great deal of encouragement, cajoling and assistance from others. Even if their names are omitted here, their deeds and words shall remain etched on my heart. When this work was perilously close to being shelved because of work and family demands, friends and relatives inspired me to finish.

Of course, I have been blessed with a truly extraordinary committee. All members helped me to think about larger issues, without forsaking the smaller inconsistencies, which abound in meta-analytic projects. Each delivered clear and cogent comments on short notice, a task not unnoticed in a day when everyone is overloaded with demands.

I owe much to David Weisburd, now with Hebrew University and the Police Foundation. He took me under his wing early in my graduate studies, and I learned a great deal working on his research projects. His earlier work on experimentation in criminal justice inspired this larger meta-analysis. He has been my professor, employer, advisor, mentor and friend. His comments and criticisms were crucial to this study.

Todd Clear, now with Florida State University, provided leadership early on, helping me to get this project off the ground when my first dissertation project was aborted. Like David, he has been both a mentor and friend for 10 years. Todd forced me to think not only about the mechanics of doing the research, but the underlying philosophical issues as well.

I am also greatly indebted to my Chairperson, Jim Finckenauer, who graciously agreed to take over when David and Todd left Rutgers. Jim guided a much stronger and policy-relevant product than if I had been left to my own devices. Few people provide better advice than he does, from improving the draft to getting through the dissertation bureaucracy.

Dean Ronald Clarke and Dr. Paul Lerman both provided sharp insight from their prior research on criminal justice interventions, and shaped a much better product. I take all the blame, however, for the mistakes contained in this draft.

The National Institute of Justice supported part of this research through their dissertation fellowship grant, and I am particularly grateful to a very patient Program Manager, Dr. Richard Rau. Julie Schnobrich, Michael Gordon and John Lavender worked as research assistants at various stages of this project, and went beyond their paid hours to help out. Phyllis Schultze of the Rutgers University NCCD/Criminal Justice Collection constantly shipped me experimental evaluation reports and pressured me modestly for their return. Dean Alan Futernick, Ms. Shirley Parker and

Ms. Gene Webster were always helpful in assisting with the administrative details of both the graduate program and the NIJ grant.

This work owes much to all the original investigators and reviewers before me, who supplied me with studies, reports, documents, citations, and bibliographies not available through normal channels. It was very refreshing to find an invisible college of researchers who went through files, data and computer archives to try and find information requested for this meta-analysis. Research synthesis can never take place without original studies, and I am grateful to those who persevered and conducted field experiments in difficult settings.

Dr. Blair Johnson of Syracuse University, the creator of *DSTAT* meta-analytic software, spent considerable time answering my questions and helping me work out the glitches in converting my data to *DSTAT* format. In particular, Blair assisted me in the more complicated statistical analyses and interpretations, always explaining in simple quantitative terms what the output meant. He was, in essence, my informal committee member.

Dr. Christine M. Boyle (NJ) and Dr. Rhiana Kohl (MA), supervisors of state Statistical Analysis Centers where I have been employed, provided flexibility in work hours when possible to allow opportunity to complete this work. Dr. Michael Buerger, Research Director for the Jersey City Police Department, could be counted on to continuously harass me with e-mails and phone calls to "just get it done."

From a personal standpoint, I am gratefully indebted to two loving and caring parents, Anthony and Edna Petrosino, and four terrific brothers (Bob, Mike, Joe, Jim), who always let me know how proud they were--long before the degrees. Many thanks to a man who never stopped praying for me, Dr. Guy Finch, my pastor. He is perhaps the finest example of true Christianity I know.

I could not ask for a better partner to go through life with than my wife, Dr. Carolyn Petrosino. She not only assisted with the project, but was my support line throughout the project. She remains my best friend, my soulmate, and my confidante; there's no way I could exist without her. There were many nights and weekends when she had to handle everything alone, while I worked. We have a five year old son, Elliot, and no words ever written would do justice for how much I love this little boy. I am most grateful for his warmth and compassion. My stepdaughter, Erica, is already an exceptional high school student and on her way to a great career as a "real doctor."

Most of all, I must give praise to the One who directs my paths, the Lord Jesus Christ. It would be a great honor to be used by Him to accomplish something good in this world, no matter where that takes me. I am reminded of Proverbs, which admonishes that the '*fear of the Lord is the beginning of wisdom.*' Help me, God, to be wise and not a fool in this world.

CHAPTER I. LAYING THE FOUNDATION FOR THE CURRENT STUDY: THE NEED FOR BETTER EVIDENCE AND REVIEWING METHODS

The question of what works in reducing crime is a central one for criminologists and practitioners. While political and ideological arguments are influential (e.g., Logan and Gaes, 1993; von Hirsch and Maher, 1992; Rotman, 1990; Cullen and Gilbert, 1982), knowledge also plays an important role. Specifically, the treatment effectiveness issue requires scholars and policymakers to attend both to the evidence used in the debate, and the method by which that evidence is accumulated and reviewed.

There are literally thousands of treatment program evaluations contained in reports, articles, books, and other types of documents. Despite this overwhelming amount of evidence, most reviewers attempting to summarize what works have offered one consistent finding: the research evidence considered is generally so poor that sound conclusions can not be reached (e.g., Sechrest, White & Brown, 1979).

Indeed, even the earlier pessimistic offender rehabilitation reviews contained these design caveats (e.g., Wright and Dixon, 1977; Martinson, 1974; Logan, 1972; Bailey, 1966). Additionally, the more positive reviews of the past decade still call attention to the methodological shortcomings of the evaluation reports sampled (e.g., Basta and Davidson, 1988). It is disconcerting that the recent University of Maryland crime prevention report to the United States Congress—which the New York Times called the most comprehensive report on criminal justice ever written—echoed the conclusions of reviewers made three decades earlier (Sherman, et al., 1997). In fact, Sherman, et al., write that (1997:10-1):

> The effectiveness of most crime prevention strategies will remain unknown until the nation invests more in evaluating them. That is the central conclusion of this report...Using "rigorous and scientifically recognized standards and methodologies"...the review of over 500 impact evaluations reveals only a handful of conclusions that can be generalized from those studies to similar programs around the nation..."

In addition to the problem of the evidence, the traditional methods used to review and synthesize information are problematic. The information explosion in criminal justice has produced a dramatic increase in the amount of research which must be accumulated and reviewed. For example, there are now scores of journals which publish criminology articles, compared to the few which existed two decades ago. An international data base of justice documents has been created (NCJRS) and is easily accessed. Many other sources exist to locate and track down evaluation reports. Finally, the Internet provides rapid access to research documents, without the usual lag time associated with publishing in mainstream outlets. The reverse side of the information technology advance is that it has produced too much information for the treatment effectiveness reviewer–using traditional, narrative methods–to comprehend (e.g., Cooper, 1989).

Two methodologies have surfaced, however, during the past ten years which may provide solutions to these concerns. First, the re-emergence of the *randomized field experiment* within criminal justice settings indicates that the plea to conduct better controlled studies is being heeded. Moving beyond the prison walls and the therapeutic settings where experiments often took place during the rehabilitation model era (Clarke and Cornish, 1972),[1] randomized field tests have now been reported with suspects facing police arrest, with defendants at sentencing, with offenders on probation or parole and with high risk individuals in the community. These experiments potentially provide valuable and more stable information on the efficacy of criminal justice and social interventions to reduce crime (e.g., Farrington, Ohlin & Wilson, 1986; Reicken and Boruch, 1978). The recently released University of Maryland Report to the Congress is also replete with recommendations for controlled experimentation (Sherman, et al., 1997).

Second, while the controlled experiment is an older design being used with renewed vigor in the "what works" debate, *meta-analysis* emerged within criminal justice in 1984 to provide researchers with a systematic and replicable method for quantitatively reviewing and synthesizing the information we learn from individual empirical studies (e.g., Lipsey, 1988). Meta-analysis allows researchers to cull together the evidence provided by hundreds of evaluations which examined the effects of some intervention on criminal behavior. The continuing debate over the accuracy of Martinson's (1974) landmark qualitative review of 231 studies–reported ten years

[1] For example, Phillip Cook (1980:250) writes that "A controlled experimental design with random assignment is generally viewed as the most reliable source of information about the effects of social innovations...The use of this technique in criminal justice research has largely been limited to correctional programming studies, focused on rehabilitation effects..."

before the first meta-analytic contribution in criminal justice–is sufficient evidence alone that better research integration and literature review methods are needed (e.g., Izzo and Ross, 1990).

The Importance of Knowing "What Works?"

Despite the reemergence of retributive sentencing, utilitarian goals of deterrence, rehabilitation and incapacitation continue to be emphasized.[2] In fact, though the move toward sentencing philosophies which emphasize *just desert* is well documented (e.g., Rhine, 1992), Burton and his colleagues found (1993) that 41 state penal codes legally mandate rehabilitation as a primary goal for corrections. Moreover, the resulting increase in prison populations and financial costs–without a parallel reduction in crime rates or public fear–has resulted in criticism of the *Justice Model* (e.g., Clear, 1994) and renewed interest in rehabilitation techniques (e.g., Rotman, 1990).[3]

[2] For example, the New Jersey Code of Criminal Justice (1990) adopted a desert based sentencing code in 1979, but retains all four primary goals of punishment: retribution, rehabilitation, deterrence and incapacitation. Massachusetts has recently enacted a 'Truth in Sentencing' law which emphasizes just desert but retains the goal of rehabilitation through education and vocational training (Massachusetts General Court, 1993:432).

[3] It should be noted that there is nothing inherent in retributive philosophy–or desert–that mandates long prison terms. In fact, its chief advocates call for a sparing use of incarceration (e.g., von Hirsch, 1976; Friends of the American Service Committee, 1971) and point to desert-oriented states which used prison sentences more rationally (e.g., von Hirsh and Maher, 1991).

While the philosophy of retribution requires no crime control evidence (e.g., Feinberg and Gross, 1983), simply returning offenders to the streets with the caveat that they were punished solely according to desert–and nothing more–may engender little support. American citizens certainly want punishment to meet retributive aims, but they also want punishment to accomplish something more, preferably rehabilitation (e.g., Cullen and Gendreau, 1989; Cullen and Gilbert, 1982).

Some have argued that focusing on reducing offender recidivism is misplaced since most crime is committed by first time offenders (e.g., Walker, 1994; van den Haag, 1983). This argument, however, must be prefaced by several remarks: adults who commit felony offenses are very likely to have had contact with the juvenile justice system; many serious adult offenders have had one or more contacts with the criminal justice system as adults; and many who come in contact with the criminal justice system for the first time were referred at an earlier point to social or public welfare agencies for services. All of these observations accentuate the importance of determining which social policy interventions are effective.

Even if sentencing was based on a strict just desert model, with no goal of reducing recidivism, crime reduction would still be pursued at other points. Both policymakers and scholars would still want to know if certain police actions are more effective than others in deterring crime, if prevention programs make a difference in the lives of high risk children, and if there are is anything which can smooth the reintegration of the punished offender back into the community. Moreover, many treatment providers across different settings–including prisons–want feedback on the long-term effects of strategies and techniques they utilize professionally. In addition, it is likely that all but strict desert theorists would be interested in voluntary rehabilitation outcomes, provided treatment considerations did not jeopardize proportionality and equity at sentencing (e.g., von Hirsch and Maher, 1992).

Additionally, although the juvenile justice system was heavily influenced by the Justice Model (e.g., Fogel, 1975), it never completely abandoned its original goal of rehabilitating criminal youth (e.g., MacCalliar, 1993). As with the adult system, some positive treatment outcome studies, and disfavor with the results of 'getting tough,' has inspired some to reaffirm juvenile rehabilitation (e.g., Macallair, 1993). The emotional charge over the proper role of the juvenile justice system has created a climate where information about treatment effectiveness has never been more important. The ominous warning about a coming tidal wave of juvenile violence—proffered by those in academia and the general media (e.g., Thomas, 1995)–due to an anticipated population increase in high-risk age groups underscores the urgency of determining which programs, policies and interventions are effective.

In addition, times of great fiscal restraint result in smaller budgets, less staff and decreased resources for all agencies. The monetary constraints and public loss of confidence in social programs puts enormous pressure on policymakers to make informed decisions. Knowledge concerning effective interventions can help those charged with making difficult policy choices.

These observations focus attention on the importance of knowing what works in the area of *specific or special crime reduction*.[4] This domain of studies is comprised of programs and policies delivered to individuals already identified with the intent of reducing their subsequent criminal behavior. For the purpose of this study, we categorize specific crime reduction programs into three broad areas;

> *(1) specific or special deterrence*: the threat of some or additional punishment will decrease subsequent offending (e.g., arrest versus police mediation, intensive surveillance-oriented supervision versus traditional parole or probation supervision).
>
> (2) *offender rehabilitation*: treatment programs designed to reduce the individual's risk for reoffending, sometimes provided within the context of traditional sanctions (e.g., group counseling for prisoners or probationers).
>
> (3) *delinquency prevention*: the provision of special services to particular youths before official contact with the criminal justice system, to reduce their likelihood of criminal behavior onset (e.g., providing vocational counseling to minority youths in an impoverished, high crime area).

While these groups are not mutually exclusive (i.e., some programs may achieve deterrent or rehabilitative effects), they provide a 'good-enough' set of categories to compare broad intervention types. It is important to note that these groups are comprised solely on individual-level programs and policies, and as such, do not include the wide range of place or area-level interventions implemented in the interests of situational crime prevention, general deterrence or other public safety goals (e.g., Sherman and Weisburd, 1995).

[4] This work was originally titled "Experiments in Special Deterrence," but that term is used frequently in the literature to denote punitive sanctions (e.g., Sherman and Berk, 1984). The more inclusive phrase, *specific crime reduction*, was opted for here.

Knowing 'What Works': Problems With the Evidence

Unfortunately, policymakers who seek information about crime reduction programs will be frustrated by the ambiguous and often conflicting evidence in the criminological literature. The MacArthur Foundation's Justice Study Group, following years of review, stated that:

> Policymakers who wish to put in place new programs to reduce crime, or to expand the scope or effectiveness of programs already in place, will quickly discover that the knowledge necessary to do this responsibly does not exist except in fragmentary and unsatisfactory form.
>
> Whether we wish to prevent delinquency or rehabilitate offenders, whether we seek to strengthen families or improve schools, whether we think that juvenile courts should get tougher or provide better services, we will be forced to admit, if we are honest, that we only have scattered clues and glimmers of hope (and sometimes not even that) on which to base our actions (Farrington, Ohlin & Wilson, 1986:17).

While the aforementioned University of Maryland report (Sherman, et al. 1997) was less pessimistic about existing research, the lack of rigorous evaluation research led the team to urgently recommend a drastic infusion of funding into researching federal justice program outcomes. Both the University of Maryland Report and the MacArthur Foundation Justice Study Group echoed critical conclusions reached by earlier broad surveys of the treatment effectiveness literature (e.g., Sechrest, White & Brown, 1979; Wright and Dixon, 1977; Lipton, Martinson & Wilks, 1975; Logan, 1972; Bailey, 1966). In fact, reviewers broadly examining social program evaluations in fields as diverse as education, criminal justice, mental health and organizational development found a literature base which they rated as neither reliable or valid (e.g., Prather and Gibson, 1977).

Logan (1972) found, for example, that none of the 100 offender treatment evaluations he reviewed met ten minimum methodological requirements for an adequate scientific test. Bailey's (1966) earlier synthesis of 100 correctional outcome reports found only 22 used a control group of any kind, and he concluded by characterizing the research evidence supporting treatment efficacy as "slight, inconsistent and of questionable reliability." Wright and Dixon's (1977:57) found the 96 juvenile delinquency prevention and treatment programs they reviewed to be of such low scientific validity that "few of them should have seen the light of day."

Martinson's (1974:25) summary conclusion that "with few and isolated exceptions, the rehabilitative efforts reported so far have had no appreciable effect on recidivism" became one the most widely cited statements in criminology. Few remember, however, his critical appraisal regarding the quality of the research evidence. He said (1974:49) that "it is just possible that our treatment programs are working to some extent, but that our research is so bad it is incapable of telling."

The lack of sound evidence was also noted by reviewers focusing on a single type of intervention or program. Farrington (1979) found that behavior modification program evaluations were plagued by internal validity problems, stemming from poorly controlled research designs. Sarri and Vinter (1965) reviewed 110 papers on group counseling techniques with juvenile offenders and concluded that rigorous evaluation was needed; research on effectiveness consisted of descriptive, anecdotal and generally unreliable reports.

While noting some improvement in the quality of treatment research, recent reviewers still find evaluations largely inadequate. Basta and Davidson (1988) reviewed 37 juvenile treatment programs reported in the literature from 1980 to 1987 and found several consistent methodological problems, including the failure to use appropriate control or comparison groups in the analysis. Even more recently, the United States Government Accounting Office (1996:3) reviewed the literature on sex offender treatment, finding that 22 prior literature reviews on the topic "identified methodological problems with sex offender research as a key impediment to determining the effectiveness of treatment programs."

Even reviews of treatment effectiveness with particular offender subtypes have noted the weak quality of the evaluations. This is true across a wide variety of literature reviews, including those focusing on: *adolescent drug abusers* (e.g., Catalano, Hawkins & Wells, 1991); *spouse abusers* (e.g., Gondolf, 1997); *adult sex offenders* (e.g., Government Accounting Office, 1996; Furby, Weinrott & Blackshaw, 1989); and *juvenile sex offenders* (e.g., Camp and Thyer, 1993). These and many other syntheses usually conclude with compelling arguments for better designed studies.[5]

[5] It is interesting to note that even researchers conducting a single program evaluation cautiously present findings and add that a controlled study is needed to further test conclusions (e.g., Roundtree, Grenier & Hoffman, 1993; MacKenzie, 1991; Larson, 1990).

Better Evidence: Randomized Field Experiments

One potential solution to the problem of inadequate evidence is the randomized field experiment. This is certainly not a new idea, given that arguments for the use of randomized experiments have been made for decades to test various interventions, including: *social programs* (e.g., Berk, Boruch, Chambers, Rossi & Witte, 1985; Saxe and Fine, 1981; Reicken and Boruch, 1978; Campbell, 1969); *general deterrence policies* (e.g., Zimring and Hawkins, 1973; Andenaeus, 1966); *offender treatment programs* (e.g., Quinsey, 1983; Empey, 1980; Glaser, 1971; 1965); *legal innovations* (e.g., Federal Judicial Center, 1981; Zeisel, 1968); and *diversion programs* (e.g., Roesch, 1978). In fact, the influential Justice Study Group strongly urged the increased use of randomized experimentation nearly ten years ago, to evaluate criminal justice programs and policies conceived to reduce crime (e.g., Farrington, Ohlin and Wilson, 1986).[6]

Even scholars less than enthusiastic about experimentation in social settings acknowledge the scientific wisdom of the randomized design (e.g., Mitroff, 1983; Twain, 1983; Harre and Secord, 1972). It is clear that if the experiment is carried out with full integrity, changes in the outcome variable can be causally linked to changes in the independent variable. No other design permits such a strong connection between variables of interest (e.g., Weisburd, 1993; Farrington, 1983).

[6] The Justice Study Group also recommended the increased use of longitudinal designs to study crime (e.g., Farrington, Ohlin & Wilson, 1986).

It is the ability of randomization to remove selection bias and produce equivalent groups–prior to introduction of the independent variable–that distinguishes the experiment from even the strongest quasi-experimental studies (e.g., Jaeger, 1990; Kerlinger, 1964). Assigning subjects at random to two or more groups should balance extraneous variables, such as individual traits, that often cloud the interpretation of criminal justice outcome evaluations (e.g., Farrington, 1983).

This is because randomized experiments rely on the strength of statistical probability. As long as each unit in the experiment has the same chance probability as the next unit to be assigned to each condition, then groups–within chance fluctuation–should be comparable on *all* individual characteristics (e.g., Cochran and Cox, 1992; Brown and Melamed, 1990; Farrington, 1983).[7] Investigators can also combine matching, stratification or blocking techniques with random assignment to reduce chance probability of differences between experimental conditions (e.g., Sherman and Weisburd, 1995; Gelber and Zelen, 1985).

This strength separates the randomized study from designs where subjects were matched on particular individual characteristics (e.g., age, race, prior record, etc.). Investigators using matched designs are generally only able to insure equivalence on a

[7] The laws of probability also mean that the possibility of chance differences between experimental study groups declines with larger sample sizes (e.g., Jaeger, 1990; Kerlinger, 1964). No optimal number for randomized group samples is suggested in the literature, although Farrington (1983) uses 50 subjects per condition as part of his inclusion criteria for reviewing experiments (e.g., Weisburd, 1993).

few selective variables (e.g., Farrington, 1983; Kerlinger, 1964).[8] Therefore, selection bias remains a potential explanatory factor for results (e.g., Campbell and Stanley, 1966).

Selection bias is perhaps the most frequent threat present in crime reduction evaluations.[9] In most of the positive quasi-experimental studies reported, principal investigators generally conclude that the results are equivocal; positive outcomes may have been due to pre-existing differences rather than any distinct intervention effect. Even quasi-experimental studies which find no difference in recidivism rates can not rule out selection bias (e.g., MacKenzie, 1991).[10]

Therefore, randomized experiments provide great improvement over designs where outcome changes due to selection bias and other internal validity threats are difficult to rule out (e.g., Farrington, 1983). In fact, only the randomized experiment

[8] Taylor (1994:285) notes that matched designs assume that researchers know the important variables on which subjects should be matched and that data on those variables is available before the project starts.

[9] This is particularly true, in the author's experience, within state government justice research. Most—if not all—evaluations are quasi-experimental and involve the selection of a post-hoc comparison group. In the final analysis, differences may be found, but the treatment effect is often hopelessly confounded with selection bias.

[10] MacKenzie's (1991) evaluation of shock incarceration ("boot camp") raises an interesting point. It is easy to see where selection bias can lead to positive results in program evaluation, as in the case of the practitioners who select best risk cases for the treatment group. Of course, no difference or negative findings may also be due to selection bias (e.g., Cook and Campbell, 1979; Campbell and Stanley, 1966).

potentially can counter every internal validity threat (Babbie, 1983).[11] While most major internal validity threats are canceled out through the design (e.g., selection bias), other threats can be eliminated through effective implementation of the experiment (e.g., effective monitoring can insure that experimental and control groups receive differential treatment). In fact, Reicken and Boruch (1974) refer to several instances where an experimental program evaluation reported dramatically different results than earlier quasi-experimental studies testing the same intervention. More detailed methodological and statistical presentations on the advantages of randomized experiments are discussed elsewhere (e.g., Brown and Melamed, 1990; Cook and Campbell, 1979; Campbell and Stanley, 1966).

One frequent concern raised regarding the classic experimental design is its lack of external validity or generalizability to other settings (e.g., Jupp, 1989). This criticism is largely irrelevant to the criminal justice experiments considered here.

First, the external validity criticism rose in response to the frequent use of college students in laboratory experiments, a phenomenon so common that one famous psychologist remarked that "ours is largely the science of sophomores" (e.g., Rosenthal, 1991). Randomized experiments in crime reduction occur in field settings, with actual offenders or other individuals and involve tangible interventions.

[11] Babbie states that "only the classical experiment...if coupled with proper selection and assignment, handles each of the twelve problems of internal invalidation (1992:248)."

Second, external validity becomes a concern if the subjects chosen to *participate* in the experiment are selected in ways that would not occur if the program was operating with no evaluation. Many of the reasons why subjects are excluded from participating in an experiment would apply even if no field study was being conducted (e.g, high risk individuals are excluded from probation experiments to reduce community risk). Although few experimenters in the studies considered here randomly selected individuals from the larger population of interest to participate, it appears that the external validity threat is not a major problem facing randomized studies in criminal justice.

Finally, Jupp (1989) pointed out that rigorously controlled experiments may produce an artificiality which makes the field setting unique, threatening external validity. This is an inherent risk nonetheless for all obtrusive methods, irrespective of the design rigor, leading some to advocate inconspicuous forms of research (e.g., Webb, et al., 1966). In addition, there is nothing about field experimentation which precludes the use of good process measures to identify changes in study conditions during the experiment (e.g., Clarke and Cornish, 1972). While these criticisms must be taken into account, the problem with primary study evidence within criminal justice are largely internal validity ones.

Difficulties in Implementing and Conducting Experiments

While the unparalleled methodological strength of randomized experiments is acknowledged, there are many practical issues which can hamper attempts to implement and conduct a field experiment (e.g., Dennis, 1988; Lemert and Visher, 1987; Farrington, 1983; Clarke and Cornish, 1972). For example, treatment practitioners may initially agree to the experiment, only to later covertly manipulate the randomization process in order assign certain subjects to treatment conditions (e.g., Dunford, 1990; Dennis, 1988; Connor, 1977). Maintaining the integrity of the random assignment process–the crucial element in the design–is the major obstacle to criminal justice field experiments (e.g., Petersilia, 1989; Farrington, 1983).

There are other problems, less common perhaps, which still require investigator attention. As mentioned earlier, these dilemmas can be ameliorated through effective implementation. For example, in experiments where the treatment group is receiving a more intensive or larger amount of contact than the control group, monitoring must be done to insure that the two conditions actually receive differential treatment. For example, in a Minneapolis general deterrence experiment, investigators monitored the experimental and control sites to insure that each condition received the designated levels of police patrol (e.g., Sherman and Weisburd, 1995).

Another problem facing experimenters is attrition or loss of subjects from the experimental groups during the follow-up period (e.g., experimental mortality). This is a frequent occurrence in crime reduction experiments, where subjects are followed for some specified period of time in the community to ascertain reoffending behavior (e.g., Farrington, 1983).

For example, if there is a loss of subjects from the study after one year in the community (through death, inability to locate, etc.), the groups–which were assumed to be equivalent after randomization–may no longer be so. This is even more problematic if the subjects who drop out from treatment differ in some unique way from subjects who drop out of the control group. Fortunately, there are statistical techniques which allow experimenters to compensate for subject mortality and differential attrition (e.g., Yeaton, Wortman and Langberg, 1983).

Despite the complications involved in conducting field experiments, these problems should not be overstated (e.g., Berk, Boruch, Chambers, Rossi & Witte, 1985; Boruch, 1975). First, sound randomized experiments have been carried out in a variety of social and criminal justice settings, providing evidence that rigorous designs can be done well in sensitive surroundings (e.g., Petersilia, 1989; Garner and Visher, 1988; Farrington, 1983).

Second, many impediments to a good, randomized design can be countered through sound planning and effective communication with practitioners involved in the study (e.g., Dunford, 1990; Petersilia, 1989; Riecken and Boruch, 1974). Moreover, listings of minimum threshold conditions are available to guide scholars and practitioners before deciding to conduct a randomized field study (e.g., Garner and Visher, 1988; Dennis and Boruch, 1989; Federal Judicial Center, 1981).

While practical obstacles can often be circumvented, ethical considerations preclude the use of randomized designs in many instances (e.g., Erez, 1986; Farrington, 1983; Zeisel, 1973). One might easily see the ethical and legal objection to randomly assigning offenders guilty of the same offense to drastically different dispositions, particularly if one condition is considered much "harsher" than the other (e.g., Erez, 1986). Green (1976) and other legal scholars, however, have provided guidelines to minimize the unfairness of random allocation to different sanctions.[12]

Even in cases where sentencing disparity is not an issue, practitioners may object to random assignment if they determine that a potentially helpful treatment is being withheld from needy individuals (e.g., Farrington, 1983; Boruch, 1975). Again, modifications can be made in the design which satisfy these ethical considerations and still retain experimental rigor (e.g., Powers and Alderman, 1979).

While it is true that the contentions to using randomized tests are generally exaggerated (e.g., Boruch, 1975), it is also the case that labeling the study an "experiment" does not guarantee methodological rigor. True experiments represent potential improvement over other designs, but they must be done well. Investigators should monitor the experiment carefully, particularly random assignment, to take full advantage of the strengths offered by the design (e.g., Weisburd, 1993). Moreover, host

[12] Zimring and Hawkins (1973:43) make a much ignored point referring to this issue, stating that "failure to test policies while continuing to penalize offenders in the name of deterrent beliefs becomes morally obnoxious."

agencies and officials need to be receptive to the experiment and informed about the importance of randomization integrity.

The Growth of Interest in Criminal Justice Experiments

While the obstacles facing field experimenters are sizable, they have not affected the exponential growth in the use of randomized tests over the past decade. Spurred on by funding from the National Institute of Justice [NIJ], experiments have been used in a variety of field settings with success (e.g., Weisburd, 1993; Petersilia and Turner, 1993; Weisburd and Garner, 1992).

For example, the influential Minneapolis domestic violence experiment (Sherman and Berk, 1984) has now been replicated in six additional cities (Sherman, 1992). The Rand Corporation completed a Bureau of Justice Statistics [BJS] funded experimental evaluation of community-based intensive supervision at fourteen sites across the nation (e.g., Petersilia and Turner, 1993). Indeed, Garner and Visher (1988) noted that NIJ funded over two dozen experiments during the 1988-1989 program cycles, most of which have now produced final reports. The MacArthur Foundation and NIJ are collaborating on a major longitudinal project on criminal development which includes randomized experiments at various stages (e.g., Tonry, Ohlin & Farrington, 1991; National Institute of Justice, 1990; Sherman, 1989).

Further underscoring the popularity of the design, an issue of the influential *Journal of Research in Crime & Delinquency* was dedicated in 1992 to experimentation in criminal justice (Weisburd and Garner, 1992). Since 1989, field experiments have been reported in crime reduction programs as diverse as: *drug monitoring for pretrial releasees* (e.g., Britt, et al., 1992); *multisystemic therapy for adolescent sex offenders* (e.g., Borduin, et al., 1990); *relapse prevention treatment for adult sex offenders* (e.g., Marques, et al., 1989); *comprehensive services for violent juvenile felons* (e.g., Fagan, 1990); and *vocational rehabilitation programming for offenders* (e.g., Lattimore, et al., 1990). All of these evaluations, only a handful of those conducted since 1989, highlight the increased use of the classic experimental design in criminal justice settings.

The increase of randomized studies in criminal justice field settings parallels a scholarly interest in experimentation. To illustrate, there have been several recent attempts to exclusively identify and collect controlled studies in criminal justice (e.g., Weisburd, Sherman and Petrosino, 1990; Dennis, 1988; Lemert and Visher, 1987; Farrington, Ohlin and Wilson, 1986; Farrington, 1983; Boruch, McSweeney and Soderstrom, 1978; Reicken and Boruch, 1974).

Boruch and his colleagues (1978) were able to locate over 300 randomized studies, including 75 conducted in criminal justice. While their main goal was to provide an available bibliography for prospective investigators to consult, the lengthy list of studies provided strong corroborating evidence for their assertion that experiments are feasible in social settings.

Despite stringent inclusion criteria, Farrington (1983) was able to locate 42 field experiments that tested interventions designed to "help people in the natural environment," or were conducted with the police, courts or in correctional institutions.[13] Farrington, Ohlin and Wilson (1986) updated Farrington's the first author's earlier work to include more recent experiments in their review for the Justice Study Group, noting their surprise with how many influential experiments they located.

Dennis (1988) located 41 experiments in criminal and civil justice conducted in the United States since 1972.[14] Since he was interested in the elements of a successfully implemented field experiment, Dennis used the evaluation reports to form a sample from which the original principal investigators would later be interviewed for further information about the study.

Weisburd, Sherman and Petrosino (1990) were able to locate 76 controlled studies for their *Registry of Experiments in Criminal Sanctions, 1950-1983*. While this project initially began in earnest to locate, acquire and reanalyze the *original data* from past experiments, the practical difficulties and ethical concerns encountered in

[13] Farrington (1983) imposed a strict set of eligibility criteria for experiments to be considered in his state of the art review: the report had to have a clear statement of randomization; the study used individuals and not aggregates as the unit of analysis; a minimum of 50 subjects must have been assigned to each group; and the results must have been available in a journal or book.

[14] Dennis (1988) required that sampled experiments be conducted in the United States and have at least 13 subjects in study groups.

obtaining the data sets forced the investigators to focus their efforts on collecting and examining experiments from the summary evaluation reports (e.g., Weisburd, 1993; Weisburd, Sherman and Petrosino, 1990; Sherman and Weisburd, 1987).

The scholarly clamor for controlled studies in criminal justice, particularly with regard to crime reduction, appears to have taken root over the past decade. The increase in crime reduction experiments and the growth of criminological interest in experimentation suggests that potentially stronger evidence is being produced and can be utilized in the 'what works?' debate.

A Need to Systematically Accumulate the Evidence

It is well-documented that experiments–if they are implemented with full integrity–represent the best evidence to consider in the area of program and policy evaluation (e.g., Campbell and Stanley, 1966). Learning what works, however, requires more than examining the isolated results of single experiments. Knowing what interventions are effective requires a way to systematically review prior evaluations and collectively analyze the evidence contained in those reports.

Certainly, a single experiment can be influential and important, providing answers to a practitioner in a jurisdiction or spurring criminologists to reexamine old notions and retest relationships. Stand-alone projects generally provide insufficient information for a contribution to a discipline's knowledge base (e.g., Schmidt, 1992). There are several reasons for this. First, experiments can be done poorly. While influential criminal justice experiments have been reported, later examination often revealed that there were enough methodological questions to recommend serious replication before making hasty conclusions or implementing policy (e.g., Binder and Meeker, 1988; Mrad, 1979; Lerman, 1975). Certainly, the more studies we consider on a single issue, the more confident we can be that something does or does not work. This was a point reemphasized by the University of Maryland Report (Sherman, et al. 1997).

Secondly, single studies are often influenced by time, place and population; similar studies often report different results when replicated at later dates, initiated across various jurisdictions, or when conducted with different subjects (e.g., Farrington, Ohlin and Wilson, 1986).[15] For example, the significant deterrent effect of arrest for spousal abuse found in a Minneapolis experiment (Sherman and Berk, 1984) was not found in most of the other sites (Sherman, 1992). While these differences may be due to experimenter-specific characteristics (e.g., instrumentation, etc.), variance may also be due to other moderating variables operating across the locations (e.g., Garner, Fagan & Maxwell, 1995).

[15] Farrington, Ohlin and Wilson state that "in the history of scientific endeavor, no single study, however well designed, can be conclusive. Every research project is limited to some extent by the particular setting in which it is conducted, by the particular operational definitions of the theoretical variables, by the particular time period, and so on..." (1986:172).

Concurrent with these observations is the growing recognition that the social sciences need to do a better job of accumulating knowledge (e.g., Cooper, 1989; Light and Pillemer, 1984). In fact, Rosenthal (1991) remarked about psychology that "it seems like we start over again with every new journal volume published." Cooper (1989:11) adds that there has never been a more important time for orderly knowledge building, given the explosion of research information and the increased specialization within disciplines.[16]

Similar concerns have been raised within criminology (e.g., Sherman, 1988; Farrington, Ohlin & Wilson, 1986). While recommending experimentation as a solution to the knowledge building process, they foresee the production of studies as the first part of the process; some type of accumulation or synthesis of the experiments is inferred. For example, Farrington, Ohlin & Wilson (1986:69) write that:

> Ideally, each experiment should be one link in a chain of cumulative knowledge guided by theory...Hypotheses tested in experiments are usually isolated ideas rather than part of a program of systematic testing of a larger theory. There is little attempt to see how far these hypotheses are true over different operational definitions of independent and dependent variables or different boundary conditions. In order to establish the causes of crime and to determine the best methods of preventing and treating it, criminology needs to move farther along the road of scientific progress.

Knowledge building takes place when prior studies examining a particular relationship are systematically analyzed for the knowledge they reveal (e.g., Cooper, 1989). It is this orderly process that is envisioned by advocates for criminological field

[16] For a similar reference in psychology, see Garvey and Griffith (1971).

experiments (e.g., Sherman, 1988; Farrington, Ohlin & Wilson, 1986; Farrington, 1983). In fact, Sherman and Weisburd (1987) suggested that prior experiments represented a mine of information which is unnoticed by criminal justice researchers (1987:15):

> But we can also make better use of experiments already completed...Compared to the great expense and time of conducting experiments, the most cost-effective, immediate investment NIJ could make in advancing knowledge of sanctioning effects on recidivism might be to systematically examine all of the available experiments on the question.

A number of randomized field tests have already been reported in the literature and collected by investigators (e.g., Weisburd, Sherman and Petrosino, 1990; Sherman, 1988; Dennis, 1988; Farrington, Ohlin and Wilson, 1986; Farrington, 1983; Boruch, McSweeney and Soderstrom, 1978; Reicken and Boruch, 1974). There have been few systematic attempts, however, to exclusively analyze experimental evidence. When these analyses have been done, they have generally focused on methodological issues rather than program effectiveness.[17]

For example, Connor (1977) limited his study to the randomization process in 12 social experiments, finding that integrity is jeopardized when practitioners or officials have control of assignment. Boruch and his colleagues (Riecken and Boruch, 1974; Reicken and Boruch, 1978; Boruch, McSweeney and Soderstrom, 1978) have

[17] The one exception was Kaufman's (1985) unpublished meta-analysis of randomized delinquency prevention experiments, which will be discussed in Chapter II.

presented a strong case for the suitability of social program experiments but have not systematically examined the criminal justice or legal studies.

Farrington (1983) offered only methodological conclusions in his seminal review of criminal justice experiments, and did not attempt to make any generalizations about program effectiveness. Despite the practical and ethical problems faced by experimenters, Farrington (1983:291) concluded that "...because of their high internal validity, attempts should be made to test hypotheses and evaluate technologies using randomized experiments wherever possible."

Dennis (1988) investigated implementation factors in successful criminal and civil justice experiments through written reports and scheduled telephone interviews with principal investigators.[18] Like Connor (1977), he found that the control of the randomization process by researchers rather than practitioners resulted in significantly lower rates of "covert manipulation," the discretionary overriding of the random assignment procedure without proper guidelines or researcher knowledge.

Weisburd and Sherman (1988) applied sophisticated quantitative techniques to a preliminary sample of 48 experimental evaluations of coercive criminal justice sanctions.[19] First, Weisburd and Sherman grouped the sanctioning experiments into

[18] For eleven criminal and civil justice experiments, the principal investigators either could not be contacted or would not participate in the telephone interview.

[19] Weisburd and Sherman (1988) presented this preliminary analysis at NIJ's

three groups: (1) those which reported a statistically significant effect on recidivism which favored the harsher sanction group (a "deterrent effect"); (2) those which reported no statistically significant effect on recidivism; and (3) those which reported a statistically significant effect on recidivism which favored the lesser sanction group (or "backfire effect"). Second, logistic regression was used to determine the effect of study features on membership in any one group (compared to the other two).

They reported, for example, that controlled studies were more likely to "backfire" (i.e., the harsher sanction increased recidivism) when the study included large numbers of female subjects. While Weisburd and Sherman used some meta-analytic techniques (e.g., collected written reports, codified results and attempted an analysis of study features on those results), their analysis is more akin to a sophisticated vote-counting review rather than the meta-analyses which will be discussed in Chapter II (e.g., Lab and Whitehead, 1988; Hedges and Olkin, 1980). Despite the uniqueness of their quantitative application to the traditional vote-counting method, no overall assessment of sanction effectiveness was reported.

Weisburd (1993) found in his later analysis of 74 research reports that experimental designs with a large number of subjects (N > 200), while statistically powerful in theory, were negatively correlated with effect size.[20] In other words,

Crime Control Theory Conference in New Orleans.

[20] Weisburd's research was based on the sample of experiments in criminal sanctions, gathered by Weisburd, Sherman and Petrosino (1990). Several experiments

smaller experiments (N < 50) produced higher effect sizes, suggesting that subject heterogeneity and lack of researcher control in large scale experiments overrides the increase in statistical power that larger samples bring.[21]

Narrative or Qualitative Assessments of Evidence

Some limited narrative attempts to examine program efficacy under experimental conditions have been reported. Similar to qualitative analyses of juvenile justice programs (e.g., Basta and Davidson, 1988; Romig, 1978), Farrington, Ohlin and Wilson (1986) carefully grouped field experiments into broad intervention categories and made some clinical assessments about their effectiveness. For example, they found promise for preschool prevention programs, although that category is based on a single Head Start experiment (e.g., Schweinhart, 1987).[22]

Sherman (1988) also provided further qualitative analysis of 18 randomized experiments reported in Farrington, Ohlin and Wilson (1986) which tested the effects of criminal justice sanctioning on recidivism (i.e., special or specific deterrence). Sherman (1988) found that only one experiment, the Minneapolis Domestic Violence study, reported a deterrent effect for the conceptually harsher sanction condition.

in the original sample did not provide enough data for Weisburd's later analysis.

[21] See Tanur (1983) for similar concerns in conducting large-scale experiments in other social settings.

[22] This highlights one of the problems with conducting a broad review with a small sample of studies. When you begin to create specific categories, some cells are left with one or two cases upon which conclusions are drawn.

Similar to the landmark Lipton, Martinson & Wilks (1975) review, The University of Maryland Report to the United States Congress (Sherman, et al. 1997) rated evaluation reports on a five-point methodological scale (with 5 being the highest, reserved for randomized studies conducted with full integrity). Using this scale, the researchers assessed the evidence for various crime prevention programs, indicating which programs seemed to work, which did not, and those where the scientific evidence was inconclusive.[23] They also concluded that—despite the millions of funds expended by the federal government on criminal justice programs—there is still a paucity of evaluation evidence to inform policymakers. They did not focus exclusively on randomized designs, however, in reaching any tentative conclusions.

While qualitative reviews of program evaluations can be influential (e.g., Lipton, Martinson and Wilks, 1975; Martinson, 1974), they are considerably more difficult to conduct as the number of studies under review increases. First, reaching conclusions from the results of multiple studies is risky when the populations, settings, study characteristics and interventions vary widely across research reports (e.g., Wolf, 1986). As Glass and his colleagues noted (Glass, McGaw and Smith, 1981), accurately summarizing a considerable number of outcome studies is just as difficult without quantification as a large number of survey responses or case files.

[23] Sherman, et al. (1997) did report quantitative estimates of the experimental effect in some sections of their report (e.g., school-based prevention), but attempted no formal quantitative synthesis.

Secondly, since rules of scientific rigor and explicitness are not applied with equal force to the narrative review, the reviewer runs the risk of selectively including and excluding studies (e.g., Cooper, 1989; Wolf, 1986). Selection bias in literature reviews can lead to different published conclusions, as illustrated by the vast differences across sex offender treatment outcome studies:

> Vernon Quinsey's (1984:101) conclusion in his review of recidivism studies of rapists applies to this broader review as well: 'The differences in recidivism across these studies is truly remarkable; clearly by selectively contemplating the various studies, one can conclude anything one wants' (Furby, Weinrott and Blackshaw, 1989:22).

While narrative reviews can be done quite well, they are best conducted when the number of studies is small (e.g., Sherman, Milton and Kelly, 1973)[24] or when the purpose is to present the 'state of the art' rather than an assessment of program effectiveness.[25] In contrast, the need for a more rigorous research methodology to be applied when reviewing evidence was underscored by Cooper (1989:145):

> Because of the growth in empirical research, the increased access to information, and the new techniques for research synthesis, the conclusions of research reviews will become less and less trustworthy unless something is done to systematize the process and make it more rigorous. Because of the increasing role that research reviews play in our definition of knowledge...adjustments in procedures are inevitable if social scientists hope to retain their claim to objectivity.

[24] Sherman, Milton and Kelly (1973) present an excellent qualitative analysis of seven team policing case studies. Imagine their difficulty in attempting to analyze one hundred studies, without the use of quantification.

[25] The University of Chicago Press, through funding by the National Institute of Justice, provides the annual volume entitled Crime & Justice: An Annual Review of Research. These are often state of the art reviews, which use studies to highlight current issues, problems and prospects.

Many scholars have written in detail about additional problems with the narrative review (e.g., Wolf, 1986; Glass, McGaw & Smith, 1981). While the practice of qualitatively reviewing studies is improving (e.g., Cooper, 1989), many threats to reliability and validity remain, including: (a) the failure to report literature search method and study inclusion criteria; (b) the differential, subjective weightings of studies selected, with no empirical or rational explanation for the decision; (c) misleading and subjective interpretations of study findings; (d) generalizing the review beyond the studies considered; (e) the failure to report research study characteristics, including sample sizes, settings and treatment components; (f) dependence on whether statistically significant findings were reported; and (g) a lack of attention to statistical power issues.

These last two points have drawn increased attention from scholars in recent years, stimulated by Cohen's landmark work on statistical power (e.g., Cohen, 1977; 1962). Lipsey notes (1990:22) that most treatment effectiveness research is designed to detect large effects only; this finding has remained consistent across many areas of inquiry (e.g., medicine, gerontology, education, sociology, etc.).[26]

While the power to detect a significant result–when there is one–is based on a wide range of design sensitivity issues (e.g., Weisburd, 1993; Lipsey, 1990), most evaluation research is so severely restricted by sample size that the probability of

[26] Cohen's (1977) definition of small, moderate and large effects would translate into r's of .10, .25 and .40.

detecting small or moderate treatment effects is too small to be considered satisfactory. When one adds the generally low reliability of crime outcome measures, the inadequate strength and intensity of experimental treatment programs and so forth, there is little surprise that past reviews relying on statistical significance would find little evidence of program efficacy (e.g., Lipsey, 1990).

Within criminal justice, Brown (1991) investigated the statistical power of 3689 significance tests reported across 53 empirical studies published in the eight leading criminological journals. Although criminal justice, on average, was substantially more powerful than fields where similar analyses were done, Brown reported that two-thirds of his studies had less than a 50% probability of detecting a small effect.

This finding takes on great importance when considering that treatment, on average, tends to have small to moderate effects (e.g., Lipsey and Wilson, 1993; Lipsey, 1990). In Lipsey and Wilson's (1993) collection of meta-analyses, criminal justice programs generally hover around Cohen's small effect category. Some may argue that small effects are unimportant, but as Lipsey points out, a difference of .2 standard deviation units between experimental groups sounds trivial, but it would represent an 18% reduction in deaths for a medical experiment with 100 subjects in each group (Lipsey, 1990:24). Most social science research, particularly criminal justice, would be powerless to detect a difference of this magnitude.

This factor, along with those mentioned earlier, has caused many to criticize the traditional reviewing method in social science (e.g., Rosenthal, 1991; Hunter and Schmidt, 1990; Cooper, 1989; Light and Pillemer, 1984; Glass, McGaw and Smith, 1981; Jackson, 1980; Feldman, 1971). The futility of the narrative review is best illustrated in the debate over findings reported in reviews of the 'what works?' literature. Martinson (1974), while noting the poor quality of the evaluations sampled, found little cause for optimism regarding the efficacy of correctional treatment. While his review has been incorrectly characterized as claiming that 'nothing works,' Martinson was clearly pessimistic regarding the promise of rehabilitation.[27]

Over the years, a series of critiques of Martinson's (1974) research have been published. Palmer (1978; 1975) observed that nearly half (48%) of the programs reviewed by Martinson reported positive outcomes. It is interesting to note that both Bailey (1966) and Logan (1972) characterized the research evidence as not supportive of rehabilitation although more than half of their sample evaluations reported positive outcomes.[28] The continuing debate over these literature surveys highlights the

[27] Recent characterizations of Martinson's (1974) work might be considered unfair. Tonry and Morris (1985) hint that his publication of the findings was irresponsible. Other recent remarks refer to Martinson's work as "notorious" rather than influential. Nonetheless, Martinson conducted the largest and most exhaustive review of correctional outcomes ever reported (until Lipsey's more inclusive juvenile intervention study in 1992). His work was thoroughly supported by the National Academy of Science panel in 1979 (e.g., Sechrest, White and Brown, 1979) and by several later reviews. Martinson (1979) himself would later change--not recant--his original findings based on another exhaustive research project.

[28] For example, Bailey (1966) found that 60% of the "experimental" studies reported treatment success. Logan (1972) found 70% of his sample claimed at least "fair

problems with solely using qualitative methods to review evidence (e.g., Izzo and Ross, 1990; Cullen and Gendreau, 1989). It is clear that better methods for handling evidence must be utilized.

A Better Method for Analyzing Evidence: The Rise of Meta-Analysis

Developed in response to the shortcomings of the narrative review, meta-analysis represents a significant development in social science, particularly in the area of treatment effectiveness (e.g., Lipsey and Wilson, 1993; Durlak and Lipsey, 1991; Lipsey, 1990; Lipsey, 1988). While several meta-analytic techniques have been developed (e.g., Wachter and Straf, 1990), the underlying rationale of these methods remains similar. Meta-analysis involves:

> ...the application of quantitative methods to the problem of combining results from different analytic studies. Meta-analysis is not a statistical method per se, but rather an orientation toward research synthesis that uses many techniques of measurement and data analysis (Wachter and Straf, 1990:xiv).[29]

While methods for combining the results of studies have been around since the early 1900s, historians credit Gene Glass (1976) with originating meta-analysis following his development of a standardized effect size measure {d} that could be used to express the difference between experimental and control groups in standard deviation units (e.g., Olkin, 1990). Using this numeric effect size as a dependent variable, Smith and Glass (1977) were able to quantify over 400 psychotherapy

success" on outcome variables.

[29] The Wachter and Straf volume was the result of The National Research Council's workshop on the future of meta-analysis.

experiments. They concluded, in contradiction with some of the notable narrative reviews on the issue (e.g., Eysenck, 1961), that subjects exposed to psychotherapy had–on average–a strong, beneficial effect when compared to control group subjects.

Using the standardized effect size measure–or common metric–moved the emphasis of the review from statistical significance, which can be misleading, to the actual magnitude of effect the experimental treatment achieved. The common metric expresses the difference between the groups in a manner that is independent of statistical significance.[30]

The Smith and Glass (1977) findings led to extensive use of meta-analysis in the fields of psychology and education. Its popularity soon spread to other fields, particularly medicine and business, with the technique receiving national press coverage (e.g., Mann, 1994; Strauss, 1991; Chronicle of Higher Education, 1990; New York Times, 1990).

