Journal of Criminal Law and Criminology

Volume 70
Issue 4 *Winter*

Article 13

Winter 1979

The Policy Implications of Evaluation Research: Some Issues Raised by the Fishman Study of Rehabilitation and Diversion Services

Ronald