Judging by earlier estimates (e.g., Durlak and Lipsey, 1991), probably several thousand meta-analyses have been reported in the social and medical science literature.[31] While harsh criticisms are periodically made (e.g., Murray, 1992; Slavin,

[30] Jacob Cohen wrote (1992) that "I have long believed that psychology's preoccupation with significance testing and p values has distracted it from attending to the magnitude of phenomena and has thereby retarded its growth."

[31] For example, the Educational Resource Information Center (ERIC) data base provides over 1,000 references to the word "meta-analysis" for the years 1987-

1984; Eysenck, 1978), it is apparent that meta-analysis will continue to be heavily utilized (e.g., Cohen, 1992; Hedges, 1992; Schmidt, 1992).

While several different meta-analytic techniques exist (e.g., Bangert-Drowns, 1986), five essential steps are performed in conducting most quantitative syntheses (e.g., Abrami, Cohen and d'Apollonia, 1988). These are: (a) specifying the inclusion criteria, i.e., which studies will be included and excluded from the sample?; (b) locating studies (data collection phase); (c) coding study features; (d) calculating individual outcomes (and developing a common metric effect size); and (e) data analysis.[32] Statistical applications following these procedures are numerous and have been expounded on elsewhere (e.g., Rosenthal, 1991; Hunter and Schmidt, 1990; Wolf, 1986; Hedges and Olkin, 1985; Glass, McGaw and Smith, 1981).

While new research is exciting and important, meta-analytic inquiries are necessary links in the process of acquiring information about what has been tried, what has failed, and what has succeeded in a specified area (e.g., Wolf, 1986; Glass, McGaw and Smith, 1981).[33] The ability of meta-analysis to provide a rigorous

1993 alone.

[32] Some methods of meta-analysis use combined probabilities to assess statistical significance of several studies on the same treatment. Those syntheses predominantly do not include information on study characteristics or information regarding the magnitude of effect (e.g., Rosenthal, 1991; Bangert-Drowns, 1986).

[33] Glass (e.g., Glass and Kliegl, 1983) also envisioned meta-analysis as a tool that in studying how scientists have thought about a particular content area over time.

method for synthesizing prior evaluations may provide a technique for orderly knowledge building (e.g., Schmidt, 1992; Cooper, 1989; Light and Pillemer, 1984).

Aside from the potential of meta-analysis in assisting scientists in their task of accumulating knowledge, there is some evidence that the technique may be influential with policy and program decisionmakers (e.g., Cordray, 1990; Chelimsky and Morra, 1984). Practitioners at this level, looking for more conclusive evidence before taking action, may find a meta-analysis of many studies bearing on the issue more persuasive than a single one.

Moreover, meta-analysis may not only provide benefit to scholars and practitioners, but it could also generate guiding information for funding agencies (e.g., Hunter and Schmidt, 1990). The growing frustration over the failure of social science to recommend anything–but more research–has also led to an increased enthusiasm for meta-analysis.[34] Cordray (1990:117) writes about the potential of meta-analysis:

> Reductions in federal funding for research and evaluation have forced attention to summing up what is known about the effects of interventions. These syntheses–if properly conducted–can not only reveal what is known about the effects of interventions, but also help to identify gaps in knowledge and serve as a rational basis for the development of subsequent programs and

[34] In fact, Hunter and Schmidt (1990) write that "With as many as a hundred or more studies on a relationship, one might think that there would be a resolution of the issue. Yet most review studies traditionally have not concluded with resolution, but with a call for more research on the question. This has been especially frustrating to organizations that fund research in the behavioral and social sciences. Many such organizations are now questioning the usefulness of the research in the social sciences on just this ground. If research never resolves issues, then why spend millions of dollars on research?"

investigations. Despite its relatively short history, the meta-analytic perspective appears to have left a rather distinctive mark on basic and applied research.

Criticisms of Meta-Analysis

One of the major criticisms of meta-analysis is the so-called "garbage-in, garbage-out" problem, or the inclusion of methodologically inferior studies with rigorous designs (e.g., Bangert-Drowns, 1986; Wolf, 1986; Eysenck, 1978). Including only experimental studies with random assignment can mitigate this criticism; however, randomized experiments experience different integrity problems, including the breakdown of the random assignment proces (e.g., Dunford, 1990).

Methodological flaws can affect the outcome of studies, usually by underestimating treatment effects (e.g., Lipsey, 1990). One way to counter the effect of experimental breakdowns is to code them when they are reported, and analyze those studies later to determine if the magnitude of effect differs from those reported in the other experiments (e.g., Glass, McGaw and Smith, 1981). Another technique is to assign each experiment a methodological score so that the best conducted experiments receive the most weight in the analysis. Unfortunately, there is no consensus on how to score and weight methodological factors for a meta-analysis (e.g., Durlak and Lipsey, 1991).

Another criticism of meta-analysis is that it is totally dependent upon written reports for information. This is a precarious situation, since there are writing and editorial decisions about what to include in a research report. Meta-analysts will find some information of interest is missing from final reports. Moreover, the pertinent data coded for the meta-analysis may be in error.

Some have suggested using telephone interviews or mail surveys to retrieve information missing from study reports, but those strategies are especially difficult when meta-analyses are comprehensive, covering studies outside the United States and those published before 1980. Unless investigators are contacted in adjacent jurisdictions, phone costs are likely to be prohibitive (e.g., Durlak and Lipsey, 1991).

Perhaps the best strategy is to develop coding instruments that rely as much as possible on information universally reported, and eliminate items which are never reported or have to be subjectively inferred by coders. Walter (1992) presented such a restricted coding scheme in his meta-analysis of gene-crime relationship; only four moderating variables were coded, nearly all of which were universally contained in the studies he reviewed.

Meta-Analysis in Criminal Justice

Like randomized field experiments, meta-analysis is not without problems or critics. Nonetheless, both methodologies have advanced to the point where strategies have been developed to deal with implementation and other process problems. As the case with experiments, the use of meta-analysis in criminology and criminal justice research is rapidly increasing.

Meta-analysis is a new development in criminal justice, having first been reported in the literature in 1984. However, it has since been applied to examine several important theoretical relationships, including*: the influence of family factors on juvenile conduct problems* (e.g., Wells and Rankin, 1991; Loeber and Stouthamer-Loeber, 1986); *the effect of higher education on indicators of police performance* (e.g., Hayeslip, 1989); *the influence of economics, poverty or social class on violent behavior* (e.g., Sederstrom, 1987); *the role of genetics in the etiology of crime* (Walter, 1992); *and the longterm sequelae of childhood sexual abuse in women (Neumann, Houskamp, Pollock and Briere, 1996).*

The most frequent use of meta-analysis has been in the area of treatment effectiveness, where Lipsey and Wilson (1993) found over 300 quantitative syntheses in the applied social sciences alone. Within criminal justice, several meta-analyses examining the efficacy of crime reduction programs have now been reported. While some of these prior quantitative reviews focused solely on a particular treatment modality such as residential treatment of delinquents (e.g., Garrett, 1985), others have included all types of adult and juvenile interventions (e.g., Pearson, Lipton & Cleland, 1996).

The most comprehensive meta-analysis of treatment effectiveness may soon be available. Lipton (1995), one of the original authors of the controversial literature survey, *The Effectiveness of Correctional Treatment* (Lipton, Martinson and Wilks, 1975), is now conducting a meta-analysis of offender treatment evaluations reported since 1968.[35] While they have produced preliminary project reports at annual conferences, the analysis of over 1,200 rehabilitation evaluations has yet to be completed.

Putting it all Together: The Current Project

Given the limitations of the evidence considered by reviewers using qualitative or quantitative techniques and the shortcomings of the narrative research review, this project uses two rapidly developing tools to revisit the 'what works?' literature. Randomized experimental studies will comprise the evidence to be considered; meta-analytic techniques will be used to assess the evidence and answer the questions posed at the end of Chapter II.

No prior systematic analysis of experiments has used meta-analytic techniques to assess program effectiveness, aside from Kaufman's (1985) analysis of 20 delinquency prevention studies. Even the narrative reviews have been somewhat limited; for example, Farrington, Ohlin & Wilson (1986) considered only 50 experiments in their analysis, while Sherman (1988) examined 18 controlled criminal sanctioning studies. The exponential growth in the number of experiments, concurrent with the growing information technology for searching and retrieving studies, resulted in a much larger group of studies to be considered here (N=150). In fact, outside of the three meta-analyses which received massive federal funding (e.g.,

[35] *The Effectiveness of Correctional Treatment* reviewed evaluations published before 1968.

Lipsey, 1992a; Wells-Parker, et al., 1995; Pearson, et al., 1996), the study sample analyzed herein represents one of the largest 'what works?' projects in criminal justice.[36]

Moreover, except for Kaufman's (1985) limited analysis of 20 delinquency prevention experiments, none of the prior meta-analyses in criminal justice exclusively examined treatment effects under experimental conditions. All of the others, including Lipsey's (1992a) influential analysis of juvenile interventions, included all types of quasi-experimental designs, provided some type of comparison or control group was used (e.g., Roberts and Camasso, 1991; Izzo and Ross, 1990; Garrett, 1985).

While this meta-analysis of a wide range of experimental evaluations in the what works area may be novel, the idea of merging the two methodologies is not. In fact, in a special issue of *Evaluation Review* on social policy experimentation, a group of scholars recommended that randomized experimentation be conducted with the later goal of meta-analyzing the results (e.g., Berk, Boruch, Chambers, Rossi & Witte, 1985). Besides its obvious relevance to the treatment effectiveness debate, this project was conducted in response to the solicitations of these and other investigators long interested in criminal justice experimentation (e.g., Weisburd, 1993; Dennis, 1988; Sherman, 1988; Farrington, Ohlin & Wilson, 1986; Boruch, 1975).

[36] In fact, of 302 treatment effectiveness meta-analyses collected by Lipsey and Wilson (1993), this meta-analysis of 150 experiments would be among the ten largest.

The next chapter places this research in the context of the 'what works?' debate, with special emphasis on the prior meta-analyses, and focuses this study on several questions that need to be answered. The following chapters detail the research project, including: the criteria used to select experiments for the sample (Chapter III); search and retrieval techniques utilized to collect the experimental reports (Chapter IV), and the coding process (Chapter V). The concluding chapters present the results of descriptive analyses (Chapter VI) and several focused statistical tests (Chapter VII). Final conclusions and recommendations are presented in Chapter VIII.

CHAPTER II: FROM MARTINSON TO META-ANALYSIS: THE 'WHAT WORKS?' LITERATURE

By the 1960s, rehabilitation was firmly established as the primary justification for punishment (e.g., Logan, Gaes, Harer, Innes, Karacki & Saylor, 1991; Allen, 1959). In fact, most of the major developments in the criminal justice system during the time leading up to 1960 were established to enhance rehabilitation. For example, indeterminate sentencing was designed to permit judges wide latitude in 'fitting the punishment to the offender' (e.g., Cullen and Gendreau, 1989; Allen, 1959). Parole boards were given discretionary power to make clinical judgements about which offenders had been rehabilitated and those who needed further restraint (e.g., von Hirsch, 1985).

Allen (1959) noted how the rise of the rehabilitative ideal in criminal justice completely dominated criminology, retarding research in non-treatment areas. Allen (1959:227) pointed out that the prevailing attitude amongst scholars was that "...matters of treatment and offender reform were the only questions worthy of serious attention in the whole field of criminal justice and corrections." Deterrence and retribution, though popular ideologies with some practitioners and laypersons, were viewed as uncivilized and non-scientific approaches to dealing with criminals (e.g., Menninger, 1966).

However, the 1960s witnessed the decline of the rehabilitative ideal in the United States (e.g., Logan, et al., 1991; Cullen and Gilbert, 1982; Allen, 1981; Clear, 1978). Cullen and Gilbert (1982) have detailed how the social turbulence of that decade moved America, ideologically, away from rehabilitation toward a hybrid which emphasized deterrence, incapacitation and retribution, referred to as the *Justice Model.* Other writers have noted their opposition to the treatment ideal because of the unfairness which resulted from discretionary judgments made by criminal justice officials and treatment providers (e.g., Clear, 1978; von Hirsch, 1976; Morris, 1974; American Friends Service Committee, 1971).

In addition, many noted that the underlying theory of rehabilitation was flawed: the origins of crime are located in social processes, not individual defect. Some scholars posited that attempts to reform the individual were misguided at best, and unfairly coercive at worst (e.g., Reiman, 1985; Clear, 1978; Tittle, 1974).

Moreover, scholars noted that rehabilitation philosophy rested on assumptions that underlying deficits responsible for criminal behavior could be accurately diagnosed and effective treatment administered (e.g., Di Gennaro and Vetere, 1974). Not only did this belief lack strong supportive scientific evidence, but there was also no validation for the assertion that parole boards or clinicians could tell who was rehabilitated and who was not before making a parole decision (e.g., Cullen and Gilbert, 1982; Morris, 1974; Di Gennaro and Vetere, 1974). Additionally, institutional behavior measures and attitudinal test outcomes seemed to bear little relationship to post-release success (e.g., Morris, 1974; Tittle, 1974); some charged that the parole decision was comprised of uninformed guesswork (e.g., von Hirsch, 1976; Morris, 1974).

While the decline of rehabilitation was influenced by social events, political climate and ideological opposition, it is clear that the discrediting of the treatment model was also based on scientific evidence (e.g., Logan, et al., 1991; Clear, 1978). It would be hard to imagine a successful attack on treatment if rehabilitation programs were demonstrably effective. While proponents of rehabilitation occasionally emphasized its humanistic and compassionate aspects (e.g., Menninger, 1966), the primary rationale for treatment was its utility.[38]

Like deterrence and incapacitation, the philosophy of rehabilitation is primarily utilitarian. Treating offenders–even coercively–is justified if it serves to control crime and protect society. Utilitarian theories ultimately rest on scientific evidence; if they are shown to be ineffective, then it is difficult to defend the injustice or inequity they invoke.[39]

[38] Note that some recent scholars have argued for rehabilitation on non-utilitarian grounds, namely that it is the only justification for punishment which requires the state to care about the offender (e.g., Rotman, 1990; Cullen and Gilbert, 1982).

[39] The utilitarian crime control goal holds for incapacitation and deterrence as well. For example, incarcerating offenders to reduce the general crime rate is (incapacitation) indefensible if that policy has no effect on crime (e.g., Walker, 1994). The same can be said of deterrence and harsher sanctions (e.g., Zimring and Hawkins, 1973). As mentioned earlier, perhaps the attraction of retribution to some is that it is non-utilitarian, at least with regard to crime control.

The Role of Evidence: Early Reports

Unfortunately for rehabilitation advocates, the evidence of rehabilitative efficacy was notably absent. Beginning in 1966, a series of narrative literature surveys began to examine treatment effectiveness by amassing prior evaluations and qualitatively analyzing the results. Although several earlier writers inferred that treatment might not be effective (e.g., Glaser, 1965), Bailey (1966) provided one of the first published reviews of correctional outcomes.

Bailey collected 100 correctional evaluation reports published between 1940-1960, as long as empirical data was used in the research. He noted the poor quality of the evidence; only 21% of the evaluations included a control or comparison group of any kind. Despite finding that a majority reported success on outcome variables, Bailey questioned the veracity of the findings. For example, the number of negative results, i.e., treatment groups doing worse, increased with the rigor of the research design. Since most evaluations were conducted and written by the treatment providers themselves, using weak and unreliable designs, the evidence for treatment success was, in Bailey's words (1966:157), "quite discouraging."

These generally pessimistic findings were echoed by other independent reviews published shortly thereafter (e.g., Robison and Smith, 1971).[40] Logan (1972) also collected 100 correctional evaluations published after 1940 and found that none

[40] There were a few research reviews which concluded more optimistically, or perhaps less skeptically, about the promise of rehabilitation (e.g., Adams, 1967).

of the research designs met ten essential criteria for a successful test. Logan (1972:381) also found as the rigor of the design increased, claims of rehabilitative success declined.

Like Bailey, Logan (1972) did not claim that 'nothing works'–in fact, most of his sample reported at least "fair" success–but that the evaluations conducted were so lacking in design rigor that there is simply no evidence of treatment effectiveness. Slaikeu (1973) reported similar findings in his analysis of 23 institutional group counseling programs published during 1945-1970.[41]

Despite the skeptical nature of these large surveys of rehabilitation programming, it was Martinson's (1974) *Public Interest* article, a narrative review of 231 correctional rehabilitation studies, which became the watershed of evidence used against the treatment model. Martinson's paper was a precursor of the larger work, *The Effectiveness of Correctional Treatment* (Lipton, Martinson & Wilks, 1975),[42] and

[41] Slaikeu (1973:88) writes that "even though the evaluative studies report a variety of positive results...they fall short of the criteria of scientific research...This makes it impossible to conclude that group treatment in correctional institutions is an effective rehabilitation mode."

[42] Despite its seminal stature in the field, this work leaves a sad personal legacy. In the late 1970s, at the height of his professional career, Robert Martinson committed suicide. In an unrelated event, co-author Judith Wilks disappeared from New York around the same time.

was the largest survey of evaluation studies published in the literature until Lipsey's (1992a) analysis of 443 juvenile interventions.[43]

The Martinson Report

In 1966, New York commissioned a group of scholars to examine the prior evidence on rehabilitation programs, presumably to guide New York's plan to implement a treatment-oriented philosophy (Martinson, 1974:23). However, when researchers found little evidence of rehabilitation effectiveness, the state's criminal justice planning agency attempted to squelch the report (Martinson, 1974:23). It was only after an attorney subpoenaed the research during a case that the state released the findings (Martinson, 1974:23).[44]

In *"What Works? Questions and Answers About Penal Reform,"* Martinson examined correctional evaluations reported in English before 1968, provided the research design used a control or comparison group and included some outcome measure of crime. Martinson examined these studies across several broad intervention areas (e.g., vocational and educational training), concluding that there was little evidence that any particular strategy reduced criminal behavior.

[43] Although Lipton's (1995) current CDATE project, which will collect and meta-analyze correctional evaluations since 1968, will likely double this work in sample size.

[44] Sanchez (1990), a former student of Martinson, has written an interesting piece on Martinson's efforts to publish the report.

Although often wrongly credited with claiming 'nothing works,' Martinson was clearly pessimistic regarding the effect of treatment, writing (1974:49):[45]

> ...I am bound to say that these data, involving over two hundred studies and hundreds of thousands of individuals as they do, are the best available and give us very little reason to hope that we have in fact found a sure way of reducing recidivism through rehabilitation. This is not to say that we found no instances of success or partial success; it is only to say that these instances have been isolated, producing no clear pattern to indicate the efficacy of any particular method of treatment.

Despite the earlier surveys, it was Martinson's report that is most remembered and cited in the literature. The comprehensiveness of the survey certainly added to its weight, but the ideological and political climate in which it arrived enhanced its publicity (e.g., Sanchez, 1990). The rising crime rate and social upheaval of the 1960s had focused considerable public and official attention on crime control (e.g., Cullen and Gilbert, 1982).

Conservatives, for example, generally saw rehabilitation as a philosophy which coddled criminals (e.g., Sanchez, 1990; Cullen and Gilbert, 1982: Clear, 1978). Martinson's report was persuasive evidence for conservatives that the system needed to get tougher; many writers began to espouse classical deterrent themes.[46] In fact, James Q. Wilson (1975:172) recommended--following the discouraging results

[45] *The Effectiveness of Correctional Treatment* echoed these findings, but provided annotated summaries so other investigators could replicate the review (e.g., Lipton, Martinson & Wilks, 1975).

[46] Interesting that the NAS Panel on Deterrence (e.g, Blumstein, Cohen & Nagin, 1978), which concluded that there was a glimmer of evidentiary support for the general deterrence hypothesis, did not stem the rush toward harsher penalties.

reported by Martinson and others–that the criminal justice system focus on isolation and punishment and move away from rehabilitation altogether.[47]

Even academic liberals, who had provided the major support base for the treatment ideal, underwent significant change during this period. The growing mistrust in the integrity and benevolence of government–confirmed some argued by Watergate, the Vietnam War and campus unrest–generated a close examination by liberals of treatment in practice (e.g., Logan, et al., 1991; Cullen and Gilbert, 1982). The coercive nature of treatment and the disparity of punishment due to discretionary judgments made by treatment providers and officials moved some toward a position that punishment–if it does not reform–at least can be fair.

Martinson's article affirmed the academic liberal's worst fears: injustice was being perpetuated in the name of an ineffective model. It was in this ideological camp that the modern view of just desert originated (e.g., von Hirsch, 1976; Fogel, 1975; American Friends Service Committee, 1971). By focusing on the harm caused by the offense and individual culpability, discretionary judgments can be reduced and punishment levels between different offenders convicted of similar offenses can be made more equitable (e.g., von Hirsch, 1976).

[47] However, Wilson (1980) later noted that some treatments might work with some offenders ("differential treatment effects").

Apart from ideology, something else was happening during the 1970s that fostered an atmosphere receptive to Martinson's report. The decade, often referred to as the age of accountability, was a period when programs in all areas were under great scrutiny (e.g., Fischer, 1977). The fiscal problems of the early part of the decade not only meant that unlimited federal investment in social programming was over (i.e., The Great Society), but that existing programs would have to be demonstrably effective (e.g., Fischer, 1977). Correctional treatment, when examined as Martinson did, failed to measure up (e.g., Sanchez, 1990).[48]

It should be pointed out that offender rehabilitation was only one of many areas of social intervention examined and found lacking by research reviewers (e.g., Prather and Gibson, 1977; Fischer, 1977).[49] No clear effects were found for programs in education, psychology, social work and other human service areas (e.g., Prather and Gibson, 1977; Fischer, 1977).

The Response

Martinson's report had lasting consequences on the way scholars, officials and the public thought about correctional rehabilitation. Palmer (1992), the Chief of

[48] Conrad (1975:7) writes: "The disillusion in criminal justice is not an isolated frustration. It occurs at a time when we must reexamine all of our assumptions about the functions of government and its responsibilities to and for those governed."

[49] Adams (1974) pointed out that even if Martinson were correct in his assessment, correctional treatment research shows the same success rate as research and development in other fields, notably medicine and business.

Research for California's Department of Corrections, noted that the paper triggered a period of widespread pessimism in corrections until the early part of the 1980s. "Nothing works" became the slogan of the public, the media, policymakers and correctional practitioners, erroneously crediting Martinson with fully proving the futility of rehabilitation programs (e.g., Palmer, 1992; Cullen and Gendreau, 1989), an attribution Martinson himself objected to (1979:254).

The article also brought a strong response from the academic community (e.g., Sanchez, 1990). Perhaps one important result of the Martinson Report was the thoughtful analysis it generated, as some writers attempted to account for why treatment failed to demonstrate success.

Some of the reactions to Martinson's report, however, took the form of rebuttal rather than reflection. For example, Klockars (1975) pointed out the discrepancies between Martinson's article and the *Effectiveness of Correctional Treatment* book, also criticizing the organization and writing errors in the larger text. Chaneles (1975) disputed Martinson's role and contribution in the research. Some of these exchanges were confrontational and acrimonious (e.g., Martinson, 1976; Martinson, Palmer and Adams, 1976; Chaneles, 1975; Klockars, 1975).[50]

[50] Klockars (1975:54) even noted that John Conrad, former editor of the *Journal of Research in Crime and Delinquency*, invited Martinson to reply to Palmer's criticisms of his work. Martinson apparently wrote a personal attack on Palmer which Conrad refused to publish. There also seemed to be antagonism toward Martinson for "marketing" his results in the media, including an appearance on the CBS news program, *60 Minutes*.

Another point of rebuttal was to attack the credibility of Martinson's paper, claiming that evidence of positive outcomes was ignored (e.g., Palmer, 1975; Adams, 1976). For example, in "*Martinson Revisited*," Palmer (1975) found that nearly half of the evaluations reviewed in "*What Works?*" reported success on the outcome criterion. Palmer (1978) later published a full length monograph which essentially was a refutation of the Martinson Report and the 'nothing works' mentality.

Other scholars put the question of veracity aside in an attempt to explain why treatment did not demonstrate effectiveness. Quay (1977), in his review of a group counseling experiment in a California prison (e.g., Kassenbaum, Ward & Wilner, 1971), found that the treatment program had not been implemented with full integrity. Treatment staff were not provided enough training, they were unenthusiastic about the program and the original theoretical principles of effective treatment were never followed. Quay (1977) noted that group counseling--in theory--could be effective with offenders, but that the disappointing result should have been expected given the collapse of therapeutic integrity.

Sechrest and West (1983) urged scholars to attend to the *strength and integrity* of treatment interventions before concluding that rehabilitation failed. While Sechrest and West's (1983) concept of integrity was similar to Quay's (1977), strength implies something different. Treatment strength refers to the amount and intensity of the intervention, quite analogous to the term dosage in medicine. For example, one contact per month with a probation officer may not be strong enough to impact recidivism, but five contacts per month may be effective.

Another important reflection on "*What Works?*" was the growing emphasis which scholars placed on differential treatment effects (e.g., Wilson, 1980; Brody, 1976). Several writers used evidence from the Pico experiment to make their point. In the experiment, boys were classified as amenable or non-amenable to treatment and then randomly assigned to group counseling or no-group counseling conditions. Adams (1970) found that group counseling was beneficial for some boys (those classified as "amenable to treatment"), had no effect on others, and was even harmful to a minority of them (e.g., Adams, 1970). It could be that treatment works for some subjects–some of the time–but success is masked in global comparisons between experimental and control groups (e.g., Gendreau and Ross, 1983-1984; Wilson, 1980).[51]

Some additional literature reviews reported shortly after Martinson's paper added considerable weight to pessimism regarding treatment programs. The first of these was published by Brody (1976), who failed to find any evidence of correctional effects after reviewing 65 evaluations of differential sentencing. However, Brody

[51] However, von Hirsch and Maher (1991) hinted at the difficulty in transferring differential treatment effects into policy or practice. For example, if a particular strategy works well with white females, can it be withheld from males or non-white females?

focused on the poor quality of the research and concluded that the question of effectiveness had yet to be tested.

David Greenberg (1977) reviewed program evaluations reported in English through 1975. He found little evidence to reverse Martinson's position, concluding that "the assertion that 'nothing works' is an exaggeration, but not by much" (1977:141).

Romig (1978) reviewed 179 juvenile intervention program evaluations which used randomized or matched control designs. Similar to Martinson (1974), Romig found little evidence for a single effective treatment approach with juvenile offenders. This was a critical conclusion, given the quality of the designs in Romig's sample. These pessimistic findings were consistent across several reviews of juvenile treatment and delinquency prevention programs, albeit with less rigorous evidence (e.g., Wright and Dixon, 1977; Lundman and Scarpitti, 1978).

The National Academy of Science Panels

As a response to the conflict over the Martinson Report, the increasing pessimism in correctional settings and the sweeping sentencing reforms being instituted across the nation, the National Academy of Science (NAS) convened a prestigious panel of scholars to examine research on rehabilitative techniques (e.g., Lipsey, 1988; Sechrest, White and Brown, 1979). The panel's first report strongly corroborated the findings in *The Effectiveness of Correctional Treatment*, even noting that Lipton, et al., were 'overly lenient' in their assessment of the evaluations they reviewed (e.g., Lipsey, 1988; Sechrest, White & Brown, 1979).

However, the NAS panel was clearly concerned with the quality of the evidence rather than the substantive issue of whether treatment worked. Sechrest, et al. (1979) asserted that the question of rehabilitative efficacy could not be sufficiently answered, given the paucity of unequivocal research evaluations.[52] They made several recommendations for improving correctional evaluations, including the use of randomized experiments and greater attention to statistical power issues (e.g., Lipsey, 1988).

The second NAS report (e.g., Lipsey, 1988; Martin, Sechrest & Redner, 1981), while acknowledging the poor evidentiary base, recommended new directions for rehabilitation in the 1980s. They directed program evaluation toward five important areas, four of which had yet to be sufficiently researched: family interventions, school-based programs, workplace approaches, and community strategies.[53]

[52] The NAS Panel stated that "The one positive conclusion is discouraging: the research methodology that has been brought to bear on the problem of finding ways to rehabilitate criminal offenders has been generally so inadequate that only a relatively few studies warrant any unequivocal interpretations" (Sechrest, White & Brown, 1979:3).

[53] Individual level interventions were the most commonly evaluated treatment strategies, and receive less emphasis in the second NAS report.

Lipsey (1988) noted the impact of the NAS reports on framing the treatment effectiveness debate. First, although the panels concurred with the early reviews that there was no evidence of treatment efficacy, neither could rehabilitation be readily dismissed. The reports stressed that the jury was still out with regard to treatment effectiveness due to the poor evidence accumulated (e.g., Sechrest, White & Brown, 1979).

Second, it appears as though the NAS reports inspired better program evaluation research. While evaluation evidence is still plagued with problems, Lipsey (1992a; 1988) and others note the higher quality of research when compared with the reports sampled by Martinson and other earlier reviewers (e.g., Basta and Davidson, 1988; Gendreau and Ross, 1987). The use of control or comparison groups is a more frequent practice, and advanced statistical analyses are now occasionally applied to treatment evaluation data (e.g., Lipsey, 1992a; 1988).

New Evidence and the Revival of the Rehabilitative Ideal

Several observers have noted a renewal of optimism in the rehabilitative ideal over the past 15 years (e.g., Logan and Gaes, 1993). Unlike the period which witnessed its decline, this revival is more the product of evidence than social events, ideological shifts or political undercurrent. Certainly, these other phenomena have played a role; for example, ideological arguments for rehabilitation have been well articulated in recent publications (e.g., Rotman, 1990; Cullen and Gilbert, 1982).[54]

Yet, "getting tough" remains politically smart and popular with the public, and the call for more punitive measures is still frequently raised (e.g., Gibbons, 1992). Despite this, the rehabilitative ideal no longer draws the same skeptical response, particularly from scholars, due to evidence being amassed on two fronts (e.g., Rhine, 1992).

First, there is strong evidence that the Justice Model—emphasizing retribution, deterrence and incapacitation—is a failed philosophy (e.g., Clear, 1994; Gibbons, 1992). While the number of persons under correctional supervision–especially prison–skyrocketed due to harsher and more certain sentencing, the increase has brought little relief in the war on crime or fear of victimization (e.g., Clear, 1994; Gibbons, 1992).[55]

In addition, the Justice Model appears to have had limited impact on reducing the inequity prevalent in the criminal justice system (e.g., Cullen and Gilbert, 1982).

[54] In fact, Rotman (1990) writes that "Rehabilitation offers a constructive way to improve the criminal justice system. Its concern for offenders as whole human beings enriches the state's reaction to crime with a higher notion of justice and leads to a better law."

[55] Some would argue, however, that recent well-publicized drops in crime are the result of tougher crime policies (e.g., Justice Research and Statistics Association, 1997).

For example, incarceration policies under the Justice Model have dramatically impacted the percentage of African-American males behind bars (e.g., Maurer, 1992). While there is some evidence that less disparity occurs between persons *sentenced* after being convicted of similar crimes (e.g., D'Alessio and Stolzenberg, 1995), there is also data to suggest that disparity has been largely displaced to other stages of the criminal justice process, particularly at the prosecutorial decision phase (e.g., Turpin-Petrosino, 1993; Cullen and Gilbert, 1982).[56]

Secondly, while evidence of the Justice Model's futility is amassing, there is some indication that rehabilitation programs work–at least some of the time. While negative reviews continue to appear (e.g., Lab and Whitehead, 1988), offender treatment reviews have become profoundly more optimistic since the first NAS panel report (e.g., Logan and Gaes, 1993). The more recent reviews of rehabilitative techniques have emphasized positive outcomes using rigorous designs in an effort to counter the prevailing "nothing works" mentality, and there is evidence that this strategy helped lift the prevailing pessimism amongst academicians and correctional practitioners (e.g., Palmer, 1992).

[56] One argument used by rehabilitation advocates was that the treatment model was never fully implemented in practice throughout the criminal justice system. This also applies to the Justice Model, which stressed some of the principles of just desert but–in some jurisdictions–left discretionary powers of the parole board, judge, prosecutor and so on intact, influencing equity (Turpin-Petrosino, 1993).

It is interesting to note that Martinson, even while his earlier work was being corroborated by the NAS panel, changed his view based on new research information. In a massive collection and synthesis of recidivism rates reported in 555 correctional studies, Martinson (1979:244) concluded that "...contrary to my previous position, some treatment programs do have an appreciable effect on recidivism."[57] Based on this data, Martinson (1979) urged policymakers to use caution in adopting the Justice Model sentencing reform measures which were occurring nationwide. It is perhaps unfortunate that this *Hofstra Law Review* article went largely unnoticed.

Also at the turn of the decade, Canadian researchers Paul Gendreau and Robert Ross, motivated by the popularity of the 'nothing works' position, countered this pessimism by publishing reviews of research which demonstrated that some programs were effective (e.g., Gendreau and Ross, 1987; Ross and Gendreau, 1980; Gendreau and Ross, 1979). Their reviews (one was entitled "*Bibliotherapy for Cynics*") focused on primary studies which employed rigorous research designs.

In their first review, they accumulated 95 treatment peer-reviewed, published evaluations from the years 1973-1978 (e.g., Gendreau and Ross, 1979). These evaluations had to employ some type of control, and must have had a six month

[57] Martinson (1979) provided one of the earliest precursors, though unrecognized at the time, of meta-analysis. He accumulated 555 research studies, created an effect size measure (the percentage difference in recidivism) for each outcome variable, and analyzed the variability in that effect size across several independent variables.

follow-up measure of crime in the community. Although there was some overlap with Greenberg's (1977) earlier review, they concluded that several intervention programs were successful with offender populations.[58] One of their observations in this early review was that multi-method approaches seem to be more successful than programs relying on one treatment modality (1979:485).[59]

Their later review used the same criteria, but focused on research published during 1981-1987 (e.g., Gendreau and Ross, 1987). Although no final number of studies they reviewed is provided, they similarly conclude that some treatment programs work across a variety of settings, with diverse types of offenders.[60]

They state (1987:395) that it "is downright ridiculous to say 'Nothing works'...The principles underlying effective rehabilitation generalize across far too many intervention strategies and offender samples to be dismissed as trivial." Their view that something works was offered in several other literature surveys during this time frame (e.g., Cullen and Gendreau, 1989; Basta and Davidson, 1988; van Voorhis, 1987; Palmer, 1983). However, a careful reading of both the earlier "pessimistic"

[58] Although Greenberg (1977) and Gendreau and Ross (1979) cover much similar ground, their conclusions are quite disparate, highlighting the need for a more replicable and overt method of synthesizing studies.

[59] Interesting that Yin (1986) also found that successful community crime prevention programs were those which relied on multimodal approaches rather than a single response.

[60] A hand count of studies mentioned in the review put the number over 100.

reviews and these later syntheses indicates a difference on emphasis, rather than substantive content.

Earlier reviews noted that successful outcomes were reported, but they were isolated results based on poorly done evaluations (e.g., Bailey, 1966, Logan, 1972; Martinson, 1974). These reviewers emphasized that the evidence for treatment efficacy was weak and spurious, a point reinforced by the NAS panel (e.g., Sechrest, White & Brown, 1979).

These later reviews, which have appeared in criminological literature consistently over the past 15 years, have generally stressed successful outcomes (e.g., Cullen and Gendreau, 1989; Basta and Davidson, 1988; Gendreau and Ross, 1987; Ross and Gendreau, 1980). One important result of these later reviews was to demonstrate that rehabilitation programs were far from being an historical oddity, having been lost in the rush to enact Justice Model sentencing. Indeed, Cullen and Gendreau (1989) noted that despite the entrenchment of the 'nothing works' mentality with politicians and scholars, over 200 rehabilitation program evaluations were reported in peer-reviewed publications alone from 1973-1987.

Despite the optimism generated by these reviews, however, some caveats are in order. Some of the reviewers, while noting improvement in research design, still emphasized the methodological problems in treatment evaluation research, including the failure to develop better control groups through randomization (e.g., Basta and

Davidson, 1988). More importantly, some recent reviewers have stated something quite similar to what their predecessors concluded: some programs work, at least some of the time, but there is still no consistent pattern of success which can be used as a basis for policy recommendations (e.g., Palmer, 1992).

In fact, about the only recommendation scholars can make--that more treatment program research is needed--had already been urged by the NAS panels (e.g., Sechrest, White & Brown, 1979). It is compelling that the recent University of Maryland Report similarly concluded, after reviewing over 500 crime prevention studies, with a call for more funding and research (Sherman, et al. 1997).

While the narrative reviews certainly contributed to the 'revivification of rehabilitation,' the findings from several meta-analyses have provided the strongest evidence used to promote the treatment ideal (e.g., Logan and Gaes, 1993; Lipsey, 1992a; Lipsey, 1992b; Logan, et al., 1991; Lipsey, 1988). Twenty-two prior statistical reviews of the 'what works' literature have now been reported, beginning with the research on juvenile interventions reported by Michigan State University researchers (e.g., Davidson, Gottschalk, Gensheimer & Mayer, 1984). It is interesting that the most negative quantitative review about the promise of rehabilitation (e.g., Whitehead and Lab, 1989) still found an overall effect for treatment rivaling the more optimistic meta-analyses (e.g., Lipsey, 1992a). Table 1 lists the treatment effectiveness meta-analyses reported to date.

Table 1.

Prior Meta-Analyses of Treatment Effectiveness In Criminal Justice (N=22)
Study Characteristics

Author(s) & Year	Type of Subjects	Years Covered	Literature	N of Studies	Designs Included	Common Metric Used	Outcome Measures Coded	Design Effect?	Type of Treatments
Davidson, et al. (1984)	Juveniles	1967-83	Pub/Unpub	90	Exp/Q-Exp*	Glass' d	All measures	Yes	All
Garrett (1984, 1985)	Juveniles	1960-83	Pub/Unpub	111	Exp/Q-Exp	Glass' d	All measures	No	Residential Treatment
Kaufman (1985)	Juveniles	Unk**	Pub/Unpub	20	Exp	Glass' d	Delinquency	N/A	Preventive Programs
Gensheimer, et al. (1986)	Juveniles	1967-83	Pub/Unpub	44	Exp/Q-Exp	Glass' d	All measures	Yes	Diversion
Mayer, et al. (1986)	Juveniles	1967-83	Pub/Unpub	39	Exp/Q-Exp	Glass' d	All measures	Yes	Social Learning Treatments
Gottschalk, et al. (1987)	Juveniles	1967-83	Pub/Unpub	90	Exp/Q-Exp	Glass' d	All measures	Yes	Community-Based Treatments
Gottschalk, et al. (1987)	Juveniles	1967-83	Pub/Unpub	25	Exp/Q-Exp	Glass' d	All measures	Unk	Behavioral Treatments
Losel & Koferl (1989)	Adults	1978-87	Pub/Unpub	16	Ezp/Q-Exp	Rm	Recidivism	No	Sociotherapeutic prison (Germany)
Whitehead & Lab (1989)	Juveniles	1975-84	Published[61]	50	Exp/Q-Exp	Phi	Recidivism	Yes	All
Andrews, et al (1990)	Adults/Juvs	1950-89	Published	79	Exp/Q-Exp	Phi	Recidivism	No	All
Izzo & Ross (1990)	Juveniles	1970-85	Published	46	Exp/Q-Exp	Glass' d	Recidivism	Unk	All
Roberts & Camasso (1991)	Juveniles	1980-90	Published	46	Exp/Q-Exp	Cohen's d	Recidivism	Yes	All
Lipsey (1992a)	Juveniles	1950-87	Pub/Unpub	443	Exp/Q-Exp	Cohen's d	All measures	Yes	All
Cox, et al. (1995)	Juveniles	1966-93	Pub/Unpub	57	Exp/Q-Exp	r	All measures	Yes	Alternative Schools
Hall (1995)	Adults	1988-94	Published	12	Exp/Q-Exp	r	Recidivism	No	Sex offender treatment
Losel (1995)	Adults	1978-87	Pub/Unpub	18	Exp/Q-Exp	r, phi	Recidivism	No	Sociotherapeutic prison (Germany)
Wells-Parker, et al. (1995)	Adults	1955-92	Pub/Unpub	215	Exp/Q-Exp	Cohen's d	All measures	Yes	DWI Interventions
Pearson, et al. (1995)	Adults/Juvs	1989-94	Unk	43	Exp/Q-Exp	Phi	Recidivism	No	All
Gendreau & Goggin (1996)	Adults/Juvs	1950-95	Unk	Unk	Exp/Q-Exp	Phi	Recidivism	Unk	All
Gendreau & Goggin (1996)	Adults/Juvs	1950-95	Unk	Unk	Exp/Q-Exp	Phi	Recidivism	Unk	"Punishing Smarter" Sanctions
Redondo, et al. (1996)	Adults/Juvs	1980-93	Pub/Unpub	49	Exp/Q-Exp	r	All measures	Yes	All (European treatment programs)
Pearson, et al. (1996)	Adults/Juvs	1968-94	Pub/Unpub	508	Exp/Q-Exp	varied	Recidivism/Drug	Yes	All

* Exp = Experimental, Q-Exp = Quasi-experimental

** Unk=Unknown

[61] Published refers to studies found in refereed journals.

New Technique for Reviewing Evidence: Meta-Analysis Enters the 'What Works?' Debate

While meta-analytic investigators have not always concluded that "treatment works," the overall results of these quantitative reviews have been that the average person receiving treatment--as opposed to the person receiving no treatment or handling as usual--performs about .25 standard deviations [SDs] better on subsequent outcome measures (e.g., Lipsey, 1992a). This statistic translates into an approximate 12% reduction in the recidivism rate, a figure which is modest but would be considered non-trivial (e.g., Rosenthal, 1991; Lipsey, 1988).

The first of these meta-analyses was reported by William S. Davidson and his colleagues at Michigan State University (e.g., Davidson, Gottschalk, Gensheimer & Mayer, 1984). Using a computerized search of *Psychological Abstracts* and a mail campaign with prominent research investigators, they were able to locate 91 juvenile treatment studies published or available between 1967-1983.[62] While the investigators took pains to cautiously present their results, they reported that treated subjects performed an average .35 SDs better on all outcomes (e.g., recidivism, attitudinal, etc.) than control subjects in 58 comparison group designs.[63]

[62] All of the meta-analyses focused on studies reported in English.

[63] Comparison group designs include randomized groups, matched groups, or other evaluation designs which allow for comparison between a treated group and untreated group.

For recidivism outcomes only, experimental subjects performed .32 SDs better than untreated controls, a finding which translates into a 16% reduction in recidivism rates (e.g., Lipsey, 1988). It is interesting to note that experimental subjects performed .75 SDs better than control group subjects in the methodologically inferior pre/post designs, causing some later meta-analysts to dismiss such designs from their samples (e.g., Lipsey, 1992a). The most promising interventions, when all research designs were included in the analysis, were academic and vocational rehabilitation programs (e.g., Lipsey, 1988).

Four subsequent meta-analyses have actually been subset investigations of this larger data set. Gensheimer, Gottschalk, Mayer & Davidson (1987) sampled 44 juvenile diversion programs, finding that such programs achieved an average .40 effect size on all outcomes relative to usual processing through the criminal justice system.[64] While effects were smaller when only recidivism outcomes were considered (.26 for comparison studies), it still supported an overall positive treatment effect. Interestingly, Gensheimer and her colleagues do not conclude that diversion programs are effective and cite the relatively poor quality of the studies and reports in the sample.

Mayer, Gottschalk, Gensheimer & Davidson (1986) also conducted a meta-analysis of 34 studies which examined the effects of social learning treatments with

[64] Interestingly enough, both comparison studies and pre/post studies achieved the same effect of .40.

juveniles.[65] They found large effect sizes for treated subjects; on average, experimental groups achieved effect sizes of .64 on all outcomes and .50 on recidivism when compared with non-treated controls in comparison studies. Again, Mayer and his Michigan State colleagues cautiously stressed the poor quality of the evidence and the insufficient reporting by original investigators.

Gottschalk, Gensheimer, Maher & Davidson (1987) examined community-based interventions with juvenile offenders, analyzing the results from 90 research designs.[66] Again, treated subjects achieved a .37 effect size on all outcome measures and .33 on recidivism, when examining comparison studies only.

Smaller effects were reported by Gottschalk & his colleagues (1987) in a meta-analysis of 25 research reports which examined behavioral treatment programs.[67] They found treated subjects achieved a .25 effect size on all outcome measures and a .13 effect size on recidivism when examining comparison group studies alone. It should be noted that only 14 behavioral treatment comparison studies included any data on recidivism.

[65] Actually, Mayer, et al., (1986) have 39 research designs in their sample, since some studies involved the comparison of two treated groups with a single control group.

[66] It is unknown how many studies or reports these 90 evaluation designs came from.

[67] Curiously, Gottschalk, et al. (1987) report on a meta-analysis of behavioral treatments which is very similar to the earlier study they reported on social learning program evaluations (Mayer, et al., 1986). How studies were chosen for each of these meta-analyses is not explained, but since the N of cases and results are different, it must be assumed that they represent different meta-analyses.

An interesting part of the Michigan State meta-analyses was the inclusion of a study rating scale. Judges were asked to rate each study–before the computation of effect size–on whether they thought the program had a positive, negative or null effect. In several of the meta-analyses, the conclusions from the qualitative ratings were different than the quantitative results, again highlighting the need for more rigorous reviewing techniques.

Independent of the Michigan State researchers, Garrett (1985) reported on her meta-analysis of juvenile offender residential treatment program evaluations, which were published or available between 1960-1983.[68] Using computerized searches of relevant data bases and the reference citations of located studies, Garrett was able to locate and retrieve 111 studies which used quasi-experimental or randomized designs.

Again, Garrett reported an overall positive effect for treatment programs. She found that treated juveniles performed .37 SDs better than control subjects on all outcome measures (e..g, psychological, behavioral, recidivism), and .13 SDs better on recidivism. Unfortunately, only 18% of her sample included follow-up measures of criminal behavior.

[68] Garrett's (1985) article is based on her 1984 dissertation at the University of Colorado, where one of her committee members was the originator of modern meta-analytic techniques, Dr. Gene V. Glass.

It is important to add that Garrett found a strong design effect, i.e., a dramatic difference in effect size between rigorous and less rigorous research designs. Randomized or matched controls had much smaller average effects–generally half or one-third in size–than the other designs (e.g., pre/post).[69]

Kaufman's (1985) analysis is the only prior treatment effectiveness meta-analyses to focus exclusively on randomized experimental designs. Supplementing computerized searches of the Educational Resources Information Center (ERIC) and Juvenile Justice Clearinghouse data bases with manual searches of relevant psychological, sociological and education indexes, Kaufman located 20 studies which tested some program with preadjudicated youths available through 1983.[70]

He found that experimental treatment subjects performed .20 SDs better than controls on subsequent measures of delinquency.[71] Kaufman found, like some of the other meta-analyses, that increased treatment exposure and intensity was related to effect

[69] Mayer, et al. (1986) also found that randomized studies reported smaller effect sizes than other designs, although the correlation (-.12) was not statistically significant. Gottschalk, et al. (1987a) found randomized designs had higher effect sizes than other research methodologies.

[70] Kaufman's (1985) report was a required paper for the Claremont Graduate School of Psychology program.

[71] Since studies often report more than one outcome, Kaufman also averaged outcomes within each study, producing a higher effect size (d=.25).

size; when treatment was increased to 2.1 contacts or more per week, the average effect size increased from d=.15 to d=.63.[72]

Losel and Koferl (1989) reported on a smaller meta-analysis of sociotherapeutic prison treatment effects in the Federal Republic of Germany through 1985. While the investigators were only able to gather 16 government evaluation reports, their unique sample allowed for more intensive follow-up research than broader meta-analyses allow. They conducted interviews with the original research evaluators, collected unpublished data not available in the reports, and gathered further information on treatment and prison context. They found that adult subjects exposed to the sociotherapeutic prison performed .22 SDs better on recidivism and personality outcomes than prisoners released from regular institutions.[73]

An interesting aspect to their study was the coding of 39 items related to five areas of research validity (internal, external, statistical, construct and descriptive). They found that the 16 research reports in their sample could not be easily categorized into good or poor studies; some evaluations which scored high on internal validity were problematic in other categories. Design rigor–the PIs concluded–depends on which type of validity

[72] Some earlier meta-analyses used Glass' d (Glass, et al. 1981) as the common metric effect size. This is expressed as **d=Mt-Mc/s** where **Mt** is the treatment group mean, **Mc** is the mean of the control group and s is the standard deviation of the control group.

[73] Losel and Koferl (1989) used Freidman's rm statistic, finding the overall effect of treatment being rm=.110 or d=.22. To compute rm, the equation $\sqrt{\frac{Q2}{Q2+S}}$ is used, where Q is the inferential statistic used in the study, and S=total sample size.

the reviewer wishes to emphasize. Losel (1995) updated this meta-analysis with two additional studies but did not alter these earlier findings.

Whitehead and Lab (1989) published a meta-analysis of juvenile offender treatment studies appearing in peer-reviewed journals during 1975-1985. The investigators used a manual search of indexes for the *National Criminal Justice Reference Service* (NCJRS), *Psychological Abstracts*, *Sociological Abstracts*, *Criminal Justice Abstracts* and *Abstracts on Criminology and Penology* to locate 50 evaluation reports. The studies must have had a control group (only 18 studies included randomization) and at least one outcome measure of recidivism.[74]

Whitehead and Lab computed 2 x 2 classification tables for each study and computed a phi correlation for the association between group membership (experimental or control) and the failure-success proportions.[75] Although Whitehead and Lab conclude

[74] Whitehead and Lab (1989) excluded alcohol and drug studies, as well as those focusing on punishment.

[75] A phi correlation is a measure of association that involves dividing the chi-square statistic by the sample size and taking the square root of the result. For example, phi for the following table would be .12 ($\sqrt{2.79 / 200}$):

	EXPERIMENTAL (N)	CONTROL (N)
% SUCCESS	72% (72)	61% (61)
% FAILURE	28% (28)	39% (39)

that the evidence for treatment efficacy is slight, their average phi for experimental treatment translates into d=.27, quite comparable to the other meta-analyses reported.[76]

Whitehead and Lab (1989) cited little support for behavioral interventions, and found non-system diversion to be the most effective "treatment" approach. Similar to prior reviews of the 'what works?' literature, they reported that randomized experimental studies had a higher proportion of negative findings, i.e., where the treatment group did significantly worse than the control subjects.

Inspired by the negative tone of the Whitehead and Lab (1989) analysis, Andrews and his colleagues (Andrews, Zinder, Hoge, Bonta, Gendreau & Cullen, 1990) computed phi correlations on 154 2 x 2 comparisons derived from 70 offender treatment evaluations. Their meta-analytic sample included 45 studies from Whitehead and Lab's analysis, plus a supplementary 35 studies which were "located in the investigators' files."

These additional 35 studies included some adult program evaluations reported between 1980-1989, but the sample was largely comprised of juvenile treatment studies. All of the studies in the Andrews, et al. sample had at least one outcome measure of recidivism.

[76] One of the major problems with the Whitehead and Lab study is that no information is provided on how they handled studies which provided more than one outcome measure of recidivism. Did they use an average phi correlation? Or did they compute separate analyses for each?

Andrews, et al., only extracted seven items of information from each study. The most crucial variable of interest was the type of treatment, which they categorized into four groups: (a) criminal sanctions; (b) inappropriate correctional services; (c) appropriate correctional services; and (d) unspecified correctional services. A treatment was categorized as appropriate if any of the following conditions were met (1990:379);

- included service delivery to high risk cases
- treatment was comprised of some behavioral program (unless subjects were low risk cases)
- comparisons reflected specific responsivity-treatment principles
- non-behavioral programs which clearly stated that criminogenic needs were targeted and that structured intervention was employed

Andrews and his colleagues found that appropriate correctional services had an average phi correlation of .30 (which is converted to d=.63); criminal sanctions and inappropriate correctional service had negative effects (i.e., the control groups did better). Unspecified correctional services had a smaller effect that was positive in direction (.10). The more positive finding for appropriate correctional services as opposed to the other three groups was statistically significant ($p < .05$).[77]

[77] If one combines the effects of interventions delivered from all four categories, the average d=.20, lower than Whitehead and Lab's (1989) average effect size. This supports the need for differentiating categories and effect sizes instead of reporting a single global comparison.

These findings were similar for adults and juveniles separately; however, only 15% of the classification tables were comprised of adult program comparisons (N=23). Similar to some other meta-analyses, Andrews and his co-investigators found a slightly smaller effect for rigorous designs (i.e., matched or randomized), larger effects for studies in the 1980s as opposed to the 1970s, and more positive findings for behavioral and community-based interventions.

While the other meta-analyses were more cautious, the optimistic conclusions of the Andrews, et al. (1990) analysis sparked a debate in the literature somewhat reminiscent of the response to the Martinson (1974) report. Though Whitehead and Lab (1990) took umbrage to Andrews, et al.'s attempts to claim that their research supported a very firm version of 'nothing works,' they did not back away from their negative conclusions about juvenile treatment programs.

They noted how Andrews, et al., (1990:409) provided little detail on how they defined their terms (e.g., "high risk cases" was not defined, a crucial part of their categorization of studies), sometimes used changing definitions to categorize cases, and ignored some recidivism outcome data in certain studies in favor of others without explanation. In conclusion, Whitehead and Lab (1990:414) argue that the Andrews, et al. research is tautological; prior evaluations are defined as being appropriate if they have a positive effect–therefore supporting the conclusion that appropriate correctional services are more effective than the other categories. They infer that the best method for testing assertions about appropriate services would be to adequately define such categories in advance and apply it to a *prospective* sample of correctional studies.

Logan and others have also sharply criticized the Andrews, et al. paper for comprising a circular and tautological argument–particularly since investigators non-blindly applied the "appropriate" label to studies whose effects are already known (e.g., Logan and Gaes, 1993; Logan, et al., 1991). Logan and Gaes (1993) go on to assert that the effectiveness of treatment–even if conclusively proven by meta-analysis, of which they are skeptical–does not matter, since the appropriate rationale for handling offenders is punishment. It appears, however, that their argument is limited to offenders who receive an adult prison term.[78] They state (1993:246) that "We still do not know what works in correctional treatment, but it really wouldn't matter even if we knew, because the fundamental purpose of imprisonment is not the correction but the punishment of criminal behavior." Despite the controversy, the National Institute of Correction's (NIC) Advisory Board voted to utilize the Andrews, et al. (1990) research to inform subsequent policy recommendations.[79]

Concurrent with the Andrews, et al. (1990) work, Izzo and Ross (1990) also published an optimistic meta-analysis of juvenile treatment evaluations which used an

[78] In fact, Gerald Gaes was the Director for the Federal Bureau of Prisons.

[79] In fact, this author was asked to present to the NIC Advisory Board in November, 1995, in an attempt to convince some of the more skeptical members that meta-analysis was not voodoo science.

experimental or quasi-experimental design. They used standard search techniques to uncover 46 reports published in a refereed journal during 1970-1985; these 46 studies yielded 68 effect sizes of recidivism outcomes which they analyzed.

While the investigators reported no average effect size–nor do they differentiate effect size for any specific category–they found that two coded items were statistically significant in a regression analysis with effect size (using Glass' d) as the dependent variable. First, effect sizes were higher if the treatment program had a cognitive component (r^2=.06, F=5.54, p=.02),[80] or if the treatment setting was in the community (r^2=.13, F=6.11, p=.00).

The investigators, much like Andrews, et al. (1990), do not support a blanket assertion that everything works–or that nothing works. Rather, they state (Izzo and Ross, 1990:141) that "Whether a program works depends on who does what to whom, why and where." Nonetheless, they clearly interpret their results as supporting a cognitive model of offender rehabilitation–programs which attend to *how the offender thinks*.

Roberts and Camasso (1991) also focused on juvenile treatment programs, locating 46 peer-reviewed evaluations published during 1980-1990 through a manual hand

[80] Programs were considered "cognitive" if they employed at least one of the following modalities (Izzo and Ross, 1990:139): problem-solving, negotiation skills training, interpersonal skills training, rational-emotive therapy, role-playing and modeling, or cognitive behavior modification.

search of relevant journals. Quasi-experimental and randomized designs were included; 35% of their studies did not include a separate comparison group. Like earlier quantitative reviews, Roberts and Camasso (1991:433) reported an average effect size of .35 in favor of experimental treatment programs, concluding that "...intervention with juvenile offenders typically has small, positive effects."[81]

Like some earlier meta-analyses, Roberts and Camasso did find that effect size was affected by methodological factors. For example, research studies with short follow-up periods or small sample sizes had higher effect sizes. Moreover, the size of the effect decreased with increases in statistical design rigor. The meta-analysts concluded by strongly recommending family therapy with juvenile offenders to practitioners, since its large effects on recidivism–an average of .55–held up in rigorously designed studies.[82]

One of the more extensive reviews–quantitative or qualitative–was conducted by Lipsey (1992a, 1992b) with support from the Sage Foundation and the National Institute of Mental Health. Lipsey used computerized searches of 23 electronic data bases and a variety of manual search methods to locate 443 quasi-experimental or randomized juvenile intervention studies which reported at least one quantifiable outcome measure of

[81] Roberts and Camasso (1991) used Cohen's *d*, which can be expressed as *mt-mc/s*, where *mt* is the mean of the treatment group, *mc* is the mean of the control group, and *s* is the pooled standard deviation of both the experimental and control groups.

[82] Roberts and Camasso noted that group counseling achieved an average effect size of .81, but observed that the effect was inflated by the poor designs used to evaluate the method.

delinquency. This exhaustive search strategy paid off; over 60% of his sample came from sources other than journals or academic books. A most impressive part of Lipsey's study is that pre/post studies or comparison studies without group equivalence pretests were excluded from his meta-analysis; he focused only on studies which employed at least a fairly rigorous quasi-experimental design.

Lipsey's data base ranges from 1945 through 1986, covering a wide range of interventions delivered to persons age 21 or younger. While prior meta-analyses focused on a few coding variables, ranging from seven (Andrews, et al., 1990) to nearly 50 (Losel and Koferl, 1989), Lipsey conducted an exhaustive analysis of 154 items. He found that treatment interventions, on average, have an effect size of .17, when considering the first effect reported.[83] The exclusion of less reliable quasi-experimental designs might be the reason for this lower average effect when compared to earlier meta-analyses, which sometimes report average effects twice as high.

Lipsey did find that randomized studies, with no appreciable attrition at follow-up, reported slightly higher effect sizes than all studies combined. Again, Lipsey noted that small sample size studies–even after applying Hedges and Olkin's (1985) correction for bias–still produce larger effects.[84] Moreover, smaller effects were associated with

[83] While Lipsey coded all effect sizes reported, his analysis in 1992 only considered the first effect size (i.e., the first delinquency outcome measure at the first follow-up period).

[84] The coefficient for adjusting effect size is created by using the formula 1-[3/(4nt + 4nc - 9)].

designs that had a large number of delinquency outcomes, long follow-up periods, and great attrition from the original sample.

Similar to earlier studies, Lipsey found larger effects for behavioral methods, skills training, and multi-method approaches regardless of whether the interventions were delivered in the criminal justice system or as part of non-justice system preventive services. While Lipsey's average effect was the smallest reported across the prior meta-analyses (including the more pessimistic Whitehead and Lab [1989] study), he concludes that (1992b:142-143):

> The meta-analysis results summarized here support a different interpretation of the body of research on delinquent treatment than that conveyed by traditional research reviews. Quantitative aggregation and statistical analysis of study effect sizes revealed that the overall means were positive and statistically significant. The meta-analysis work summarized in this paper confirms and extends the pattern of results found in prior meta-analyses. All show positive mean treatment effects, disproving the 'nothing works' interpretation of the literature.

In response to Furby, et al.'s (1989) narrative review of sex offender treatment, which concluded somewhat negatively about rehabilitative efforts, Hall (1995) conducted a meta-analysis of 12 evaluations published since 1988. He found a small, positive effect for treatment across the studies (d=.24). Diverging from other meta-analyses in criminal justice, Hall reported that effect sizes were larger in studies which had follow-up periods longer than five years. However, cognitive-behavioral and hormonal treatments were again found to be the most effective treatments, although the investigator urged caution due to the small sample of studies considered

Cox, et al. (1995) reported the results of a meta-analysis of 57 alternative school program evaluations. Studies were found through standard electronic searches (e.g., ERIC), and covered the years 1966-1993. Only nine of the evaluations included both a comparison group and some outcome measure of delinquency. For these most rigorous studies, the intervention demonstrated nearly no effect (d=.03). Again, pre/post designs had higher effect sizes (d=.23) than comparison group studies, reinforcing decisions by other meta-analysts to exclude them since they appear to inflate positive outcomes. Only 40% of the sample used random assignment, and only three studies included a follow-up of criminal behavior in the community. It is also interesting that Cox, et al. found higher effects on attitudinal measures in comparison designs in contrast with the much smaller effects on behavioral outcomes (e.g., delinquency, school performance, etc.).

In one of the most sophisticated meta-analyses conducted to date, Wells-Parker and her colleagues (1995) at Mississippi State University focused on remedial interventions with driving while intoxicated [DWI] offenders. Using multiple search strategies, researchers found 215 independent evaluations of DWI interventions published between 1955-1992. As part of their innovative strategy, they contacted experts in the area to develop a comprehensive coding instrument of 71 critical elements of program success, which they used to extract information from each evaluation report.

Similar to prior meta-analyses, Wells-Parker, et al. reported a small, positive effect (d=.19) for treatment across their sample. The investigators also found that more rigorous studies, as indicated by group equivalence, were associated with smaller effects. As recommended by Gendreau and Ross (1983-84), Wells-Parker, et al. also found that combinations of modalities—such as education, psychotherapy, counseling and follow-up/contact probation—were more effective than other methods for handling DWI offenders.

Gendreau and Goggin (1996) provided an update of the earlier Andrews, et al. (1990) work , with evaluations from 1990-1995 included in the analysis. Once again, with nearly twice the number of evaluation reports in the sample (and 215 effect size comparisons), studies which tested the effects of "appropriate correctional services" on criminal offending had an effect size nearly double the average es (.25 vs. .13 overall). In contrast, studies which tested the effects of "punishing smarter" (e.g., drug testing, electronic monitoring, fines, ISPs, Restitution, Scared Straight, Shock Incarceration) showed no effect on recidivism (phi=0).

As mentioned earlier, researchers at the National Development and Research Institute (NDRI), headed by Douglas Lipton, are conducting a comprehensive meta-analysis of correctional treatment evaluations available since 1968 (Lipton, 1995). Funded by the National Institute on Drug Abuse (NIDA), the Correctional Drug Abuse Treatment Effectiveness Project or CDATE is synthesizing reports—in any language and with varying degrees of methodological rigor–available through 1994.

In a preliminary analysis of 43 studies published between 1989-1994 (47 effect sizes included), Pearson, et al. (1995) could not replicate the Andrews, et al. (1990) finding for appropriate correctional service with this new sample of studies. They found a substantively smaller phi (.19) than Andrews, et al. (1990) found (.69) for 'better' services. Pearson, et al. speculated that this conflict may be the result of coding unreliability between the two meta-analytic studies.

In a later analysis of 508 published and unpublished reports—constituting less than 50% of the eligible documents found in their worldwide search for correctional treatment evaluations—Pearson and his colleagues (1996) found that two-thirds of their sample reported outcomes favoring treatment over control on crime or substance abuse outcomes. This was true of both adult and juvenile studies.

When examining effect size, Pearson, et al. (1996) reported small, positive effects for treatment over controls; however, the weighted effect for treatment programs with adults was considerably smaller than that reported for juveniles (d=.035 for adults; d=.125 for juveniles). They also found a design effect, i.e., that randomized designs had smaller effects than quasi-experimental ones.

As with Pearson et al.'s (1995) earlier analysis, they did not replicate the powerful effects that Andrews, et al. (1990) found for 'appropriate correctional services.' However, the NDRI researchers caution that an additional 700 reports wait to be coded and entered into the analysis.

Redondo, Garrido and Sanchez-Meca (1996) reported on the first European 'what works?' meta-analysis. Using three search techniques, they located 47 studies conducted in Europe during 1980-1993 which reported the effects of a treatment program on subsequent outcomes. Studies were retrieved from six countries (Britain, Spain, Germany, Netherlands, Sweden, Israel).

Redondo and his colleagues also reported a positive effect for treatment (d=.24) on recidivism. Again, consistent with some earlier meta-analyses, cognitive-behavioral treatments were the most effective. They also noted a design effect: randomized experimental designs had the lowest effect size (d=.13) when compared to other methods. Moreover, effect size was inversely correlated with age, as interventions with juveniles achieved greater reductions in crime than those with adults.

What Do These Prior Meta-Analyses Tell Us?

These meta-analyses conclusively demonstrate that treatment has a small, positive effect on recidivism measures, although making strong recommendations to policymakers about which programs to employ with which offenders remains problematic (e.g., Palmer, 1992). Certainly, the increased rigor of the meta-analytic approach renders these results more persuasive than the optimistic narrative reviews of the same period. These prior studies show that doing something to an offender is--on average--better than treatment as usual or doing nothing at all.

Certainly, the finding of small, positive effects is nearly uniform across each of the 22 prior meta-analyses. This poses great concern, as Glass (Mann, 1994) noted, since it indicates that these results may simply be the result of 'intervening in people's lives' rather than the substance of treatment. Such a finding was hinted at by Lipsey and Wilson (1993), who found small to moderate positive effects to be the norm across 302 treatment effectiveness meta-analyses in the social sciences. They also left open the possibility that there was something about the meta-analytic method itself which artificially produced such consistently positive results.

While these prior results consistently demonstrate positive effects, they are modest in nature. Some might argue that the findings are disappointing, since it is assumed that evaluations are generally conducted on the best criminal justice programs. Thus, small effects would indicate that the sum result of all treatment programming is not very effective. As some scholars have noted, the findings from meta-analysis are open to pessimistic or optimistic interpretations, depending on whether one wants to see the "glass as half-empty or half-full" (Whitehead and Lab, 1990; Andrews, et al., 1990).

While the evidence for a particular strategy with offenders is lacking, cognitive-behavioral based approaches appeared to be most effective across studies. This has, however, traditionally been a broadly defined category in meta-analysis, encompassing a number of programs and therapies. Support has also been found for multimodal rather than molar treatments (e.g., Lipsey, 1988), as well as community-based rather than institutionally based programs.

The most controversial finding in meta-analysis has been the very large effects for 'appropriate correctional services' in a single synthesis (Andrews, et al., 1990). Subsequent analyses have not replicated this finding, although the Gendreau and Goggin (1996) meta-analysis found larger effects for programs labeled appropriate. Yet, their average effect size for appropriate correctional service was nearly 74% smaller than that reported by Andrews, et al. (1990).

In only one study did researchers attempt to examine the effects of different philosophical categories on effect size. Gendreau and Goggin (1996) actually conducted two meta-analyses and found that treatment programs achieved considerably better effects on recidivism than the new 'punishing smarter' sanctions ushered in during the past decade.

Most of the other studies lump together prevention, rehabilitation and deterrence programs in global categories of "treatment" or "intervention." A few meta-analyses have tried to separate out the effects of "punishment" or "criminal sanctions," but these have been quite limited (e.g., Lipsey, 1992a; Andrews, et al., 1990). Certainly, the Pearson, et al. (1996) database contains the kind of detail where such comparisons can be made but those analyses have not been reported. Lipsey (1992a) is the only researcher who also collected prevention program evaluations along with sanctioning and treatment study reports.

Perhaps an unsettling finding from meta-analysis is that design factors, such as sample size or methodological rigor, influence effect size as much as substantive treatment categories (e.g., Lipsey, 1992a). For example, design rigor generally has some impact on magnitude of effect; in many of the prior meta-analyses, randomized designs report smaller effects than other designs. The nearly universal large effects for treatment produced by pre/post designs clearly highlight the need to exclude them from subsequent meta-analysis studies.

Moreover, some quantitative reviews have reported a higher percentage of negative results--where the treatment group does worse--when isolating the results of well-controlled studies, a finding concurrent with several narrative literature reviews. Lipsey's (1992a, 1992b) sample of fairly well-controlled studies reported the lowest average effect of all prior quantitative reviews, again suggesting that variability in effect size is at least partially related to design rigor.

Kaufman's study (1985) focused exclusively on randomized designs, but he was only able to procure a limited sample of 20 delinquency prevention experiments. However, Kaufman's sample did yield an average effect size quite comparable to meta-analyses which included quasi-experimental designs. However, it is unknown what effects a sample of quasi-experimental prevention studies using Kaufman's inclusion criteria would have obtained.

It should be noted that design rigor as a variable has been generally categorized as "rigorous/less rigorous," and has almost always been applied with regard to internal validity rigor. Experimental designs are nearly always classified as rigorous, although the problems alluded to in Chapter I can subvert the internal validity strength achieved by randomization. Only Losel and Koferl (1989), Lipsey (1992a), Pearson, et al. (1996) attempted to code more detailed information regarding design factors and their influence.

In addition, recidivism outcomes have been demonstrably more difficult to influence than other measures of programmatic effectiveness. In nearly every meta-analysis, the effect sizes for reoffending behavior are lower than those reported for attitudinal, psychological or other non-behavioral outcome measures. Moreover, effect size has been found to decrease with increases in sample size or follow-up periods (but see Hall, 1995 for an exception).

The problem of multiple outcomes (more than one measure or more than one follow-up) has been handled in a variety of ways in these meta-analyses. Most frustrating are those studies where multiple outcomes are not clearly discussed. However, where procedures have been mentioned, they generally focus on reporting all effects contained in the report (e.g., Andrews, et al., 1990; Izzo and Ross, 1990) and/or an average effect for each study (e.g., Garrett, 1985). Lipsey (1992a), due to the voluminous and preliminary nature of his investigation, simply focused on first effects.

It also appears that effect size is higher for juvenile interventions than adult treatments. This is an important policy finding, since it indicates that policymakers can achieve a 'bigger bang for the buck' by focusing on programs for children and adolescents. This seems to fit the common sense notion that children are more amenable than adults, at least with regard to reducing criminal offending.

Only the more recent meta-analyses were able to use the exponentially growing literature on research synthesis to utilize weighting and statistical techniques in handling data. Most now routinely apply small sample size bias correction formula, and weighted effect sizes (using the inverse-variance method) to take advantage of the more stable effects that larger samples have on outcome measures (e.g., Hedges and Olkin, 1985).

Focused Questions

No individual study, despite its comprehensiveness, can hope to answer every question (e.g., Dennis, 1988). However, based on the literature review discussed here and the results found in prior treatment effectiveness meta-analyses, five initial questions are raised that should be addressed by this study:

(1) *What kind of effect size do offender treatment programs achieve--on average--under the most rigorous internal validity conditions?* That is, if we exclude all studies except those using randomized experimental designs, how well does offender treatment perform? Does it provide further support for the revival of the rehabilitative ideal?

(2) *How does this average effect size for offender treatment programs compare to prior meta-analyses of treatment effectiveness?* Using only the most rigorous evidence, how well does the average effect for treatment line up with prior results?.

(3) *How does this average effect size for offender treatment compare to special deterrence interventions and special prevention programs?* These provide philosophical groups to compare the effects reported in offender treatment evaluations.

(4) *Does the consistent finding that small sample size studies achieve larger effects hold in this sample?* Using the latest statistical correction and weighting formulas, do we find that smaller samples still achieve larger effects?

(5) *Do juvenile programs achieve higher effects than those for adults?* This would be an important policy finding, since it would reinforce the notion that resources should be directed toward the most amenable persons early in their development.

CHAPTER III: INCLUSION CRITERIA: THE SAMPLE 'IN/OUT' DECISION

One of the most important stages of a meta-analysis is specifying the inclusion criteria. In other words, which research studies will be included in or excluded from the quantitative review? Once deciding on the criteria, how well do they work in practice? What rules are adopted to handle troublesome studies, the ones that are neither readily included or excluded? After discussing the importance of specification, this chapter lists the eight criteria for including studies in the study, discusses problematic studies which were confronted using the criteria, and details the rules used to maintain research consistency.

The Importance of Specifying Inclusion Criteria

Abrami, Cohen and d'Apollonia (1988:155) list five steps in conducting a meta-analysis: (a) specifying the inclusion criteria, (b) locating studies, (c) coding study features, (d) calculating individual outcomes, and (e) data analysis. Wanous, Sullivan and Malinak (1989) call "defining the domain of research" and "establishing the criteria for including studies in the review" as two of the most crucial decisions made in a meta-analysis. Cooper (1989) stresses that the study domain must be adequately defined for narrative reviews as well as for meta-analyses. The general inference from these writings and others is that unsound practice at this early stage can result in a misleading and invalid quantitative or qualitative review.

Wanous and his colleagues at Ohio State (1989) demonstrated that many of the discrepancies between meta-analyses bearing on the same issue are due to different, non-quantitative decisions made by researchers at various stages, including judgment calls about which studies to include or exclude from the sample.

Abrami, Cohen and d'Apollonia (1988) examined, among other things, the different inclusion criteria used by several meta-analyses of the student evaluation validity research; they showed how the selection of inclusion criteria affected the magnitude of the effect found across individual meta-analyses. Their point was clear: if reviews examining the same issue are analyzing different studies due to dissimilar inclusion criteria, that fact alone could account for the differences in effect magnitude across reviews.

In a different light, Bryant and Wortman (1984) presented in detail the process by which they excluded over seventy percent of the original studies they located on the effect of desegregation on the educational achievement of black students bused to white schools. By making criteria and rules explicit, they account for why only 31 studies (or 26%) from an original sample of 118 research reports were analyzed.

The Bryant and Wortman (1984) work is important on a number of fronts. They encourage the replication of meta-analysis and illustrate their point by noting the frequent reanalyses of the seminal meta-analysis of psychotherapy effects reported by Smith and Glass (1977). They even encourage the creation of a depository where original studies used in the meta-analysis can be kept for other researchers to duplicate and use. Replication of any review, however, is only possible if inclusion criteria are explicitly defined so they can be followed.

Another important aspect of inclusion criteria is the information it provides the reader on the scope of the project. For example, the "what works?" debate in criminal justice is a continuous and often volatile argument among scholars and practitioners. Often, the conflict is ideological in nature (e.g., Cullen and Gilbert, 1982; Clear, 1978), but it is also about evidence and the way that evidence is reviewed (e.g., Lipsey, 1988). By providing the inclusion criteria for this study, different investigators interested in what works can assess how important this meta-analysis is to that debate.

Study Inclusion Criteria

The criteria for study inclusion were first developed by Weisburd, Sherman and Petrosino (1990) and were modified for this collection process.[85] Study reports were collected using the following eight criteria;

[85] The original inclusion criteria used by Weisburd, Sherman and Petrosino (1990) were: individuals as units of analysis; random assignment to conditions; coercive sanction as delivered by an agent of the criminal justice system; included one outcome measure of crime [including official records or self-report data]; and a minimum of 15 subjects in at least two of the experimental conditions.

(1) if the report contained a discernible statement of random assignment to experimental and control groups

(2) if the random assignment process was conducted under the auspices of the research investigator(s)

(3) if individuals were the units of analysis in the experiment

(4) if the experimental report included at least one official and quantifiable outcome measure of crime in the community

(5) if the experimental report was published, printed or otherwise available between 1950-1993, inclusive

(6) without regard to type of publication or manuscript

(7) without regard to type of administering agency

(8) if the report was available in English

These research criteria distinguish this study from nearly all prior meta-analyses of intervention efficacy in criminal justice. Andrews and his colleagues (1990) were the only other meta-analysts to include both adult and juvenile program evaluations in their sample of correctional treatment studies; most other prior meta-analyses included juvenile interventions only (e.g., Lipsey, 1992a; Whitehead and Lab, 1988; Garrett, 1985). This project posed no inclusion criteria based on the age of the subjects.

In addition, only Kaufman (1985) excluded quasi-experimental designs in his meta-analysis of delinquency prevention experiments. All of the other works included quasi-experiments, generally meaning that the research design include some type of control or comparison group, although Lipsey (1992a) excluded the less rigorous quasi-experimental designs from his sample (e.g., simple posttests, etc.). This study excluded all designs except those which employed random assignment, which is discussed next.

A Discernible Statement of Random Assignment to Conditions

While the word "experiment" has often been used to describe any planned social intervention (e.g., Farrington, 1983; Twain, 1983), this study only included experimental reports with a clear statement of random assignment to conditions.[86] This inclusion criteria was easily applied in many instances; investigators used the terminology expected (e.g., "subjects were randomly assigned to experimental and control groups") in a frequent number of experimental studies.

Quasi-experiments were also identified and excluded for the most part. For example, Hamm and Kite (1991) reported the results of an evaluation of a treatment program for spouse abusers. They wrote (1992:232) that since "it became impossible to assign men randomly to experimental conditions," they developed two different designs to test program impact.

While most evaluation reports were easily categorized studies (either "black or white"), problems with categorization occurred when studies did not clearly

[86] This was particularly a problem during the search process, where a large number of "hits" for the term experiment turned out to be project or policy descriptions where no evaluation had taken place.

explicate the type of assignment process used in creating study groups ("gray area"). Terms like experiment, control group, comparable groups, and the like were found in a great many quasi-experimental designs.

To illustrate a gray area evaluation, Adams and Vetter (1971) reported on the results of a Maryland study which tested the impact of reduced probation caseloads on the recidivism rate of probationers. They state (1971:391) that "cases were assigned to officers in both units prior to the original court hearing by a court director so that various types of cases were equally distributed between the high and low caseload officers." There was no other information in the abbreviated research note about the assignment process. There were numerous reports which met the other inclusion criteria but contained insufficient information on the assignment process to determine if it was truly a randomized design (e.g. Vigdal, et al., 1980; Lerner, 1953).

It was sometimes the case that an additional report by the investigator(s) on the same study would provide the information needed. Shore and Massimo (1979; 1966; Massimo and Shore, 1963) published extensively on the effects of vocational counseling on ten boys as compared to a control group of ten boys who did not receive the treatment; it was not until a 1966 report was located that a clear statement of random assignment to experimental and control groups was found. In other cases, particularly for more recent evaluations, some investigators were contacted directly for further information.

Some studies were gray area experiments because the type of assignment process they used is not considered by all scholars as comprising a truly random assignment method. For example, Ross and Blumenthal (1974) had judges in Denver sentence all cases each week to one of three sanctions: fine, probation, or probation plus driver education. The judges were instructed to alternate their sentences each week.

Not only did the judges violate the alternate sequence assignment procedure (which is discussed later) but assigning subjects to experimental groups on alternate days or weeks may not be considered a truly random design, particularly if there is some bias in the way cases enter the court (e.g., Farrington, 1983). Others have included these alternate day/week designs in their samples of randomized field experiments (e.g., Weisburd, Sherman and Petrosino, 1990; Dennis, 1988; Boruch, McSweeney and Soderstrom, 1978).

What about cases, like the aforementioned Denver Drunk Driving Experiment (Ross and Blumenthal, 1974), where randomization has broken down? As Dunford (1990) implied, does the violation of the randomization process turn the design into a quasi-experimental one? One of the fortunate developments in the literature has been the increased attention given by experimenters to randomization, particularly when practitioners override random assignment and selectively place a subject in a particular group. When the violation rate is high, however, as is the case in many criminal justice experiments where subjects are receiving drastically different conditions (e.g., probation instead of prison)–should that study be included?

There were also studies which were designed in the proposal stage as randomized experiments, but before the project started, the design was modified out of necessity to a quasi-experimental one. For example, Kobrin and Klein (1983) reported that the nationwide Deinstitutionalization of Status Offender [DSO] experiments were originally designed as random assignment studies but in nearly every case, random assignment was never implemented in the study jurisdiction. Should these studies be included?

Rules for handling the random assignment criteria were developed following the examination of several of these and other evaluation reports. They were:

(1) There must be a discernible statement of random somewhere in the evaluation report(s).

(2) Studies which use assignment methods based on alternate days or weeks would be included since they essentially remove the kind of selection bias which hinders evaluation research (e.g., Farrington, 1983; Campbell and Stanley, 1966). According to Glass, McGaw and Smith (1981), study features like the assignment process can be coded and examined in a meta-analysis for their impact on implementation and outcome.

(3) Studies which begin with a randomized design were included despite the override or violation rate. Again, the dominant theme is that violations can be codified and examined in meta-analytic research.

(4) Proposed field experiments which were never implemented at any time during the study were not included (e.g., most of the DSO studies reported by Kobrin and Klein, 1983).

There were some other implications to the research by focusing solely on randomized designs. It is obvious that a meta-analysis which excludes quasi-experimental evaluations eliminates the majority of evaluation studies from consideration. For example, Whitehead and Lab (1989) found less than 40% of 50 studies in their sample of juvenile correctional treatments used random assignment (n=18).

Randomization Conducted under Auspices of Researcher(s)

The classic or true experiment implies that random assignment and the delivery of the intervention to the experimental group is under the control of the researcher. This is true for many medical and pharmaceutical experiments, and laboratory studies conducted in settings with college students, but it is rarely the case in criminal justice field settings.

In most cases, random assignment will be conducted by practitioners and the treatment delivered by agents of the criminal justice system or social agency. The problems of conducting true experiments in real world settings should not be underestimated; sound implementation of an experiment can be affected by practitioner control of randomization, particularly when there is vested interest in seeing subjects assigned to certain conditions (e.g., Glaser, 1995; Dennis, 1988).

While the inclusion criteria did not mean that the researchers actually had to perform the randomization (although in some cases they did), they had at least be aware of the process while it was going on, even if they did not monitor it in any fashion.

Several important gray area experiments were confronted using this criteria. Martin, Annan and Forst (1986) reported on the deterrent effect of jail time for drunk drivers by conducting a retrospective study of judicial sentencing patterns. Two judges were found who consistently gave persons convicted of driving while intoxicated [DWI] different sanctions. One judge sent the majority of his cases to jail (75%), while the second judge was more lenient (25% were jailed). Cases in Hennepin County, Minnesota were already randomly assigned to judges. Researchers ingeniously went back and got a representative sample of defendants sentenced in that court and checked to see if any were subsequently arrested. They found no difference in recidivism between defendants sentenced by the two different types of judges. Should this study be included?

Fagan (1990) and Zeisel (1973) have detailed how 'natural' and 'indirect' experiments often have the unplanned effect of removing selection bias, permitting better comparison groups. However, while recognizing the strength of these designs, the following rule was used:

(1) The randomization process must be under the control of the researcher or accomplished with researcher awareness. The randomization of

subjects to the study groups could not be haphazard, accidental or retrospective in nature.

While there is much to learn from indirect and natural experiments, the focus of this study remained classic experimental designs. Including these studies in a later meta-analysis could make for an enlightening comparison of natural and indirect designs versus randomized ones to determine if effect size fluctuates across these different studies.

Individuals as the Unit of Analysis

The meta-analytic sample was limited to experiments which involved individuals as the units of analysis. Therefore, experiments that involved random assignment of larger, aggregate units such as patrol beats (e.g., Police Foundation, 1981; Kelling, et al., 1974) or hot spots (e.g., Weisburd, Maher and Sherman, 1992) to determine the effect of police patrol deployment were not included.

Outcomes in experiments like these are usually reported for the larger aggregate units and not for the individuals living in those units (e.g., Barrow, 1978). For example, in the Minneapolis Hot Spots Experiment (e.g., Sherman and Weisburd, 1995), geographical clusters of addresses which generate large numbers of police calls were randomly assigned to two different police patrol intensity levels. Outcomes were reported for calls generated to the police from the 'hotspot' locations during the experimental period. Experiments like these are clearly more suited to answer questions about general deterrence than the more specific or individual level effects of intervention which this meta-analysis is focused upon.

Yet, what about interventions which involved the assignment of families of delinquent subjects? These experiments generally included an outcome measure for individuals exposed to the therapy (e.g., the delinquents in those families). For example, Alexander and Parsons (1973) report on a successful family therapy program which showed demonstrable impact on outcome measures of delinquency. In the study, families referred to the project by the juvenile court were assigned to the experimental conditions. This was true for other delinquent counseling interventions, which often involved some type of family level therapy. These experiments were allowed through the following rule:

> (1) Only experiments which assigned subjects to experimental conditions were permitted. Subjects within families were included if the intervention is targeted toward reduction of crime for individual subject(s) within those families.

One Official and Quantifiable Outcome Measure of Crime in the Community

The randomized experiment must have included at least one official and quantifiable outcome measure of crime, including but not limited to rearrest, reconviction, revocation of parole or probation, and reincarceration. Experiments that solely evaluated the effects of an intervention on 'paper and pencil tests' (e.g., attitudinal and educational tests) were not collected. While some prior meta-analytic works have included these measures (e.g., Lipsey, 1992a; Garrett, 1985), others have directed their efforts only toward outcome measures of criminal behavior (e.g., Andrews, et al., 1990; Whitehead and Lab, 1989).

Obviously, a study focused on crime reduction must collect studies with some outcome measure of offending behavior (e.g., arrests, citations, revocations, etc.). From a practical research basis, excluding these other outcome measures obviously reduces the amount of coding, analysis and number of studies that are reviewed. No meta-analytic review, no matter how comprehensive, can consider every issue.

There is additional rationale for excluding the other information. First, some integrative reviewers have noted that the relationship between psychometric or attitudinal measures and subsequent offending is often unclear (e.g., Furby, Weinrott and Blackshaw, 1989). In addition, scholars have consistently noted the weak relationship between institutional measures of success and success in the community (e.g., Morris, 1974). Perhaps most pertinent, crime control remains the ultimate concern of practitioners and the public (e.g., Wilson, 1975). In fact, Palmer (1992:25) writes that;

> Despite its complexities and the differing ways it is measured, this index [recidivism] is widely accepted by researchers, practitioners, policymakers, and the public itself, and is usually considered a key element in any outcome evaluation...Without this index, program evaluation would not just be incomplete, it would miss the main point.

Following Palmer's (1992) point, experiments which tested the impact of intervention on institutional measures of misbehavior were excluded from consideration. Although prison or other residential infractions range from minor violations to serious assaults, it was decided to remain focused on measures of crime in the community and leave institutional behavior to a later meta-analysis.

Therefore, every randomized experiment included in this meta-analysis examined the effect of an intervention on at least one official outcome measure of criminal behavior in the community. While the outcome measure of crime did not have to be the main focus of the randomized test–for example, the experimenter may have been more interested in social skill improvement–it was a necessary and sufficient criteria for inclusion into this sample.

Of course, these crime outcomes had to be quantifiable, in that enough statistical information was present to compute an effect size, which is discussed in more detail later. Studies which report "no significant difference between the groups was found" with no further information were not included. Generally, in all but a few cases, information to complete the computation of effect size was obtained from the reports or from follow-up contact with investigators or other reviewers.

The gray area studies which were confronted here had more to do with the type of outcome measure. Crime in the community denotes a legal, official reaction to behavior by an individual. By using the "official outcome of crime in the community" criterion, the study was essentially focused upon criminal justice system reactions to behavior by an individual. But what about self-reported acts? What about anti-social behavior that would be criminal if police were called (and decided to take action), but is only observed by teachers or other observers and thereby escapes the eye of police?

This was a difficult call, particularly since many experiments were conducted in the substance abuse prevention and addiction treatment area, where self-reported illicit drug use remains a major source for outcome data used in experimental evaluations. For example, Freidman (1989) reports on an experiment which evaluated the effect of two family therapy counseling treatments, finding both groups reported significant reductions in self-reported drug use. The rule for handling these gray area cases was as follows:

> (1) There must have been at least one *official* outcome measure of crime in the community reported. Experiments relying solely on observations of anti-social behavior, or on self-reported drug abuse or crime were collected when located but not included in this meta-analysis.

This rule still allowed me to consider some experiments which evaluated drug and alcohol interventions, or juvenile anti-social behavior interventions, since they occasionally reported on an official outcome like police arrests or juvenile petitions. For example, Dole, et al's (1967) study of methadone maintenance was included since it involved a six month follow-up of police arrests. Again, another meta-analysis can include studies which used alternative outcome measures to assess intervention efficacy.

Available Between January 1, 1950 and December 31, 1993

One of the most important things the meta-analyst can do with regard to inclusion criteria is establish the time frame for the study domain. Only experimental reports published, presented or circulated in some manner between January 1, 1950 and December 31, 1993 were collected.

This is valuable from a practical standpoint, since it provided a measure of consistency for search efforts. Setting the earlier date allowed for search efforts to have a definitive starting point; manual checks of the social science indexes (e.g., Psychological Abstracts) were started in the 1950 volumes. Having a set end date for experiments assisted the search as well. For example, once a thorough manual search was made of leading criminal justice journal volumes through 1993, there was no need to return to the stacks when the new volumes in 1994 were published.

Restricting the sample by time frame can reduce the number of studies included, and also increase the potential relevance to policymakers, who are probably going to be more interested in recent programs and effects. For example, Antonowicz and Ross (1993) analyzed 44 offender rehabilitation studies published in journals or edited books from 1970-1991 in their review. While the present study includes older experiments (back to 1950), it remains concentrated in a period of time Lipsey (1992a) referred to as the "relatively modern era."

One problem which must be noted with this inclusion criteria is that there is a considerable time lag between project start and dissemination of findings through publication or informal circulation. In fact, in the earlier *Registry of Randomized Experiments in Criminal Sanctions, 1950-1983*, (Weisburd, Sherman & Petrosino, 1990), the investigators listed the chronological data when the experiment started rather than the publication date, since it was not uncommon to find published reports a decade after the experiment started. By using the December 1993 cut-off, many of the field experiments completed in the last several years which have not produced an available report were unfortunately excluded.

Consistency can be difficult to maintain. For example, an experiment in pretrial release conducted in Costa Rica (Carranza, Houved and Mora, 1994) was not available until the edited volume it appeared in was published. Despite much agonizing over this case, particularly since it represented the first Latin American experiment in an international search, the rule was followed with the following modification;

(1) The experimental report(s) must be available before January 1, 1994. Nonetheless, an unpublished manuscript, a conference paper, or other document can be obtained from an investigator, as long as it was available to others before the December 31, 1993 cut-off date.

This allowed—in rare cases—contact with principal investigators in an attempt to retrieve unpublished reports of experimental studies, which otherwise would have been excluded.

Without Regard to Type of Publication or Manuscript

Smith (1980) has noted that research studies published in academic journals consistently demonstrate a higher program effect than unpublished research found in government reports, dissertations, conference papers and masters theses.[87] To offset this possible sample selection bias, experiments reported in the "fugitive literature" (e.g., Sechrest, White and Brown, 1979) were collected simultaneously with those contained in academic journals and books.

This provided more studies for the sample. For example, Lipsey (1992a) found that only 38% of his total sample came from journal articles and book chapters. We found that some of the *Registry* experiments were located in government reports issued by the California Youth Authority or the British Home Office (e.g., Weisburd, Sherman and Petrosino, 1990). To focus, as other meta-analysts have done, on published articles would have restricted the number of experiments considered and the potential scope of the findings.

[87] There is some debate over this issue. Some contend that the difference is due to the higher quality of the research reported in the academic journals. Others argue that the question of 'quality' is an empirical one that needs to be examined before making ad-hoc decisions about sample exclusion (e.g., Cook, et al., 1992; Wachter and Straf, 1990) .

Establishing this inclusion criteria did not come without a price. Efforts to track down government reports, conference papers, unpublished manuscripts in academic file drawers, master's theses and old dissertations, take up a considerable amount of time and effort, and become costly when phone calls and mailings are included. The number of experiments, however, outside the scope of peer review journals was considered too large to exclude them.

Without Regard to Administering Agency

As mentioned earlier, this research embraced a larger set of experiments than that examined by Weisburd, Sherman and Petrosino (1990). Weisburd, et al (1990) were only interested in gathering randomized experiments which were delivered in coercive fashion by agents of the criminal justice system. By coercive, the intervention either could not be refused (the subject was arrested regardless of his or her desires), or if the experiment did permit refusals (such as Lamb and Goertzel's [1974] Ellsworth House experiment), the subject who refused the treatment condition received a harsher sanction (in the Ellsworth House, subjects assigned to the residential home could refuse the assignment and return to jail).

While most of the experiments collected here were delivered under the auspices of the criminal justice system, many experiments in this sample lacked a coercive element. Some experiments were controlled studies of the impact of social work or counseling interventions delivered by community organizations, universities or other non-coercive agencies (e.g., Powers and Witmer, 1951). Subjects could refuse to take part, and there was no penalty for not participating.

Some gray area experiments were confronted here. Green (1985) reported on a randomized experiment which tested the effect of a threatening letter from a cable company on persons illegally receiving cable services. A follow-up of the effect of the letter on illegal transmission was reported. However, although a few experiments like these were found, I adopted the following rule;

> (1) The experimental manipulation (independent variable in the experiment) must have tested the effects of a tangible program or policy on criminal behavior. The effect of warning letters, moral pleas or reminders about civic duty (e.g., Schwartz and Orleans, 1967) were excluded.

Again, the experimental effects of warning letters and the like are important and need to be comprehensively analyzed. In fact, they may be effective enough in certain areas of minor violation (e.g., parking tickets, overdue books, tax reporting, etc.) to be recommended as a standard policy or program. Nevertheless, while retrieved when found, these studies were excluded them from the present analysis.

Available in English

Similar to prior treatment effectiveness meta-analyses in criminal justice, the experimental evaluation reports had to be originally written in or translated into English. This restricted the experiments to settings in North America and

England.[88] This exclusion may have provided, as Lipsey (1992a) argued, some control over cultural bias, i.e., that the definitions of crime and the workings of the criminal justice system are similar across the United States, Canada and England.

The extent of experimentation in criminal justice in other countries is unknown, although Schumann (1994) was working on a chapter for a German publisher on the topic.[89] As mentioned earlier, an experiment on pretrial release was reported in Costa Rica (Carranza, Houved and Mora, 1994), and it is possible that randomized studies have been reported in non-English journals. Again, future meta-analyses could make use of the growing electronic data bases, which are including more international citations, to recover reports, provided that translation was possible.

[88] Weisburd, Sherman and Petrosino (1990) found that 75 of the 76 experiments in their sample were conducted in the United States, Canada or England; one was conducted in Denmark.

[89] However, Redondo, et al. (1996) and Losel (1995) found few randomized studies in their European meta-analyses. Two sets of independent researchers in Sweden and Germany are currently collecting rehabilitation studies for future meta-analyses as well (Pearson, et al. 1996).

CHAPTER IV. THE HUNT FOR RANDOMIZED EXPERIMENTAL REPORTS: DOCUMENT SEARCH AND RETRIEVAL [90]

One area which needs more attention is the way in which the research reports, or documents, are located and retrieved. In effect, *these efforts represent the data collection phase for the meta-analysis.* How does the meta-analyst decide where to search? What techniques should be utilized? What kinds of problems are encountered when retrieving documents? Are there solutions to the problems of gathering this data? This chapter describes the search and retrieval process for meta-analysis.

Obviously, there are editorial and other constraints when reporting a meta-analytic study. Prior meta-analyses generally contain little information about the type of searching that was done (e.g., "relevant publications for the years 1970-1990 were searched"). Many lack a thorough description of the search and retrieval process, the rationale behind the search techniques used, and the problems confronted in document retrieval (but see Wells-Parker, et al., 1995 and Lipsey, 1992a for important exceptions).

Some scholars have addressed some of the issues within the context of larger discussions (e.g., Cook, et al., 1992; Wachter and Straf, 1991; Hunter and Schmidt, 1990; Cooper, 1989; Light and Pillemer, 1984). There is a need for more detail of this

[90] Many thanks to research assistants Julie Schnobrich of Westfield State College and Michael Gordon of Northeastern University with their help in locating and retrieving documents.

important phase when conducting a meta-analysis (e.g., Hunter and Schmidt, 1990; Glass, McGaw and Smith, 1981). To provide such detail, the search and retrieval strategies used in this meta-analysis are presented and discussed.

The Importance of Search and Retrieval Techniques

One of the most important goals of the primary researcher–the investigator conducting an original piece of research–is to develop a complete, random or representative set of observations (e.g., Hagan, 1993). This is no less true of the reviewer, who also must insure that a representative, if not complete, set of primary research reports is obtained before making generalizations (e.g., Cooper, 1989; Light and Pillemer, 1984).

There is no doubt that meta-analysis has brought attention to systematizing the review process (e.g., Glass, McGaw and Smith, 1981; Glass, 1976), forcing researchers who synthesize information to detail their inclusion criteria and the process of searching, locating and retrieving primary research reports which fit the criteria. This applies with equal force to narrative or qualitative reviewers; the entire validity and reliability of the review rests on how well these initial stages are carried out (e.g., Cooper, 1989). As with primary research, it would be difficult to trust a final report when data have been collected with substantial biases and are obviously inadequate.

The search and retrieval process represents the data collection phase for the meta-analyst. Many techniques now exist which broaden the scope of search efforts and quicken the process by which they are conducted. To illustrate, Lipton and his colleagues at NDRI are using the Internet to communicate an international request for correctional treatment studies (O'Kane, 1995). Even with advances in information technology, however, there will always be a need for hand searches and manual checks; this is especially true when conducting a comprehensive meta-analysis of offender treatment studies, since many evaluations exist outside the scope of electronic search capabilities (internal government evaluation reports, unpublished masters theses, etc.).

The Goal in Collecting the Sample

This project had as its goal the acquisition of the complete set of experiments reported in the English world, provided they met the eight criteria detailed in Chapter III. Certainly, these studies do not represent a randomly selected sample from some larger population (e.g., Cooper, 1989). Unfortunately, given the amount of fugitive literature, an unknown universe of experimental studies exists. Despite all of the techniques described in this chapter, many randomized studies will remain in researcher files, inaccessible to various search strategies.

Beginning the Sample

This research began with the initial list of randomized field experiments in criminal sanctions identified in *The Registry of Experiments in Criminal Sanctions, 1950-*

1983 (e.g., Weisburd, Sherman & Petrosino, 1990), since all 76 studies met the broader inclusion criteria used here. While two of the earlier experimental reports contained insufficient data on the outcome measure to be included in earlier analyses (Weisburd, 1993), supplemental reports obtained from the "fugitive literature" provided the quantified outcome measures needed to be included in this meta-analysis (e.g., Blumenthal and Ross, 1973; Kirby, 1970).[91]

The Search Methods

While some prior meta-analyses have relied solely upon one search technique to locate research reports, this is considered an unwise practice (e.g., Hunter and Schmidt, 1990; Cooper, 1989). Even if the search is limited to academic journals over a recent time period, using the electronic data base searches alone will miss considerable literature unless keywords are precisely specified. It has been recommended elsewhere that investigators supplement the computerized technology with some other search method to insure that eligible studies are retrieved (e.g., Durlak and Lipsey, 1991; Cooper, 1989).

[91] There is considerable debate about how to proceed with studies that include no quantified outcome data. Some argue that if the investigator reports "no significant result was obtained," then the meta-analyst should simply set the effect size at zero. Others argue against this approach, since studies rarely have zero effects; even if effects are small, the intervention will have some positive or negative impact on the outcome variable. Given this information, there are some who argue for using the mean or median effect size for the category or simply excluding the study from consideration.

For a comprehensive "what works?" meta-analysis, using multiple channels for retrieving empirical studies is the only way to collect a representative sample of studies (e.g., Glass, McGaw and Smith, 1981). This is especially true, as stated earlier, when the investigator is expanding the sample to include fugitive reports.

There may be value in some instances of relying on academic journal reports, but it would be hard to justify such a decision when conducting a comprehensive meta-analysis of treatment effectiveness. As mentioned in the preceding chapter (CH III), not only does this present a potential bias, but it also excludes a majority of evaluation reports from the sample (e.g., Lipsey, 1992a; Weisburd, Sherman & Petrosino, 1990).[92]

Using the Electronic Search Techniques

The advancement in information technology has made millions of reports and articles accessible to researchers through computers (e.g., Cooper, 1989). Lipsey (1992a) Pearson, et al. (1996; 1995) and Wells-Parker, et al. (1995) made the most extensive use of electronic searches in their meta-analyses; most of the other searches used manual search indexes. Two technologies, the CD-ROM and DIALOG systems,

[92] Of course, some would argue that randomized experiments are so rare in criminal justice that they would be published regardless of positive or negative treatment results. The Weisburd, et al. (1990) research, however, found many experiments were contained in government reports or other non-journal sources, which would be excluded by persons focusing their meta-analysis solely on published, refereed articles.

were utilized during this process.[93] Both CD-ROM and DIALOG searches were considerably more productive when the data bases included searches of abstracts, rather than titles alone.

CD-ROM Technology.[94] CD-ROM allows millions of information units to be stored on disks through laser imprinting. For example, one laser diskette holds two decades of Sociological Abstracts (or Sociofile), which includes citations and abstracts for 250,000 documents. CD-ROM machines are readily available without charge at college and university libraries.[95]

[93] The CD-ROM technology (particularly with the SilverPlatter system) is fairly easy to learn and use, even for the computer novice. Thousands of citation titles and their abstracts can be searched using a variety of search types (e.g., author, classification codes, type of document, years, language, descriptors or identifiers, etc.). The key is to develop a consistent search strategy that produces the largest number of hits and reduces the amount of citations which have to be checked out manually. Most of the hits were sifted out by simply scanning the abstracts, without having to manually examine the original reports.

[94] Another advancement in CD-ROM technology which should be noted is the ability to download information onto a floppy diskette. This means that investigators can run the searches, save the information on a personal floppy, and scan it at a convenient time. More importantly, using the download feature allows the meta-analytic investigator to build several separate files of document citations. A file can be created to store "leads" or citations which should be checked for eligible studies. Another file can hold the rejected studies. This becomes invaluable, as investigators readily lose track of what has been checked when dealing with thousands of citations. This file allows one to check a citation before attempting to locate it. Used in conjunction with bibliographic software (e.g., ProCite), a meta-analyst can better manage the massive collection of citations.

[95] CD-ROM technology was utilized primarily at Northeastern University and University of Massachusetts, Lowell. The Dialog searches were run at Rutgers University's Dana Library, Rutgers University's Alexander Library and Northeastern University.

To track down randomized experiments, data base searches were made of available CD-ROM holdings at area universities. Four were determined upon close inspection to have relevance to the topic: *Criminal Justice Abstracts* was searched for documents from 1968-1993; *Sociofile (i.e., Sociological Abstracts)*, which encompasses 1,600 journals in 55 countries, was searched for the years 1974-1993; *PsycINFO (i.e., Psychological Abstracts)*, which covers 1,300 journals produced in 50 countries, was searched for documents from 1974-1993, and *ERIC (i.e., Educational Research Information Center)*, which represents 750 journals worldwide, was searched for documents from 1966-1993. All four contain references to crime and delinquency literature, although information in ERIC tend toward correctional educational programs and school based prevention programs.[96]

A broad search of the CD-ROM holdings was conducted, since past experience with the experimental literature demonstrated that many evaluators who employed random assignment did not use classic experimental keywords (randomized field experiment, controlled study, classic experimental design, etc.) in the title or abstract of the report. After several trials, a command statement containing the following

[96] It should be noted that *Criminal Justice Abstracts* provides comprehensive coverage of journals, books, dissertations and government reports. *Sociofile* abstracts journal articles and dissertations. *PsychINFO*, beginning in 1987, began to index books along with journal articles, but only abstracts the articles. *ERIC* contains titles and abstracts for journal articles, conference papers, government reports and other fugitive documents.

keywords (and their derivatives) produced the most relevant citations across the three data bases: random, experiment, controlled, evaluation, impact, effect, and outcome.

These keywords were searched in combination with some identifier of criminal justice in three of the CD-ROM data bases; *Criminal Justice Abstracts* exclusively focuses on crime and delinquency and therefore needed no identifier. In Sociofile and PsycINFO, classification codes exist which clearly identify the topic area where experiments are likely to be uncovered (e.g., offender rehabilitation, penology and corrections); the keywords identified above were used in conjunction with the classification code. For the ERIC system, no classification code exists, so a large number of descriptor terms related to crime and delinquency were used in combination with the keywords (e.g., crime, delinquency, sanctions, law, justice, etc.).

Using such a broad search produced a large number of hits which had to be sifted. For example, Criminal Justice Abstracts searched nearly 60,000 documents, providing 3,025 hits. Sociofile searched 167,281 documents, yielding 873 hits.[97] PsycINFO searched 607,711 documents, yielding 4,681 hits. The ERIC system checked 463,106 documents, producing 1,915 hits.[98]

[97] An additional 7,492 citations and abstracts were read using other keywords in Sociofile. Those keywords included analysis, data, delinquency, deviance, behavior, law, juvenile, justice, enforcement, legal, methodology, offender, offenses, parole, probation, prisons, sanctions, rehabilitation, treatment, program, recidivism, crime, criminal, correction and study.

[98] The totals for these three CD-ROM holdings represent the documents in the

Upon closer inspection of the abstracts, most of the citations and abstracts did not provide leads to randomized studies. Ironically, while some investigators who conduct a randomized test do not use the expected terminology (e.g., experiment), many evaluation reports contain the term "experiment" but are not randomized studies.

As Farrington (1983) and others have noted, the word "experiment" is colloquially used to describe any planned intervention into a social setting, regardless of whether randomization was used. Reading the abstracts helped eliminate many of these. Using broad keywords such as "random" meant that a number of studies using random samples or randomly selected populations were also retrieved; these were also easily excluded by reading the abstracts.[99]

DIALOG Technology. The other technology utilized during the data collection process was a search through the Dialog system. Dialog allows the user access to over four hundred literature and indexing data bases, some of which are not easily accessible to researchers on available CD-ROM machines (e.g., Hunter and

data base available in English before January 1, 1994.

[99] Investigators should be prepared to invest time learning and using CD-ROM technology. Trial searches generally need to be run before one can focus on keywords and a search strategy that is most effective. Most time consuming will be sifting through citations and abstracts, particularly if a broad search is developed. However, one must remember that the hours of time spent working with CD-ROM pales in comparison to a manual search of bound indexes to the periodicals.

Schmidt, 1990; Cooper, 1989). DIALOG searches are usually ordered through the university or college, conducted by information specialists, and printed citations and abstractions are delivered to the investigator within several days. Information technology has also permitted users to access Dialog through modems on personal computers via the Internet or online services such as CompuServe (e.g., Cooper, 1989).

Three separate Dialog system searches were conducted during the project. This system allowed information specialists to search 18 relevant data bases [see Appendix A], including *NCJRS* (National Criminal Justice Reference Service) and *Dissertation Abstracts Online*. While information specialists are quite proficient in the technology and search methods, they vary in their ability to run a distinctive criminal justice search. To illustrate, an information specialist was informed about the specific type of information needed during the earlier *Registry* project (e.g., Weisburd, Sherman and Petrosino, 1990). Following some trial searches and a personal meet the specialist conducted a search of the NCJRS data base and produced 556 document hits for the years 1972-1989.

During this project, another Dialog search was ordered, exploring several additional data bases. To provide a test of consistency, the NCJRS search was reordered with a different information specialist. The specialist in this case was given the same request and was even provided with the keywords used successfully by the first specialist. This time, the Dialog search of NCJRS–covering a wider time period (1972-1992)–yielded only 162 document hits.[100]

Another search was conducted later in the project, with the goal of investigating several data bases. Again, the NCJRS search was reordered–using the same keywords and commands—but this time the researcher assisted the information specialist in the search. Over 700 document hits were recorded! While the searches cost money, it demonstrated the variability in results across information specialists when using the Dialog system. It reemphasized the importance of using a variety of search methods to uncover reports and underscored the importance of the investigator conducting the detailed criminal justice search alongside the information specialist.

Examining Prior Narrative Reviews and Meta-Analyses

Another profitable source for locating eligible studies was an examination of the works cited in prior qualitative and quantitative reviews. For this meta-analysis, 32

[100] It is important to note that costs can accumulate rapidly when using the Dialog system. Most universities and colleges will charge faculty for the search; costs depend on the time spent online, and the number of hits produced. To reduce costs, prospective meta-analysts should not order the Dialog search until they have determined if relevant data bases such as NCJRS are available on CD-ROM. If a Dialog search is considered necessary, then the meta-analyst should spend time examining documents in the content area, conduct some trial searches on the available CD-ROM technology, and at least complete some hand checks of relevant periodical volumes. This will help the investigator specify the search parameters more carefully to reduce costs.

narrative reviews and 22 meta-analyses were compiled to provide leads to other experiments (see Appendix B).

While this search technique was very productive, it was hampered by the fact that some reviewers simply list their sampled studies in the reference section. This resulted in a large number of citations that needed to be checked to determine if the study fit the inclusion criteria laid out here.

One godsend for meta-analysis, however, is that many reviewers present their sampled studies in a concise table; type of design is ordinarily one of the major characteristics included in the table (e.g., Lab and Whitehead, 1988; Basta and Davidson, 1988; Wright and Dixon, 1977; Logan, 1972). These reviews allow randomized experiments reviewed to be exclusively targeted for retrieval. In some cases, reviewers grouped the citations to their sampled studies under broad design categories; Bailey (1966) provided such a categorization which allowed me to eliminate 79 of the 100 studies he reviewed from consideration. Lipton, Martinson, and Wilks (1975) provided annotated summaries of 231 studies in their famous monograph, which also allowed for quick identification of eligible experiments.

There are scores of "what works?" reviews in journals, books, government documents and unpublished papers. While reviewers often cover similar ground and citations may overlap across reviews, each review generally contained new citations to potentially eligible experiments and was one of the most productive sources for locating randomized studies.[101]

Conducting a Manual Hand Search of Major Journal Volumes

Since electronic searches and indexes often missed experiments because specified keywords were not contained in report titles or abstracts, a manual hand search of 29 periodicals likely to publish criminal justice experimental reports was conducted.[102] Most publish criminal justice articles exclusively, but a few periodicals represented other disciplines (e.g., sociology-*American Sociological Review,* social work-*Journal of Social Service Research*, or law-*Journal of Legal Studies*). A complete list of the journals searched manually can be found in Appendix C.

A hand search meant that the investigator scanned the titles and articles of each volume's contents. While most articles could be discarded immediately based on the title (e.g., "Prediction Methods in Criminal Justice") or some other indicator (e.g., article was essay or book review), the empirical articles dealing with interventions had

[101] While a meta-analyst should try to be comprehensive in gathering reviews, the main goal of the search and retrieval process (and time and resources) must be in getting the primary reports. An investigator can spend too much time attempting to track down literature reviews; it is best to focus on the available and most comprehensive reviews which are readily accessible through journals or books.

[102] Vaughn and del Carmen's (1992) annotated list of periodicals likely to publish criminal justice research was used to form a targeted list of journals. The prestige rankings used by Sorenson, Patterson and Widmayer (1992) were also checked to insure that most important journals were not omitted.

to be inspected a bit more closely. While invaluable in uncovering experiments missed by the computer searches, this was a very time consuming endeavor.[103] Libraries generally had only the most recent volumes and some of these were invariably missing. Older volumes were sometimes on microfiche, but that added even more time to scanning the articles.[104]

Published Bibliographies

Another search technique utilized during data collection was a manual search of eight published bibliographies in criminal justice (e.g., Berens, 1987; Hewitt, Poole and Regoli, 1985). While some of these are available in the form of books or government reports, others are published in the scholarly journals (e.g., Goyer-Michaud, 1974). These bibliographies were predominantly useful when they included annotations of the documents referenced (e.g., Cooper, 1989).

[103] The following university and college libraries were utilized during the search: Boston College's main, social work and law libraries, Harvard University's main and law libraries, the NCCD Criminal Justice Collection at Rutgers University, Northeastern University's main and law library, the Westfield State College library, and the University of Massachusetts at Lowell' two main libraries. Unfortunately, there were no libraries within reasonable travel of this investigator which housed complete sets of important criminal justice periodicals, and the interlibrary loan became crucial part of this project.

[104] While it is recommended that meta-analysts undertake such a hand search, it should be done very early in the project when it would help the investigator learn the craft and permit sharper inquiries when using the computerized technology. In addition, the searches should be planned by obtaining the lists of serials held by area libraries, so that searching for volumes does not become time prohibitive. For a treatment effectiveness meta-analysis, hand searches should include more psychology, social work and sociology journals than were checked here.

The bibliographies retrieved and examined ranged from specialized lists of experiments (e.g., Boruch, McSweeney and Soderstrom, 1978) to broader lists of deterrence works (e.g., Beyleveld, 1980) to even more expansive lists of criminal justice documents (e.g., Hewitt, Poole and Regoli, 1985). Appendix D provides a complete list of the bibliographies searched to date during this meta-analysis. Although a great many of the citations from most bibliographies examined could be excluded, some were crucial in providing leads. Boruch, McSweeney & Soderstrom's (1978) annotated bibliography of experiments provided 75 leads alone to randomized studies in criminal justice.

Using Published Solicitations

Given the fact that this study was seeking published and unpublished reports, unconventional search methods were utilized. One such technique was the publication of solicitations in major social science association newsletters. The advantage of using solicitations–which informed newsletter readers about the project and the type of reports wanted–is that large memberships of persons who have conducted or are aware of randomized field experiments are contacted.

The two major associations for criminologists, the American Society of Criminology [ASC] and the Academy of Criminal Justice Sciences [ACJS] were targeted first. Both produce newsletters (ASC's *The Criminologist*, *ACJS Today*) which are mailed to all members on a quarterly or more frequent basis.

While using ASC and ACJS meant that over 4,500 scholars, practitioners and educators in criminal justice were potentially notified about the study, wider coverage within criminal justice and across other disciplines was desired. Solicitations were placed in the newsletters for the major associations in psychology [American Psychological Association's *The Monitor*] and sociology [American Sociological Association's *Footnotes*]. These newsletters conceivably communicate the goal of the study to over 100,000 scholars worldwide, many of whom conduct criminal justice research.[105]

"Calls for experiments" were also run in some smaller, related association newsletters. These ranged from newsletters with a scholarly constituency (e.g., Section on Criminal Justice Administration, American Society for Public Administration's *The Key*) to those that reach a large community of practitioners (e.g., American Correctional Association's *On the Line*).[106]

[105] APA's *The Monitor* is distributed to 95,000 persons across the globe; ASA's *Footnotes* has a circulation of 13,000.

[106] The other solicitations were published in the following; *Justice Research* (circ. 1,500), which is published bimonthly by the National Criminal Justice Association; *CJ Update* is a newsletter printed and circulated by Anderson Publishing Company, an academic criminal justice press, predominantly to the criminal justice faculty; the *AJA Newsletter* (circ. 10,000), published by the American Jail Association; *Perspectives*, which is printed by the American Probation and Parole Association (circ. 3,000); and *The Forum*, which is published by the Justice Research and Statistics Association and distributed to a variety of federal, state and local government justice agencies. Finally, two broadly circulated journals also published the call for experiments: the *National Institute of Justice Journal*, which is published by NIJ, and *CJ International*, which is published by the Office of International Criminal Justice in Chicago.

Despite the comprehensiveness of these published solicitations, there were some problems using this method. First, there are other associations which may be familiar with the content area that were not contacted (e.g., evaluation societies, social work associations, statistical associations, law associations, etc.). Second, while most newsletters published the calls as a professional service without charge, significant costs were incurred when a fee was imposed, such as the case when the American Psychological Association charged over $300 for a obscure classified ad. Finally, responses sometimes border on the absurd, such as one investigator who wanted a detailed explanation of the how to do a meta-analysis over the telephone.[107]

Contacting Major Investigators

It is also true that an invisible college exists in academia, where researchers with similar interests are aware of each other's work and may even exchange reprints and other scholarly communications (e.g., Cooper, 1989). It is important that meta-

[107] Some modest improvements can be made in the future. First, solicitations should be geared toward publicizing the study. They should be run as early as possible during the project, to allow for information about the project to filter through the academic network and to allow for communication between meta-analysts and other researchers. Second, the solicitation should be run to promote the return of fugitive documents. Investigators generally responded to the broad call for experiments by sending published reprints, which were already in the sample. Future meta-analysts should focus the solicitations on literature that is not easily retrieved, urging investigators to send conference papers or other documents which are difficult to acquire. Finally, timing should be considered when publishing requests in newsletters. Summer month issues may not engender the same response, particularly since faculty may be away from offices where publications are stored or can be xeroxed. It may be more prudent to consider an early fall month (September and October issues), when colleges and universities have started the traditional academic school year.

analysts communicate the goals of the project in the network of scholars with similar interests.

For this meta-analysis, a mail campaign was undertaken with persons familiar with the "what works?" area. The most recent published membership directories of the Academy of Criminal Justice Sciences and the American Society of Criminology were used to compile a list of scholars and practitioners who were considered influential in the treatment effectiveness area. A cover letter and information about the project, including criteria for study inclusion, were mailed out to each identified member to solicit more experiments.

The importance of networking in this invisible college can not be overstated. This method generated numerous mailings from scholars sending their studies they believed to fit the inclusion criteria.[108] In some cases, correspondence about this project resulted in respondents sending bibliographies used in class or recent books.[109] Prior reviewers often mailed their complete set of citations to studies in their sample.[110]

[108] For example, Dr. Daniel Glaser, Professor Emeritus at the University of Southern California, sent me a publication of a recent electronic monitoring study.

[109] For example, Nathaniel J. Pallone, distinguished professor of criminal justice at Rutgers University and editor of the *Journal of Offender Rehabilitation*, sent me his annotated bibliographies of offender treatment materials. Contact with other authors usually resulted in reprints and other useful materials being sent.

[110] For example, Dr. Mark Lipsey of Vanderbilt University generously supplied me with an entire bibliography, but the data base was unable to exclusively identify the randomized studies.

Sometimes an offer of reciprocal assistance can help one within this informal network. After mailing citations and abstracts of located studies to other scholars conducting similar quantitative reviews, they generally responded by forwarding helpful citations to randomized experiments from their review samples.[111] It is also true that a number of investigators could not locate their original reports, and more frequently could not supply answers to questions about missing crucial data in those documents (e.g., sample size, etc.).

In addition, public and private research centers which conduct criminal justice studies were identified using *A Guide to Research Centers.* A mailing was undertaken, similar to that described above, requesting eligible studies. This method was not as successful, perhaps due to the lack of a particular contact person and personal touch in the mailing.[112]

Compiling References from Other Literature

Another search technique used during this study was to examine the references contained in primary research reports and other related literature. Checking and

[111] To illustrate, after mailing the *Registry* to Douglas Lipton of the National Research Institute, he agreed to forward citations to experiments uncovered in his CDATE meta-analysis and provided information regarding the whereabouts of the studies reviewed in *The Effectiveness of Correctional Treatment* (e.g., Lipton, Martinson, and Wilks, 1975).

[112] To illustrate, many of the contact persons listed in the *Guide* were not the actual research staff but foundation presidents and directors. Perhaps the requests were not forwarded to those familiar with designs used.

compiling citations off retrieved studies is a time consuming and tedious process. While some citations can be excluded from the title, many can not. Most of the leads turn out to be non-experimental studies, but a few were located in this manner. Unfortunately, later experiments rarely cite or consult earlier experimental studies, a point noted earlier by Dennis (1988) in his research on implementation problems encountered in randomized field research.

While primary research studies can sometimes provide an interesting lead, texts dealing exclusively with the 'what works?' literature were also checked to uncover experimental studies (e.g., Palmer, 1992; Cullen and Gilbert, 1982). When these larger works cite experiments, they are usually influential and well known experiments. This was also time consuming, given the hundreds of citations for well-researched texts, although again, some of them can be eliminated by title.

The indexes for *Sociological Abstracts*, *Psychological Abstracts*, and *A Guide to Periodicals in the Social Sciences and Humanities* were searched back to 1950, to account for years not covered earlier electronic searches. Since NCJRS does not exhaustively cover the journals, and Criminal Justice Abstracts were not available before 1968, hand searches were done of *Abstracts on Criminology and Penology* and *Police Science Abstracts* back to 1950. This method yielded few experimental reports.[113]

[113] Another check for earlier experiments was conducted by thoroughly inspecting the earlier Lipton, Martinson & Wilks (1975) review, since they exhaustively covered the literature before 1968.

Searching is Not Always Retrieving

Some of the techniques discussed earlier not only result in hits or potential eligible studies, but they also result in retrieval of the document. For example, when searching the periodicals manually, one can easily duplicate the eligible report. When writing to investigators or placing a solicitation, reprints or other articles are sent, meaning that document retrieval has been accomplished.

However, some of the methods, such as the computer searches and the use of prior reviews produce the citations to "potential" studies. Some of these were tracked–with moderate difficulty–at the university or college library. A check of the library's serial holdings determined whether the reports were in volumes accessible in the stacks (recent volumes usually were) or had to be copied off the microfiche (older volumes were usually housed there).

Moreover, since the source documents for eligible studies were likely to be scattered across many libraries, states and countries, interlibrary loans became an essential part of the data collection process. In rare cases, obscure journal articles or reports were not located by library staff and could not be included in the meta-analysis.

Importance of Terminating Search and Retrieval Efforts

As mentioned last chapter, including time limits as part of the inclusion criteria greatly assists the search process. For example, limiting publications in this study to the years 1950-1993 meant that citations to earlier or later studies could be ignored. Otherwise, one runs the risk of never completing a study, as new journal volumes and books are being added to the shelves continuously.

Nonetheless, despite the time frame criteria, at some point the investigator must make a decision about when to terminate the search and retrieval efforts. Particularly when conducting a comprehensive meta-analysis such as this one, documents meeting the time frame and other inclusion criteria are continuously being received by the meta-analyst.

With every new article, particularly a review or anthology of research works, there will be references to potential leads that need to be checked out. Even retrieved experiments will contain a citation to a past study that may or may not be applicable until it is retrieved and read. In short, searching and retrieving is never over until the investigator terminates it.

There is certainly no set answer for deciding when to terminate the data collection phase in meta-analysis. Certainly, as the search process continues, potential leads to eligible studies will begin to dwindle. However, in a search as massive as this one, the investigator always runs the risk of checking the same citations more than once, particularly when there is more than one person involved in the searching.

For this project, as time progressed, the number of new leads was not only dwindling, but they were more frequently comprised of older citations to unpublished, fugitive literature that would be quite difficult to obtain. In addition, despite stringent criteria, the size of the meta-analytic sample had eclipsed nearly all prior reviews except those who used similar search techniques, increasing confidence that there was good coverage of the published and unpublished literature (e.g, Wells-Parker, 1995; Lipsey, 1992a).

Following a final wave of mailings and loan requests, particularly from the NCJRS and NCCD collections, search and retrieval efforts were terminated during the early part of 1997. While documents keep coming in, they are no longer coded or added them to the data base.

The Final Sample

These search techniques led to over 300 randomized experimental studies being located and retrieved for the meta-analysis, the largest collection of such evaluations reported to date in the justice literature. The data collection process uncovered about twice as many experiments as originally expected. While the generalizability and power of the sample would undoubtedly increase by doubling the sample, the resources needed to code and keypunch over 150 additional experiments would also be sizable.

To meet the competing needs of sample rigor and research costs, a random subset of 150 randomized experimental studies was selected for the subsequent analysis. Random selection in this instance was accomplished by alternate sequencing: every other experiment was chosen until 150 were entered into the sample. While the sample represents less than 50% of the eligible documents collected during this phase, it is the fourth largest treatment effectiveness in criminal justice, behind the heavily funded Lipsey (1992a), Wells-Parker, et al. (1995) and Pearson, et al. (1996) studies. Randomly choosing the subset of 150 studies provided a barrier to selection bias.[114]

Unfortunately, since the larger collection of 307 randomized field experiments was not coded and keypunched into a data set, there is no scientific way to insure that this 150 meta-analytic sample is representative of the larger set it was drawn from. Further complicating this sampling issue is the fact that the 307 retrieved studies were drawn non-randomly from a larger unknown universe of all eligible experiments. Though not unique to this study, it is a persistent dilemma in meta-analysis research and requires some cautious interpretations when generalizing results.

Pipeline

As expected, the experiments file kept by Weisburd and his colleagues (1990) was the most productive method for retrieving experiments for this meta-analysis.

[114] For example, it is much easier to code journal articles than books or dissertations. Furthermore, if left to the meta-analyst's own devices, articles with two-group designs

Table 2 also demonstrates that electronic techniques, manual searches of journals, and prior review work were also relatively productive. Formal solicitations in newsletters and informal requests through the invisible college rarely uncovered any new experiments.

Table 2.

Which Search Methods Were Most Productive?
Analysis of How 150 Experiments in Meta-Analysis Were Located

Search Method	Number Retrieved	Percentage Retrieved
Registry Files	73	44.6%
Electronic Searches	42	28.0%
Manual Hand Searches	25	16.6%
Prior Reviews	7	4.6%
Bibliographies	1	0.6%
Solicitation	1	0.6%
Original Study Citation	1	0.6%
Total	150	100%

and simple experimental statistics would have been selected over more complex multiple group experiments.

CHAPTER V. THE CODING PROCESS: EXTRACTING INFORMATION FROM THE ORIGINAL EXPERIMENTAL REPORTS

As eligible documents are retrieved by the meta-analyst, he or she must extract from those reports the information which will answer the research question (e.g., Cooper and Hedges, 1994). This is crucial, as information that is not coded and entered into the database can not be considered, unless one is willing to recode (e.g., Cooper and Hedges, 1994). This chapter will review the coding process for the study, with special attention on the following issues: the type of variable information extracted from study documents; the decision rules used in coding the reports; the computation of effect size; and intercoder reliability.

Coding Schemes in Prior Criminal Justice Meta-Analyses

As expected, the coding process varied considerably across the 22 prior meta-analyses reviewed for this study. In some projects, the coding scheme was simple, limited to ten or less variables (e.g., Andrews, et al., 1990; Garrett, 1985). In some of the later meta-analyses, a great amount of information was extracted (e.g., Pearson, et al., 1996; Lipsey, 1992a); it appears that the number of variables in the meta-analysis was directly linked to the presence or absence of federal funding for the project.[115] This makes sense, since extensive coding schemes require more time and therefore

[115] Funding agencies often require that projects produce a database and codebook to be archived by use for other researchers. Investigators in such instances are probably more likely to exhaustively code to capture data that secondary researchers may wish to examine.

more resources to conduct, literally impossible with outside research support. Given the limited federal funding available here,[116] the coding system was more extensive than most meta-analyses, but not as detailed as the research conducted by Lipsey (1992a), Wells-Parker, et al. (1995) or Pearson, et al. (1996).

Developing a Coding Instrument

The initial coding instrument was greatly influenced by earlier research. The codesheet used by Weisburd, Sherman and Petrosino (1990) was most helpful, since it was developed for a project specifically for randomized experiments. Their earlier instrument contained 99 items,[117] most of which were incorporated into this work. Additions to the coding instrument were made following reviews of extensive works on randomized experiments (e.g., Dennis, 1988; Farrington, 1983) and prior treatment effectiveness meta-analyses (e.g., Pearson, et al. 1995; Wells-Parker, et al. 1995; Lipsey, 1992a; Davidson, et al. 1984). In addition, committee members critiqued both the initial and revised instrument, and these comments were incorporated into the initial version, which contained approximately 210 items.

Wells-Parker, et al. (1995) described their use of the invisible college of experts in the field of drunk driving intervention to develop a list of items for coding

[116] The National Institute of Justice Graduate Research Fellowship, which provides support for doctoral dissertations in criminal justice, was awarded for this project in 1993 and allowed $17,007 in support

[117] All 99 items were coded if there was available data for three distinct outcome measures at three different follow-up periods.

treatment programs. While this method increases the costs and time needed to conduct a meta-analysis, it does insure that the coding instrument captures items perceived to be theoretically and practically important. It also serves as a check against the literature, since prior researchers may have missed factors which are important to programmatic success or failure. Nonetheless, given that there were 22 prior treatment effectiveness meta-analyses, and an extensive literature on randomized studies, the expense of using an invisible college to create a coding instrument outweighed any potential benefit.

Since coding can be a dynamic process, a pretest of the instrument was conducted on a small subset of documents (N < 10). This pretest indicated that a few items solicited such remote information that they were not likely to be contained in any report. For example, data for the item "how were estimates of caseflow obtained?" were not found. A few items were also found to be too confusing to be useful. These were dropped, leaving a final coding instrument which solicited information for 196 variables. Appendix E represents the final version of the codesheet used to extract information from the eligible experimental reports. For ease of illustration, final coding items can be grouped into the following eleven categories:

Document Information	year, type, total used, percentage contribution by primary document, number of experiments in document, coder, pipeline

Investigator Information	affiliation and field,[118] relation to research setting
Experiment Information	year started, length, multisite, region, scope, id, name, funding, point in criminal justice system
Randomization Information	method, blocking, matching, stratification, number of groups, equivalency pretests, overrides, how overrides handled
Other Methodological Information	attrition problems, how attrition handled, caseflow problems, how caseflow handled, statistical power
Sample Selection Information	voluntary/consent, payment to subjects, eligibility or exclusion criteria
Subject Information	N of subjects, percent white, percent male, average age, average education completed, average IQ, prior record, instant offense
Program Information	type of treatment, agency delivering treatment, how delivered, contact, monitoring, treatment problems
Outcome General Information	total follow-ups, minimum/maximum follow-up, total crime outcomes, types, total non-crime outcomes, types
Outcome Specific Information	follow-up in months, crime measure, data source, direction of effect, statistical significance, test used, number of tails, actual probability level of test score, test value, small sample/statistical power, crime effect (for 3 follow-up periods) each group's sample size for outcome tests

[118] Affiliation and field is solicited for up to two investigators.

Subgroup Effect Information	subgroup analyses, differences found, type of effects

In developing the instrument, a balance was desired between extracting detailed information to be used in this meta-analysis and later secondary analyses, and in reducing the resources expended on coding. Thus, eliminating remote or confusing items reduced some coding time with no perceived drawback for the project. In addition, some freeform comment fields were added to the codesheet to allow coders to expand upon items or note anything else of interest about the study.

Coding Guidelines

To assist with coding during the project, rules for making decisions about data extraction were developed and used during the project. The full text of the coding guidelines is presented in Appendix F. While the purpose of the guidelines was to instruct the contracted coders, it also helped to maintain consistency during the life of the project, i.e., that items would be coded reliably across time by the same coder. It was important, given the fluid nature of the coding process, to insure that the same items were coded the same way at the end of the project as in the beginning. When a variable was changed during the coding process, the investigator went back to insure that all preceding documents were similarly coded to insure reliability.

The most important coding decisions have to do with handling multiple data sources or elements. During this study, multiple documents, experiments, groups, follow-up periods, outcome measures and effect size data were confronted frequently. Handling their occurrence is part of the research decisionmaking that greatly affects a meta-analysis, and it has been recommended that the guidelines used to resolve these difficult issues be clearly explicated (e.g., Matt, 1989).

Multiple Documents. In the case of multiple documents, where more than one report was available on a single experiment, the manuscript providing the most information to the meta-sample was designated as the "primary document." Other reports were used, when applicable, to supplement the primary document; this usually took the form of subsequent follow-ups, additional information on the intervention, subgroup analyses, or practical research papers. When conflicting information on the same experiment was found, which was rare, the primary document was used unless supplementary reports indicated that the data was in error.[119]

For nearly 100 experiments, the entire information for the meta-analysis came from a single document (N=96). For 36% of the sample, two or more documents were used to extract information; the primary document rule was invoked for those cases (N=54). Most primary documents contributed 50% or more of the data on the study to the meta-analysis; approximately nine in ten contributed 80% or more information to the data base.

[119] Most frequent was when different documents reported conflicting sample sizes; again, the primary numbers were used unless later reports indicated those numbers were inaccurate.

Multiple Experiments. It was occasionally the case that a single document contained the results from several randomized studies. Indeed, 123 primary source documents accounted for the 150 randomized studies in the sample. This occurred most often in reports describing national, multisite studies, or local experiments which followed two individual study cohorts. If the investigator presented the outcome data distinctively for each site (or cohort), the experiments were individually coded as separate studies. In the rare instance where the investigator presented only a combined analysis across all sites, the studies were coded as a single experiment. Thus, as in other meta-analyses, there are more studies (or experiments) than primary documents, since some documents contribute two or more experiments to the sample.

Seven in ten experiments were reported as single studies. Three in ten experiments (N=45) were part of multisite or multiwave research programs, which reported the results of 2-4 experimental field tests. For example, the Arizona Pretrial Drug Testing Experiments actually reported the results of four separate field studies across two counties in Arizona (e.g., Britt, Gottfredson and Goldkamp, 1992).

Multiple Groups. Most randomized field tests were two group designs, comprised of a treatment and control condition. The treatment group was comprised of subjects receiving the intervention under investigation, while the control group normally received treatment as usual or no contact at all. In essence, treatment effectiveness meta-analyses were designed for such simple experiments. A minority of studies in the sample, however, were multiple group designs, comprised of three or more groups, which presents unique problems for meta-analysis. Since effect size computations are based on a statistical comparison between two groups only, handling the information from the additional groups is problematic.

Some prior investigators (e.g., Garrett, 1985) simply coded multiple group designs as separate studies; for example, an evaluation with two treatment groups and one control condition was coded as two separate designs (Treatment 1 v. Control; Treatment 2 v. Control). Others have argued that partitioning one study into separate designs and including them all in the meta-analysis violates the assumption of statistical independence of the effect sizes (e.g., Gleser and Olkin, 1994; Lipsey, 1992a).

Following the caution of Lipsey (1992a) and others about statistical independence, only one group comparison per experiment was selected for the sample. This resulted in two methods for handling multiple group experiments. For those studies where groups theoretically could be grouped together into two logical conditions, collapsing was done. For example, the misdemeanor domestic violence experiments were essentially concerned with the effects of formal arrest against informal handling by police (mediation, etc.). Thus, these other treatments could be collapsed into an informal processing group, a comparison often performed by some of the original investigators (e.g., Dunford, 1990).

When groups could not logically be grouped for a comparison test, the strongest contrast between two groups was utilized. In nearly every case where this rule was applied, the strongest intervention was compared to the control condition. In the rare case where this could not be applied, the treatment in which the original investigator was most interested was compared to the control group.

As expected, most experiments were simple two group designs (71%). The remaining 44 experiments ranged from three to eight group designs (29%). It was only in the latter type of study where decision rules had to be utilized. In 31 cases, the strongest versus weakest contrast was utilized (21%). In the remaining 13 experiments, eight were collapsed into two comparison groups by original PIs and five were collapsed by this investigator.

Multiple Follow-ups. Most experiments provided results for only one follow-up time interval (e.g., one year). Clearly, federal funding plays a role in the lack of multiple follow-up intervals, since investigators often do not have resources to conduct lengthier studies (e.g., Sherman, et al. 1997). To handle those experiments with multiple follow-ups, however, up to three distinct time intervals were coded. Few studies had more than three, but when they did, the earliest, latest and middle follow-up periods were used.

Monitoring the changes in effect size over time was an initial focus of this meta-analysis and could only be performed if multiple time interval data were captured. This type of analysis has been lacking in earlier meta-analyses, which have either reported a global effect across all follow-ups (ignoring the time interval) or first effects (ignoring later follow-up data). Unfortunately, the lack of second and third follow-up data prevented this analysis from taking place (Chapter VII).

Multiple Outcomes. A more common dilemma were experiments which reported multiple measures of criminality, such as rearrest and reconviction, at a single follow-up period. Although correlations across measures are presumed to be strong, it is also the case that effect size could be inflated by selecting the most positive outcomes reported (and reduced by taking the least positive outcomes). It is important that consistent rules be applied to selecting outcome measures for effect size estimation in meta-analysis. A description of methods for handling this problem in prior meta-analyses is lacking, although some have simply included all effects (e.g., Andrews, et al. 1990) or reported an average effect in their study (e.g., Garrett, 1985).

For this project, the 'Sellin' rule was invoked: the broadest measure of criminality from the earliest point in the criminal justice system was coded. As noted first by Sellin (e.g., Senna and Siegel, 1995), the more removed crime data is from its initial occurrence, the more it reflects practitioner decisionmaking rather than actual offending conduct. Using this logic, police contact or rearrest data are probably more accurate reflections of reoffending conduct—despite their attendant problems–than reconviction or reincarceration measures. Thus, arrests were preferred over conviction, and total arrests were preferred over arrest subtypes (e.g., felony arrests, robbery arrests, etc.). Since rearrest data was the predominant outcome measure used by experimental researchers, these rules increased consistency for the meta-analysis.

The only exception to the Sellin rule for coding outcome measures was applied to parole and probation data, even though revocation data is broader than arrest or other measures. In a few cases, only revocation data was available and had to be coded. However, in those cases where rearrest data also existed, the police data was selected, since revocation measures are more susceptible to differential effects due to organizational pressures than actual criminality (e.g., Lerman, 1975).

For nearly half the sample (45%), the outcome information used was the only data reported in the document. For 82 cases (55%), rules for selecting the most inclusive and standardized official outcome data were invoked.[120] This most frequently resulted in rearrest or other police data being used. Whether any differences are found in effect size across police and non-police outcome measures are explored in Chapter VII.

Multiple Effect Size Data. Finally, even after applying the rules for selecting follow-ups and outcome measures, decisions still had to be made about which data to use for effect size estimation. It was the case that some experiments contained different statistical expressions of the outcome measure. For example, a single experiment may have reported rearrest data at six months. However, that data may have been expressed in a number of ways, all of which could conceivably be used to yield an effect size: percentage rearrested (percentage of individuals rearrested), mean arrest rate

[120] There were at least two experiments where the outcome data was so uninterpretable that most of it had to be discarded in favor of one decipherable measure.

(number of arrests per person), time to first arrest, statistical test values (chi square, t, or f test values), or probability value of the test statistic. Which statistical expression should be used for effect size estimation?

While statistical test values, exact probability levels of the test statistic and the data on which it were based (e.g., means) are going to be nearly identical, that is not the case across means, percentages or survival analyses. It is true that these three data can provide different results in a single experiment. One must use caution with mean arrest rates, since they are inflated by a few individuals with multiple post-program arrests, a problem which does not similarly affect failure rates (since a person who is arrested once is considered a 'failure'). In addition, failure percentages were the predominant outcome data reported in the sample documents. Time to failure or offense severity indices were used infrequently by primary researchers.

Thus, in the interests of both methodology and uniformity, differences in group failure proportions were used to compute the effect size. If failure percentage or frequency data was not available, then means were selected provided the standard deviations were also available (group standard deviations are needed to compute the pooled standard deviation). If neither failure percentages or mean offending rates were available, then the test statistic (from chi, t, f, z or r) was transformed into the effect size.

Effect Size Estimation Procedures

In order to compute effect size correctly, a specialized software package entitled D*STAT: Software for the Meta-Analytic Review of the Literature* was purchased.[121] *DSTAT* automatically computes effect sizes for outcome data entered into it, a major improvement over the hand caculated meta-analyses prevalent before 1990. Having an automated and valid effect size generator saved time and removed the risk of incorrect calculation. With the program, provided the outcome data are reliably selected, the effect sizes will be correct.

The common metric utilized in this meta-analysis was Cohen's *d*, which is computed by dividing the mean difference between the experimental and control groups by the pooled standard deviation ($(\mathrm{x}_E - x_C / S_{pooled})$. As mentioned earlier, however, the overwhelming majority of reports compared group failure proportions; means were infrequently reported. DSTAT derives *d* from the difference in proportions by treating proportions as the mean of a distribution of 0's and 1's (Johnson, 1989:105). Thus, *d* can be estimated by using the following formula:

$$d = (p_E \cdot P_C) / S_{pooled},$$

[121] *DSTAT* was created by Dr. Blair Johnson, Psychology Department, Syracuse University and is available from Lawrence Erlbaum Associates (Hillsdale, NJ). The investigator expresses his gratitude to Dr. Johnson for his assistance in using the software, and overcoming normal start-up glitches.

where p_E and p_C are failure proportions for the experimental and control groups, and S_{pooled} is the pooled standard deviation of the sample of 0's and 1's.[122] Other transformation formulae generally convert test statistics to Pearson's r and then to *d* (Johnson, 1989).[123]

Intercoder Reliability

Intercoder reliability is the chief manner by which data extraction schemes are validated. Essentially, intercoder reliability refers to the agreement or disagreement between two independent persons extracting data from the same documents. If the data changes from coder to coder, then it is not reliable and less faith is placed in findings. In addition, if only a single investigator can code the reports, then the study is not replicable, thereby frustrating an important goal of science.[124] Most prior meta-analyses in criminal justice have not discussed coding or reported an intercoder reliability estimate, but those who did usually provided one for all items collectively.

[122] S_{pooled} in this case is $\{[(n_E - 1) * s^2_E + (n_C - 1) * s^2_C] / [n_E + n_C - 2]\}^{1/2}$, where n_E and n_C are the total observations for the experimental and control groups, and s^2_E and s^2_C are the variances for the experimental and control groups. S^2 is computed by the formula $p * (1 - p)$, where $p =$ the group failure proportion (Johnson, 1989).

[123] To convert r into *d*, the formula $d = 2r /(1 - r^2)^{1/2}$ was utilized.

[124] Weisburd (1995), however, makes a provocative observation. An investigator who has worked intimately with the documents and has coded them for past projects has an expertise that subsequent coders will not possess. Thus, low reliability estimates may reflect a second coder's inexperience rather than problems with the coding instrument. This inexperience is heightened when the second coder is simply contracted for that assignment (as in this study) and has no time to acquire an expertise in the literature.

For example, Davidson and the Michigan State University meta-analyses reported a reliability coefficient of .86 across all five studies.

Generally, the few global reliability coefficients reported have been acceptable, with most ranging in the .80-.95 range. Therefore, prior treatment effectiveness meta-analyses in criminal justice appear to have generated coding schemes which are reliable across independent coders. The reliability scores are even more impressive when one considers how missing or deficient reporting—a consistent complaint of all prior meta-analysts—greatly reduces coding reliability (Orwin and Cordray, 1985). In Orwin and Cordray's (1985) study, they showed how deficient reporting results in different coders incorrectly inferring data from study reports.[125]

However, Yeaton and Wortman (1993) have illustrated how global intercoder reliability estimates are inaccurate, since they likely mask the problematic items which greatly influence the interpretation of meta-analytic results. For example, two independent coders may achieve nearly perfect agreement on study descriptive items such as the year of publication or type of document (e.g., book , journal, etc.), but experience low agreement on which outcomes and follow-ups to select to compute effect size. The high reliability coefficients are generated by the large number of descriptive items with perfect agreement; however, it is the effect size information which is most influential in interpreting meta-analytic results.

Yeaton and Wortman (1993) also discuss the hierarchical nature of coding decisions; clusters of important variables are often dependent on the reliability of earlier coding decisions. For example, effect size—even if computed correctly–will be in error if the experimental treatment is inaccurately coded. They urge meta-analysts to assess reliability for these clusters (which they define as levels) rather than relying on a single estimate.

Several safeguards were instituted to increase coding reliability for this meta-analysis. First, given the low reliability for judgment and rating items in prior meta-analyses, no ratings were used in this study (e.g., Wells-Parker, 1995). Items where information was to be inferred rather than clearly collected were also avoided. Second, confusing or remote items were eliminated after the pretest and before formal coding commenced, reducing the number of difficult variables to be extracted. Third, a set of decision rules was developed and given to both independent coders contracted on the project; the investigator spent approximately two hours with each coder to discuss the coding instrument. Fourth, only post-graduate level coders familiar with criminal justice research literature were hired as coders.[126] However, the investigator–with long experience in coding experimental reports–was the sole coder for 74% of the sample documents (N=111). However, despite these procedures, intercoder reliability must be demonstrated and not assumed.

[125] This leads to another point: coding reliability can be high on certain items, with both coders being inaccurate. This happened at least twice on this project, and forced the investigator to review all of the coding.

In line with the work of Yeaton and Wortman (1993) and Weber (1990), intercoder reliability was assessed for clusters of variables. Intercoder reliability, as Weber (1990) urged in his discussion of content analysis, was determined *before resolving coding discrepancies.*

To perform the reliability check, eight documents (5% of the sample) were randomly selected for coding. These were initially completed by the investigator, and then coded independently by one of the contracted individuals. The data were entered into a statistical software (SPSS) program. Freeform comment fields were eliminated from the comparison. The actual coding value given by the investigator (A) and the other individual (B)–for each of the eight records–was entered into the SPSS data base. For each item, the percentage agreement was computed across the eight records. Thus, if the investigator and the additional coder exactly coded an item across all eight records, the percentage agreement was 100%; if there was consensus on four records, the percentage agreement was 50%. This was a stringent test: an item was only considered in agreement if it was identical across both coders; even variables which were misspelled or were not rounded off properly were defined as unreliable.

Overall, the mean rate of agreement across all instrument items was 80% (or .80), which is consistent with prior treatment effectiveness meta-analyses (e.g., Yeaton

[126] The author thanks Dr. Carolyn Petrosino (Ph.D., Criminal Justice, 1993) and Mr. John Lavender (M.A., Criminal Justice, 1994) for their assistance with coding.

and Wortman, 1993). It is true, though, that the global rate masks the variation across items; agreement ranged from 25% to 100%.

To determine which areas were problematic, items were grouped by their substantive content area (document, investigator, experiment, randomization, etc.) and the agreement rate compared across these categories. As Table 3 indicates, there were some surprising results. For example, outcome information had the second highest rate of agreement (86%); prior meta-analyses had indicated that this was the most problematic area for intercoder reliability (e.g., Wells-Parker, et. al. 1995). This indicates that the guidelines for handling outcome measures, follow-up periods and effect sizes had their intended effect and removed variation across independent coders.

Table 3.

Intercoder Reliability Check
Rate of Agreement Across Item Categories

Item Category	Rate of Agreement
Experiment Information	88%
All Outcome Information	86%
Subject Information	80%
Document Information	78%
Sample Selection Information	75%
Randomization Information	71%
Methodology Information	67%
Program Information	46%
Investigator Information	44%
TOTAL FOR ALL ITEMS	80%

Unfortunately, both program items (46%) and investigator items (44%) were highly unreliable. For the program items, the coders conflicted most often on treatment contact, monitoring and problems. For investigator information, the coders disagreed most often on the relationship of the researchers to the setting.

A review of coding discrepancy was undertaken to determine the source of conflict, for all items, including those with high agreement. Reasons for disagreement were varied and generally took the following forms: a) incomplete reading of primary source document; b) errant reading of primary source document; c) typographical errors; d) rounding errors; and e) misinterpretation of the coding instrument. Coding conflict was resolved for the eight cases, and problem items were inspected and corrected, where possible, for the entire sample.

It is important to note that although corrections were made to the data base when errors were found, program and investigator items were used for descriptive purpose only. They were not included as independent variables in any statistical analysis of intervention effect. This was also true of individual items which had low reliability scores within the other categories.

Data Base Management and Analysis

Since the database software used to manage the sample (dBase III)[127] was not equipped to handle extremely large data files, four partitioned files were created for data entry purposes. These files were then merged together using statistical software

(SPSS for Windows, 6.1), which is capable of handling a data set of this size. The data base was then cleaned and analyzed for two purposes: (a) descriptive analyses were run to present the characteristics of this meta-analytic sample (Chapter VI); and (b) focused analyses were run to answer the five questions posed at the end of Chapter II (Chapter VII).

[127] The investigator first attempted to link the four files using other data base management software (Microsoft Access), but the records were too large to combine more than any two at one time.

CHAPTER VI. RANDOMIZED FIELD EXPERIMENTS: DESCRIPTION OF THE META-ANALYTIC SAMPLE

This chapter is comprised of a census of information on the 150 experiments in the sample. As mentioned in the preceding chapter, these studies represent less than half of the eligible experiments retrieved between 1950-1993. Although the sample of experiments was selected without bias from the larger set, chance fluctuation alone could account for some differences between this sample and the population. Nonetheless, despite this compromise, the 150 experiments considered here represents the largest study of experimentation reported in criminal justice to date, and the fourth largest treatment effectiveness meta-analysis in the literature.

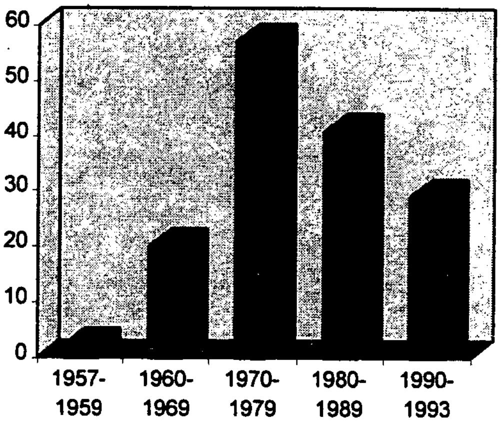
It should be noted that a considerable amount of methodological and other information for this chapter was missing from the experimental reports. This is a frequent problem faced in meta-analysis, but it does not necessarily mean that PIs failed to attend to them. It could be that these sections were edited out of journal articles or other documents in the interests of economy.

It should also be stated that this chapter has a purely descriptive goal. As such, considerations of effect size are not discussed until the next section, when the focused analyses are reported. In all figures, the number of experiments in the category--and not the percents--are provided unless otherwise indicated.

Document Information. It is clear that the multiple search methods described in Chapter IV were successful in retrieving a fairly diverse sample of experiments. This meta-analysis contains studies which were published between 1957-1993, available in seven different types of documents (e.g., journal, book, etc.), and obtained from 14 fields of study (criminal justice, psychology, medicine, etc.).

As Figure 1 indicates, while randomized studies in this sample were reported as far back as 1957, experiments were conducted at a rate of about two per year until 1970, when they began to increase to about five per year. It is interesting that the rate of experimentation is going up; although the 1990s only account for four years (1990-1993), 29 studies have already been reported (about seven per year).[128] The increase during the 1990s might well be due to the efforts by NIJ and other agencies to expand the use of experimentation in criminal justice during the mid-1980s (Weisburd and Garner, 1992); publication time lag likely accounts for the recent surge.

FIGURE 1.
Sample Experiments
By Decade of Publication



[128] Of course, these numbers are based on a random split of the total randomized studies eligible for this meta-analysis. If probability means anything, these numbers could safely be doubled.

Time lag simply refers to the time between the start of an experiment and when the findings are made available. The average delay for randomized experimental studies was five years, although the range was from six months to 25 years. Thus, even with a major effort to increase randomized field studies beginning in 1984, the time lag meant that many funded study reports were not published until 1990.

Clearly, as Table 4 indicates, criminal justice was the predominant discipline in which experimental studies meeting the criteria outlined in Chapter III were located. In all likelihood, these evaluations were cited in one of the field's abstracting or indexing services (e.g., Criminal Justice Abstracts, Criminal Justice Periodicals Index, etc.). However, nearly 41% of the sample came from publication outlets in other fields of study (e.g., psychology). This underscores the need for broad search techniques when conducting a treatment effectiveness meta-analysis.

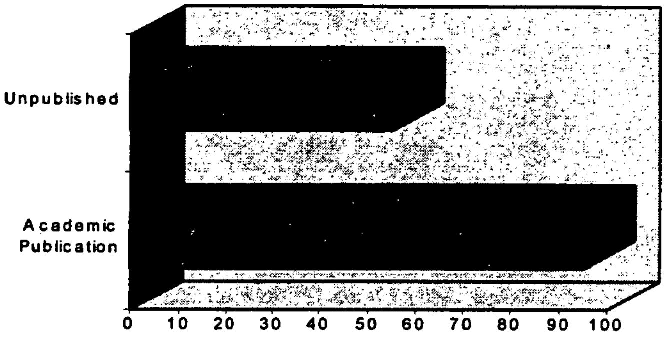
Table 4.

Experiments Located by Document Field of Study

Field of Study	Number of Experiments	Percentage of Total
Criminal Justice	91	60.7%
Psychology/Mental Health	22	14.7%
Social Work	11	7.3%
Social Science-General	5	3.3%
Alcohol/Substance Abuse	5	3.3%
Other	16	10.7%
TOTAL	150	100.0%

To guard against publication bias—that studies showing a significant treatment effect are more likely to be published in peer reviewed journals—a wide variety of documents were retrieved for this meta-analysis. While Figure 2 shows that most experiments in criminal justice were located in academic journals and books, nearly four in ten documents located were found in the 'fugitive literature' (37%). This was an appropriate strategy, given that Lipsey and Wilson (1993) found that peer-reviewed publications were considerably more effective in their sample of 302 meta-analyses than unpublished works.

FIGURE 2.
Experiments Found in Published and Unpublished Documents



Investigator Information. Approximately 60% of sample experiments were conducted by a team of two or more principal investigators (PIs). An analysis of available data on their positions and fields of study (for the 221 first and second PIs only) indicates that most experimenters were academicians within the field of criminal justice (Table 5). However, there was considerable variation; 19 distinct fields of study were represented.

Table 5.

Principal Investigators in Randomized Experiments By Position and Field of Study

PI Position	Number and Percent	PI Field of Study	Number and Percent
Academic	116 (53%)	Criminal Justice	93 (42%)
Government/Internal	48 (22%)	Psychology/Counseling	39 (18%)
Private Research	26 (11%)	Sociology	26 (11%)
Practitioner	21 (9%)	Social Science	15 (7%)

PI Position	Number and Percent	PI Field of Study	Number and Percent
Student/Post-Doc.	10 (5%)	Other	48 (22%)
TOTAL	221 (100%)	TOTAL	221 (100%)

A more crucial issue is whether the relationship of the PIs to the research setting may have influenced experimental results. It is often assumed that experimenters who developed the intervention and are heavily invested in its success may be less objective than outside evaluators hired to assess a program. Where information was available (N=122), exactly half of the experiments were conducted by external researchers and half by internal investigators. This distinction may be more imagined than real, as outside evaluators may lose objectivity for a host of reasons, including the desire to renew grant contracts with the host agency (e.g., Weiss, 1972).

Experiment Information. Based on this sample of experiments, randomized field tests have been conducted at nearly every stage of the criminal justice process. Most frequent are experiments at the latter end of the system; controlled studies with subjects under the formal supervision of the criminal justice system (prison, jail, probation and parole) accounted for 51% of all sample studies (Table 6). Glaser (1995) speculated that this should be expected. Random assignment studies within prisons are easier to maintain, since subjects are under the complete control of institutional authorities.

Table 6.

Experiments at Each Stage of the Criminal Justice System

Criminal Justice System Point	Total and Percentage
Pre-system	10 (7%)
Police arrest stage	4 (3%)
Post-arrest or police contact	8 (5%)
Pretrial supervision	8 (5%)
Juvenile/family court intake	14 (9%)
Prosecutorial stage	2 (1%)
Adjudication/Sentencing	6 (4%)
Post-sentencing/adjudication	18 (12%)
Probation	29 (19%)
Prison/Jail	39 (26%)
Parole/Release	12 (8%)
TOTAL	150 (100%).[129]

Information on how the experiment was funded was not frequently present in the research reports.[130] However, for 57 experiments where such data was available, 67% were funded solely by the federal government. Private, state and multiple sources comprised alternate funding types. The United State Department of Justice was the most frequent source for funding criminal justice experiments through its program offices (e.g., NIJ, BJA, OJJDP), providing grant support for 23 true experiments.

As noted in Figure 3, The English-only language requirement restricted experiments to four countries: United States (N=131), England (N=11), Canada (N=7) and Denmark (N=1). Research by Redondo, et al. (1996), Schumann (1997;

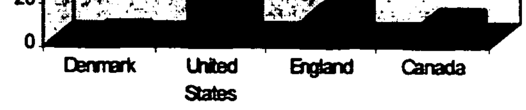
[129] Numbers do not add up to 100% due to rounding.

[130] Funding agencies normally request that grant support be acknowledged in any subsequent publications. However, it is possible that a number of experiments were supported without external funding.

1994) and Losel (1995) have shown that few experimental evaluations have yet to be conducted in other countries.[131] Within the United States, California was the most frequent site for experimental field studies, comprising 28% of all American-based randomized tests (N=37). This was likely due to a very ambitious research program during the 1960s-1970s within its state justice agencies (e.g., California Youth Authority). Michigan was the second most frequent setting, accounting for ten studies (8%), driven by the ambitious program of juvenile diversion experiments conducted by Michigan State University (Davidson, et al. 1993).

Two other characteristics of experiments are the length of the study and its intervention scope. The experiment duration was measured in months, from the beginning of subject randomization to the end of treatment (i.e., when the last randomized subject finished treatment). The average length of a randomized field experiment—excluding the follow-up period–was approximately two years (23 months), ranging from one month to 96 months.

Figure 3.
Experiments by Country of Origin



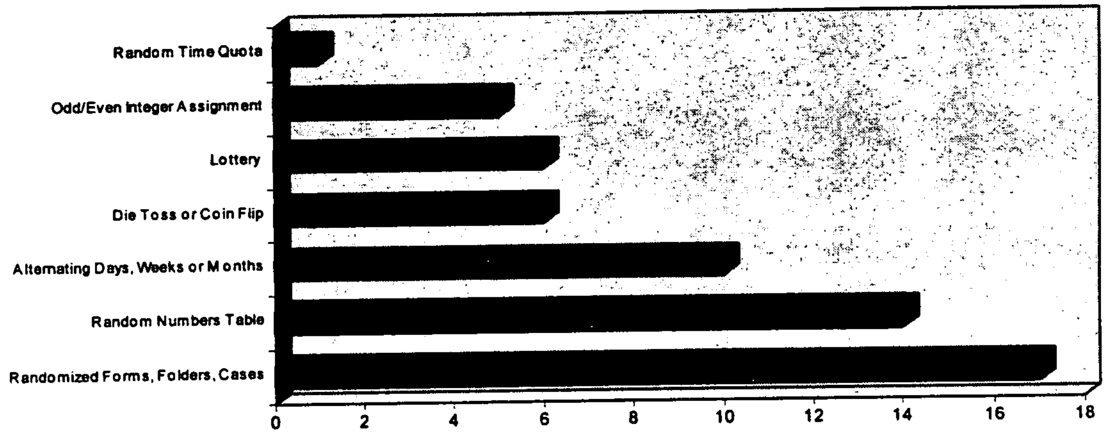
[131] In fact, Schumann (1997) noted in his personal correspondence that he located only one randomized study in Germany, while Redondo and his colleagues (1996) found only two randomized tests which included a follow-up measure. Losel (1995) found two randomized tests of West German penal programs.

The interventions tested by these experiments had generally wide geographic scope: eleven were statewide interventions, 43 were countywide, 35 were citywide and 26 were conducted in multiple counties, cities or institutions. Only 22% of the sample studies tested an intervention limited in scope to a single institution (e.g., prison, probation office, etc.). This may be an important methodological factor, particularly if wide scale experiments are more difficult to control than single institution studies.

Fifty-two experiments (35%) were considered multisite experiments, where subjects were randomly allocated at more than one setting to an experimental and control group condition. For example, the Arizona Pretrial Drug Testing Experiments (e.g., Britt, Gottfredson & Goldkamp, 1992) were actually four separate randomized studies conducted across two Arizona counties. These were certainly multisite studies, a designation that was helpful in capturing the complexity of some of the sample evaluations. In most cases, the PIs analyzed and reported on the experiments as distinct, and only in one case was a combined analysis of the sites reported (e.g., Fagan, 1990).

Randomization Information. Despite the importance of random assignment to the experimental design, information on how randomization was accomplished was generally lacking in this sample. This was a factor in Dennis' (1988) decision to use telephone interviews rather than published reports to analyze experimental studies. In fact, the method of randomization was not reported in 61% of the cases. The reported data indicates, however, that PIs were generally resourceful in the methods they utilized to randomize subjects, ranging from die toss and coin flips to randomized time quotas (Figure 4).

Figure 4.
Methods of Random Assignment



Only 28 randomized tests (19%) employed either blocking, matching or stratification techniques to increase the likelihood of equivalent study groups. The PIs generally performed these techniques using few variables (mean=3), although this ranged from one factor to as many as eight. While a number of different variables were used in these methods, the most common were age, race, sex, and criminal history.

Matching was used in nine studies (6%). Matching was also implemented as a control against differential attrition effects; if the experimental subject dropped out, the matched control would be deleted from the analysis also. Stratification was used in 18 studies (12%), generally to insure that groups would have similar proportions of individuals with certain racial, age, gender or other characteristics. Stratification does not attempt to link individuals on a 1:1 basis, but is designed to produce conditions with similar proportions of desired subjects (e.g., experimental and control groups with a minimum of 25% female subjects each). Finally, blocking was used in only one experiment, although its use has been urged to reduce heterogeneity across settings in multisite designs (Weisburd, 1996).[132]

While the use of matching and other techniques was infrequent in this sample, group equivalency pretests were reported for 71% of the experiments. Perhaps the infrequency for which matching and other methods are used in combination with randomization is appropriate. It does not appear that post-randomization differences–which lead to a bias toward one group or the other–is a problem for criminal justice experiments (Table 7). For the 107 studies reporting such a measure, group differences favoring one condition over the other were infrequent (N=19). It is important to note that the number of variables pretested ranged considerably across experiments; some PIs only pretested one or a few factors, while others looked at several dozen or more.

Table 7.

Results of Pretests for Group Equivalency

Pretest Result	Total and Percentage
No differences on measured variables	71 (66%)
Differences, but not substantive	7 (7%)

[132] Blocking on a key variable would allow the PI to statistically reduce the noise introduced by heterogeneity on some third factor in the experimental design.

Pretest Result	Total and Percentage
Ambiguous, differences can favor either group	10 (9%)
Differences clearly favor experimental group	10 (9%)
Differences clearly favor control group	9 (8%)
TOTAL	107 (100%)

Principal investigators reported a breakdown in the random assignment process in only 24 cases (16%). This coincides with the high percentage of null findings on group equivalency pretests. Unfortunately, only 13 studies reported the percentage of random assignment error. Randomization breakdown resulted in an average misassignment rate of 14%, ranging from 1%-48% contamination. When overrides or contamination of the random assignment process did occur, PIs generally took one of four strategies (e.g., Gartin, 1995):[133]

(1) *analyzed the groups as randomly assigned and ignored the actual treatment delivery* (N=10);

(2) *analyzed both groups as treatment assigned and treatment delivered and determined differences* (N=5);

(3) *analyzed the groups as treatment delivered and ignored treatment assigned* (N=1);

(4) *deleted misassigned cases from the experiment* (N=1).[134]

Other Methodological Information. Aside from randomization, other methodological issues are important in executing a randomized experiment with full integrity (e.g.,

[133] Unfortunately, in seven cases, it was not stated what remedies the PI(s) implemented to address random assignment errors.

[134] It was rare for any PIs to report that specialized statistical corrections were made to adjust for any differences found in treatment-assigned and treatment-delivered analyses.

Farrington, 1983). Three crucial factors are attrition, caseflow and statistical power. Attrition refers to the number of subjects lost from the experiment, post-randomization. As noted by Gartin (1995) and others, attrition poses unique problems for the investigator, particularly in how to handle treatment drop-outs in the statistical analysis. Unfortunately, there is still no consensus among scholars on how to proceed.

Caseflow is the number of subjects received into the experimental program, and becomes critical when the number of cases is far below that initially expected, forcing the PI to modify the study. Finally, statistical power refers the ability of the analysis to avoid Type II error, i.e., failing to detect a significant difference when one is truly present. As mentioned earlier, most social science evaluations are not powerful enough to detect the small to moderate treatment effects found in meta-analytic studies (e.g., Lipsey, 1990).

According to PI reports, attrition was a slightly more common problem than randomization breakdown or misassignment. Approximately one-third of the sample studies experienced more than a minor loss of subjects from the experiment following random assignment.[135] The reasons for attrition were varied, but for those experiments where data was available, treatment drop-outs were the most common factor in producing a loss of experimental subjects (Table 8).

[135] Minor attrition was the loss of less than 5% of the total sample, spread out equitably across the groups (i.e., no differential attrition).

Table 8.

Reasons for Attrition in Randomized Experiments

Reason for Attrition	Number and Percentage
Treatment drop-out or no-show	26 (58%)
Could not track subject in follow-up	9 (20%)
Practitioner or administrative problems	5 (11%)
Insufficient treatment exposure	4 (9%)
Refusal to participate after randomization	1 (2%)
TOTAL	45 (100%)

When attrition did occur, PIs reported several methods for analyzing subjects in the remaining analysis. Some of these methods overlap with earlier techniques used by PIs to handle random assignment failures. Including minor attrition cases (N=19), experimenters used one of the following five techniques to handle analysis problems after loss of subjects:

(1) *deleted lost cases and analyzed available subjects only* (36 cases);

(2) *analyzed as originally assigned and ignored treatment delivery* (19 cases);

(3) *analyzed treatment-assigned and treatment-delivered cases to determine if attrition affected results* (4 cases);

(4) *compared treatment drop-outs to treatment subjects to see if different* (3 cases);

(5) *replaced subjects in experiment* (2 cases).

Caseflow problems, unlike attrition, were less common in this sample of experiments (18%). Only 26 experiments experienced an insufficient or lower number of study subjects. The most common rationale for caseflow inadequacy was an inaccurate estimate on how many subjects would be eligible for the experiment; these pre-experimental plans sometimes overestimated potential clientele (Table 9).

Table 9.

Reason for Caseflow Problem

Reason for Caseflow Problem	Number and Percentage
Inaccurate estimate of eligibility pool	6 (23%)
Low referral rate to experimental group	5 (19%)
Practitioners increased exclusion rate	5 (19%)
Treatment program administration problems	4 (16%)
Low proportion of cases from certain regions	2 (8%)
All other reasons	4 (15%)
TOTAL	26 (100%)

While insufficient caseflow problems are not as critical to the experimental analysis as randomization breakdown or attrition, they pose practical threats to the project. Generally, funding or agency resources are devoted to a program, staff are hired and money allocated based on an estimate on the number of clientele expected. When caseflow is insufficient, modifications in the design or operation must be made to keep the experiment running. In this sample, PIs handled caseflow problems using one of the following nine adjustments:

(1) *accepted less subjects than originally designed (N=7)*;

(2) *adjusted randomization of subjects to experimental group from .5 to higher proportions (N=6)*;

(3) *disbanded random assignment and put all subjects in experimental group (N=3)*;

(4) relaxed eligibility criteria and accepted more subjects into experiment (N=3);

(5) extended experiment length to acquire more subjects (N=2);

(6) created specialized unit to increase cases (N=2);

(7) pressured practitioners to increase caseflow (N=1);

(8) collapsed multiple treatment groups into single group (N=1);

(9) dropped second experiment in multisite design (N=1).

Statistical power was an infrequently discussed topic in experimental reports; in fact, only eleven studies alluded to it at all (7%). This is discouraging, given the renewed attention to the problems of low power in the social sciences (e.g., Weisburd, 1993; Lipsey, 1990). In each case, mention of statistical power was only made in response to the failure to find statistically significant results, leading PIs to note that "results would have been significant" if the sample size had been larger (e.g., Stratton, 1975; Venezia, 1973). However, not a single PI mentioned the use of statistical power analysis—in the planning stage– to design a more powerful study, a practice becoming more common in experimental research (e.g., Weisburd, 1993; Lipsey, 1990; Weisburd, 1989).[136]

Sample Selection Information. The fact that only 50 randomized studies reported using voluntary subjects (35%) hints at the coercive nature of many sample interventions.[137]

[136] In fact, in Weisburd's (1989) proposal to evaluate the Project Muster restitution experiment, he includes estimates of statistical power with each potential sample size.

[137] Experiments where the subject could refuse an intervention and face a harsher sanction were considered coercive.

Some standard research texts would not classify voluntary subject experiments as true controlled studies, and they are open to attack on external validity grounds, if volunteers differ from non-volunteers on important dimensions (e.g., Weisburd, 1995).

On the other hand, the voluntary nature of an experiment only affects external validity if it reflects a difference in how subjects would normally be handled. Many of the experiments that allow individuals to volunteer or refuse assigned treatment are no different than actual policy. For example, targeted youths can opt not to receive prevention program services, while those under the supervision of the criminal justice system may sometimes be coerced to undergo particular interventions.[138]

None of the sample 150 experiments reported using a random sampling procedure from the eligibility pool to select study subjects. This was because PIs nearly always fashioned an elaborate set of inclusion and exclusion criteria (or combination of both) to select subjects (92%);[139] all subjects meeting these criteria were then randomly assigned to the conditions of the experiment. Given the need to insure adequate caseflow, random rather than total selection into the experiment would have posed operational problems. Sample selection criteria generally focused on:

- *limiting study to specific subjects likely to receive intervention in practice (e.g., age groups, instant offenses, criminal histories, criminal justice system stage)*

[138] American courts have generally recognized the prisoner's right to refuse treatment (e.g., Winnick, 1981).

[139] Only 13 randomized tests did not report the eligibility criteria for selecting subjects into the study (8%).

- *decreasing danger to the community (e.g., no serious or habitual offenders)*
- *decreasing administrative problems in experiment (e.g., no transfers, parental consent, non-English speaking, etc.)*
- *decreasing geographic range of study (e.g., limit to certain areas)*
- *decreasing number of inappropriate treatment subjects (e.g., psychopaths, low I.Q., low-risk cases unlikely to benefit, etc.)*
- *decreasing number of controversial cases (e.g., high-profile)*

Only 14 experiments reported the use of financial incentives for subjects to participate in some aspect of the study (9%). However, in eleven cases, nominal fees (usually $5-$10) were paid for completed interviews or questionnaires. In the other three cases, individuals were given monetary rewards for showing up every day on time for the intervention (Ostrom, et al. 1971), paid to work on teaching machines and evaluate their operation (Hackler and Hagan, 1975), or given merit pay to participate in the different institutional regimes (Craft, 1964).

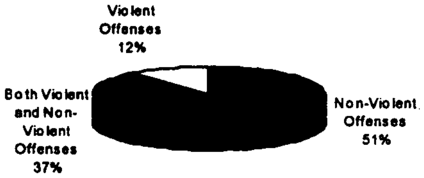
Subject Demographic Information. The 150 sample experiments were comprised of a total of 385 randomized groups (median=2 groups, mean=2.6 groups per study)[140]; the preceding chapter elaborated on the guidelines used to handle multiple group design problems in creating a common metric effect size. These experiments included 82,825

[140] A few experiments also included a non-randomized comparison group in their studies (e.g., Ku, 1976).

subjects who were randomly assigned to conditions[141]; the median total sample size per randomized field test was 193, ranging from 19 to 11,976.[142]

Experiments generally included mostly male subject samples (mean=86% per study), with diverse racial backgrounds (mean=45% white per study) and a mean age per experiment of 21.[143] Nearly half of the experiments focused on adult subjects (47%), while the remaining studies included only juveniles or a combination of teenagers and young adults (under 21). Subject education levels, mean IQs sand socioeconomic status were so frequently missing that statistical analyses were not run using these factors.

Figure 5.
Experimental Samples: Subject Instant Offenses



These experiments tested interventions for a wide range of individuals, from toddlers in specialized daycare facilities to serious violent offenders in prison (Figure 5). Excluding prevention experiments, the types of instant offenses (the crime of arrest,

[141] Interesting that ten experiments did not include the number of subjects assigned to conditions, and two only reported the subjects allocated to the experimental condition. Outcome totals were reported, however, or else effect size could not be computed without an estimation of sample size.

[142] The median is reported instead of the mean, which is heavily influenced by several huge outliars (mean=600).

[143] Unfortunately, missing data was common for subject demographic data.

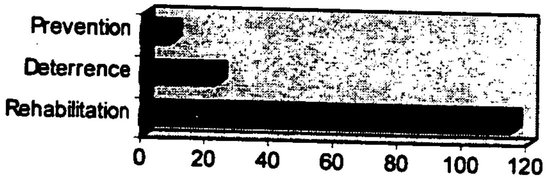
conviction or incarceration) which experimental subjects committed ranged from motor vehicle infractions to murder. As Figure 5 indicates, most experimental samples included only non-violent offenders; however, 49% of the experiments included at least some subjects who committed crimes against persons.

The prior criminal histories of experimental subjects were often not explicated in the reports, and similar to instant offense descriptions, had to be coded in a broad manner (extensive versus low/moderate). Using this categorization, 48% of the randomized studies included subjects with extensive prior records; generally, experiments with extensive prior record samples were comprised of individuals with three or more prior arrests.

Most experiments dealt with a heterogeneous pool of instant offenders (e.g., property and person offenders); only 26% of the studies focused on a single offense type (e.g., driving while intoxicated, sex offender, truant, status offender, etc.). Arguments for homogeneous pools for treatment considerations continue to be made (e.g., Pallone, 1990), but the utility of this approach is rooted in whether there is crime-specific or crime-general motivations which drive offending (e.g., Weisburd, Sherman and Maher, 1992). If the crime-general approach is valid, then one underlying cause such as low self-control or impulsivity is motivating all types of offenses (e.g., Gottfredson and Hirschi, 1990).

Program Information. As mentioned in Chapter I, randomized experiments were found in three major philosophical categories: rehabilitation, deterrence and delinquency prevention.[144] Coders classified interventions into one of these categories based on the major rationale provided by PIs for why the program should affect crime outcomes. When the reason provided was punitive, such as the Minneapolis Domestic Violence Experiment (e.g., Sherman and Berk, 1984), the experiment was coded as "deterrence." Controlled experiments that evaluated interventions with juveniles before they came into official contact with the criminal justice system were categorized as "prevention" studies. This would include studies such as the containment theory classroom program conducted in Ohio middle schools in the 1960s (e.g., Reckless and Dinitz, 1972). Finally, studies of interventions with a treatment focus were considered "rehabilitation" experiments, such as Kassenbaum, et al.'s (1971) evaluation of a group counseling program in the California penal system.

Figure 6. Randomized Experiments by Broad Philosophical Category



As Figure 6 indicates, most studies were categorized as rehabilitation experiments (78%). However, this category included 20 randomized tests of interventions which could arguably be considered

[144] There were three experiments, however, which could not be classified into one of these categories and were not included in this analysis.

deterrent studies.[145] This would include the reduced caseload/increased contact on probation and parole experiments of the 1950-1985 period, when the focus–according to PIs–was rehabilitative rather than deterrent. Of course, the move toward punishing smarter has led to reinvestment in these increased contact interventions, with an eye toward surveillance and control rather than meeting offender needs (e.g., Gendreau and Goggin, 1996). The Intensive Supervision Probation and Parole (ISP) experiments of the 1980-1990 period were clearly deterrent in focus and coded as such. [146]

These broad categories mask a myriad of intervention strategies designed to reduce criminal behavior. As other meta-analysts have noted, creating a smaller group of intervention types is challenging, given the eclectic nature of the programs under evaluation (e.g., Lipsey, 1992a; Whitehead and Lab, 1989). To further specify the type of interventions studied in these experiments, the independent variable was coded in detail. Unfortunately, a paucity of information describing treatment exists, a problem that affected coding of information on staff, training, treatment duration and intensity.

[145] The analysis strategies for handling this problem are described in Chapter VII.

[146] The analyses in Chapter VII were conducted both with and without these 20 experiments in the rehabilitation category, and the substantive findings did not alter.

This may explain the low rate of intercoder reliability on program information noted in Chapter 5.[147]

Figure 7. Interventions Tested by Sample Randomized Experiments (N) In Alphabetical Order

Arrest or Warrant	(5)	Individual-based Psychological Treatment	(4)
Casework	(6)	Institutional Change/Therapeutic Communities	(4)
Cognitive/Behavioral	(14)	Intensive Supervision Probation or Parole (ISP)	(3)
Community-Based/Residential	(12)	Juvenile Tours or "Scared Straight" Programs	(5)
Community Service	(2)	Medical/Pharmacological Treatment	(6)
Crisis Intervention	(2)	Payment and Renumeration	(2)
Diversion, With/Without Service	(14)	School-based Prevention Programs	(5)
Drug Urine Testing	(6)	Supervision and Treatment	(6)
Education-based Treatment	(2)	Vocational/Educational	(9)
Family-based Counseling	(4)	Volunteers as Counselors	(4)
Forestry/Wilderness Programs	(3)	Other Sanctions	(4)
Group Counseling	(8)	Other Treatments	(4)
Increased Contact/Reduced Caseload Supervision	(15)	Not Classified	(2)

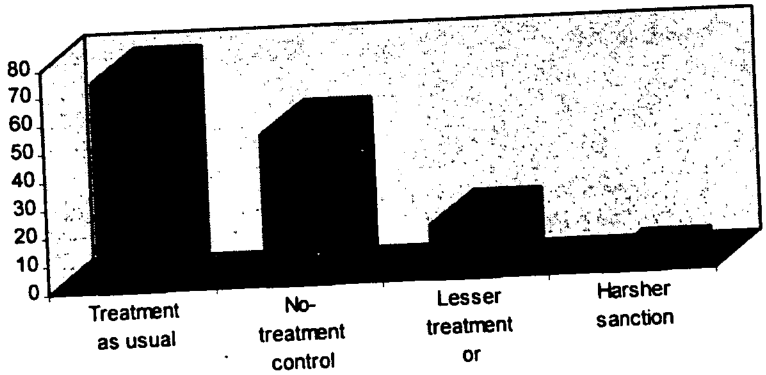
As Figure 7 denotes, even after grouping similar programs together, 26 intervention categories are still represented in this meta-analysis. It appears that reduced caseload/increased contact (N = 15), diversion with or without services (N = 14),[148] cognitive/behavioral treatment (N = 14), and community-based/residential programs (N = 12) are the most frequent intervention types under experimental evaluation.

[147] In fact, even after corrections to the data, so little useful information from the treatment delivery, contact and agency items was available that they were deleted from the analysis.

[148] A diversion experiment reported by Klein (1986) included four groups, two of which received services. However, using the two group rule reported in Chapter 5, only the strongest/weakest conditions were compared, which was diversion without service versus petition to juvenile court.

Approximately 29% of sample experiments were conducted in an institution such as a prison or jail; the remainder were community-based, with subjects who were not confined at all. This has been an important factor in prior meta-analyses, where several have found larger effect sizes for community programs rather than institutional ones (e.g., Lipsey, 1992a).

Figure 8. Randomized Experiments by Type of Control Condition



It is also important to note that randomized experiments in criminal justice do not often test an intervention against a no-treatment control group. As Figure 8 shows, the control group is frequently the standard or normal treatment subjects received before the program (50%). Thus, it is often the case that control subjects are receiving some type of intervention, which is an important contrast with the no-treatment or placebo conditions more frequent in psychological or medical research (e.g., Lipsey, 1990).

Most PIs reported conducting some attempt at monitoring study conditions during the experiment (69%). Techniques ranged from simple written logs for parole and probation officers to record the number of contacts with clientele to formal observation of treatment delivery. Though not common, major treatment problems were noted in 47 randomized studies (31%).[149] Most common were difficulties in implementing a part of the planned treatment program (Table 10). Except for the crossover, misassignment and drop-out problems which were mentioned in earlier sections on randomization and attrition, PIs did not note any strategies for dealing with these issues after they occurred.

Table 10.
Major Treatment Problems in Randomized Field Experiments

Major Treatment Problem	**Number and Percentage**
Treatment component never implemented as planned	16 (34%)
Insufficient dosage for all or some subjects	13 (28%)
Crossover, misassignment or drop-outs	9 (19%)
Administrative or internal project conflict	5 (11%)
Indistinguishable experimental and control conditions	4 (9%)
TOTAL	47 (100%)

General Outcome Information. Sample experiments generally tested the impact of intervention on four measures of programmatic success: two criminal outcomes (e.g., rearrest) and two non-crime (e.g., psychological or attitudinal tests).[150] While crime

[149] The most serious treatment problem was used if the PI listed more than one.

[150] There were 366 total crime outcomes and 445 non-crime measures utilized; of course, many of them were the same (e.g., rearrest).

measures used in a single experiment ranged from 1-7, non-crime measures ranged from 0-38. Non-crime outcomes included employment or public assistance, education or school measures, lifestyle stability, attitude change, various psychological tests, ratings by staff or significant others, accidents, fines and restitution paid, institutional disciplinary infractions or incident reports (non-criminal); and program completion or use of services.[151]

PIs were resourceful in developing crime outcomes; while rearrest was the most common, 97 different constructs of subsequent offending were used. These included variations of arrests, police contacts, rebookings, new charges, offense severity, court referrals, petitions, probation and parole violations and revocations, adjudications or convictions, dispositions, disposition severity, incarcerations in jail or prison, drug use, self-reported criminality, unfavorable parole discharges, moving violations and traffic tickets, abscondings or escapes, victim reports, hotline calls, and illegal income. As Chapter V noted, there were a multitude of ways PIs chose to quantify this data, and rules had to be developed to systematically choose the most important and standard outcome information.

PIs reported 238 total follow-ups, ranging from one to eight distinct time intervals in an experiment. Repeated measurement was a rare occurrence in this sample; only six studies had more than three follow-up periods (4%). The guidelines

[151] These do not include the process or system impact analyses which PIs sometimes reported. For example, cost/benefit estimates, since they pertain to a

discussed in Chapter V for selecting a maximum of three time intervals were therefore not invoked in nearly all sample cases (96%).

This is a major deficit of experimental outcome research in general, leading Sherman, et al. (1997) to advocate funding to sustain additional follow-ups. In essence, this has resulted in some meta-analyses such as this one being labeled as 'first-effects' studies (e.g., Pearson, et al., 1996; Lipsey, 1992a).

Specific Outcome Information. It was clear that a narrative or more sophisticated vote-counting review of this data that relied upon statistical significance as the criterion for success would argue that 'little or nothing works.' Indeed, only 28 experiments reported a statistically significant difference between the experimental and control groups (19%), with nearly all of these favoring the experimental group. It is precisely this type of result which has led many to advocate for meta-analysis in treatment effectiveness research over traditional reviewing methods, and effect size over null hypothesis tests (e.g., Cohen, 1992; Rosenthal, 1991).

The Sellin rule for handling multiple outcome measures selected the most inclusive criterion reported; almost six in ten experiments used police contact or arrest as an impact variable (Table 11). Nonetheless, if court data or other outcomes were the only available, then they were utilized.

program rather than individual measurement, were excluded in this discussion.

Table 11.

Type of Outcome Construct Used for Common Metric Effect Size For Each Follow-up Period

Type of Outcome Construct Used for Effect Size	First Follow-up: Total and Percentage	Second Follow-up: Total and Percentage	Third Follow-up: Total and Percentage
Arrest or Police Contact	88 (59%)	29 (60%)	14 (66%)
Conviction or Court Contact	24 (16%)	8 (16%)	3 (14%)
Parole or Probation Violation/ Revocation	15 (10%)	5 (10%)	2 (10%)
Incarceration	8 (5%)	2 (4%)	1 (5%)
Other Outcomes	15 (10%)	5 (10%)	1 (5%)
Total	150 (100%)	49 (100%)	21 (100%)

While Sherman, et al. (1997) and others have urged greater attention be paid to longer more frequent follow-ups, most experiments had a single follow-up period (67%), ranging from three to 134 months. The median first follow-up period was a year. For those experiments which reported a second time interval (33%), the median was 18 months. Only 21 experiments had a third follow-up period (14%); again, the median was 18 months.[152]

Proportion and frequency data were the most frequently reported and utilized data in this meta-analysis (Table 12). However, when raw percentages or frequencies were not available, test values from significance tests or the means (and standard deviations) were utilized to create a common metric effect size.

[152] The median is used since the mean was heavily affected by huge outliars (7 studies had field tests between 120-134 months).

Table 12.

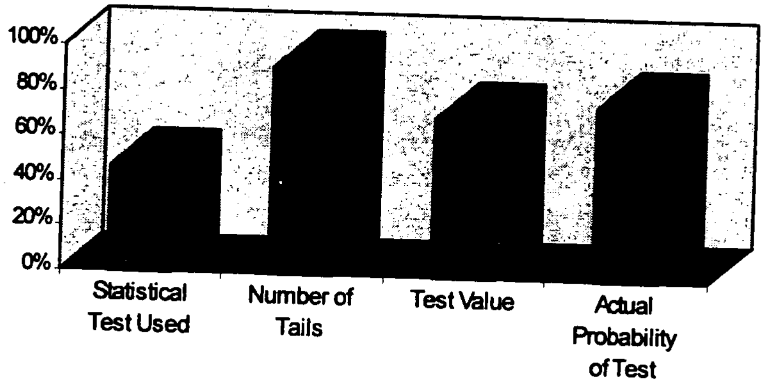
Type of Quantified Data Used for Computing Common Metric Effect Size For Each Follow-up Period

Type of Data Used For Common Metric Effect Size	First Follow-up: Total and Percentage	Second Follow-up:Total and Percentage	Third Follow-up: Total and Percentage
Proportions/Frequencies	138 (92%)	45 (92%)	19 (90%)
Chi, T, F, or Z Test Value	8 (5%)	2 (4%)	1 (5%)
Means & Standard Deviations	4 (3%)	2 (4%)	1 (5%)
Total	150 (100%)	49 (100%)	21 (100%)

Although not needed for common metric effect size, data on the significance test used by PIs, the test value, probability level and number of tails was also extracted from each experimental report. It was hoped that this information could be used to reanalyze the experiment, and verify the statistical findings.[153] Unfortunately, this data was plagued so greatly by missing values that it could not be used (Figure 9). When tests were reported, chi square was the most frequently reported statistical test used (75%). It should also be noted that assumptions about experimental tests being two-tailed tests might be wrong; of the ten PIs reported the number of tails in the significance test, two used more liberal one-tail tests to analyze their data. The use of one-tailed tests might be appropriate, given the low power of some experiments to detect small or moderate effects.

[153] For example, the chi value provided in one report was computed incorrectly, and was redone by this investigator. Although statistical significance did not change, the effect for the intervention did.

Figure 9. Percentage of Sample with Missing Outcome Information, For First Follow-up Period



Subgroup Effect Information. It is also surprising that despite the number of scholars who have written influentially on differential effects (e.g., Palmer, 1992; Wilson, 1980; Adams, 1970), few subgroup analyses were reported from sample reports.[154] In fact, only 55 randomized studies reported any statistical information on the effect of intervention for certain categories of individuals (37%). Determining 'what works with whom'—which has been called the next frontier of meta-analytic research (e.g., Palmer, 1994; Lipsey, 1992a, 1992b)–is going to be difficult without additional studies with explicit subgroup information.

When PIs did test subgroup effects, they generally focused on traditional variables: age, race, gender, prior record, treatment exposure, personality or other psychological scales, and behavioral classification systems. While there is a scattering

[154] A subgroup analysis was defined as statistical information on the effects of treatment or intervention on certain categories of individuals (e.g., race, gender, age, etc.). Other statistical analyses which did not involve an examination of differential effects, such as risk prediction for the entire sample (i.e., regardless of group assignment) were not coded.

of statistically significant findings, the results are so diverse that it would be difficult to find one factor significant across any two studies.

It is unfortunate that some subgroup analyses were handled poorly. Some PIs analyzed scores of variables, not attending to the possibility that several would be significant by probability alone, a strategy referred to as "capitalizing on chance" (e.g., Cooper and Hedges, 1994). When a few factors are significant out of a hundred, it is hard to make a case for a finding of practical or theoretical import. Other PIs only tested one variable, such as treatment exposure, and did not explore differential effects on other dimensions (e.g., race, age, etc.). It was also the case that subgroup analyses were sometimes ambiguous, resulting in a few significant findings favoring experimental subjects and others favoring controls with no attempt by the PIs to describe why this would occur.

CHAPTER VII. 'What Works?' Revisited Again: Answers to Focused Questions

In this chapter, the results of several focused analyses are reported to shed additional light on the 'what works?' debate. As with any research endeavor, prior steps build toward this final phase. Research is similar to a house of cards, with each stage precariously reliant upon prior ones. Meta-analysis adds an additional layer, since it utilizes the original reports of primary investigators. In neither primary research or meta-analysis can exemplary statistical methods control for bad decisions made at earlier points.

Given that logic, it makes sense to summarize what has been done to this point. Specific inclusion criteria were developed to collect a series of randomized experimental studies, all of which tested the effect of some individual-level intervention on a quantified measure of official crime. A variety of search techniques were utilized to develop a representative sample and control for certain biases (e.g., publication bias). The descriptive statistics presented in Chapter VI indicate that these methods were successful in gathering a large and diverse set of experiments, from both published and unpublished sources.

A coding instrument was developed in accordance with prior literature and data was extracted from each experiment. As explicated in Chapter V, guidelines were developed to handle troublesome coding decisions, to increase reliability and replication. A specialized software program was utilized to perform the necessary computations for effect size. The coding reliability test indicates that the most crucial items (e.g., outcome data) had the highest rate of agreement across coders. Item groupings with low reliability–even after corrections (program, investigator)–were not used in any of the statistical analyses of effect size described in this chapter.

Also elaborated upon in Chapter V, the common metric effect size selected was Cohen's *d*, which is simply the experimental effect divided by the pooled standard deviation. It should be noted that the use of all other quantified data, such as proportions and test values are estimates of *d*. This is necessary, particularly in criminal justice, since so few experiments report the means and standard deviations necessary to compute*d*[154]. While these estimation procedures are, from a statistical purist's vantage point, less precise, they have been found to estimate *d* accurately, with a margin error of approximately .01 or less (e.g., Johnson, 1989). This error can be bi-directional and should be negligible with a sample of 150 studies.

The effect sizes in this meta-analysis are expressed in three ways. In rare cases where there was absolutely no treatment effect, $d=0$. Otherwise, where the treatment group outperformed the control group, *d* is expressed as a negative value (e.g., -.50) to indicate a decrease in crime. Where the experimental group performed worse than the controls, *d* is expressed as a positive value (e.g., .50) to indicate an increase in criminal

[154] Psychology and education research often utilizes test score research, where means and deviations are universally reported.

behavior. While this is different than the way prior meta-analysts described results, it appears to be intuitively correct to express a decrease in crime as a negative.[155]

It should also be noted that this meta-analysis, like many before it (e.g., Pearson, et al., 1996, 1995; Lipsey, 1992a), represents a "first-effects" study. As discussed in Chapter VI, a minority of experiments reported outcome results beyond a single follow-up period. Averaging the effects across multiple follow-up periods seemed unwise, particularly since it is assumed that experimental effects dissipate over time. To the extent that treatment effects eventually weaken, studies reporting several time interval measures would be penalized by averaging across all outcomes.[156]

A review of statistical methods in meta-analysis

Quantitative techniques for synthesizing the results of independent studies have been available since the 1930s (e.g., Hedges and Olkin, 1985). However, these methods were developed and utilized in the agricultural and physical sciences and were not readily adopted by social sciences (e.g., Hedges and Olkin, 1985). It was not until the work of Gene Glass (1976) that statistical techniques were applied with regularity to fields like

[155] Lipsey (1992a), for example, expressed decreases in crime as a positive value. When comparing the results here with prior meta-analyses, the signs in prior meta-analyses are reversed to standardize them with results reported here.

[156] If it is true that treatment effects are reduced over time, studies with multiple follow-ups would have lower mean effects than studies with only one follow-up measurement.

education and psychology. This section briefly reviews some of the more popular methods and the rational for selecting the Hedges and Olkin (1985) technique.

One meta-analytic technique is known as the 'omnibus' or combined significance test. While many omnibus test methods are available, they have the same goal. The omnibus null hypothesis is that there is no effect in any study in the sample; the reviewer combines statistical significance or p-values from several independent studies to determine whether the null hypothesis is rejected or affirmed (e.g., Hedges and Olkin, 1985).

A different type of omnibus test was later popularized by Rosenthal (1991). Using this technique, the investigator computes the one-tail probability for each study and its corresponding Z-score. The cumulative Z-score is used to test whether the null hypothesis of no statistically significant treatment effect across all combined subjects in the studies is supported (e.g., Bangert-Drowns, 1986). Omnibus tests have not been as widespread as other techniques since they do not indicate the magnitude or consistency of an effect.[157]

[157] Frequently, the investigator knows that a series of experiments produces some statistically significant positive and negative results, as well as zero results. The research question is not whether there is an effect in any one study, since it is clear that there was, but what is the average size of the effect across studies and how they vary--if they do--across study characteristics.

Glass and his colleagues (Smith, Glass and McGaw, 1981) applied a standardized measure of effect which was essentially scale-free,[158] so that studies utilizing different outcome constructs could be compared (i.e., common metric). Essentially, effect size revealed how many standard deviations the experimental group performed better or worse on outcome measures than the control group. Once the common metric was developed, Glass, et al. (1981) argued that the data could be handled by conventional statistical methods; they recommended multivariate regression be used to explore the relationship between study-level characteristics and effect size.

Glassian meta-analysis came under attack by methodologists, who argued that their method violated the assumptions of parametric techniques (e.g., regression). For example, the homoscedascity assumption about variance has been shown to be seriously violated in Glassian meta-analysis (e.g., Hedges and Olkin, 1985). Glass replied that these techniques should be used in exploratory fashion instead of attempting to estimate true population effects (Glass, 1995). In addition, the Glassian technique does not include a method for assessing effect size variation to determine if regression analyses of study characteristics are even necessary (e.g., Hedges and Olkin, 1985).[159]

[158] While they advocated the use of Cohen's *d*, Glass and his colleagues (1981) developed their own version of *d*, which is the difference of experimental and control groups divided by the standard deviation of the control group only.

[159] As explained later in the chapter, if the average effect is based on a sample of homogenous effect sizes, then exploratory analyses may not needed to determine where the source of variance is, since there will is little.

Parallel to the work of both Glass and Rosenthal, Hunter and his colleagues (Hunter, Schmidt & Jackson, 1982) developed a meta-analytic technique which applied corrections for statistical artifacts (e.g., sampling error, outcome measure unreliability, invalidity, etc.) to the data. However, these researchers were working in an area (organizational psychology) where several outcome measures (e.g., tests) had reliability and validity estimates which could be used in their formulas to make corrections.

In large measure, the Hunter and Schmidt (1990) methods have been viewed as unrealistic in social science settings, where the true reliability of the outcome variable is often unknown (Hedges, 1992).[160] As noted by Bangert-Drowns (1986), their corrective formulas for dependent measures are impractical because they require information that is rarely available, such as the reliabilities of the criterion and a valid response variable (and the correlation between the two measures to assess validity). In fact, if one applied the corrections advocated by Hunter and Schmidt (1990) using Lipsey's (1990) estimate of .30 reliability for official arrest data, the resulting effect sizes would be much larger than the small to moderate effects normally seen in criminal justice. On its face, it does not seem possible that criminal justice programs could achieve such enormous effects–only attenuated by poor criterion variables.

[160] We know arrest is an insensitive measure, since many persons who reoffend are never arrested by police. Yet, how much of this insensitivity should be corrected in meta-analysis? In addition, randomized experiments should result–though not always–in groups which are measured equally with insensitive measures.

Hedges and Olkin (1985) developed one of the more elaborate statistical treatments of research synthesis to date. Their meta-analytic technique produces an average effect size, which is corrected for small sample bias, referred to as the n-adjusted effect size. The n-adjusted effect is then weighted according to sample size, since larger samples produce more precise estimates than smaller samples. Homogeneity tests are used to determine how well the average effect size represents the full set of studies; if large heterogeneity exists, the investigator searches for explanatory variables using study-level characteristics.

This method was considered to be best-suited for the set of data synthesized here. Since the five focused questions guiding this analysis required information on effect magnitude, omnibus tests were ruled out. Glassian meta-analysis was not a viable option, given the statistical criticisms of the techniques. Hunter and Schmidt's (1990) methodology was impractical, given the lack of information about statistical artifacts in any criminal justice study, let alone 150.[161]

Question 1: What is the Effect of Rehabilitation Programs Under Experimental Conditions?

As stated in Chapter VI, an examination of statistically significant findings would be misleading, since many experiments are not powerful enough to detect the small to moderate effects some social programs achieve. Less than 25% of this sample reported significant findings at first follow-up. As discussed, narrative reviews or vote-counting methods which relied upon statistical significance as a criteria for success would conclude that these justice interventions were not demonstrably successful. This research supports the view that null findings dominate the outcome literature in criminal justice, and social science in general (e.g., Lipsey, 1990), leading to widespread pessimism for intervention (e.g., Palmer, 1994).

A helpful way of analyzing effect size, the binomial test, was suggested by both Lipsey (1992a) and Pearson, et al. (1996). If rehabilitative programs were not demonstrably effective, we would see a distribution of effect sizes around zero and to the right of zero. Any decreases in crime outcomes would be distributed to the left of zero.

This is not the case. Using effect sizes (*d*) at the first follow-up period for 115 rehabilitation experiments, Figure 10 indicates that a majority of treatment effect sizes are negative and to the left of zero, providing evidence that–on average–intervention reduces official recidivism.

Even if we accept that treatment will randomly produce an even proportion of positive and negative effects (i.e., 50% above or below zero),

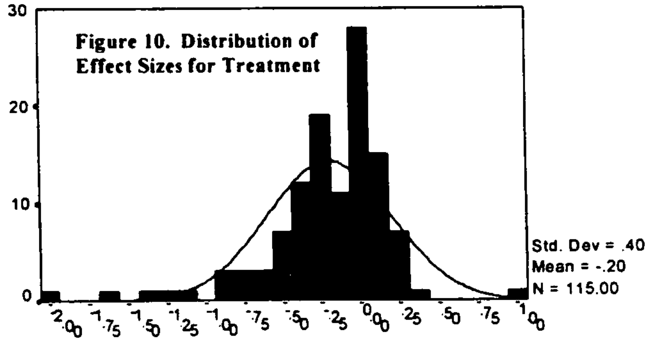


Figure 10. Distribution of Effect Sizes for Treatment

[161] In fairness, Hunter and Schmidt (1990) argue that if we had statistical artifact data from 20-50 studies, we could use it to estimate the other studies in the sample.

the binomial test clearly rejects that hypothesis. Approximately two-thirds (63%) of all first outcomes are negative in sign, a result which is significant at the .01 level.[162] This was nearly identical to earlier binomial distributions reported by Lipsey (1992a) and Pearson, et al. (1996).

The Magnitude of Effect for Rehabilitation Programs

While the binomial test indicates that the direction of effects consistently favors treatment, with 63% of the studies demonstrating a reduction in official crime, it is essential that the magnitude of that effect be computed. For this stage of the analysis, the equal-weighted, uncorrected mean effect size across studies is provided, which is simply the total sum of *d*s, divided by N(115). This yields an average *d* of -.20. On average, the treatment group outperforms the control group by 2/10 of a standard deviation on outcome measures of official criminality. One common way of interpreting meta-analytic findings like these is by stating that the results show that "doing something is better than doing nothing" (Losel, 1995).

It is true, however, that the mean effect size *d* is affected by extreme positive and negative values (e.g., Hedges and Olkin, 1985). If the distribution of effect sizes included large positive and negative outliars of equal size, then their effect on the mean would be negligible. This will rarely be the case in meta-analytic research, and caution must be used when reporting mean results (e.g., Johnson, 1989).

[162] This means that the observed distribution of effect sizes was significantly different than the expected distribution of 50% above and below zero.

An inspection of the median *d* for the 115 rehabilitation experiments showed that some very effective experiments (1-2 SDs from zero) were inflating the mean (median =-.12). Removing all outliars over -1.0 (or 1.0) reduced the mean *d* to -.12 and the median *d* to -.06. While the removal or addition of outliars from the distribution has advocates on both sides, the variability between the mean and median underscores the importance of caution in interpreting global means from meta-analysis.

It has been noted before that Cohen's *d* is the uncorrected, equal-weighted estimate of effect (e.g., Laird and Moseteller, 1991; Johnson, 1989). Hedges has previously shown that small samples, particularly those with total sample sizes of less than 30 subjects, consistently overestimate *d.* He developed a correction formula to compensate for this (e.g., Hedges and Olkin, 1985), which was referenced earlier. This correction is negligible for individual studies with total samples greater than 30. Since most criminal justice experiments are considerably larger than this, the impact is also negligible. As Figure 11 demonstrates, applying the small sample bias correction formula, the n-adjusted effect size—referred to as Hedges' *g*—is now -.197 or -.20 rounded.[163]

Figure 11. **Comparison of effect size and n-adjusted effect size estimates for Rehabilitation Experiments**

Cohen's *d*	-.20
Hedges' *g*	-.197 or -.20

[163] The corrective formula would not have changed Weisburd's (1993) point that smaller samples consistently achieve higher effects, since its impact on studies larger than 30 is negligible and Weisburd had an inclusion criteria of no less than 15 subjects in any one group.

Stopping at this point in the analysis would result in a meta-analysis with findings similar to nearly every prior synthesis in criminal justice. With the exception of Lipsey (1992a) and Pearson, et al. (1996), all prior researchers have ended their statistical analyses here. A *g* of -.20, even with the mean inflation due to outliar effects, would likely be interpreted positively to support the renewed optimism over rehabilitative programs.[164] While a d of -.20 would be considered small using Cohen's (1977) classification, it could represent a finding with important policy considerations (e.g., Lipsey, 1990).

Rosenthal (1991) developed a binomial effect size display (BESD) to quickly translate meta-analytic effects into raw percentages for policymakers, to assist them in interpreting statistical findings. Using the BESD table, a *g* of -.20 corresponds to a correlation coefficient (*r*) of -.10; the decrease in proportion rearrested would be equivalent to the value of *r*. To illustrate this finding, if it is assumed that the control group baseline recidivism rate is 50%, treatment on average would result in a recidivism rate of 40%. In other words, a reduction of this magnitude (20% reduction in recidivism rate or 50%÷10%) would likely be considered an important treatment effect and worth the social investment.

However, all things being equal, larger samples are more precise than smaller samples. Intuitively, one trusts a finding from an experiment with 1,000 subjects more

[164] It should be noted that median effect sizes have not been reported in prior justice meta-analyses, although the results here should encourage that practice.

than a study with 10 subjects. In most prior meta-analyses, a study with 10 subjects is given equal weight with a study of 1,000. The additional precision with larger samples is ignored.[165] This has been considered by methodologists to be an unwise practice, and many urge that weighting procedures which take sample size into account be used (e.g., Durlak and Lipsey, 1991; Johnson, 1989; Hedges and Olkin, 1985).

Hedges and Olkin (1985) have developed a method which takes the greater precision for larger samples into account, known as the inverse-variance method. In short, each study *d* is weighted by its sample size in the analysis. In this meta-analysis, larger individual experiments are given more weight than smaller sample experiments.

Applying the sample-size weights to the data, the average effect (referred to as *g+)* for treatment programs drops from -.20 to -.03. This provides evidence that effect sizes from small sample studies were driving the equal-weighted *g* upward. It also appears that larger sample experiments were less successful, and by sample-size weighting, the global *g+* is lowered. The 95% confidence intervals for *g+* do not include zero (-.05 to -.01), indicating that there is still a statistically significant effect for treatment on official crime measures, but it is unlikely that this finding could influence policy.

[165] The precision with larger samples is demonstrated by the confidence intervals; larger samples have smaller ranges between lower and upper intervals.

Using the BESD table (Rosenthal, 1991), a g+ of -.03 corresponds to a difference between study groups in proportion rearrested of 1.5%. In other words, if we assume the control group baseline recidivism rate is still 50%, the average experimental group rate would be 48.5%. This average decrease of 3% (1.5%÷50%)in official crime would not likely reaffirm rehabilitation.

Study Quality in Rehabilitation Experiments

Several prior syntheses have found research design to be an explanatory factor in meta-analysis, i.e., random assignment studies had smaller average effects than non-randomized studies (e.g., Garrett, 1985).[166] One of the motivations for conducting this study was to determine if a sample of randomized studies would yield different results than meta-analyses that had included a range of designs.

Though randomized experiments are considered the ideal evaluation design (e.g., Weiss, 1972) and comprise the sample, they also differ on many methodological dimensions. Random assignment studies are best viewed as representing a continuum from very strong evaluations to impotent ones. There are several threats to the integrity of an experiment, and to the extent that these occur, they weaken the internal validity of the study. It was important to determine if results differed for well-controlled studies—

[166] It is also true that randomized studies can have higher average effects, since they can more precisely indicate a treatment effect by reducing the noise of other variables (e.g., Weiss, 1972).

those which reported no breakdowns in randomization or substantial attrition–when compared to problematic experiments.

Rating schemes for evaluating and scoring experiments have been recommended and at least one has been utilized in the medical field (e.g., Laird and Mosteller, 1991), but there is no agreement on the factors which should comprise the scale or their relative weights. It is true that the University of Maryland report (e.g., Sherman, et al. 1997) included a methdological rating device, but it tended to give high scores (i.e., 4-5 points) to random assignment studies.

Compounding the problem is that so few reports expound on the methodology of the experiment; missing data is one of the most crucial impediments to conducting a meta-analysis (Cooper and Hedges, 1994). Added to that, meta-analysis actually penalizes the descriptive document over the parsimonious one, since no study is perfect—and those that have the space to elaborate on methodological difficulties generally do so.

A compromise was reached in this analysis to test whether the quality of the randomized design affected the results. Rehabilitation experiments which experienced no randomization breakdown, reported no substantive group equivalence pretest differences and experienced no more than minor attrition were categorized as "strong" internal validity studies. All other experiments–which reported at least one threat to internal validity using these three factors–were collapsed into a second group which experienced internal validity threats. The g+ was then examined for the two categories (Figure 12).

Figure 12. **Comparison of weighted effects (g+) along methodological dimensions for rehabilitation experiments**

	N	g+
Randomized Experiments *With Strong Internal Validity*	64	-.04
Randomized Experiments *With Some Threats to Internal Validity*	51	-.02

These findings, while indicating a slight design effect, are in line with Lipsey's (1992a) comparison of experiments with established group equivalence and no appreciable attrition, with other random assignment studies and quasi-experimental designs. He also found a very slight increase in effect size for the well-controlled studies.

It may be the case that methodological failures in experiments are important and lead to bias at the individual study level. This bias may either act to artificially inflate the experimental effect or to reduce it. As these effects are averaged across studies, these biases are also averaged to some extent, leading to the rather negligible results found in Lipsey (1992a) and here as well.

The Type of Official Crime Outcome Data

Although the official crime data used by PIs and relied on in this meta-analysis has been sharply criticized (e.g., Lipsey, 1990), it is also true that the experimental design should result in equivalent groups exposed in equal fashion to outcome data problems.[167] Nonetheless, it is possible that certain outcome measures are more sensitive than others; for example, since police arrest is a more probable occurrence than conviction or incarceration, it may be easier to impact. To the extent that this is true, effect sizes for rehabilitation experiments which reported police data may be higher than those which utilized court or other data.

Figure 13. **Comparison of weighted effects (*g+*) for police data and non-police data outcome measures in rehabilitation experiments**

	N	*g+*
Randomized Experiments With Police Data	61	-.03
Randomized Experiments With Non-Police Data	54	-.02

To examine this relationship, all rehabilitation studies which used police data (e.g., arrests, contacts, etc.) were compared to experiments which reported non-police outcomes (e.g., convictions, incarcerations, etc.). Figure 13 presents the weighted effect sizes for these two groups. As can be seen, there is a negligible difference in *g+* when comparing police data with non-police data. It does not appear that differences in the type of official outcome data used influenced effect size.

[167] Although Lerman (1975) presents a classic study of an experiment where the outcome measure was handled differently for treatment subjects than controls.

Question 2: Where Do These Findings Compare to Earlier Treatment Effectiveness Meta-Analyses?

Unfortunately, since so few meta-analysts have applied the same statistical techniques to their data, comparisons across studies are problematic. Figure 14 below presents evidence that equal-weighted *d*s (or n-adjusted *g*s) for the studies are similar, suggesting that small to moderate effects are the norm for criminal justice interventions. Although this study excluded all but the most rigorous designs, the equal-weighted d was also in the -.15 to -.30 range (e.g., Losel, 1995).

Figure 14. **Comparison of Effects on Recidivism Found in Treatment Effectiveness Meta-Analyses in Criminal Justice, 1984-1997**

Study	*d*	Study	*d*	*Weighted g*	Winsor/ Weighted *g*
Davidson, et al. (1984)[168]	-.35	**Lipsey (1992a)**	**-.17**	**unknown**	**-.10**
Garrett (1985)	-.13	Cox, et al. (1995)	-.03		
Kaufman (1985)	-.20	Hall (1995)	-.24		
Gensheimer, et al. (1986)	-.26	Losel (1995)	-.22		
Mayer, et al. (1986)	-.50	Wells-Parker, et al. (1995)	-.19		
Gottschalk, et al. (1987a)	-.13	Pearson, et al. (1995)	-.19		
Gottschalk, et al. (1987b)	-.33	Gendreau & Goggin (1996)[169]	-.25		
Losel & Koferl (1989)	-.22	Gendreau & Goggin (1996)[170]	0		
Whitehead & Lab (1989)	-.27	Redondo, et al. (1996)	-.13		
Andrews, et al. (1990)[171]	-.20	**Pearson, et al. (1996)**	**-.19**	-.04/-.07[172]	
Izzo & Ross (1990)[173]	unknown	**Petrosino (1997)**	**-.20**	**-.03**	**-.07**
Roberts & Camasso (1991)	-.36				

[168] Effects in Davidson, et al. (1984), Gensheimer, et al. (1986), Gottschalk, et al. (1987a), Gottschalk, et al. (1987b), Mayer, et al. (1986), and Cox, et al. (1987).

[169] This main effect is only for what the authors define as appropriate correctional services.

[170] This main effect is for 'punishing smarter' sanctions.

[171] This is the main effect for all interventions. Of course, Andrews, et al. (1990) reported that appropriate correctional service achieved a phi of .32 or d=-.64.

[172] Pearson, et al. (1996) reported weighted effects for juveniles (-.07) and adults (-.04) separately.

[173] Izzo and Ross (1990) did not report a main effect.

These meta-analyses all examined a different set of studies, adding to the difficulty in comparing syntheses. In addition, the different goals of the research, the various inclusion criteria used, and the diversity of search techniques also added to the complexity of examining prior meta-analyses.[174] The congruence of findings despite these differences is remarkable. The comparison is designed to provide some insight, but should not overshadow the fact that meta-analyses may have dealt with specific subjects (e.g., only sex offenders), specific treatments (e.g., residential treatment) or specific settings (e.g., only Europe). Caution must be used when comparing this set of 115 broadly categorized rehabilitation experiments with meta-analyses which classified interventions differently.[175]

As seen in Figure 14, weighted effect sizes for Lipsey (1992a), Pearson, et al. (1996) and this project are substantially lower than equal-weighted effects. Lipsey (1992a) reported a *g*+ of -.10, while Pearson, et al. (1996) reported a *g*+ of -.04 for adult treatment and -.07 for juvenile programs. As discussed earlier, this meta-analysis of randomized experiments resulted in a *g*+ of -.03. Nonetheless, Lipsey (1992b) concluded optimistically about the effects of rehabilitation, interpreting the -.10 result as a small, but nontrivial finding.

[174] Has the proliferation of meta-analysis resulted in the need for meta-meta-analysis or meta^2-analysis (but see Lipsey and Wilson, 1993)?

[175] For example, some included sanctions and treatments in their global mean analyses (e.g., Garrett, 1985).

Exploring the Difference Between Lipsey (1992a) and This Project

Obviously, one source of the difference between Lipsey's meta-analysis of juvenile interventions and this study is that different study samples were selected. After all, Lipsey (1992a) included quasi-experimental designs, but excluded adult studies. However, the equal-weighted *gs* was higher here (-.20 to -.17), suggesting that the weighting procedure led to the differences in *g*+. It could be that Lipsey had a greater proportion of effective large-n studies, or it could be a difference in the weighting procedures used.

Upon closer inspection, Lipsey (1992a) also used Hedges and Olkin's (1985) inverse-variance method, but *winsorized* the samples at 300 in each group to prevent massive-n studies from dominating the statistical analysis. Winsorizing removes the effect of extremely large sample size studies by setting the upper limit of individual study experimental and control Ns at some number (e.g., Mosteller, 1997). In Lipsey's (1992a), he set all sample sizes above 300 in experimental or control groups to 300. It is debatable among methodologists whether such a procedure should be employed (e.g., Johnson, 1997), since winsorizing negates some of the impact of sample-size weighting. Nonetheless, this approach was used here to determine if winsorizing was the source of discrepancy between the two meta-analyses. To winsorize, individual sample sizes in the individual experiment in excess of 300 were reduced in the analysis to the 300 cut-off.

The results showed that winsorizing the individual study increased *g*+ to -.07, again demonstrating the influence that weighting procedures had on these meta-analytic findings. If the winsorizing technique is appropriate, then the -.07 effect magnitude translates into an average 7% reduction in recidivism per rehabilitation program, using the BESD table (e.g., Rosenthal, 1991). The average treatment recidivism rate would be 47.5% compared to the 50% control baseline. While closer to Lipsey's (1992a) finding, it could be argued that this is a negligible impact for treatment programs.

To summarize Figure 14, this meta-analysis is fairly congruent with prior studies when using equal-weighted effect size only.[176] Outside of the second Gendreau and Goggin (1996) meta-analysis—which dealt with 'punishing smarter' sanctions—and the Cox, et al. (1995) synthesis of alternative education programs, the main effects across all prior syntheses ranged between -.13 to -.50. The average equal-weight effect size across the meta-analyses was -.22, supporting the renewed optimism for intervention (e.g., Palmer, 1994; 1992). Yet, given the substantively lower results when weighting by sample size, this cautious enthusiasm based on prior meta-analyses must be tempered.

Question 3: Does the Experimental Effect for Rehabilitation Programs Differ from Deterrence or Prevention Programs?

Another method of analysis is to compare the performance of rehabilitation with deterrence-based interventions or delinquency prevention programs. Rehabilitation programs have shown great advantage over deterrence-based programs in those meta-analyses where they could be compared (e.g., Gendreau and Goggin, 1996; Lipsey, 1992a; Andrews, et al. 1990). As mentioned in Chapter VI, it was possible to categorize 147

[176] All common metrics were converted, where possible, to Cohen's effect size *d*

experiments into three categories: those which tested a rehabilitation program (N = 115); those which tested a deterrence-based program (N = 23) and those which evaluated a delinquency prevention program (N = 9).

As with the earlier discussion on the sole effects of rehabilitation, the type of statistical procedure utilized in the analysis influences conclusions. As Figure 15 demonstrates, the average *g* shows a powerful advantage for rehabilitation (-.20) over deterrence and prevention programs. In fact, *g* is zero across 23 deterrence experiments, coinciding with the poor findings for deterrence in prior syntheses (e.g., Andrews, et al. 1990).

Figure 15. **The Comparative Efficacy of Rehabilitation Programs: Different Effect Size Estimates for Rehabilitation, Deterrence and Prevention-Based Interventions**

Philosophical Type	N	*g*	*g*+	Winsorized *g*+
Rehabilitation Programs	115	-.20	-.03	-.07
Deterrence Programs	23	0	-.05	-.01
Prevention Programs	9	-.06	.02	-.03

Again, weighting by the inverse-variance method to account for the precision in larger study sample sizes changes the interpretation of the results. Weighted effects (*g*+*)* are very similar, with deterrence-based programs slightly more effective (-.05 to -.03), while prevention programs experience a slight average backfire effect on official crime outcomes (.02). In addition, winsorizing the individual study samples at 300 further influences the results, with rehabilitation programs slightly more effective in reducing post-program crime than deterrence or prevention programs.

These findings are troublesome, as they underscore the instability of meta-analytic findings. If methodologists are correct, and effect sizes must be weighted by sample size, than the small, non-trivial effect found for rehabilitation in prior syntheses may be a methodological artifact rather than a substantive finding. As seen in Figure 15, when this weighting is applied, deterrence programs—often considered less effective than treatment (e.g., Andrews, et al., 1990)—more than doubles treatment's experimental impact on official crime (-.07 to -.03). Given the reversal of fortune for these grand effects when weighting and winsorizing is introduced, sample size appears to be a solid candidate for explaining this heterogeneity.

Question 4: Does Sample Size Explain Effect Size Heterogeneity in Rehabilitation Experiments?

Hedges and Olkin (1985) developed a statistical procedure to assess the homogeneity of effect size. The homogeneity test indicates whether the weighted mean effect adequately describes the sample of studies (e.g., Johnson, 1989). If the homogeneity statistic, known as Q, is large and significant, that indicates that considerable heterogeneity exists and *g+* does not represent the total sample of studies well. It is then recommended that study factors be analyzed to determine the source of this heterogeneity (e.g., Durlak and Lipsey, 1991).

A small Q would mean that examining moderators such as sample size or type of subject would make little sense, since the effect sizes were fairly homogeneous. In contrast, the homogeneity test for the 115 rehabilitation experiments showed substantial heterogeneity across studies. The test for homogeneity was both large and significant ($Q=340$, $p<.00000$), lending strong support to the decision to examine the moderators aforementioned in the focused questions.

Given the earlier evidence that effect size variation may be attributable to study sample sizes, and the consistent finding in the literature that smaller studies achieve higher effects, a focused analysis of this variable on $g+$ was warranted. It is preferable that statistical tests for sources of heterogeneity in meta-analysis proceed in systematic fashion; simply fishing for explanation using a wide range of study factors is likely to produce some variables as adequate predictors simply by chance capitalization alone (e.g., Durlak and Lipsey, 1991).

For this analysis, sample sizes for experimental and control groups used to create the effect size were summed for each study to create a total sample size ($N_E+N_C=T$). The total sample size (T) was then categorized into five ranges: 10-50, 51-100, 101-300, 301-500 and 501+. A categorical homogeneity test was then used to determine if sample size was a source of $g+$ heterogeneity. Winsorizing the samples at 300 was not done for this part of the analysis, since the interest was examining variation for different size experiments.

The results in Figure 16 confirm that these sample size categories account for some heterogeneity in $g+$. The most effective rehabilitation programs generally have total samples of 100 persons or less; the weakest programs handle over 300 subjects. The relationship is nearly a perfect linear one; effect size decreases dramatically with each increase in sample size range. This confirms the earlier work by Weisburd (1993), and results found in prior criminal justice meta-analyses (e.g., Lipsey, 1992a).

Figure 16. **Average $g+$ for Sample Size Categories in Rehabilitation Experiments**

Total Sample Size Category (N)	$g+$
10-50 (22)	-.36
51-100 (23)	-.38
101-300 (43)	-.08
301-500 (8)	0
501+ (19)	.01

Q_B=84, p<.000000

The categorical homogeneity test examines the variation across different levels of the independent variable. In this case, if Q_B (heterogeneity between levels) is both large and significant, it confirms that the independent variable is a strong moderator of effect size. In this analysis, Q_B was 84, and highly significant (p=.00000).[177] Sample size is an influential variable in rehabilitation experiments.

One explanation for small sample effects is methodological. Smaller samples produce less stable findings and reduce less of the noise from extraneous variables than larger randomization procedures. If this were the case, the small sample effect should also be present in deterrence and delinquency prevention experiments also. Figure 17 examines *g*+ for the five sample size range groups across rehabilitation, deterrence and delinquency prevention programs.

Figure 17. **Average *g*+ for sample size categories across rehabilitation, deterrence and delinquency prevention experiments (N=147)**

Sample Size Category	Rehabilitation(N)	Deterrence(N)	Prevention(N)
10-50	-.36 (22)	n/a	0 (1)
51-100	-.38 (23)	.08 (3)	.12 (1)
101-300	-.08 (43)	.04 (8)	-.12 (5)
301-500	0 (8)	-.12 (5)	n/a
501+	.02 (19)	-.05 (7)	.02 (2)

As Figure 17 indicates, it does not appear as though the small sample size effect is a methodological artifact, since it does not repeat in deterrence-based or delinquency prevention programs. Although larger samples may be more difficult to control (e.g., Weisburd, 1993; Tanur, 1983), deterrence-based programs actually had higher effects in the 301-500 total subject group than in smaller studies.

It is possible that these are indications of real substantive differences between philosophical types and how they are operationalized in random assignment studies. Treatment program experiments which are small may be easier for the service provider to control and more conducive to establishing and maintaining close client contact. It may also be that deterrence experiments, while more difficult to control administratively when large numbers of subjects are involved, are able to exert both general and specific

[177] Q_B is the statistic for between group (or level) heterogeneity; when this is large and significant, the study factor explains some—though not all—of the

deterrent effects on subjects by the large numbers of persons receiving the presumably harsher sanctions (e.g., Clear, 1997).

Question 5: Does the Type of Subject (Juvenile v. Adult) Explain Effect Size Heterogeneity?

Another factor used as a moderator in this study was the type of subject included in the experimental program. In this analysis, three groups were compared: juveniles (ages 0-17); juvenile and young adult (generally studies which included a range of subjects such as 14-21); and adults (18 and over). As seen in Figure 18, the average weighted effect for juvenile treatment is -.15, which would translate into an average 15% reduction in recidivism rates (e.g., from a 50% baseline to 42.5%). The average *g*+ across 53 adult treatment program experiments was negligible (.01). Winsorizing the study sample sizes in this analysis did not greatly affect *g*+ within categories, although there was a change in sign for "adult only" treatment studies.

Figure 18. **Average *g*+ for Juvenile and Adult Rehabilitation Programs**

Type of Subjects (N)	*g*+	*Winsorized g*+
Juvenile Only (55)	-.15	-.16
Juvenile and Young Adult (7)	-.10	-.11
Adult Only (53)	.01	-.02

Q_B=47, p<.000000

Since this is one of the few meta-analyses which has combined juvenile and adult programs, this type of finding has only been elaborated upon once before. In Pearson, et al.'s (1996) analysis of juvenile and adult programs, they find more congruence between the two (adults=-.04, juveniles=-.07). This sample of experiments

heterogeneity of *g*+.

indicates that juvenile rehabilitation is modestly successful under experimental conditions and more effective than adult rehabilitation.

The categorical model test for homogeneity again was again large and significant (Q_B=47, p=.000000), supporting the hypothesis that there is considerable effect size variation across the three variable levels. Type of subject appears to be an important moderator and was also compared for each of the broad philosophical category types (Figure 19). For this analysis, only juvenile and adult categories are compared; the seven studies which included both were excluded.

Figure 19. **Average *g*+ for juvenile and adult programs across rehabilitation, deterrence and delinquency prevention experiments**

Type of Subjects	Rehabilitation(N)	Deterrence(N)	Prevention(N)
Juvenile	-.15 (55)	.12 (6)	.02 (9)
Adult	.01 (53)	-.06 (17)	n/a

Rehabilitation programs appear more demonstrably affective with juveniles (17 and younger) than either deterrence-based approaches and delinquency prevention interventions. Interestingly, deterrence has a backfire effect (.12) with juveniles, likely propelled by the unsuccessful Scared Straight experiments (e.g., Finckenauer, 1982). Yet, deterrence shows a very modest crime reduction effect with adults (-.06), which is higher than the average *g*+ for rehabilitation (.01).

Which Specific Programs Seem to Work for Juveniles and Adults?

One problem with examining effect sizes for specific programs is the incredible diversity in interventions. Even when using broad categories such as "family counseling," "casework" or "arrest/warrant," over 25 separate intervention types were found. This problem is exacerbated by the poor descriptions of treatment modalities by PIs, and the lack of specificity in the program itself.

A major problem with such a large number of categories is that some specific program types include only one experiment. Basing policy decisions or research conclusions on a cell with only a couple of studies is risky and negates some of the benefits of the meta-analytic method.

Table 13 provides weighted effect sizes for each of the specific programs listed. Program effects are listed for juveniles, the combined juvenile and young adult group, and for adults only. While the small cell Ns make any firm conclusions dangerous, it should be noted that nearly all of the juvenile interventions are negative in sign, except for the Scared Straight prison tour programs, crisis intervention and community service. Approximately half of the adult programs are positive in sign, meaning that the program type had a backfire effect (e.g., Sherman, 1988).

The major strength of Table 13 is how clear it makes a future experimental research agenda. For many of the program types, there are simply too few studies to base any conclusions. The inclusion of strong quasi-experimental designs, natural experiments and indirect experimental evaluations are probably warranted in future meta-analyses to increase these cell sizes. Lipton (1995) is currently completing such a meta-analysis of correctional rehabilitation studies, although the CDATE project is not meta-analyzing prevention studies.

Table 13.

What Works in Specific Crime Reduction?
Juvenile and Adult Programs Ranked by Weighted Effect Size (g+)[*]

Juvenile Programs (N)	g±
Family counseling (4)	-.33
Social Skills Training (2)	-.29
Community-based/Resid (9)	-.29
Casework (3)	-.26
Enhanced Supervision (1)	-.24
Diversion/Diversion w/serv (11)	-.20
Institutional Change (3)	-.17
Group Counseling (5)	-.14
Individual Psych Counsel (3)	-.09
Increased Super/Reduce Caseload (2)	-.07
Vocational Based (2)	-.04
Citizen Volunteer Programs (1)	-.02
Cognitive/Behavioral (7)	-.02
School-based Programs	-.02
Forestry/Wilderness (3)	-.01
Juvenile Prison Tour (5)	.05
Crisis Intervention (2)	.11
Community Service (2)	.33

Juvenile/Young Adult Program (N)	g±
Citizen Volunteer Programs (2)	-.63
Vocational Based (2)	-.28
Diversion with Services (1)	.07
Group Counseling (1)	.12
Institutional Change (1)	.25

Adult Programs (N)	g±
Individual Psych (1)	-.34
Cognitive/Behavioral (4)	-.28
Casework (3)	-.24
Arrest/Warrant (5)	-.15
Diversion/Divert w/serv (2)	-.14
Medical/Drug Treat (6)	-.13
Other Sanctions (4)	-.12
Payments (2)	-.03
Other Treatments (3)	-.02
Drug Testing (6)	0
Citizen Volunteer (1)	0
Vocational Based (3)	0
Enhanced Supervision (5)	.02
Education/Information (4)	.06
Increase Super/Reduce Case (13)	.06
Group Counseling (2)	.07
Community-based/Resid (3)	.08
ISP Probation/Parole (3)	.10

[*] Three experimental interventions could not be categorized and are not included.

CHAPTER VIII. CONCLUSIONS AND RECOMMENDATIONS

In this chapter, substantive findings in Chapter VII are reviewed and their implications for criminal justice policy discussed. In addition, the methodology of meta-analysis is examined, and recommendations for improvement in quantitative research synthesis are offered.

'What Works?' Revisited Again: Substantive Findings and Discussion

As mentioned in Chapter I, the debate about 'what works?' is not only about philosophy, values and politics. It is also about evidence in the form of individual studies—and how that evidence is assessed. In this project, randomized experiments, considered by methodologists to be the 'best evidence' on the question, were collected. This sample of 150 randomized evaluations is the largest collection of criminal justice experiments reported in the literature. This study also employed advanced meta-analytic techniques to review the experimental evidence.

Chapter II outlined the evolution of rehabilitation, and discussed the role of recent meta-analyses in fueling renewed optimism about treatment (e.g., Lipsey, 1992a; Palmer, 1992). It remains to be seen if these findings will be interpreted as good news or bad news by rehabilitation advocates and skeptics.

The exclusion of all evaluation designs except random assignment studies was a concerted effort to focus on the 'best evidence.' Nonetheless, this sample yielded a global equal-weighted effect size of -.20, which was comparable to nearly all prior meta-analyses (e.g., Losel, 1995). As with the findings for study quality presented in Chapter VII, random assignment is crucial at the individual study level in providing more precise estimates of treatment effect than quasi-experimental designs. Nonetheless, that increase in precision may result in larger or smaller effect sizes. In meta-analysis, the aggregate treatment of the studies results in a global effect size that may cancel out increases and decreases stemming from greater precision.[178]

The most important finding may be the instability of the meta-analytic findings when employing different statistical techniques to the data. The equal-weighted effect size of -.20 for rehabilitation could have important policy considerations; it represented a much larger experimental effect than either deterrence-based or delinquency prevention programs. Given the congruence of this finding with earlier meta-analyses, rehabilitation might be completely revivified (e.g., Gendreau and Ross, 1987).

Yet, the median value for the equal-weighted effect size was -.12, indicating that outliars were inflating the global mean effect (-.20). Even more telling, when sample-size weighting and winsorizing techniques were introduced, the overall mean effect size changed dramatically, to -.03 (*g+)* and -.07 (winsorized *g+)* respectively.

[178] I found this very interesting, since randomized experiments in this sample nearly always found insignificant differences where earlier quasi-experimental evaluations reported a significant effect for treatment.

Neither type of outcome data or experimental design rigor influenced the average sample-size weighted effect size for rehabilitation studies. In addition, the great advantage of treatment-oriented programs over deterrence and delinquency prevention was eliminated. Deterrence programs, for example, had a *g*+ that was more than twice the size of the average rehabilitation program (-.07 to -.03).

Though winsorizing study samples at 300 slightly reversed these results, treatment advocates might be less than enthusiastic. First, the winsorized *g*+ for rehabilitation was only -.07, about 1/3 of the average equal-weighted effects found in this and prior treatment effectiveness meta-analyses. Second, it is debatable whether winsorizing should even be applied to the data (e.g., Johnson, 1997).

Given the instability of the data, great caution must be exercised in interpreting the global effect sizes reported here. Adding to this instability is the recognition that it would take only a few fairly large zero-effect studies (or those which show a crime increase) to change *g*+ from -.03 to 0 (e.g., Wolf, 1986). Yet, it must be pointed out that these findings for sample-size weighted data are very close to those reported by Pearson, et al. (1996) from their massive CDATE project.

As with most analyses, the global estimates mask important differences across the sample experiments. The homogeneity tests demonstrate that the *g*+ of -.03 does not represent the entire sample of rehabilitation studies very well; there is significant heterogeneity across effect sizes. Two study factors explored as potential explanatory variables for this heterogeneity were sample size and type of subjects. In both cases, the homogeneity tests indicated both were important moderators, with treatment programs demonstrating larger effects with smaller samples and with juvenile subjects.

Small Sample Effect. It is still unclear why smaller samples should achieve consistently higher effects than larger samples. If this finding was the result of a methodological artifact, why would the effect not be found in deterrence-based or delinquency prevention experiments? The reasons for small sample effects are being hotly debated in the medical field (e.g., Johnson, Carey and Muellerliele, 1997).

Could Weisburd's (1993) assertion that experimenters are better able to control small sample experiments correct? If this is true, why were these findings not repeated for 23 deterrence-based experiments or nine delinquency prevention studies? It may be that small samples interact with treatment programs in a substantive manner, including the possibility of an experimenter expectancy effect. Rather than the particular intervention affecting subject behavior, PIs in small subject studies are able to exert influence on subjects to act the way they believe the PI wants them to act (e.g., Rosenthal, 1991).

It is possible that treatment in large programs is diluted; smaller samples should permit a stronger and more intensive 'dosage' of the intervention. Unfortunately there was so little information on treatment contact provided—and what was available was unreliably coded—that it could not be used to further investigate this. Some limited evidence, however, from Kaufman's (1985) delinquency prevention meta-analysis indicates this could be the case; he found that prevention programs with increased levels of client contact were more effective in reducing post-program delinquency.

Future investigations of this data will need to examine the correlation of sample size with other variables, to determine if the effect is spurious (i.e., the impact of sample size is actually due to some highly correlated third variable). It could be that the most effective treatments in experimental evaluations are administered with small samples only; thus, the increased effect size is the result of better treatment and not sample size.

It should be noted that the sample size analysis was reported using *g*+ for each of the range categories, a procedure that reduced each small sample study's effect size. If equal-weighted effect sizes were used in the analysis (e.g., Cohen's *d*), the powerful finding for samples under 100 total subjects would have been even larger.

Type of Subjects Effect. A potentially substantive policy finding is that juvenile rehabilitation programs were more effective under experimental conditions than those administered to adults. One could see where this finding could influence policy; scarce resources could be allocated toward programs for children. While the sample-size weighted effect is small for juveniles ($g+=-.15$), it is still demonstrably larger than that for adults ($g+=.01$). This finding parallels the recommendations of the Sherman, et al. (1997) report, where they urge increased funding for programs with high-risk juveniles. In addition, it would also seem to fit Gottfredson and Hirschi's (1990) contention that crime policy be focused at much earlier ages; since crime is a result of low self-control, focusing rehabilitation programs on criminal adults is misdirected.

The Methodology of Meta-Analysis

The proliferation of meta-analysis in the social sciences—and in criminal justice specifically—may render the impression that it is easy research. An uniformed person once told me that "all you have to do in meta-analysis is collect a few studies, punch in a few numbers and crank out the results." Conducting a comprehensive meta-analysis, however, on treatment effectiveness is an enormous endeavor. Recent works by the National Research Council and the Sage Foundation have attended to this point (e.g., Cook, et al., 1992; Wachter and Straf, 1990).

Inclusion Criteria

As noted in Chapter III, careful consideration in defining the study domain is essential to the project. If the study is restricted too greatly, it may be easier to complete, but will lack generality and scope. If the inclusion criteria are too broad, it will end up trying to be all things to all readers, probably exhausting resources at the search and retrieval stage. Just as important, decisions must be made using rules to insure consistency.

The inclusion criteria used were restrictive, given the focus on randomized experimental designs. Even with such narrow criteria, a computer search through 25 years of *Criminal Justice Abstracts* on Cd-Rom (1968-1993) produced over 3,000 titles or abstracts containing variations of key words found in experimental reports (e.g., "experiment," "random" or "controlled"). The present criteria were used to exclude over 90% of these documents using the abstracts alone, leaving less than 300 reports which had to be tracked down and reviewed. Nonetheless, this project collected over 300 randomized evaluations and included 150 of them in the statistical meta-analysis reported here.

One of the important benefits of meta-analysis—if reported with full veracity—is that it systematizes the research synthesis process. One of the major criticisms of narrative reviewing is that decisions were made in arbitrary fashion and were not subject to scientific verification (e.g., Glass, McGaw and Smith, 1981).

Fortunately, even qualitative reviews have improved in response to these criticisms, although it is becoming more difficult to publish a narrative synthesis since the advent of meta-analysis (e.g., Hamby, 1996).

Several recommendations flow from this research. First, specification of inclusion criteria should be accomplished long before the document retrieval stage, since the way the study domain is defined will affect search efforts. The criteria will also impact the amount of time and money spent in search and retrieval efforts, and later coding and keypunching. It might be wise to test the criteria against a small set of studies to insure they are not too broad or restrictive.

Even when inclusion criteria are specified, gray area studies will be confronted which compel investigators to alter the criteria. A clear lesson from this project is the dynamic nature of meta-analysis; newly retrieved documents consistently test the appropriateness of earlier formed criteria. Modifications may be made, but there should be rules so that others can learn why certain studies were included or excluded, if it is not readily apparent. Moreover, changes may mean going back in the project files to insure that prior studies still comply with the criteria, and project consistency maintained.

Like other research, meta-analysis is comprised of decisions made at many different stages. The rules for making those decisions should be explicit and made available, even if editorial constraints prevent journal space. Knowing which studies were included and why is informative to other investigators, and sheds light on discrepancies between meta-analyses examining the similar relationships.

It is important to note that many meta-analyses can be done. In primary research, access may not always be available to certain data sources for surveys, interviews and experiments. However, the data for meta-analysis is not going anywhere, and technology has made more information available with less time required to obtain it. When specifying criteria, it should be whether some studies could be used in a future meta-analysis rather than trying to fit them all into the same sample.

Search and Retrieval Efforts

As mentioned in Chapter IV, a crucial factor in search and retrieval efforts is the inclusion criteria selected for the study, since it provides boundaries for data collection. The criteria set the time frame for the search, define the type of literature to be collected and the comprehensiveness of retrieval methods. For example, setting a start period of 1970 would eliminate the need for hand searches of manual bibliographic indexes, since electronic indexing began for most content areas in 1968.

Even with broad inclusion criteria, several recommendations to improve search and retrieval efforts are made. First, published solicitations and mail campaigns of prominent authors should be done early in the project, particularly for broad treatment effectiveness meta-analyses. This would not only help in the retrieval process, but would also enable the investigator to publicize the goals of the project to the research community.

After using these communication methods, the investigator would be wise to undertake a manual search of journals that publish relevant studies. For criminal justice meta-analyses, it would be prudent to conduct a search at a specialized justice library (e.g., Rutgers University's NCCD/Criminal Justice Collection), where all relevant journals are likely to be stacked. Not only will a hand check uncover additional empirical reports, it will also assist in the development of a sharper list of keywords to use in subsequent electronic searches (e.g., Cd-Rom, online databases, etc.).

In conjunction with the manual search, prior research reviews and published bibliographies should be retrieved, particularly those which focused on the subject area. While those containing descriptive tables or annotations are most helpful, a large number of potential citations are generally found in these works.

Electronic technology is a necessary and insufficient search technique. If Dialog or other online searches are used, it is wise to invest time in learning how to electronically search designated data bases; information specialists may not have a sufficient understanding of criminal justice terminology to run an effective search. If limitations in time or money make personal searches impossible then the online search should be conducted with the information specialist, particularly in the selection of keywords and search terms. Simply relying on electronic searches is likely to result in a relatively incomplete data base (e.g., O'Kane, 1995).

Checking the reference sections of primary research articles or relevant books will provide thousands of citations, quickly overwhelming the meta-analyst. Future reviewers should consider using these citation sources for references to fugitive literature. Electronic searches and manual searches will predominantly cover academic journals, so using these cumbersome methods to find unpublished or fugitive literature is highly recommended.

Although not utilized in this study, advances in information technology have resulted in the availability of evaluation studies on the World Wide Web. Future reviewers will have to include a search of home pages for government, university and private agencies likely to make study reports available over the Internet.

Since all these search strategies generate hundreds—if not thousands--of references, record-keeping is essential. As citations are checked, two files are essential: a potentially relevant citations file and a rejected documents file. Some researchers have found specialized bibliographic software, such as ProCite, to be helpful in maintaining organization of cites (e.g., O'Kane, 1995). This assists the search process, particularly if more than one person is involved, avoiding rechecking or reordering the same documents. As an added benefit, citations to excluded studies could be used as a 'starter set' in a second meta-analysis.

An early document screening process is also worth the investment. Meta-analysts should insure that study reports have the necessary quantitative data needed to be included in the sample. This can be done with a quick review of the document after it is located or retrieved. One of the more frustrating things in this project was to code part of an experimental evaluation report only to find—at the end of the article--that it lacked quantifiable recidivism data. An early screening system could have alerted the investigator to this problem and led to contacting the original researchers or excluding the study from the sample.

Coding

As mentioned in Chapter V, several procedures were implemented to improve the reliability of the meta-analysis and to extract relevant information. While the overall rate of agreement was an acceptable 80%, there was variability across item categories. Outcome information, surprisingly, achieved a higher rate of agreement than that reported in some prior syntheses, lending credence to the effectiveness of the coding guidelines. However, treatment and investigator items were unreliable and could not be used in any substantive analyses.

There are several ways to improve reliability in the future. First, the rate of agreement should be assessed early in the project, perhaps after two independent persons coded the first ten cases. Coding reliability would be reassessed after several months to insure consistency over time, as coders get tired, lazy or cut corners. The rationale for the early trial is to correct the coding instrument before intensive extraction of data begins. It is much more difficult to go back into the data sets to make corrections than to simply code them correctly the first time. Thus, two tests for interrater reliability would be reported.

It is recommended that coding reliability be checked by including the person most familiar with the documents. Using two independent coders who are unfamiliar with treatment evaluation reports could result in a high rate of agreement, but they could both be consistently wrong. By including the expert in the reliability test (usually the PI or researcher directly involved with the data), a more realistic rate of agreement could be achieved. After all, the goal is to reliability extract valid data of interest.

It is preferred that the meta-analyst extracts all outcome data of interest from the documents. First, this prevents having to revisit the primary studies to code additional information. More importantly, it allows the investigator to explore the impact of using different indices on effect size. Although the comparison of police with non-police data did not reveal any differences in *g*+, collecting all outcome data would have allowed extensive analyses.

It is not likely that contact with original investigators would be productive in filling in the gaps in the primary evaluation literature. The experimenters contacted during this project had moved onto new endeavors, misplaced files, could not remember sample sizes, or were simply too busy to look for requested information. Since telephone interviews proved fruitful for Dennis (1988), it may useful to combine his techniques--in limited fashion--with standard coding. A small set of recent experimental evaluators could be queried first to prepare them for survey questions; each PI who volunteered would then be interviewed to determine the reliability of extracted information and to compensate for missing data.

Although there is no agreement on rating schemes for the methodological quality of an experiment, this is an area that could be developed by an expert panel of scholars on randomized studies. This rating device would necessarily extend beyond simple classification systems, such as the one utilized here ('strong internal validity' versus 'studies which experienced threats to internal validity'). It may be that randomized field experiments will vary on many methodological dimensions, and unless certain threats are weighted more than others, the net effect of the rating scheme will be minimal (e.g., Losel and Koferl, 1987).

Common Metric Effect Size

Cohen's *d* is a frequently used effect size index and was used in this study. It has the advantage of providing a scale free measure of experimental effect that is also easily converted to Pearson's correlation coefficient (*r*). Although *d* was frequently computed from group failure proportions and frequencies, an important improvement in this study would be to compare *d* for different expressions of the experimental effect (e.g., failure proportions, exact probability level, test statistic, etc.). This would have provided some idea of the variation expected when using these different methods for estimating *d*.

It might also be important to collect effect size data on non-crime outcome measures. In a few experiments, non-crime data was used to test whether the underlying program theory was successful in modifying attitudes or changing something else about the subjects. A school intervention which theorized that helping kids get stay in school would ameliorate crime could hardly be expected to reduce delinquency when program kids were absent as much as control kids (e.g., Finckenauer, 1982; Weiss, 1972).

Statistical Analysis

It is true that the categorical homogeneity tests used and reported here were designed for one independent variable (e.g., Johnson, 1989). Standard multivariate regression, using canned statistical software (e.g., using SPSS-PC), can not be used without making corrections to the data for the violation of parametric assumptions (e.g., Hedges and Olkin, 1985). The application of such techniques in criminal justice meta-analyses has been very limited to date (e.g., Lipsey, 1992a),[179] but multivariate analysis is strongly recommended to better model the influence of sample size, type of subjects and other independent variables on effect size.

It will also be important to better integrate the data on randomized experiments in subsequent statistical analyses. While sample size and type of subjects were selected because prior work suggested their inclusion, the data set contains many theoretically important variables. For example, it would be important to include 'year of study' in a homogeneity test to determine if interventions have become more effective over time. Other important factors to consider are whether weighted effect sizes are affected by the type of publication (published versus unpublished), or how subjects entered into the experiment (coercive versus voluntary).

Confidence intervals for the individual *g*+ values could also be examined as another method for gauging the instability of certain experiments. Yet, it should be noted that confidence intervals for an effect size at the study level are greatly influenced by the sample size; all things being equal, larger samples are more precise and will have narrower lower and upper 95% confidence interval limits. Since

[179] It is not clear from Lipsey (1992a) whether suggested corrections were made to the regression analyses reported.

sample size precision is already taken into account by the inverse-variance method of weighting, heavy reliance on confidence intervals may be redundant.

Future Research Agenda

While alternatively selecting 150 randomized experiments from a set of 307 retrieved studies was a compromise in the face of limited resources, this decision has set the stage for a potentially important research project. By meta-analyzing the experiments not chosen for this study, a comparison can be done between these results and those found with a second sample.[180] In essence, the second sample will serve to validate these meta-analytic findings.[181]

Additionally, randomized field tests meeting the inclusion criteria and reported in 1994 or after should also be retrieved and added to the data base. This will allow the researcher to assess the impact of newly acquired studies on meta-analytic findings. Of course, a larger sample of studies adds more precision and stability to meta-analytic findings.

It will also be interesting to see if these meta-analytic findings hold up using different synthesis techniques. This has been done to some extent in psychology

[180] Of course, when the full set of experiments is entered into the computer, a test of the similarity between the 150 alternatively selected sample and the population it was drawn from can be done.

(e.g., Johnson, et al., 1995), but these are relatively untested waters. These data can be used to test the robustness of results across various meta-analytic methods. Since this is a dynamic area, future research will undoubtedly have to include some newer techniques for research synthesis.

With a full set of experiments in the meta-analytic sample, it might become possible to examine fluctuations in effect size over time. In short, it would be valuable to examine changes in *g*+ over follow-up periods. If effect size decreases as subjects move into second and third follow-up periods, as is commonly assumed, this would suggest a need for a more effective aftercare component to be implemented with treatment. It would also suggest that a 'first-effects' meta-analysis is misleading, since first-effects are more powerful than later measurements. It may be that the time to first follow-up is a more powerful predictor than any substantive treatment factor.

Finally, advances in individual-level intervention will be more possible when the best of quantitative and qualitative syntheses are combined. One way to combine the features of both is in outliar analysis. When the full set of experiments is codified and readied for analysis, statistical outliars--those one SD above and below the mean--

[181] Validation procedures are often used in prediction research, where part of a sample is used to develop the predictive or risk instrument and the second part of the sample is used to validate it (e.g., Jones, 1995).

should be identified. Detailed qualitative analysis should then proceed to determine why these studies differ dramatically from the sample majority.

A Final Word on the Philosophy of Meta-Analysis

There is something that is troubling about the underlying rationale for meta-analysis. On the one hand, advocates for meta-analysis claim that narrative reviews rely heavily on statistical significance as a criteria for success, which in turn is influenced by sample size (e.g., Lipsey, 1990). When a finding is statistically insignificant, the traditional reviewer discounts it--essentially saying that the effect can not be trusted. Meta-analysis redirects the focus toward averaging these actual effects of the intervention rather than statistical significance, and accepts this non-significant finding from the small sample study.

Yet, during the analysis stage, the studies are weighted by sample size, essentially giving more emphasis to large sample evaluations. By introducing sample-size weighting, meta-analysis appears to be saying that the effects from small samples can not be implicitly trusted. Is not this what the traditional reviewers have been telling us? Of course, small-sample effect sizes are included, but they are diluted so much by the weighting procedure (at least in the Hedges and Olkin method) that they contribute as little as they do in traditional vote-counting or narrative reviews.

'WHAT WORKS?' REVISITED AGAIN: A META-ANALYSIS OF RANDOMIZED EXPERIMENTS IN INDIVIDUAL-LEVEL INTERVENTIONS

APPENDICES

A-18 DIALOG DATA BASES SEARCHED, IN ALPHABETICAL ORDER

B-NARRATIVE REVIEWS, META-ANALYSES SEARCHED BY YEAR OF STUDY

C-LIST OF 29 JOURNALS MANUALLY SEARCHED IN ALPHABETICAL ORDER

D-LIST OF EIGHT BIBLIOGRAPHIES SEARCHED, IN ALPHABETICAL ORDER

E-CODING INSTRUMENT

F-EXTRACTION RULES

Appendix A.

18 DIALOG DATABASES SEARCHED
IN ALPHABETICAL ORDER

Data Base	Abstracts?	Years Searched	# of Total Documents (as of)	# of Journals	Other Docs Contained?
Academic Index	Yes	1976-1993	1.6 million (2/94)	1500	No
British-HMSO	No	1976-1993	130,000 (2/90)	none	Govt. Pubs
British-Non HMSO	No	1976-1993	46,800 (5/87)	none	Reports
Child Abuse & Neglect	Yes	1965-1993	13,600 (9/91)	unknown	Yes
CJ Periodical Index	No	1975-1993	14,900 (11/88)	100	No
Dissertation Abstracts Theses	>1980	1950-1993	1 million (1/89)	None	M.A.
Family Resources	Yes	1970-1993	12,000 (9/91)	800	Yes
GPO Monthly Catalog	Yes	1976-1993	153,706 (5/83)	None	Govt. Pubs
Legal Resource Index	No	1980-1993	432,000 (11/90)	700	No
MEDLINE	Yes	1966-1993	6.5 million (12/90)	3,700	Yes
Mental Health Abstracts	Yes	1969-1993	475,000 (6/83)	1,000	Yes
NCJRS	Yes	1972-1993	95,000 (10/88)	200	Yes
NTIS	Yes	1964-1993	1.8 million (12/93)	none	Govt. Pubs
PAIS International	>1985	1976-1993	350,000 (11/91)	1200	Yes
PsycINFO	Yes	1967-1973	720,000 (4/90)	1300	Yes
Social SciSearch	No	1972-1993	2 million (5/91)	2400	No
Sociological Abstracts	Yes	1963-1973	281,252 (7/89)	1,600	Yes
U.S. Political Science	Yes	1975-1993	58,500 (4/92)	150	No

Appendix B.

NARRATIVE REVIEWS (N=32) AND META-ANALYSES (N=22) SEARCHED BY YEAR OF STUDY

Narrative Reviews	N of Studies
(1965) Sarri & Vinter	Unknown
(1966) Bailey	100
(1967) Adams	22
(1969) Berleman & Steinburn	5
(1970) Harlow	Unknown
(1971) Robison & Smith	Unknown
(1971) Vetter & Adams	Unknown
(1972) Logan	100
(1973) Slaikeu	23
(1974) Stratton	Unknown
(1975) Lipton, et al.	231
(1976) Brody	65
(1977) Ross & Price	Unknown
(1977) Greenberg	Unknown
(1977) Johnson	29
(1977) Wright & Dixon	96
(1978) Lundman & Scarpitti	Unknown
(1978) Ross & McKay	27

Narrative Reviewers	N of Studies
(1978) Romig	162
(1979) Gendreau & Ross	95
(1982) Linden & Perry	Unknown
(1983) Farrington	42
(1985) Lind	Unknown
(1986) Goldstein	30
(1987) Dillbeck & Abrams	Unknown
(1987) Gendreau & Ross	Unknown
(1988) Basta & Davidson	37
(1988) Dennis	41
(1988) Lab & Whitehead	50
(1989) Firby, et al.	42
(1991) Bazemore	20
(1992) Becker & Hunter	Unknown

Meta-Analyses	
(1984) Davidson, et al.	90
(1984) Garrett	111
(1985) Kaufman	20
(1986) Gensheimer, et al.	44
(1986) Mayer, et al.	39
(1987) Gottschalk, et al.	90
(1987) Gottschalk, et al.	25
(1989) Losel & Koferl	16
(1989) Whitehead & Lab	50
(1990) Andrews, et al.	78
(1990) Izzo & Ross	46
(1991) Roberts & Camasso	46
(1992) Lipsey	443
(1995) Cox, et al.	57
(1995) Hall	12
(1995) Losel	18
(1995) Wells-Parker, et al.	215
(1995) Pearson, et al.	43
(1996) Gendreau & Goggin	Unknown
(1996) Gendreau & Goggin	Unknown
(1996) Redondo, et al.	49
(1996) Pearson, et al.	508

Appendix C.

List of 29 Journals (and Years) Manually Searched In Alphabetical Order

American Sociological Review, 1950-1993
Australian and New Zealand Journal of Criminology, 1974-1993
British Journal of Criminology 1960-1993
Canadian Journal of Criminology, 1969-1993
Crime and Delinquency, 1955-1993
Criminal Justice and Behavior, 1974-1993
Criminal Justice Policy Review, 1986-1993
Criminal Justice Review, 1976-1993
Criminology, 1963-1993
Evaluation Review, 1974-1993
Federal Probation, 1950-1993
International Journal of Applied & Comparative Criminology, 1976-1993
International Journal of the Sociology of Law, 1973-1993
International Journal of Offender Therapy & Comparative Criminology, 1974-1993
Journal of Contemporary Criminal Justice, 1987-1993
Journal of Crime and Justice, 1982-1993
Journal of Criminal Justice, 1973-1993
Journal of Criminal Law and Criminology, 1950-1993
Journal of Legal Studies, 1973-1993
Journal of Offender Rehabilitation, 1980-1993
Journal of Quantitative Criminology, 1986-1993
Journal of Research in Crime and Delinquency, 1968-1993
Journal of Social Service Research, 1977-1993
Justice Quarterly, 1984-1993
Justice System Journal, 1973-1993
Juvenile and Family Court Journal, 1950-1993
Law and Society Review, 1972-1993
Prison Journal, 1950-1993
Violence and Victims, 1986-1993

Appendix D.

LIST OF EIGHT BIBLIOGRAPHIES SEARCHED IN ALPHABETICAL ORDER

Bibliographies	Topic	Total Cites	Leads
Berens (1987)	Government documents	1,094	23
Beyleveld (1980)	General deterrence	800+	0
Boruch, et al. (1977)	Experiments	300+	75
Cordasco & Alloway (1985)	Crime in USA	1,879	13
Goyer-Michaud (1974)	Female offenders	200+	0
Hewitt, et al. (1985)	Criminal justice in USA	813	11
Monahan, et al. (1981)	American jails	200+	720
Trudel, et al. (1976)	Recidivism	48	0

WHAT WORKS? REVISITED AGAIN: A META-ANALYSIS OF EXPERIMENTS IN INDIVIDUAL-LEVEL INTERVENTIONS

INFORMATION EXTRACTION SHEET

ID # ___

Name of EXP: ____________________

A. DOCUMENT INFORMATION

A1. PRIM DOC YEAR	
A2. TYPE OF PRIMARY DOC/FIELD	
A3. TOTAL DOCS USED	
A4. TOTAL PERCENT PRIM DOC	
A5. NUMBER EXPS IN PRIM DOC	

B. INVESTIGATOR INFORMATION

B1. PI #1 AFFILIATION/FIELD	
B2. PI #2 AFFILIATION/FIELD	
B3. WERE PIS TREAT PROVIDERS	YES / NO
B4. WERE PIS GOVT RESEARCHERS	YES / NO
B5. PIs WERE OUTSIDE RESEARCHERS	YES / NO

C. EXPERIMENT INFORMATION

C1. YEAR EXP STARTED	
C2. LENGTH OF EXPERIMENT	
C3. WAS IT A MULTISITE PROJECT	YES / NO
C4. REGION	
C5. SCOPE OF THE EXPERIMENT	* statewide * county * city/town * institution * other ("comments")

D. RANDOMIZATON INFORMATION

D1. METHOD OF RANDOMIZATION	
D2. STRATIFICATION/BLOCK/MATCH	YES / NO
D3. IF YES on D2, INDICATE....	* blocking * stratification * matching
D4. IF YES TO D2, ON WHAT VARIABLES....	LIST:
D5. # OF RANDOMIZED GROUPS	
D6. PRETESTS EQUIV REPORTED	YES / NO
D7. RESULTS...	* no significant diff * favors experimental * favors control
D8. PRACTITIONER OPPORTUNITY TO SUBVERT RANDOMIZATION	* high * med * low * none
D9. MISSASS/OVERRIDE REPORTED	YES / NO
D10. TOTAL % MISASSIGN/OVERRIDE	
D11. HOW MISSASS/OVERRIDE HANDLED	* analyze as assign * analyze as deliver * both * other ("comments")

E. SUBJECT GENERALIZABILITY INFORMATION

E1. VOLUNTARY/CONSENT	YES / NO
E2. PAYMENT TO SUBJECTS	YES / NO
E3. ELIGIBILITY CRITERIA...	LIST:

F. RANDOMIZED GROUPS

	TYPE OF CONDITION SUBJECTS ASSIGNED	N
ONE		
TWO		
THREE		
FOUR		
FIVE		
SIX		
SEVEN		
EIGHT		

G. SUBJECTS

VARIABLE INFORMATION

G1. PERCENT WHITE ______________

G2. PERCENT MALE ______________

G3. AVERAGE AGE ______________

G4. AGE RANGE ______________

G5. HIGHEST %/TYPE AGE GROUP ______________

G6. AVER GRADE COMPLETED ______________

G7. HIGHEST %/TYPE EDUC GROUP ______________

G8. AVERAGE IQ ______________

G9. PRIOR RECORD INFORMATION ______________

G10. INSTANT OFFENSE INFO ______________

USE "COMMENTS" TO RECORD OTHER PERTINENT INFORMATION.

H. CRIME REDUCTION PROGRAM CONDITION

H1.TREATMENT AT WHAT POINT IN CJ SYSTEM ______________

H2. AGENCY TYPE DELIVERING TREATMENT ______________

H3. HOW IS TREATMENT DELIVERED ______________

H4. CONTACT INFORMATION

Daily contact (hrs per day) ________
Weekly contact (days per week) ________
Total weeks ________

H5. WAS TREATMENT MONITORED YES / NO

H6. TREATMENT PROBLEMS YES / NO

H7. IF YES TO H6, WHAT TYPE.... LIST:

I. OTHER POTENTIAL PROBLEMS

I1. WERE ATTRITION PROBLEMS NOTED: YES / NO

I2. IF SO, LIST TYPE

I3. HOW WERE ATTRIT PROB HANDLED

I4. WERE CASEFLOW PROBLEMS NOTED: YES / NO

I5. IF SO, LIST TYPE

I6. HOW WERE CASEFLOW PROBHANDLED:

J. OUTCOME GENERAL

J1. TOTAL NUMBER OF FOLLOWUPS	
J2. MINIMUM FOLLOWUP IN MOS.	
J3. MAXIMUM FOLLOWUP IN MOS.	
J4. TOTAL CRIME OUTCOMES	
J5. CRIME OUTCOME TYPES	LIST:
J6. TOTAL NON-CRIME OUTCOMES	
J7. NON-CRIME OUTCOME TYPES	LIST:

K. OUTCOME SPECIFIC

Information	Crime Effect One	Crime Effect Two	Crime Effect Three
# Months Followup			
Crime Measure			
Data From?			
Direction of Effect			
Statistical Significance?			
Test Used			
Number of Tails			
Probability			
Test Value			
Small Sample or Statistical Power Mentioned?			

L. EFFECTS

Information	Crime Effect 1	N	Crime Effect 2	N	Crime Effect 3	N
TYPE DATA USED						
GROUP ONE						
GROUP TWO						
GROUP THREE						
GROUP FOUR						
GROUP FIVE						
GROUP SIX						
GROUP SEVEN						
GROUP EIGHT						

M. SUBGROUP EFFECTS

M1. Subgroup Analyses.....Any reported? Yes / No

M2. If reported...................Any differences found? Yes / No

M3. List subgroup effects

POSITIVE NEGATIVE

FREE FORM COMMENTS (Any substantive comments, criticisms of studies, etc.)

APPENDIX F.

WHAT WORKS? REVISITED AGAIN:

A META-ANALYSIS OF RANDOMIZED FIELD EXPERIMENTS IN INDIVIDUAL-LEVEL INTERVENTIONS

EXTRACTION RULES

A. DOCUMENT INFORMATION

(A1) **Prim Doc Year**

Code the publication year of the primary document.

The primary document is the document you are using to get the crime outcome information (data on the crime reduction effect of the intervention). In cases where more than one document contains crime reduction information, the primary document is the one containing the most information about the study.

NOTE:
While there may be a primary document which is relied upon, information can be extracted from all sources available on the study. If there is a discrepancy between documents, rely upon the primary document unless one of the other documents was an update or correction. Also note this in the "comments" field.

(A2) **Type of Primary Doc/Field**

For Type: What type of document category would the primary document fall under? Was it a book? Was it a chapter from an edited book? Was it an academic journal article? Was it a dissertation?

For Field: What field or discipline of study would you classify the primary document type? For example, if found in an academic journal, was the journal from psychology, social work or criminal justice?

(A3) **Total Docs Used**

What was the total number of documents, including the primary document, that you used to code information about this experiment? This includes documents containing reanalyses, longer follow-ups, critical analyses, treatment or staff descriptions, etc. If the other document was used to confirm information in the primary document, include it here. However, even if a document was available, do not include it if it was not relied upon at all.

(A4) **Total Percent Prim Doc**

Simply estimate how much information--out of all the information available on the experiment--was extracted from the primary document. If you had only one document to work with, then 100% of all information was extracted from that document.

(A5) **Number Exps in Prim Doc**

How many randomized experiments eligible for this meta-analysis were described in the primary document? In other words, if the primary document contains the results of two experiments eligible for this meta-analysis, you should enter "2" in this space and code each of those experiments separately. Each will become a distinct case in the database.

B. INVESTIGATOR INFORMATION

(B1) **PI #1 Affiliation/Field**

Indicate the affiliation of the first author of the primary document and the field or discipline of that affiliation. For example, if PI #1 is an assistant professor of psychology at a college, then affiliation = academic and field = psychology.

(B2) **PI #2 Affiliation/Field**

Repeat for PI #2.

(B3) **Were PIs Treat Providers**

If the PIs (those who conducted the research) were also the treatment providers, indicate "Yes."

(B4) **Were PIs Govt Researchers**

If the PIs conducted an evaluation of a government program in the context of their employment as government researchers, indicate "Yes."

(B5) **PIs were outside researchers**

If PIs were outside researchers, either from private research centers or academic institutions, indicate "Yes."

C. EXPERIMENT INFORMATION

(C1) **Year Exp Started**

What year did the experiment start? Generally, this means the year that random assignment began.

(C2) **Length of Experiment**

How long was the experiment? How long did assignment of cases and maintenance of treatment conditions exist? This does not include follow-up periods that extend past the end of treatment for the last case.

(C3) **Was it a Multisite Project**

If the experiment was conducted in several sites (different towns, cities, departments, prisons, etc.), indicate "Yes."

(C4) **Region**

What country or state (if USA) did experiment take place?

(C5) Scope of the Experiment

What was the scope of the experiment? Was it citywide? Statewide? Was it limited to one prison? Was it county or multiple sites?

D. RANDOMIZATION INFORMATION

(D1) Method of Randomization

How did PIs randomly assign subjects? Did they use a random numbers table? A lottery? A coin toss? A toss of die?

(D2) Stratification/Block/Match

Were any techniques used prior to randomization, such as stratification, blocking or matching, that would insure a greater probability of equivalence between groups? If so, then indicate "Yes."

(D3) If Yes to D2, indicate...

Circle the technique used (blocking, stratification, or matching).

(D4) If Yes to D2, on What Variables...

Indicate which variables the PI(s) used in stratification, blocking or matching to insure greater probability of group equivalence.

(D5) # of Randomized Groups

Indicate number of study conditions or groups involved in the randomized experiment. Do not include comparison or convenience groups that were not randomly assigned.

(D6) Pretests Equiv Reported

Indicate "Yes" if PI(s) report that they performed tests to determine the equivalence of groups post-randomization but prior to treatment.

(D7) **Results**

Circle the appropriate result of the pretest for equivalence (if tests showed difference between groups, did difference provide a potential bias in favor of the experimental group or control group?).

(D8) **Practitioner Opportunity to Subvert Randomization**

Rate the control the PI(s) had over the randomization process. If PIs completely controlled randomization and line staff were "blind" (unaware) of the assignment, then code "none." If practitioners knew the assignment and were able to tamper with it in anyway, then rate the opportunity for subversion.

(D9) **Missass/override reported**

Did PI(s) report that random assignment broke down or was violated in some way? In other words, do PI(s) report that some cases did not receive the condition as randomly assigned? If so, then indicate "Yes."

(D10) **Total % Missassign/Override**

Indicate the total percentage of missassigned or override cases.

(D11) **How Missassign/Override Handled**

Indicate how the PI(s) handled the misassignments/overrides in their outcome analysis. Did they analyze the groups as randomly assigned? Did they analyze them as actually delivered? Did they conduct both analyses? If another method for handling misassignment was used, then indicate "other" and include a description in the comment section on page 7.

E. <u>SUBJECT GENERALIZABILITY INFORMATION</u>

(E1) **Voluntary/Consent**

If subjects volunteered for the experiment, or gave their consent to be a part of the experiment, indicate "Yes."

(E2) **Payment to Subjects**

If subjects were paid for participation in the experiment, indicate "Yes."

(E3) **Eligibility Criteria**

List the criteria for including or excluding subjects from the experiment.

F. RANDOMIZED GROUPS

There should be entries for each of the conditions totaled in item D3. For example, if item D3 indicated that there were four randomized study groups, then each of the four groups should be listed here. For each study condition, state the most precise description of the condition (e.g., "returned for regular court processing," "psychotherapy," "halfway house") and the number of subjects randomly assigned to the condition.

G. SUBJECTS (Note: any pertinent information not covered by these items can be included in the Comments section on page 7).

(G1) **Percent White**

Simply indicate the total percentage of white or Caucasian subjects in the experiment.

(G2) **Percent Male**

Simply indicate the total percentage of males in the experiment.

(G3) **Average Age**

Use the mean or median age, if provided, as the average age.

(G4) **Age Range**

Provide the range of ages involved in the experimental design (e.g., ages 12-18).

(G5) **Highest %/Type Age Group**

Indicate the age bracket with the greatest percentage of subjects. For example, 54% of subjects were 16-18 years of age.

(G6) **Aver Grade Completed**

Provide the average mean grade of education completed (e.g., 10.9 years)

(G7) **Highest %/Type Educ Group**

Indicate the education level group with the highest percentage of subjects. For example, 55% of subjects completed high school or GED.

(G8) **Average IQ**

Provide the mean or median IQ score if given.

(G9) **Prior Record Information**

Provide as much detail as possible on prior record. Include average number of prior arrests, and the type of offenses committed.

(G10) **Instant Offense Information**

Provide as much detail as possible on the instant offense--which is the present offense they have been arrested or taken to court for.

H. CRIME REDUCTION PROGRAM CONDITION

(H1) **Treatment at What Point in CJ System**

At which point in the criminal justice process is this experiment taking place? For example, if the experiment involved a group counseling program in the prison, then indicate "while in prison."

(H2) **Agency Type Delivering Treatment**

How would you classify the agency delivering the treatment or intervention? For example, if the treatment was intensive probation supervision, then the agency type was "criminal justice" or "probation department."

(H3) **How is treatment delivered**

Is treatment delivered individually (1:1 staff-subject ratio) or is it administered to groups (e.g., group counseling, classroom)? Or was treatment delivered to an individual as part of a large overall caseload?

(H4) **Contact information**

Indicate the level of contact or treatment intensity by coding three sub-items:

Daily contact (hrs. per day): How many hours per day were subjects in treatment

Week contact (days per wk): How many days per wk were subjs in treat

Total weeks: How many total weeks were subjects in treatment

(H5) **Was Treatment Monitored**

Did PIs report that they monitored treatment conditions to determine the nature and quality of treatment being delivered? If so, indicate "Yes."

(H6) **Treatment Problems**

If the PIs indicate that there were problems or breakdowns in treatment integrity, then write "Yes."

(H7) **If Yes to H6, What Type....**

If you indicated "Yes" to (H6), then indicate the nature of the treatment integrity breakdown. For example, staff resistance to experiment led to changes in proposed treatment services, etc.

I. Other Potential Problems

(I1) **Were Attrition Problems Noted**

If there was a loss of subjects from randomization to the crime outcome posttest, indicate "Yes."

(I2) **If so, List Type**

Briefly describe the loss of experimental subjects. Include percentages lost from each subject condition if possible.

(I3) **How Were Attrit Prob Handled**

How did the PI(s) handle loss of subjects? Was any analysis done to determine if subjects dropped differed from subjects who remained?

(I4) **Were Caseflow Problems Noted**

If PI(s) indicated that there were caseflow problems (i.e., insufficient number of subjects eligible for the study), then circle "Yes."

(I5) If So, List Type

Briefly describe the nature of caseflow problems.

(I6) How Were Caseflow Prob Handled:

Indicate how the PI(s) handled caseflow problems. Did they alter randomization to allow more subjects into the treatment group? Did they change eligibility requirements? Did they extend the experiment length?

J. OUTCOME GENERAL

(J1) Total Number of Follow-ups

Indicate the total number of crime outcome follow-up measurements taken by the investigators. For example, if investigators measured rearrests at 1 year and 2 years, then indicate "2."

(J2) Minimum Follow-up in Months

When was the earliest crime outcome measure follow-up in months taken? In the example from (J1), you would indicate "12 months" here.

(J3) Maximum Follow-up in Months

When was the latest crime outcome measure follow-up taken? In the example from (J1), you would indicate "24 months" here.

(J4) Total Crime Outcomes

Provide the total number of crime outcome measures used by the PI(s). For example, if they reported rearrests, reconvictions and reincarcerations, then indicate "3."

(J5) **Crime Outcome Types**

List each of the crime outcome measures used by the PI(s). The total here should match the response to item (J4).

(J6) **Total Non-Crime Outcomes**

Provide the total number of non-crime outcome measures used by the PI(s). For example, if they reported suspensions, school attendance, and grade point average, then indicate "3."

(J7) **Non-Crime Outcome Types**

List each of the non-crime outcome measures used by the PI(s). The total here should match the response to item (J6).

K. OUTCOME SPECIFIC

For this table, only three crime effect outcomes are coded. The rules for selecting crime outcome effects are:

* Although most studies will only report one follow-up period (e.g., one year), crime effect one, two and three should correspond to the first three follow-up periods (one, two, five year follow-ups).

* Use only official crime measures (arrest, police contact, court petitions, convictions, jail time, summonses, etc.). Do not include self-reports, victim interviews, positive drug urines, etc.

* If two or more crime effect measures are available for the same follow-up period (e.g., six months), then use the official measure which occurs the EARLIEST in the criminal justice system. Generally speaking, this will be rearrest or police contact data.

* However, for probation and parole experiments, only use violation or revocation data if it is the only official crime outcome data available (due to their bias toward experimental subjects).

For each crime effect to be coded, the following information is requested:

# Months Follow-up: Indicate the length of the follow-up period in months being reported.

Crime Measure: What is the official crime measure being used?
Data From?: Where did the data measuring the official crime outcome come from (e.g., courts, police station, etc.)?

Direction of Effect: Was the effect of the program positive or negative in direction? Or was there absolutely no difference between groups?

Statistical Signif?: Was the difference between groups statistically significant?

Test Used: What statistical test was used to test differences between groups (e.g., chi, anova, t-test, etc.)?

Number of Tails: This will either be "2" or "1."

Probability: Indicate the actual probability level of the outcome result (e.g. p=.08).

Test Value: Indicate the actual score or value of the statistical test (e.g., F=10.25).

Small Sample or Statist. Power Mentioned? Did PI(s) mention small sample or statistical power as reason for failure to find a significant effect?

L. <u>EFFECTS</u> (Note: Group One must be the same group as Group One under Section F. This will allow us to associate the effects with the appropriate treatment or condition).

More specific information is required on crime effect 1, crime effect 2, and crime effect 3. This information will be used, if it is available, to compute the effect size needed for analysis.

For each of the three crime effects, provide the following information:

Type data used: This refers to the type of statistical data being used to create effect size. In most studies, it will be the proportion or percentage of failure for each of the study groups. In some cases, it will be the mean number of arrests or police contacts for the group.

Group One...Eight: For each study condition, provide the statistical data in the columns marked Crime Effect 1, Crime Effect 2, and Crime Effect 3. Suppose the experiment involved two groups, and failure rates were used, then indicate the failure rate for group one in the appropriate space.

. Indicate the n of subjects involved in the posttest crime effect analysis. This will probably not be the same as the N of subjects randomly assigned to the group (as indicated in Section F), given attrition rates.

M. SUBGROUP EFFECTS

(M1) **Subgroup Analyses....Any reported?**

Indicate "Yes" if PI(s) reported that they analyzed subgroups to determine if the crime reduction program had any differential effects (e.g., PIs examined effect on males or females, etc.)

(M2) **If reported, ...Any differences found?**

Indicate "Yes" if PI(s) reported that they found subgroup differences on the effect of the crime reduction program.

M3) **List Subgroup Effects**

Simply list the subgroup analyses that showed a positive impact of the crime reduction program (e.g., females, older inmates, no priors), and those which showed a negative impact of the crime reduction program (younger offenders, males, etc.).

COMMENTS

Reserved for any comments, criticisms of studies, additional information we might want to look at, etc.)

References

Abrami, Phillip C., Peter A. Cohen and Sylvia d'Apollonia (1988). Implementation problems in meta-analysis. *Review of Educational Research 58 (2): 151-179.*

Adams, Reed and Harold J. Vetter (1971). Probation caseload and recidivism rate. *British Journal of Criminology 11: 390-393.*

Adams, S. (1976). Evaluation: A way out of the rhetoric. Pgs. 75-91 in Martinson, R., T. Palmer and S. Adams (eds.) *Rehabilitation, Recidivism, and Research.* Hackensack, N.J.: National Council on Crime and Delinquency.

Adams, S. (1974). Evaluative research in corrections: Status and prospects. *Federal Probation 33 (1): 14-21.*

Adams, S. (1970). The PICO Project. Pgs. 548-561 in N.B. Johnston, L. Savitz and M.E. Wolfgang (eds.) *The Sociology of Punishment and Correction.* New York: Wiley.

Adams, Stuart (1967). Some findings from correctional caseload research. *Federal Probation 31: 48-57.*

Alexander, J.F. and B.V. Parsons (1973). Short-term behavioral intervention with delinquent families. *Journal of Abnormal Psychology 81 (3): 219-225.*

Allen, F.A. (1981*). The Decline of the Rehabilitative Ideal: Penal Policy and Social Purpose.* New Haven, CT: Yale University Press.

Allen, F.A. (1959). Criminal justice, legal values and the rehabilitative ideal. *Journal of Criminal Law, Criminology and Police Science 50 (2): 226-232.*

American Friends Service Committee (1971). *Struggle for Justice.* New York: Hill and Wang.

Andenaeus, John C. (1966). The general preventive effects of punishment. *University of Pennsylvania Law Review 114 (7): 949-983.*

Andrews, D.A., I. Zinger, R.D. Hoge, J. Bonta, Paul Gendreau, and Francis T. Cullen (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology 28 (3):369-404.*

Antonowicz, Daniel H. and Robert R. Ross (1993). *Essential components of successful rehabilitation programs for offenders.* Unpublished manuscript, University of Ottawa, Ontario, Canada.

Babbie, Earl (1992). *The Practice of Social Research. Sixth Edition.* Belmont, CA: Wadsworth.

Babbie, Earl (1983*). The Practice of Social Research. Fourth Edition.* Belmont, CA: Wadsworth.

Bailey, William C. (1966). Correctional outcome: An evaluation of 100 reports. *The Journal of Criminal Law, Criminology and Police Science 57 (2): 153-160.*

Bangert-Drowns, R.L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin 99:388-399.*

Barrow, C.R. (1978). *Field experimentation: One approach to contemporary issues concerning the deterrence doctrine.* Doctoral dissertation, University of Arizona. Ann Arbor, MI: University Microfilms International.

Basta, J.M. and W.S. Davidson (1988). Treatment of juvenile offenders: Study outcomes since 1980. *Behavioral Sciences and the Law 6: 355-384.*

Bazemore, Gordon (1991). New concepts and alternative practice in community supervision of juvenile offenders: Rediscovering work experience and competency development. *Journal of Crime and Justice 14: 27-52.*

Becker, Judith V. and John A. Hunter (1992). Evaluation of treatment outcome for adult perpetrators of child sexual abuse. *Criminal Justice and Behavior 19: 74-92.*

Berens, John F. (1987). *Criminal Justice Documents.* New York: Greenwood Press.

Berk, Richard A., Robert F. Boruch, D.L. Chambers, Peter H. Rossi and Ann D. Witte (1985). Social Policy Experimentation: A Position Paper. *Evaluation Review 9: 387-430.*

Berleman, W.C. and T.W. Steinburn (1969). The value and validity of delinquency prevention experiments. *Crime and Delinquency 15: 471-478.*

Beyleveld, Deryck (1980). *A Bibliography on General Deterrence.* London: Saxon House.

Binder, A. and J.W. Meeker (1988). Experiments as reforms. *Journal of Criminal Justice 16: 347-358.*

Blumenthal, Murray and H. Laurence Ross (1973). *Two Experimental Studies of Traffic Law: The Effect of Legal Sanctions on DUI Offenders and the Effect of Court Appearance on Traffic Law Violators.* National Highway Transportation & Safety Administration, contract DOT-HS-249-2-437.

Blumstein, Alfred, Jacqueline Cohen and Daniel Nagin (1978). *Deterrence and Incapacitation. Report of the National Academy of Sciences Panel on Research on Deterrent and Incapacitative Effects.* Washington, DC: National Academy Press.

Borduin, Charles M., Scott W. Henggeler, David M. Blaske and Risa J. Stein (1990). Multisystemic treatment of adolescent sexual offenders. *International Journal of Offender Therapy and Comparative Criminology 34 (2): 105-113.*

Boruch, R.F. (1975). On common contentions about randomized field experiments. Pgs. 107-142 in R.F. Boruch and H.W. Reicken (eds.) *Experimental Testing of Public Policy: The Proceedings of the 1974 Social Sciences Research Council Conference on Social Experimentation.* Boulder, CO.: Westview Press.

Boruch, Robert F., A.J. McSweeney, and E.J. Soderstrom (1978). Bibliography: Illustrative field experiments. *Evaluation Quarterly 4: 655-695.*

Britt III, Chester L., Michael R. Gottfredson and John S. Goldkamp (1992). Drug testing and pretrial misconduct: An experiment on the specific deterrent effects of drug monitoring defendants on pretrial release. *Journal of Research in Crime and Delinquency 29 (1): 62-78.*

Brody, S.R. (1976). *The Effectiveness of Sentencing: A Review of the Literature.* London: Her Majesty's Stationary Office.

Brown, Steven R. and Lawrence E. Melamed (1990). *Experimental Design and Analysis.* Beverly Hills, CA: Sage.

Brown, S.E. (1989). Statistical power and criminal justice research. *Journal of Criminal Justice 17:115-122.*

Bryant, Fred B. and Paul M. Wortman (1984). Methodological issues in the meta-analysis of quasi-experiments. Pgs. 5-24 in W.H. Yeaton and Paul M. Wortman (Eds.) *Issues in Data Synthesis. New Directions for Program Evaluation, No. 24.* San Francisco: Jossey-Bass.

Bullock, R.J. and Daniel J. Svyantek (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication and evaluation criteria. *Journal of Applied Psychology 70 (1): 108-111.*

Burton, V.S., R.G. Dunaway and R. Kopache (1993). To punish or rehabilitate? A research note assessing the purposes of state correctional departments as defined by state legal codes. *Journal of Crime and Justice 16:177-188.*

Camp and Thyer (1993). Treatment of adolescent sex offenders: A review of empirical research. *Journal of Applied Social Sciences 17 (2): 191-206.*

Campbell, D.T. (1969). Reforms as experiments. *American Psychologist 24: 409-429.*

Campbell, Donald T. and Julian L. Stanley (1966). *Experimental and Quasi-Experimental Designs For Research.* Chicago: Rand McNally.

Carranza, Elias, Marion Houved and Luis Paulino Mora (1994). Release on personal recognizance in Costa Rica: An experimental research study. Pgs. 439-461 in Ugljesa Zvekic (ed.) *Alternatives to Imprisonment in Comparative Perspective.* Chicago: Nelson-Hall.

Catalano, R.J., J., Hawkins and E. Wells (1991). Evaluation of the effectiveness of adolescent drug abuse treatment, assessment of risks for relapse and promising approaches for relapse prevention. *International Journal of the Addictions 25:1085-1143.*

Chaneles, Sol (1975). Review of <u>Effective Correctional Treatment</u>. *The Prison Journal.*

Chelimsky, Eleanor and Linda G. Morra (1984). Evaluation synthesis for the legislative user. Pgs. 75-89 in W.H. Yeaton and P.M. Wortman (eds.) *Issues in Data Synthesis. New Directions in Program Evaluation, No. 24.* San Francisco: Jossey-Bass.

Chronicle of Higher Education, "Footnotes," July, 1990, A4

Clarke, R.V.G. and D.B. Cornish (1972). *The Controlled Trial in Institutional Research—Paradigm or Pitfall for Penal Evaluators?* London: H.M. Stationary Office.

Clear, Todd R. (1997). *Personal communication.*

Clear, Todd R. (1994). *Harm in American Penology.* NY: University of Albany Press.

Clear, Todd R. (1978). Correctional policy, neo-retributionism, and the determinate sentence. *Justice System Journal 4 (4):26-48.*

Cochran, William G. and G.M. Cox (1992). *Experimental Designs.* New York: John Wiley and Sons, Inc.

Cohen, Jacob (1992). Meta-analysis methodically considered. Review of Wachter and Straf (Eds.) <u>The Future of Meta-Analysis</u>. *Contemporary Psychology 37 (4): 375-376.*

Cohen, Jacob (1977). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, N.J.: Erlbaum Publishers.

Cohen, Jacob (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Psychology 65:145-153.*

Connor, R.F. (1977). Selecting a control group. An analysis of the randomization process in twelve social reform programs. *Evaluation Review 1 (2): 195-244.*
Conrad, John P. (1975). We should never have promised a hospital. *Federal Probation 39 (4):3-9.*

Cook, Phillip (1980). A review of deterrrence research. In M. Tonry and N. Morris (eds.) *Crime and Justice: An Annual Review of Research.* Chicago: University of Chicago Press.

Cook, T.D. and D.T. Campbell (1979*). Quasi-experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally.

Cook, Thomas D., Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis and Frederick Mosteller (1992*). Meta-Analysis for Explanation.* New York: Russell Sage Foundation.

Cooper, Harris C. and Larry V. Hedges (1994). *The Handbook of Research Synthesis.* New York: Sage.

Cooper, Harris C. (1989*). Integrating Research. A Guide for Literature Reviews. Second Edition.* Beverly Hills, CA: Sage.

Cordasco, Francesco and David Alloway (1985*). Crime in America: Historical Patterns and Contemporary Realities.* New York: Garland.

Cordray, David S. (1990). An assessment from the policy perspective. Pgs. 99-119 in Kenneth Wachter and Miron L. Straf (eds.) *The Future of Meta-Analysis.* Beverly Hills, CA: Sage.

Cox, Stephen M., William S. Davidson and Timothy S. Bynum (1995). A meta-analytic assessment of delinquency related outcomes of alternative education programs. *Crime & Delinquency 41:219-234.*

Craft, Michael, Geoffrey Stephenson and Clive Granger (1964). A controlled trial of authoritarian and self-governing regimes with adolescent psychopaths. *American Journal of Orthopsychiatry 34: 543-554.*

Cullen, F.T. and P. Gendreau (1989). The effectiveness of correctional rehabilitation. In L. Goodstein and D. MacKenzie (eds.) *The American Prison.* New York: Plenum.

Cullen, F.T. and K.E. Gilbert (1982). *Reaffirming Rehabilitation.* Cincinnati, OH: Anderson.

D'Alessio, S.J. and L. Stolzenberg (1995). The impact of sentencing guidelines on jail incarceration in Minnesota. *Criminology 33 (2):283-302.*

Davidson, William S., R., Gottschalk, L. Gensheimer and J. Mayer (1984). *Interventions with juvenile delinquents: A meta-analysis of treatment efficacy.* Unpublished manuscript, Psychology Department, Michigan State University.

Dennis, Michael L. (1988). *Implementing Randomized Field Experiments: An Analysis of Criminal and Civil Justice Research.* Ph.D. Dissertation, Northwestern University. Ann Arbor, MI: University Microforms.

Dennis, M. and R. Boruch (1989). Randomized experiments for planning and testing projects in developing countries: Threshold conditions. *Evaluation Review 13:292-309.*

Di Gennaro, Giuseppe and Eduardo Vetere (1974). The crisis of the concept of correctional treatment. *International Journal of Criminology and Penology 2:295-314.*

Dillbeck, Michael C. and Allan I. Abrams (1987). The application of the transcendental meditation program to corrections. *International Journal of Comparative and Applied Criminal Justice 11: 111-132.*

Dole, Vincent P., J.W. Robinson, J. Orraca, and E. Towns (1969). Methadone treatment of randomly selected criminal addicts. *The New England Journal of Medicine 280 (25): 1372-1375.*

Dunford, Franklyn W. (1990). Random assignment: Practical considerations from field experiments." *Evaluation and Program Planning 13: 125-132.*

Durlak, Joseph A. and Mark W. Lipsey (1991). A practitioner's guide to meta-analysis. *American Journal of Community Psychology 19 (3): 291-332.*

Empey, L.T. (1980). Field experimentation in criminal justice - Rationale and design. Pgs 143-176 in M.W. Klein and K.S. Teilmann (eds.) *Handbook of Criminal Justice Evaluation.* Beverly Hills: Sage.

Empey, Lamar T. and M.L. Erickson (1972). *The Provo Experiment.* Lexington: D.C. Heath.

Erez, E. (1986). Randomized experiments in correctional context: Legal, ethical and practical concerns. *Journal of Criminal Justice 14: 389-400.*

Eysenck, H.J. (1978). An exercise in mega-silliness. *American Psychologist 33: 517.*

Eysenck, H.J. (1961). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology 16: 319-324.*

Fagan, Jeffrey A. (1990a). Treatment and reintegration of violent juvenile offenders: Experimental results. *Justice Quarterly 7 (2): 233-263.*

Fagan, Jeffrey (1990b). Natural experiments in criminal justice. Pgs. 108-137 in Kimberly Kempf (ed.) *Measurement Issues in Criminology.* New York: Springer-Verlag.

Farrington, David P. (1983). Randomized experiments on crime and justice. Pgs. 257-308 in Michael Tonry and Norval Morris (Eds.) *Crime and Justice: An Annual Review of Research. Volume IV.* Chicago: University of Chicago Press.

Farrington, David P. (1979). Delinquent behavior modification in the natural environment. *British Journal of Criminology 19 (4):353-372.*

Farrington, David P., Lloyd E. Ohlin and James Q. Wilson (1986*). Understanding and Controlling Crime.* New York: Springer-Velag.

Feldman, Kenneth A. (1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education 44 (Winter): 86-102.*

Federal Judicial Center Advisory Committee on Experimentation in the Law (1981). *Experimentation in the Law.* Washington, DC: Federal Judicial Center.

Feinberg, J. and H. Gross (1983). *The Philosophy of Law. Second Edition.* Belmont, CA: Wadsworth.

Fischer, Joel (1978). Does anything work*? Journal of Social Service Research 1(3):215-243.*

Fogel, D. (1975). *"...We are the Living Proof..." The Justice Model for Corrections.* Cincinnati: Anderson.

Freidman, Alfred S. (1989). Family therapy versus parent groups: Effects on adolescent drug users. *American Journal of Family Therapy 17 (4): 335-347.*

Friedman, Herbert (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin 70 (4):245-251.*

Furby, Lita, Mark R. Weinrott, and Lynn Blackshaw (1989). Sex offender recidivism: A review." *Psychological Bulletin 105 (1): 3-30.*

Garner, Joel and Christy Visher (1988). Experiments help shape new policies. *N.I.J. Reports. No. 211 (Sept/Oct): 2-8.*

Garner, Joel, Jeffrey Fagan, and Christopher D. Maxwell (1995). Published findings from the NIJ spouse assault replication program: A critical review. *Journal of Quantitative Criminology 8(1):1-29.*

Garrett, Carol J. (1985). Effects of residential treatment on adjudicated delinquents: A meta-analysis. *Journal of Research in Crime and Delinquency 22 (4): 287-308.*

Garrett, Carol J. (1984). *Meta-analysis of the effects of institutional and community residential treatment on adjudicated delinquents.* Unpublished doctoral dissertation, University of Colorado.

Gartin, Patrick R. (1995). Dealing with design failures in randomized field experiments: Analytic issues regarding the evaluation of treatment effects. *Journal of Research in Crime and Delinquency 32 (4):425-445.*

Garvey, W.D. and B.C. Griffith (1971). Scientific communication: Its role in the conduct of research and the creation of knowledge. *American Psychologist 26:349-361.*

Gelber, R.D. and M. Zelen (1985). Planning and reporting clinical trials. In P. Calabrese, P.S. Schein and S.A. Rosenberg (eds.*). Basic Principles and Clinical Management of Cancer.* New York: .MacMillan.

Gendreau, Paul and Claire Goggin (1996). Principles of effective correctional programming with offenders. *Forum on Corrections Research 8 (3).*

Gendreau, Paul and Robert R. Ross (1987). Revivification of Rehabilitation: Evidence from the 1980s. *Justice Quarterly 4 (3): 349-407.*

Gendreau, Paul and Robert Ross (1983-84). Correctional treatment: Some recommendations for effective intervention. *Juvenile and Family Court Journal (Winter): 31-39.*

Gendreau, Paul and Robert R. Ross (1979). Effective correctional treatment: Bibliotherapy for cynics. *Crime and Delinquency 25 (4): 463-489.*

Gensheimer, Leah K., Jeffrey P. Mayer, Rand Gottschalk, and William S. Davidson (1986). Diverting youth from the juvenile justice system: A meta-analysis of intervention efficacy. Pgs. 39-57 in Steven J. Apter and Arnold P. Goldstein (eds.) *Youth Violence.* New York: Pergamon Press.

Gibbons, Don (1992). The limits of punishment as social policy. Pgs. 12-28 in Clayton Hartjen and Edward Rhine (eds.) *Correctional Theory and Practice.* Chicago: Nelson-Hall.

Glaser, Daniel (1995). *Personal communication.*

Glaser, Dan (1975). Achieving better questions: A half century's progress in correctional research. *Federal Probation 39 (3):3-9.*

Glaser, Dan (1971). Five practical research suggestions for correctional administrators. *Crime and Delinquency (January):32-40.*

Glaser, Daniel (1965). Correctional research: An elusive paradise. *Journal of Research in Crime and Delinquency 2 (1): 1-11.*

Glass, Gene V. (1995). Book Review, The Handbook of Research Synthesis. *Contemporary Psychology 40:736-737.*

Glass, Gene V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher 5: 3-8.*

Glass, Gene V. and R.M. Kliegl (1983). An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology 51 (1):28-41.*

Glass, Gene V., Barry McGaw and Mary L. Smith (1981). *Meta-Analysis in Social Research.* London: Sage.

Gleser, L.J. and I. Olkin (1994). Stochastically dependent effect sizes. Pgs. 339-356 in H. Cooper and L. Hedges (eds.) *Handbook of Research Synthesis.* New York: Sage.

Goldstein, Arnold P. (1986). Psychological skill training and the aggressive adolescent. Pgs. 89-119 in Steven Apter and Arnold P. Goldstein (eds.) *Youth Violence. Programs and Prospects.* New York: Pergamon.

Gondolf, E. (1997). Batterer programs: What we know and need to know. *Journal of Interpersonal Violence, 12: 83-98.*

Gottfredson, Michael and T. Hirschi (1990). *A General Theory of Crime.* Stanford, CA: Stanford University Press.

Gottschalk, Rand, William S. Davidson II, Leah K. Gensheimer, and Jeffrey P. Mayer (1987). Community based interventions. Pgs. 266-289 in Herbert C. Quay (ed.) *Handbook of Juvenile Delinquency.* New York: Wiley and Sons.

Gottschalk, R., W.S. Davidson II, J. Mayer and L.K. Gensheimer (1987). Behavioral approaches with juvenile offenders. A meta-analysis of long-term treatment efficacy. Pages 399-423 in E.K. Morris and C.J. Braukmann (eds.) *Behavioral Approaches to Crime and Delinquency.* New York: Plenum.

Government Accounting Office (1996). *Sex Offender Treatment Research Results.* Washington, DC: Government Accounting Office.

Goyer-Michaud, Francyne (1974). The adult female offender. A selected bibliography. *Criminal Justice and Behavior 1 (4): 340.*

Green, B.H. (1976). Applying the controlled experiment to penal reform. *Cornell Law Review 62 (1):158-176.*

Green, Gary S. (1985). General deterrence and television cable crime: A field experiment in social control. *Criminology 23 (4): 629-645.*

Greenberg, David F. (1977). The correctional effects of corrections: A survey of evaluations." In David F. Greenberg (Ed.), *Corrections and Punishment.* Newbury Park, CA: Sage.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin 82: 1-20.*

Hackler, James C. and John Hagan (1975). Work and teaching machines as delinquency prevention tools: A four-year follow-up. *Social Service Review 49 (1): 92-106.*

Hagan, Frank E. (1993). *Research Methodology in Criminal Justice and Criminology. Third Edition.* New York: McMillan.

Hall, Gordon C.N. (1995). Sexual offender recidivism revisited: A meta-analysis of recent treatment studies. *Journal of Consulting and Clinical Psychology 63 (5): 802-809.*

Hamby, Sherry (1996). *Presentation on Violence in the Home,* National Institute of Justice, Data Resources Program, Violence Across Settings Workshop, June 24-28, Ann Arbor, Michigan.

Hamm, Mark S. and John C. Kite (1991). The role of offender rehabilitation in family violence policy: The Batterers' Anonymous experiment. *Criminal Justice Review 16 (2):227-245.*

Harlow, Eleanor (1970). Intensive intervention: An alternative to institutionalization. *Crime and Delinquency Literature 2: 3-46.*

Harre, R. and P.F. Secord (1972). *The Explanation of Social Behavior.* Totowa, NJ: Rothman and Littlefield.

Hayeslip, David W. (1989). Higher Education and Police Performance Revisited: The Evidence Examined Through Meta-Analysis. *American Journal of Police 8(2): 49-62.*

Hedges, Larry V. (1992). Combining evidence for scientific inference. Book Review, John E. Hunter and Frank L. Schmidt, <u>Methods of Meta-Analysis: Correcting Error and Bias in Research Findings</u>. *Contemporary Psychology 37 (4): 304-306.*

Hedges, L.V. and I. Olkin (1985). *Statistical Methods for Meta-Analysis.* New York: Academic Press.

Hedges, L.V. and I. Olkin (1980). Vote-counting methods in research synthesis. *Psychological Bulletin 88 (2):359-369.*

Hewitt, John D., Eric D. Poole and Robert M. Regoli (1985). *Criminal Justice in America, 1959-1984, An Annotated Bibliography.* New York: Garland

Hunter, J.E. and F.L. Schmidt (1990*). Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Newbury Park, CA: Sage.

Hunter, J.E., F.L. Schmidt and G.B. Jackson (1982*). Meta-Analysis: Cumulating Research Findings Across Studies.* Beverly Hills, CA: Sage.

Izzo, Rhena L. and Robert R. Ross (1990). Meta-analysis of rehabilitation programs for juvenile delinquents. *Criminal Justice and Behavior 17 (1): 134-142.*

Jackson, G.B. (1980). Methods for integrative reviews. *Review of Educational Research 50:438-460.*

Jaeger, Richard (1990). *Statistics: A Spectator Sport (Second Edition).* Newbury Park, CA: Sage Publications.

Johnson, Blair T. (1997). *Personal communication.*

Johnson, Blair T. (1989). *DSTAT: Software for the Meta-Analytic Review of Research Literatures.* Hillsdale, N.J.: Erlbaum Associates, Publishers.

Johnson, Blair T., Michael P. Carey, and Paige A. Muellerleile (1997). Letter to the editor. *Journal of the American Medical Association 277 (Feb.):377.*

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80, 94-106.*

Johnson, V.S. (1977). Behavior modification in the correctional setting. *Criminal Justice and Behavior 4: 397-428.*

Jones, Peter (1995). Risk prediction in criminal justice. Pgs.33-68 in Alan Harland (ed.) *Choosing Correctional Options That Work. Defining the Demand and Evaluating the Supply.* Thousand Oaks, CA: Sage.

Jupp, Victor (1989). *Methods of Criminological Research.* London: Unwin Hyman

Justice Research and Statistics Association (1997). *Changes in crime trends: Causes and implications.* Conference announcement, September 4-5, Miami, Florida.

Kassenbaum, G., D. Ward and D. Wilner (1971). *Prison Treatment and Parole Survival.* New York: Wiley and Sons.

Kaufman, Paul (1985). *Meta-Analysis of Juvenile Delinquency Prevention Programs.* Unpublished paper, Claremont Graduate School, Claremont, California.

Kelling, George L., Anthony Pate, D. Dieckman and C.E. Brown (1974). *The Kansas City Preventive Patrol Experiment. A Technical Report.* Washington, D.C. Police Foundation.

Kerlinger, Fred N. (1964). *Foundations of Behavioral Research, Educational and Psychological Inquiry.* New York: Holt, Rinehart and Winston.

Kirby, Bernard C. (1970). *Crofton House Final Report.* San Diego, CA: San Diego State College.

Kirby, B.C. (1954). Measuring effects of criminals and delinquents. *Sociology and Social Research 38:368-374.*

Klein, M.W. (1986). Labeling theory and delinquency policy: An experimental test. *Criminal Justice and Behavior 13 (1): 47-79.*

Klockars, Carl (1975). The true limits of correctional effectiveness. *The Prison Journal 55:53-64.*

Kobrin and Klein (1983). *Community Treatment of Juvenile Offenders: The DSO Experiments.* Beverly Hills, CA: Sage.

Ku, Richard (1976). *An Exemplary Project. The Volunteer Probation Counselor Program.* Lincoln, Nebraska. Washington, D.C.: U.S. Department of Justice.

Lab, Steven P. and John T. Whitehead (1990). From "nothing works" to "the appropriate works": The latest stop on the search for the secular grail. *Criminology 28 (3):405-417.*

Lab, Steven P. and John T. Whitehead (1988). An analysis of juvenile correctional treatment. *Crime and Delinquency 34: 17-23.*

Laird, N. and F. Mosteller (1991). Some statistical methods for combining experimental results. *International Journal of Assessment in Health Care 6:5-30.*

Lamb, H.R. and V. Goertzel (1974). Ellsworth House: A community alternative to jail. *The American Journal of Psychiatry 131 (1): 64-68.*

Larson, J.D. (1990). Cognitive-behavioral group therapy with delinquent adolescents: A cooperative approach with the juvenile court. *Journal of Offender Rehabilitation 16:47-63.*

Lattimore, Pamela K., Ann Dryden Witte and Joanna R. Baker (1990). Experimental assessment of the effect of vocational training on youthful property offenders. *Evaluation Review 14 (2): 115-133.*

Lemert, R.O. and C.A. Visher (1987). *Randomized Field Experiments in Criminal Justice: Workshop Proceedings.* Washington, D.C.: National Institute of Justice.

Lerman, Paul (1975). *Community Treatment and Social Control.* Chicago: University of Chicago Press.

Lerner, Arthur (1953). An experiment in group counseling with male alcoholic inmates. *Federal Probation 17 (3): 32-38.*

Light, Richard J. and David B. Pillemer (1984). *Summing Up. The Science of Reviewing Research.* Cambridge, MA: Harvard University Press.

Lind, E. Allan (1985). Randomized experiments in the federal courts. Pgs. 73-80 in R.F. Boruch and W. Wothke (eds.) *Randomization and Field Experimentation. New Directions in Program Evaluation, no. 28.* San Francisco: Jossey-Bass.

Linden, R. and L. Perry (1982). The effectiveness of prison education programs. *Journal of Offender Counseling, Services and Rehabilitation 6 43-57.*

Lipsey, Mark W. (1992a). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. Pgs. 83-127 in Thomas D. Cook, Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis and Frederick Mosteller (Eds.) *Meta-Analysis for Explanation.* New York: Russell Sage Foundation.

Lipsey, Mark W. (1992b). The effect of treatment on juvenile delinquents: Results from meta-analysis. Pgs. 131-143 in Losel, F., D. Bender and T. Bliesener (eds.) *Psychology and Law.* Berlin: Walter de Gruyter.

Lipsey, Mark W. (1990). *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage.

Lipsey, Mark W. (1988a). Juvenile delinquency intervention. Pgs. 63-84 in H.S. Bloom, D.S. Cordray, and R.J. Light (eds.) *Lessons from Selected Program and Policy Areas. Number 37.* San Francisco, CA: Jossey-Bass.

Lipsey, Mark W. (1988b). Practice and malpractice in evaluation research. *Evaluation Practice 9 (4):5-24.*

Lipsey, Mark W. and David B. Wilson (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *American Psychologist 48 (12): 1181-1209*

Lipton, Douglas S. (1995). CDATE: Updating the <u>Effectiveness of Correctional Treatment</u> 25 years later. *Journal of Offender Rehabilitation 22 (1/2):1-20.*

Lipton, Douglas, Robert Martinson and Judith Wilks (1975). *The Effectiveness of Correctional Treatment.* New York: Praeger.

Loeber, Rolf and Magda Stouthamer-Loeber (1986). Family factors as correlates and predictors of juvenile conduct problems and delinquency. Pgs 29-149 in Michael Tonry and Norval Morris (Eds.), *Crime and Justice: An Annual Review of Research. Volume 7.* Chicago: University of Chicago Press.

Logan, Charles H. (1972). Evaluation research in crime and delinquency: A reappraisal. *The Journal of Criminal Law, Criminology and Police Science 63 (3): 378-387.*

Logan, Charles H. and Gerald G. Gaes (1993). Meta-Analysis and the rehabilitation of punishment. *Justice Quarterly 10 (2):245-264.*

Logan, C.H., G. Gaes, M. Harer, C.A. Innes, L. Karacki and W.G. Saylor (1991). *Can meta-analysis save correctional rehabilitation?* Washington, DC: Federal Bureau of Prisons.

Losel, Friedrich (1995). Increasing consensus in the evaluation of offender rehabilitation? Lessons from recent research syntheses. *Psychology, Crime and Law 2:19-39.*

Losel, Friedrich and Peter Koferl (1989). Evaluation research on correctional treatment in West Germany: A meta-analysis. Pgs. 334-355 in Hermann Wegener, Friedrich Losel and Jochen Haisch (eds.) *Criminal Behavior and the Justice System.* New York: Springer-Verlag.

Lundman, Richard and Frank Scarpitti (1978). Delinquency prevention: Recommendations for future projects. *Crime and Delinquency 24: 207-20.*

Macallair, Dan (1993). Reaffirming rehabilitation in juvenile justice. *Youth and Society 25 (1):104-125.*

MacKenzie, Doris L. (1991). The parole performance of offenders released from shock incarceration (boot camp prisons): A survival time analysis. *Journal of Quantitative Criminology 7 (3):213-236.*

Mann, Charles (1994). Can meta-analysis make policy? *Science 266:960-962.*

Marques, Janice K., David M. Day, Craig Nelson and Michael H. Miner (1989). The Sex Offender Treatment and Evaluation Project: California's relapse prevention program. Pgs. 247-267 in Richard Laws (ed.) *Relapse Prevention With Sex Offenders.* New York: Guilford Press.

Martin, Susan, Sampson Annan and Brian Forst (1986). *Deterring the drunk driver.* Washington, D.C.: Police Foundation.

Martin, S., L. Secrest and R. Redner (1981*). New Directions in the Rehabilitation of Criminal Offenders.* Washington, DC: National Academy Press.

Martinson, Robert (1979). Symposium on sentencing. *Hofstra Law Review 7 (2):243-258.*

Martinson, R. (1976). California research at the crossroads. *Crime and Delinquency (April):*

Martinson, Robert (1974). What works? Questions and answers about prison reform. *Public Interest 10:22-54.*

Martinson, R., T. Palmer and S. Adams (1976). *Rehabilitation, Recidivism, and Research.* Hackensack, N.J.: National Council on Crime and Delinquency.

Massachusetts General Court (1993). *Chapter 432. An Act to Promote the Effective Management of the Criminal Justice System Through Truth-in-Sentencing. Acts and Resolves.* Boston: Secretary of the Commonwealth.

Massimo, J.L. and M.F. Shore (1963). The effectiveness of a comprehensive, vocationally-oriented psychotherapeutic program for adolescent delinquent boys. *American Journal of Orthopsychiatry 33: 634-642.*

Matt, Georg (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin 105 (1): 106-115.*

Mauer, Marc (1990). *Young Black Men and the Criminal Justice System: A Growing National Problem.* Washington, DC: The Sentencing Project.

Mayer, Jeffrey P., Leah K. Gensheimer, William S. Davidson II, and Rand Gottschalk (1986). Social learning treatment within juvenile justice: A meta-analysis of impact in the natural environment. Pgs. 24-38 in Steven J. Apter and Arnold P. Goldstein (eds.) *Youth Violence.* New York: Pergamon Press.

Menninger, Karl (1966). *The Crime of Punishment.* New York: Viking Press.

Mitroff, I. (1983). Beyond experimentation. Pgs. 163-177 in E. Seidman (Ed) *Handbook of Social Intervention.* Beverly Hills: Sage.

Monahan, J.M., Earle K. Stewart and Karen E. Brown (1981). An annotated bibliography of American jails. *Prison Journal 61 (1): 55.*

Morris, Norval (1974). *The Future of Imprisonment.* Chicago: University of Chicago Press.

Mosteller, Frederick (1997). *Personal communication.*

Mrad, D.F. (1979). The effect of a differential follow-up on arrests: A critique of Quay and Love. *Criminal Justice and Behavior 6: 23-29.*

Murray, Liegh W. (1992). *Personal communication.*

National Institute of Justice (1990). Massive study will trace developmental factors that cause or prevent criminality. *NIJ Reports (May/June):2-4.*

Neumann, D.A., B.M. Houskamp, V.E. Pollock and J. Briere (1996). The long-term sequelae of childhood sexual abuse in women: A meta-analytic review. *Child Maltreatment 1 (1):6-16.*

New Jersey Code of Criminal Justice (1990). St.Paul,MN: West Publishing.

New York Times, New method of analyzing health data stirs debate, August 21, 1990, C1.

O'Kane, James B. (1995). *An electronic search of the research literature: Lessons from the CDATE project.* Presentation at the Academy of Criminal Justice Sciences annual meeting, March 10.

Olkin, Ingram (1990). History and goals. Pgs. 3-10 In K. Wachter and M.L. Straf (eds.). *The Future of Meta-Analysis.* Beverly Hills, CA: Sage.

Orwin, R.G. and D.S. Cordray (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and analysis. *Psychological Bulletin 9:134-147.*

Ostrom, Thomas M., Claude M. Steele, Lorne K. Rosenblood and Herbert T. Mirels (1971). Modification of delinquent behavior. *Journal of Applied Social Psychology 1 (2): 118-136.*

Owen, Barbara (1992). Measuring success in prison drug programs. *Journal of Crime and Justice 15: 91-117.*

Palmer, Ted (1994). *The Effectiveness of Correctional Intervention.* Albany, NY: State University of New York Press.

Palmer, Ted (1992*)*. *The Re-Emergence of Correctional Intervention.* Beverly Hills, CA: Sage.

Palmer, Ted (1983). The 'effectiveness' issue today: An overview. *Federal Probation 47 (2):3-10.*

Palmer, Ted (1978). *Correctional Intervention and Research: Current Issues and Future Prospects.* Lexington, MA: Lexington Books.

Palmer, Ted (1975). Martinson revisited. *Journal of Research in Crime and Delinquency 12:133-152.*

Pearson, Frank S., Douglas S. Lipton and Charles M. Cleland (1996*)*. *Some preliminary findings from the CDATE project.* Presentation at the American Society of Criminology, Chicago, Illinois, November.

Pearson, Frank, Douglas Lipton, Charles Cleland and James O'Kane (1995). *Meta-analysis on the effectiveness of correctional treatment: Another approach and extension of the time frame to 1994—A progress report.* Presentation at the American Society of Criminology Annual Meeting, Boston, Massachusetts, November 15th.

Petersilia, Joan (1989). Implementing randomized experiments: Lessons from BJA's Intensive Supervision Project. *Evaluation Review 13:435-458.*

Petersilia, Joan and Susan Turner (1993). Intensive probation and parole. In M. Tonry and N. Morris (eds.) *Crime & Justice: An Annual Review of Research, Vol. 19.* Chicago: University of Chicago Press.

Police Foundation (1981). *The Newark Foot Patrol Experiment.* Washington, D.C.: Police Foundation.

Powers, D. and D.L. Alderman (1979). Practical techniques for implementing true experimental designs. *Evaluation Quarterly 3:89-96.*

Powers, E. and H. Witmer (1951). *An Experiment in the Prevention of Delinquency. The Cambridge-Somerville Youth Study.* New York: Columbia University Press.

Prather, J.E. and F.K. Gibson (1977). The failure of social programs. *Public Administration Review (Sept/Oct): 556-564.*

Quay, H.C. (1977). The three faces of evaluation: What can be expected to work? *Criminal Justice and Behavior 4: 341-354.*

Quinsey, V.L. (1983). Pages 27-40 in S.N. Verdin-Jones and A.A. Keltner (Eds) *Sexual Aggression and the Law.* Burnaby, B.C. Canada: Simon Fraser University Criminology Research Center.

Redondo, Santiago, Vicente Garrido and Julio Sanchez-Meca (1996*). Is the treatment of offenders effective in Europe?: The results of a meta-analysis.* Presentation at the American Society of Criminology, Chicago, Illinois, November.

Reiman, J. (1984). *The Rich Get Richer and the Poor Get Prison. Second Edition.* New York: MacMillian.

Rhine, Edward (1992). Sentencing reform and correctional policy: Some unanswered questions. Pgs. 271-288 in Clayton Hartjen and Edward Rhine (eds.) *Correctional Theory and Practice.* Chicago: Nelson-Hall.

Riecken, Henry and Robert Boruch (1978). Social experiments. *Annual Review of Sociology 4:511-532.*

Riecken, Henry W. and Robert F. Boruch (1974). *Social Experimentation. A Method for Planning and Evaluating Social Intervention.* New York: Academic Press.

Roberts, Albert R. and Michael J. Camasso (1991). The effect of juvenile offender treatment programs on recidivism: A meta-analysis of 46 studies. *Notre Dame Journal of Law, Ethics and Public Policy 5 (2): 421-441.*

Robison, J. and G. Smith (1971). The effectiveness of correctional programs. *Crime and Delinquency (January): 67-80.*

Roesch, R. (1978). Does adult diversion work? *Crime and Delinquency (January): 72-80.*

Romig, Dennis A. (1978). *Justice for Our Children.* Lexington, MA: Lexington Books.

Rosenthal, Robert (1991). *Meta-Analytic Procedures for Social Research. Revised Edition.* Beverly Hills, CA: Sage.

Ross, H.A. and M. Blumenthal (1974). Sanctions for the drinking driver: An experimental study. *Journal of Legal Studies 3: 53-61.*

Ross, Robert R. and Paul Gendreau (1980). *Effective Correctional Treatment.* Toronto: Butterworths.

Ross, R.R. and H.B. McKay (1978). Behavioral approaches to treatment in corrections: Requiem for a panacea. *Canadian Journal of Criminology (July): 279-295.*

Ross, R.R. and M.J. Price (1976). Behavior modifications in corrections: Autopsy before mortification. *International Journal of Criminology and Penology 4: 305-315.*

Rotman, Edgardo (1990). *Beyond Punishment.* Westport, CT: Greenwood Press.

Roundtree, G.A., C.E. Grenier and VB.L. Hoffman (1993). Parental assessment of behavioral change after children's participation in a delinquency prevention program. *Journal of Offender Rehabilitation 19:113-130.*

Sanchez, Jose E. (1990). The uses of Robert Martinson' writings on correctional treatment: An essay on the justification of correctional policy. *Journal of Contemporary Criminal Justice 6 (3):127-138.*

Sarri, Rosemary and Robert Vinter (1965). Group treatment strategies in juvenile correctional programs. *Crime and Delinquency 11: 326-340.*

Saxe, Leonard and Michelle Fine (1981). *Social Experiments. Methods for Design and Evaluation.* Beverly Hills, CA: Sage.

Schmidt, Frank L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist 4 (10): 1173-1181.*

Schumann, Karl (1997). *Personal correspondence.*

Schumann, Karl (1994). *Personal correspondence.*

Schwartz, R.D. and S. Orleas (1961). On legal sanctions. *University of Chicago Law Review 34: 274-300.*

Schweinhart, Lawrence J. (1987). Can preschool programs help prevent delinquency? Pgs. 137-153 in James Q. Wilson and Glenn C. Loury (eds.*) From Children to Citizens. Families, Schools and Delinquency Prevention.* New York: Springer-Verlag.

Sechrest, L. and S.G. West (1983). Measuring the interventions in rehabilitation experiments. *Annales Internationales de Criminologie 21 (1): 11-19.*

Sechrest, Lee, Susan O. White, and E.D. Brown (1979*). The Rehabilitation of Criminal Offenders: Problems and Prospects.* Washington, D.C.: National Academy of Sciences.

Sederstrom, John D. (1987). *Social class and violent behavior: A meta-analysis of empirical findings.* Doctoral dissertation, University of Washington, Seattle, Washington.

Senna, J.J. and Larry J. Siegel (1995). *Introduction to Criminal Justice. Seventh Edition.* Minneapolis: West Publishing.

Sherman, Lawrence W. (1992*). Policing Domestic Violence. Experiments and Dilemmas.* New York: Free Press.

Sherman, L.W. (1989). *Effects of Sanctions on Criminal Careers: The Case for a Longitudinal Experiment.* Unpublished paper prepared for the Working Group on Continuation and Desistance in Crime. Program on Human Development and Criminal Behavior. Castine Research Corporation.

Sherman, Lawrence W. (1988). Randomized experiments in criminal sanctions. Pgs. 85-98 in H.S. Bloom, D.S. Cordray and R.J. Light (eds.) *Lessons from Selected Program and Policy Areas. New Directions for Program Evaluation, Number 37.* San Francisco: Jossey-Bass.

Sherman, Lawrence W., Denise Gottfredson, Doris MacKenzie, John Eck, Peter Reuter and Shawn Bushway (1997). *Preventing Crime: What Works, What Doesn't, What's Promising. A Report to the United States Congress.* College Park, MD: University of Maryland, Department of Criminology and Criminal Justice.

Sherman, L.W. and D.L. Weisburd (1995). General deterrent effects of police patrol in crime 'hot spots': A randomized, controlled trial. *Justice Quarterly 12 (4):625-648.*

Sherman, Lawrence W. and Richard Berk (1984). The deterrent effects of arrest for domestic assault. *American Sociological Review 49:261-272.*

Sherman, Lawrence W., Catherine H. Milton and Thomas V. Kelly (1973). *Team Policing. Seven Case Studies.* Washington, DC: Police Foundation.

Sherman, Lawrence W. and David L. Weisburd (1987). *The Experimental Effects of Criminal Sanctions.* Unpublished grant proposal to the National Institute of Justice, Crime Control Institute, Washington, D.C.

Shore, M.F. and J.L. Massimo (1966). Comprehensive vocationally oriented psychotherapy for adolescent delinquent boys: A follow-up study. *American Journal of Orthopsychiatry 36 (4): 609-615.*

Shore, M.F. and J.L. Massimo (1979). Fifteen years after treatment: A follow-up study of comprehensive vocationally-oriented psychotherapy. *American Journal of Psychotherapy 49 (2): 240-245.*

Slaikeu, K.A. (1973). Evaluation studies on group treatment of juvenile and adult offenders in correctional institutions. A review of the literature. *Journal of Research in Crime and Delinquency (Jan): 87-100*

Slavin, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher 13:6-15.*

Smith, Mary Lee and Gene V. Glass (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist 4 (September): 753-760.*

Smith, Mary Lee (1980). Publication bias and meta-analysis. *Evaluation in Education 4: 22-24.*

Sorenson, Jonathan R., Amy L. Patterson and Alan Widmayer (1992). Publication productivity of faculty members in criminology and criminal justice doctoral programs. *Journal of Criminal Justice Education 3 (1): 1-34.*

Stratton, John G. (1975). Effects of crisis intervention counseling on predelinquent and misdemeanor juvenile offenders. *Juvenile Justice (Fall): 7-18.*

Stratton, John G. (1974). Crisis intervention counseling and police diversion from the juvenile justice system: A review of the literature. *Juvenile Justice (May): 44-53.*

Strauss, Stephen (1991). Meta-analysis: Lies, damned lies and statistics. *Globe and Mail, November 2nd, p. D10.*

Tanur, Judith M. (1983). Methods for large scale surveys and experiments. Pgs. 1-78 in Samuel Leinhardt (ed.) *Sociological Methodology 1983-1984.* San Francisco: Jossey-Bass.

Taylor, Ralph B. (1994). *Research Methods in Criminal Justice.* New York: McGraw-Hill.

Thomas, A.P. (1995). Woodstock: A family picnic. *Investor's Business Daily, December 29, 1995.*

Tittle, Charles R. (1974). Prisons and rehabilitation: The inevitability of disfavor. *Social Problems 21 (2):385-395.*

Tobler, Nancy S. (1986). Meta-anslysis of 143 adolescent drug prevention programs--quantitative outcome results of program participants compared to a control or comparison group. *Journal of Drug Issues 16: 537-567.*

Tonry, M., L.E. Ohlin and D.P. Farrington (1991). *Human Development and Criminal Behavior. New Ways of Advancing Knowledge.* New York: Springer-Verlag.

Tonry, M. and N. Morris (1985). Introduction. *Crime and Justice: An Annual Review of Research, Vol. 6.* Chicago: University of Chicago Press.

Trudel, Robert J., Marvin Marcus and Robert J. Wheaton (1976*). Recidivism. A Selected Bibliography.* Washington, DC: National Institute of Law Enforcement and Criminal Justice.

Turpin-Petrosino, Carolyn (1993). *Exploring the effects of plea bargaining on parole decision-making.* Unpublished doctoral dissertation, Rutgers University, Newark, New Jersey.

Twain, David (1983). *Creating Change in Social Settings.* Beverly Hills: Sage.

van den Haag, Ernst (1983). How not to cut crime. Rehabilitating criminal offenders cannot cut crime. *Policy Review 26:53-58.*

Van Voorhis, P. (1987). Correctional effectiveness: The high cost of ignoring success. *Federal Probation 51 (1):56-62.*

Vaughn, Michael and Rolando V. del Carmen (1992). An annotated list of journals in criminal justice and criminology: A guide for authors. *Journal of Criminal Justice Education 3 (1): 93-144.*

Venezia, P.S. (1972). Unofficial probation: An evaluation of its effectiveness. *Journal of Research in Crime and Delinquency 9 (2): 149-170.*

Vetter, H.J. and R. Adams (1971). Effectiveness of probation caseload sizes: A review. *Criminology 8: 333-343.*

Vigdal, Gerald L., Donald W. Stadler, David D. Goodrick and Denis J. Sutton (1980). Skills training in a program for problem drinking offenders. A one year follow-up evaluation. *Journal of Offender Counseling, Services and Rehabilitation 5 (2): 61-73.*

Von Hirsch, A. (1985). *Past or Future Crimes—Deservedness and Dangerousness in the Sentencing of Criminals.* New Brunswick, NJ: Rutgers University Press.

Von Hirsch, A. (1976). *Doing Justice. The Choice of Punishments.* New York: Hill and Wang.

von Hirsch, A. and L. Maher (1992). Should penal rehabilitationism be revived? *Criminal Justice Ethics 11 (1):25-30.*

Wachter, Kenneth W. and Miron L. Straf (1990*). The Future of Meta-Analysis.* Beverly Hills, CA: Sage.

Walker, Samuel (1994). *Sense and Nonsense About Crime. Third Edition.* Monterrey, CA: Brooks/Cole.

Walters, Glenn D. (1992). A Meta-Analysis of the Gene-Crime Relationship. *Criminology 30 (4): 595-613.*

Wanous, John P., Sherry E. Sullivan and Joyce Malinak (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology 74 (2): 259-264.*

Webb, Eugene, D.T. Campbell, R.D. Schwartz and L. Sechrest (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences.* Chicago: Rand McNally.

Weber, Robert Phillip (1990). *Basic Content Analysis. Second Edition.* Newbury Park, CA: Sage.

Weisburd, David L. (1996). *Personal communication.*

Weisburd, David L. (1995). *Personal communication.*

Weisburd, David L., with Anthony J. Petrosino and Gail Mason (1993). Design sensitivity in criminal justice experiments: Reassessing the relationship between sample size and statistical power. *Crime & Justice: An Annual Review of Research. Vol. 17.* Chicago: University of Chicago Press.

Weisburd, David (1989). *Project MUSTER: Proposal for External Evaluation Submitted to the State Justice Institute.* Newark, NJ: Rutgers, School of Criminal Justice.

Weisburd, David and Joel Garner (1992). Experimentation in criminal justice: Editors' introduction. *Journal of Research in Crime and Delinquency 29 (1): 3-6.*

Weisburd, David L., Lisa Maher and Lawrence Sherman, with Michael E. Buerger, Ellen Cohn and Anthony J. Petrosino (1992). Crime-specific versus crime-general theory: The case of hotspots. *Advances in Criminological Theory 4: 44-70.*

Weisburd, David L., Lawrence Sherman and Anthony Petrosino (1990). *Registry of Randomized Experiments in Criminal Sanctions, 1950-1983.* Los Altos, CA: Sociometics Corporation, Data Holdings of the National Institute of Justice.

Weisburd, D.L. and L.W. Sherman (1988*). The effects of sanctions on recidivism: Experimental Evidence.* Paper presented at the July National Institute of Justice Crime Control Theory Conference, New Orleans, Louisiana.

Weiss, Carol H. (1972). *Program Evaluation.* Englewood Cliffs, NJ: Prentice-Hall.

Wells, Edward L. and Joseph H. Rankin (1991). Families and delinquency: A meta-analysis of the impact of broken homes. *Social Problems 38 (1): 71-93.*

Wells-Parker, Elisabeth, R. Bangert-Drowns, R. McMillen and M. Williams (1995). Final results from a meta-analysis of remedial interventions with drink/drive offenders. *Addiction 90:907-926.*

Whitehead, John T. and Steven P. Lab (1989). A meta-analysis of juvenile correctional treatment. *Journal of Research in Crime and Delinquency 26:276-295.*

Wilson, James Q. (1980). 'What works' revisited: New findings on criminal rehabilitation. *The Public Interest 61:3-17.*

Wilson, James Q. (1975). *Thinking About Crime.* New York: Basic Books.

Winnick, B.J. (1981). A preliminary analysis of legal limitations on rehabilitative alternatives to corrections and on correctional research. Pgs. 328-375 in Susan Martin, L. Sechrest and R. Redner (eds.). *New Directions in the Rehabilitation of Criminal Offenders.* Washington, DC: National Academy Press.

Wolf, Frederic (1986). *Meta-Analysis. Quantitative Methods for Research Synthesis.* Beverly Hills, CA: Sage.

Wright, W.E. and M.C. Dixon (1977). Community prevention and treatment of juvenile delinquency. *Journal of Research in Crime and Delinquency (January): 35-67.*

Yeaton, William H. and Paul M. Wortman (1993). On the reliability of meta-analytic reviews. *Evaluation Review 17 (3): 292-309.*

Yeaton, William, Paul Wortman and N. Langberg (1983). Differential attrition. Estimating the crossovers on the evaluation of a medical technology. *Evaluation Review 7 (6):831-840.*

Yin, Robert K. (1986). Community crime prevention: A synthesis of eleven evaluations. Pgs. 294-308 in Dennis Rosenbaum (ed.) *Community Crime Prevention: Does it Work?* Beverly Hills, CA: Sage.

Zeisel, Hans (1973). Reflections on experimental techniques in the law. *The Journal of Legal Studies 2 (1): 107-124.*

Zeisel, H. (1968). The indirect experiment. *Law and Society 2 (3):504-508.*

Zimring, Franklin E. and Gordon J. Hawkins (1973). *Deterrence: The Legal Threat in Crime Control.* Chicago: University of Chicago Press.

Citations to the 150 Studies Used in the Meta-Analysis: Primary Documents Only (Number of Experiments in Parentheses)

Adams, Reed and Harold J. Vetter (1982). Social structure and psychodrama outcome: A ten-year follow-up. *Journal of Offender Counseling, Services and Rehabilitation 6: 111-119.*

Adams, Stuart (1965). An experimental assessment of group counseling with juvenile probationers. *Journal of the California Probation, Parole and Correctional Association 2 (Spring): 19-25.*

Alexander, James F. and Bruce V. Parsons (1973). Short-term behavioral intervention with delinquent families: Impact on family process and recidivism. *Journal of Abnormal Psychology 81 (3): 219-225.*

Annis, Helen M. (1979). Group treatment of incarcerated offenders with alcohol and drug problems: A controlled evaluation. *Canadian Journal of Criminology 21 (1): 3-15.*

Arbuthnot, Jack and Donald A. Gordon (1986). Behavioral and cognitive effects of a moral reasoning development intervention for high-risk behavior-disordered adolescents. *Journal of Consulting and Clinical Psychology 54 (2): 208-216.*

Arthur Young & Company (1983). *Final Report for "Cost Effectiveness of Misdemeanant Probation" Presented to Hamilton County Municipal Court.* Cincinnati, OH: Arthur Young & Company.

Austin, James F. (1980). *Instead of Justice: Diversion.* Ph.D. dissertation, University of California, Davis. Ann Arbor, MI: University Microfilms International.

Baker, Sally Hillsman and Susan Sadd (1981). *Diversion of Felony Arrests. An Experiment in Pretrial Intervention. An Evaluation of the Court Employment Project. Summary Report.* Washington, DC: National Institute of Justice.

Bank, Lew, J. Hicks Marlowe, John B. Reid, Gerald R. Patterson and Mark R. Weinrott (1991). A comparative evaluation of parent-training interventions for families of chronic delinquents. *Journal of Abnormal Psychology 19 (1): 15-33.*

Baron, Roger and Floyd Feeney (1972). *Preventing Delinquency Through Diversion. The Sacramento County Probation Department 601 Diversion Project. A First Year Report.* Davis, CA: University of California at Davis, Center on Administration of Criminal Justice.

Baron, Roger and Floyd Feeney (1976). *Juvenile Diversion Through Family Counseling. A Program for Status Offenders in Sacramento County, California.* Washington, DC: U.S. Department of Justice.

Berecochea, John E. and Dorothy R. Jaman (1981). *Time Served in Prison and Parole Outcome: An Experimental Study. Report Number 2.* Sacramento, CA: Department of Corrections, Research Division.

Berg, Ian, Alison Goodwin, Roy Hullin and Ralph McGuire (1986). A randomly controlled trial of interviewing children with severe school attendance problems and their families. *International Review of Applied Psychology 35: 443-451.*

Berleman, William C., James R. Seaberg and Thomas W. Steinburn (1972). The delinquency prevention experiment of the Seattle Atlantic Street Center: A final evaluation. *Social Service Review 46 (3): 323-346.*

Berman, John J. (1978). An experiment in parole supervision. *Evaluation Quarterly 2 (1): 71-90.*

Bernsten, Karen and Karl O. Christiansen (1965). A resocialization experiment with short-term offenders. Pgs. 35-54 in Karl O. Christiansen (ed.) *Scandinavian Studies in Criminology. Volume 1.* London: Oslo.

Blumenthal, Murray and H. Laurence Ross (1973). *Two Experimental Studies of Traffic Law: The Effect of Legal Sanctions on DUI Offenders and the Effect of Court Appearance on Traffic Law Violators.* National Highway Transportation & Safety Administration, contract DOT-HS-249-2-437.

Britt III, Chester L., Michael R. Gottfredson and John S. Goldkamp (1992). Drug testing and pretrial misconduct: An experiment on the specific deterrent effects of drug monitoring defendants on pretrial release. *Journal of Research in Crime and Delinquency 29 (1): 62-78.* **(FOUR)**

Burkhart, Walter R. (1969). The parole work unit programme: An evaluation. *British Journal of Criminology 125-147.*

Byles, J.A. and A. Maurice (1979). The Juvenile Services Project: An experiment in delinquency control. *Canadian Journal of Criminology 21: 155-165.*

Chandler, Michael J. (1973). Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills. *Developmental Psychology 9 (3): 326-332.*

Clay, T.R. (1978). *Rehabilitation of the Drunken Driver in the United States: An Evaluation of the Approach Used by the Phoenix, Arizona Alcohol Safety Action Project.* Phoenix: Arizona Alcohol Safety Action Project **(FOUR)**

Collins, James J., Charles L. Usher and Jay R. Williams (1984). *Research on alternative probation strategies in Maryland. Final Report.* Research Triangle Park: North Carolina.

Connors, Gerard J., Stephen A. Maisto and Seth M. Ersner-Hershfield (1986). Behavioral treatment of drunk-driving recidivists: Short-term and long-term effects. *Behavioral Psychotherapy 14: 34-45.*

Cook, Douglas Stevens (1990). *The effects of an experimental intervention on juvenile female recidivism and drug relapse.* Ph.D. dissertation, University of Washington. Ann Arbor, MI: University Microfilms International.

Cornish, Derek B. and Ronald V.G. Clarke (1975). *Residential Treatment and its Effects on Delinquency. Home Office Research Study No. 32.* London: Her Majesty's Stationary Office.

Craft, Michael, Geoffrey Stephenson and Clive Granger (1964). A controlled trial of authoritarian and self-governing regimes with adolescent psychopaths. *American Journal of Orthopsychiatry 34: 543-554.*

Davidson II, William S., Edward Seidman, Julian Rappaport, Phillip L. Berck, Nancy A. Rapp, Warren Rhodes and Jacob Herring (1977). Diversion program for juvenile offenders. *Social Work Research & Abstracts (Summer): 40-49.* **(TWO)**

Davidson II, William S., Robin Redner and Richard L. Amdur (1990). *Alternative Treatments for Troubled Youth: The Case of Diversion From the Justice System.* New York: Plenum.

Ditman, Keith S., George G. Crawford, Edward W. Forgy, Herbert Moskowitz and Craig MacAndrew (1967). A controlled experiment on the use of court probation for drunk arrests. *American Journal of Psychiatry 124 (2): 64-67.*

Dole, Vincent P., J. Waymond Robinson, John Orraca, Edward Towns, Paul Searcy and Eric Caine (1969). Methadone treatment of randomly selected criminal addicts. *New England Journal of Medicine 280 (25): 1372-1375.*

Dunford, Franklyn W., David Huizinga and Delbert S. Elliott (1990). The role of arrest in domestic assault: The Omaha Police experiment. *Criminology 28 (2): 183-206.*

Edwards, Dan W. and George A. Roundtree (1982). Assessment of short-term treatment groups with adjudicated first offender shoplifters. *Journal of Offender Counseling, Services and Rehabilitation 6: 89-102.*

Elrod, H. Preston and Kevin I. Minor (1992). Second wave evaluation of a multi-faceted intervention for juvenile court probationers. *International Journal of Offender Therapy and Comparative Criminology 36 (3):247-262.*

Empey, Lamar T. and S.G. Lubeck (1971). *The Silverlake Experiment.* Chicago, IL: Aldine.

Empey, LaMar T. and Maynard L. Erickson (1972). *The Provo Experiment.* Massachusetts: Lexington Books.

Emshoff, James G. and Craig H. Blakely (1983). The diversion of delinquent youth: Family focused intervention. *Children and Youth Services Review 5: 343-356.*

Ericson, Richard and David Moberg (1967). *The Rehabilitation of Parolees. The Application of Comprehensive Psycho-Social Vocational Services in the Rehabilitation of Parolees.* Minneapolis, MN: Minneapolis Rehabilitation Center.

Fagan, Jeffrey A. (1990). Treatment and reintegration of violent juvenile offenders: Experimental results. *Justice Quarterly 7 (2): 233-263.* **(FOUR)**

Feis, Carolyn Little (1990). *Community Service for Juvenile Offenders: An Experimental Evaluation.* Ph.D. dissertation, Michigan State University. Ann Arbor, MI: University Microfilms International.

Folkard, M.S., A.J. Fowles, B.C. McWilliams, W. McWilliams, D.D. Smith, D.E. Smith and G.R. Walmsley (1974). *IMPACT: Intensive Matched Probation and After-Care Treatment. Volume II. The Results of the Experiment.* London: Her Majesty's Stationary Office. **(FOUR)**

Ford, David A. and University Research Associates, with Mary Jean Regoli (1993). *The Domestic Violence Prosecution Experiment. Final report submitted to the National Institute of Justice.* Indianapolis, IN: Indiana University. **(TWO)**

Fowles, A.J. (1978). *Prison Welfare: An Account of an Experiment at Liverpool. Home Office Research Study No. 45.* London: Her Majesty's Stationary Office.

Gallant, D.M., M. Faulkner, B. Stoy, M.P. Bishop and D. Langdon (1968). Enforced clinic treatment of paroled criminal alcoholics. A pilot evaluation. *Quarterly Journal of Studies on Alcohol 29 (1-A): 77-83.*

Goldkamp, John S. and Peter R. Jones (1992). Pretrial drug-testing experiments in Milwaukee and Prince George's County: The context of implementation. *Journal of Research in Crime and Delinquency 29 (4): 430-465.* **(TWO)**

Gottfredson, Denise C. (1986). An empirical test of school-based environmental and individual interventions to reduce the risk of delinquent behavior. *Criminology 24 (4): 705-731.*

Gottfredson, Denise C. and Gary D. Gottfredson (1992). Theory-guided investigation: Three field experiments. Pgs. 311-329 in Joan McCord and Pierre Tremblay (eds.) *Preventing Antisocial Behavior: Interventions from Birth to Adolescence.* New York: Guilford Press. **(TWO)**

Greater Egypt Regional Planning and Development Commission (1979*). Evaluation Report. Menard Correctional Center Juvenile Tours Impact Study.* Carbondale, IL: Greater Egypt Regional Planning and Development Commission.

Greenwood, Peter W. and Susan Turner (1993). Evaluation of the Paint Creek Youth Center: A residential program for serious delinquents. *Criminology 31 (2): 263-279.*

Guttman, Evelyn S. (1963). *Effects of Short-Term Psychiatric Treatment on Boys in Two California Youth Authority Institutions.* Sacramento, CA: Department of Youth Authority, Division of Research. **(TWO)**

Hackler, James C. and John Hagan (1975). Work and teaching machines as delinquency prevention tools: A four-year follow-up. *Social Service Review 49 (1): 92-106* **(TWO)**

Haskell, Martin R. and H. Ashley Weeks (1960). Role training as preparation for release from a correctional institution. *Journal of Criminal Law, Criminology and Police Science 50 (5): 441-447.*

Henggeler, Scott W., Gary B. Melton, Linda A. Smith, Sonja K. Schoenwald, and Jerome H. Hanley (1993). Family preservation using multisystemic treatment: Long-term follow-up to a clinical trial with serious juvenile offenders. *Journal of Child and Family Studies 2 (4): 283-293.*

Hintzen, Rachel, Keith Inouye and Beryl Iramina (1979). *Research Report. A Three Year Follow-up of Project '75.* Manoa, HI: School Welfare Development and Research Center, University of Hawaii-Manoa.

Hirschel, J. David and Ira W. Hutchinson, III (1992). Female spouse abuse and the police response: The Charlotte, North Carolina Experiment. *Journal of Criminal Law & Criminology 83 (1): 73-119.*

Holden, Robert T. (1983). Rehabilitative sanctions for drunk driving: An experimental evaluation. *Journal of Research in Crime and Delinquency (Jan): 55-72.* **(TWO)**

Homant, Robert J. (1986). Ten years after: A follow-up of therapy effectiveness. *Journal of Offender Counseling, Services and Rehabilitation 10 (3): 51-57.*

Jesness, Carl F. (1971-72). Comparative effectiveness of two institutionalized treatment programs for delinquents. *Child Care Quarterly 1 (2): 119-130* **(TWO).**

Jesness, Carl F. (1975). Comparative effectiveness of behavior modification and transactional analysis programs for delinquents. *Consulting and Clinical Psychology 43 (6): 758-779.*

Johnson, Bertram M. (1962). *Parole Performance of the First Year's Releases. Parole Research Project: Evaluation of Reduced Caseloads.* Sacramento, CA: Department of the Youth Authority, Division of Research.

Kadell, Daniel J. and Raymond C. Peck (1982). *An Evaluation of the Alcohol Reexamination Program for Drivers with Two Major Traffic Convictions.* Sacramento, CA: Department of Motor Vehicles, Research and Development Office.

Kassebaum, Gene, David A. Ward and Daniel M. Wilner (1971). *Prison Treatment and Parole Survival. An Empirical Assessment.* New York: Wiley and Sons.

Kelley, Thomas M. (1972). *Student Volunteer Effectiveness in a Delinquency Prevention Experiment. II. Validation of a Selection Device for Volunteer Probation Officers.* Ph.D. dissertation, Wayne State University. Ann Arbor, MI: University Microfilms International.

Klein, Malcolm W. (1986). Labeling theory and delinquency policy. An experimental test. *Criminal Justice & Behavior 13 (1): 47-79.*

Koch, J. Randy (1985). *Community Service and Outright Release as Alternatives to Juvenile Court: An Experimental Evaluation.* Ph.D. dissertation, Michigan State University. Ann Arbor, MI: University Microfilms International.

Krenek, Richard F. (1979). *Short Term Rehabilitation. Analytic Study No. V/VI.* Norman, OK: OMEC, Inc.

Ku, Richard (1976). *An Exemplary Project. The Volunteer Probation Counselor Program.* Lincoln, Nebraska. Washington, D.C.: U.S. Department of Justice.

Kurtzberg, Richard L. (no date). *The Project: A Three-Year Experimental Investigation.* Staten Island, NY: Montefiore Hospital (Located in NCJRS data base).

Lamb, H. Richard and Victor Goertzel (1974). Ellsworth House: A community alternative to jail. *American Journal of Psychiatry 131 (1): 64-68.*

Land, Kenneth C., Patricia L. McCall and Jay R. Williams (1990). Something that works in juvenile justice. An evaluation of the North Carolina court counselors' intensive protective supervision randomized experimental project, 1987-1989. *Evaluation Review 14 (6): 574-606.*

Latessa, Edward J. and Melissa M. Moon (1992). The effectiveness of acupuncture in an outpatient drug treatment program. *Journal of Contemporary Criminal Justice 8 (4): 317-330.*

Lattimore, Pamela K., Ann Dryden Witte and Joanna R. Baker (1990). Experimental assessment of the effect of vocational training on youthful property offenders. *Evaluation Review 14 (2): 115-133.*

Lee, Robert and Nancy McGinnis Haynes (1978). Counseling juvenile offenders: An experimental evaluation of Project Crest. *Community Mental Health Journal 14 (4): 267-271.*

Lewis, Roy V. (1983). Scared straight--California style. Evaluation of the San Quentin Squires program. *Criminal Justice and Behavior 10 (2): 209-226.*

Lichtman, Cary M. and Sue M. Smock (1981). The effects of social services on probation recidivism: A field experiment. *Journal of Research in Crime and Delinquency (January): 81-100.*

Linden, Rick, Linda Perry, Douglas Ayers and T.A.A. Parlett (1984). An evaluation of a prison education program. *Canadian Journal of Criminology 26 (1): 65-74.* **(TWO)**

Marques, Janice K., David M. Day, Craig Nelson and Michael H. Miner (1989). The Sex Offender Treatment and Evaluation project: California's relapse prevention program.

Pgs. 247-267 in Richard Laws (ed.) *Relapse Prevention With Sex Offenders.* New York: Guilford Press.

McPherson, Susan J., Lance E. McDonald and Charles W. Ryer (1983). Intensive counseling with families of juvenile offenders. *Juvenile and Family Court Journal (Feb): 27-32.*

Moloff, Martin J. (1967). *Forestry camp study: comparison of recidivism rates of camp-eligible boys randomly assigned to camps or institutional programs. Division of Research, California Department of the Youth Authority. Research Report Number 53.* Sacramento, CA: Department of the Youth Authority.

Moore, Richard H. (1987). Effectiveness of citizen volunteers functioning as counselors for high-risk young male offenders. *Psychological Reports 61: 823-830.*

Nath, Sunil B., David E. Clement and Frank Sistrunk (1976). Parole and probation caseload size variation: The Florida Intensive Supervision project. *Criminal Justice Review 1: 61-71.*

Odell, Brian Neal (1974). Accelerating entry into the opportunity structure: A sociologically-based treatment for delinquent youth. A sociologically-based treatment for delinquent youth. *Sociology and Social Research 58 (3): 312-317.*

O'Donnell, Clifford R., Tony Lydgate and Walter S.O. Fo (1979). The Buddy System: Review and follow-up. *Child Behavior Therapy 1 (2): 161-169.*

Orchowsky, Stan and Keith Taylor (1981). *The Insiders. Juvenile Crime Prevention Program: An Assessment of a Juvenile Awareness Program.* Richmond, VA: Virginia Department of Corrections.

Ostrom, Thomas M., Claude M. Steele, Lorne K. Rosenblood and Herbert T. Mirels (1971). Modification of delinquent behavior. *Journal of Applied Social Psychology 1 (2): 118-136.*

Owen, Greg and Paul W. Mattessich (1987). *Community Assistance Program. Results of a Controlled Study on the Effects of Non-Residential Corrections Services on Adult Offenders in Ramsey County.* St. Paul, MN: Wilder Foundation.

Palmer, Ted (1971). California's Community Treatment Program for Delinquent Adolescents." *Journal of Research in Crime and Delinquency 8 (1):74-92* **(TWO)**

Pennsylvania Board of Probation and Parole (1969). *Resocialization of the Paroled Non-Aggressive Predatory Offender.* Allentown, PA: Pennsylvania Board of Probation and Parole.

Pennsylvania Prison Society (1980). Employment research project. Executive summary. *Prison Journal (Spring/Summer): 2-67.* **(TWO)**

Persons, Roy W. (1967). Relationship between psychotherapy with institutionalized boys and subsequent community adjustment. *Journal of Counseling Psychology 31 (2): 137-141.*

Petersilia, Joan and Susan Turner (1990). *Diverting Prisoners to Intensive Probation: Results of an Experiment in Oregon.* Santa Monica, CA: Rand Corporation.

Prentice, Norman M. (1972). The influence of live and symbolic modeling on promoting moral judgment of adolescent delinquents. *Journal of Abnormal Psychology 80: 157-161.*

Quay, Herbert C. and Craig T. Love (1977). The effect of a juvenile diversion program on rearrests. *Criminal Justice and Behavior 4 (4): 377-396.*

Reckless, Walter C. and Simon Dinitz (1972). *The Prevention of Juvenile Delinquency. An Experiment.* Columbus, OH: Ohio State University Press.

Reimer, Ernest and Martin Warren (1957). Special intensive parole unit. Relationship between violation rate and initially small caseload. *National Probation and Parole Association Journal 3 (3): 222-229.*

Reimer, Ernest and Martin Warren (1958). *Special Intensive Parole Unit Phase II. Thirty-Man Caseload.* Sacramento, CA: Department of Corrections, Division of Adult Parolees.

Reinarman, Craig and Donald Miller (1975). *Direct Financial Assistance to Parolees: A Promising Alternative in Correctional Programming.* Sacramento, CA: Department of Corrections, Division of Research.

Reis, Raymond E. and Lewis A. Davis (1980). *First Interim Analysis of Multiple Offender Treatment Effectiveness.* Sacramento, CA: Sacramento Health Department, Office of Alcoholism. **(TWO)**

Rose, Gordon and R.A. Hamilton (1970). Effects of a juvenile liaison scheme. *British Journal of Criminology 10 (1): 2-20.*

Ross, Robert R., Elizabeth A. Fabiano and Crystal D. Ewles (1988). Reasoning and rehabilitation. *International Journal of Offender Therapy and Comparative Criminology 32 (1): 29-35.*

Sarason, Irwin G. and Victor J. Ganzer (1973). Modeling and group discussion in the rehabilitation of juvenile delinquents. *Journal of Counseling Psychology 20 (5): 442-449.*

Schneider, Anne L. (1981). Effects of status offender deinstitutionalization. A case study. Pgs. 122-142 in Ronald Roesch and Raymond R. Corrado (eds.) *Evaluation and Criminal Justice Policy.* Beverly Hills, CA: Sage.

Schweinhart, Lawrence J. (1987). Can preschool programs help prevent delinquency? Pgs. 137-153 in James Q. Wilson and Glenn C. Loury (eds.) *From Children to Citizens. Families, Schools and Delinquency Prevention.* New York: Springer-Verlag.

Seckel, Joachim P. (1965). *Experiments in Group Counseling at Two Youth Authority Institutions. Research Report No. 46.* Sacramento, CA: Department of the Youth Authority, Division of Research. **(TWO).**

Seckel, Joachim P. (1967). *The Freemont Experiment: Assessment of Residential Treatment at a Youth Authority Reception Center.* Sacramento, CA: Department of the Youth Authority, Division of Research.

Shaw, Margaret (1974*). Social Work in Prison. An Experiment in the Use of Extended Contact with Offenders.* London: Her Majesty's Stationary Office.

Sherman, Lawrence W., Janell D. Schmidt, Dennis P. Rogan, Patrick R. Gartin, Ellen G. Cohn, Dean J. Collins, and Anthony R. Bacich (1991). From initial deterrence to long-term escalation: Short-custody arrest for poverty ghetto domestic violence. *Criminology 29 (4): 821-849.*

Sherman, Lawrence W. and Richard A. Berk (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review 49 (1): 261-272.*

Shivrattan, Jacob L. (1988). Social interactional training and incarcerated juvenile delinquents. *Canadian Journal of Criminology (April): 145-163.*

Smith, Charles L., Pablo Martinez and Daniel Harrison (1978). *An Assessment: The Impact of Providing Financial or Job Placement Assistance to Ex-Prisoners.* Huntsville, TX: Texas Department of Corrections, Research and Development Division.

Spence, Susan H. and John S. Marzillier (1980). Social skills training with adolescent male offenders--II. Short-term, long-term and generalized effects. *Behavior Research & Therapy 19: 349-368.*

Star, Deborah (1979). *Summary Parole. A Six and Twelve Month Follow-up Evaluation.* Sacramento, CA: Department of Corrections, Research Unit.

Stratton, John G. (1975). Effects of crisis intervention counseling on predelinquent and misdemeanor juvenile offenders. *Juvenile Justice (Fall): 7-18.*

Sullivan, Clyde E. and Wallace Mandell (1967). *Restoration of Youth Through Training. A Final Report to Office of Manpower Policy, Evaluation and Research.* Staten Island, NY: Wakoff Research Center.

Sweet, Ronald P. (1975). *Recidivist Felons in the Community. Final Evaluation Report of the Community Treatment of Recidivist Felony Offenders Project.* Davis, California: National Council on Crime and Delinquency **(TWO).**

Truax, Charles B., Donald G. Wargo and Leon D. Silber (1966). Effects of group psychotherapy with high accurate empathy and non-possessive warmth upon female institutionalized delinquents. *Journal of Abnormal Psychology 71 (4): 267-274.*

Turner, Susan and Joan Petersilia (1992). Focusing on high-risk parolees: An experiment to reduce commitments to the Texas Department of Corrections. *Journal of Research in Crime and Delinquency 29 (1): 34-61.* **(TWO)**

Venezia, Peter S. (1972). Unofficial probation: An evaluation of its effectiveness. *Journal of Research in Crime and Delinquency 9 (2): 149-170.*

Vreeland, Allan D. (1981). *Evaluation of Face-to-Face: A Juvenile Aversion Program.* Ph.D. doctoral dissertation. Dallas, TX: The University of Texas.

Webb, Allen P. and Patrick V. Riley (1970). Effectiveness of casework with young female probationers. *Social Casework 51 (9): 566-572.*

Winterdyk, John and Ronald Roesch (1982). A wilderness experiential program as an alternative for probationers: An evaluation. *Canadian Journal of Criminology 24: 39-49.*

Worzella, Charles (1992). The Milwaukee Municipal Court Day-Fine Project. Pgs. 61-95 in Douglas C. McDonald (ed.) *Day Fines in American Courts: The Staten Island and Milwaukee Experiments.* Washington, DC: US Department of Justice.

Yarborough, James C. (1979). *Evaluation of JOLT as a Deterrence Program.* Lansing, MI: Michigan Department of Corrections, Program Bureau.

ANTHONY J. PETROSINO
VITA

1960 Born in New York City, New York on November 7th.

1978 Graduated from Madison Central High School, Old Bridge, New Jersey.

1978-82 Attended Glassboro State College, Glassboro, NJ; majored in Law & Justice

1982 B.A., Law & Justice, Glassboro State College (now Rowan University of NJ)

1982-84 Employed by Bamberger's Department Store as a Detective

1984-86 Employed by Coca-Cola Bottling Company of NY as a supervisor.

1986-97 Graduate work at Rutgers, School of Criminal Justice, Newark, NJ.

1987-88 Employed as a Field Researcher, Vera Institute of Justice, New York, NY.

1988-90 Employed as Research Associate, Rutgers, School of Criminal Justice.

1989 M.A., Criminal Justice, Rutgers, School of Criminal Justice, Newark, NJ.

1990-93 Employed as a Research Specialist by the New Jersey Division of Criminal Justice, Trenton, NJ.

1993-97 Received Graduate Research Fellowship Grant from the National Institute of Justice, Washington, DC.

1993-present Employed as an Adjunct Professor, Continuing Education Studies, University of Massachusetts, Lowell.

1993-94 Employed as an Assistant Professor of Criminal Justice, Westfield State College, Westfield, MA

1994-95 Employed as a Visiting Assistant Professor, Northeastern University, College of Criminal Justice, Boston, MA.

1995-1997 Employed as a Research Associate, Massachusetts Committee on Criminal Justice, Boston, MA

1997 Employed as a Research Associate, Criminal History Systems Board, Chelsea, MA.

1997 Ph.D. in Criminal Justice, Rutgers University, Newark, NJ.

1997-current Post-Doctoral Fellow in Evaluation, Harvard University, Project on Schooling and Children, Evaluation Task Force.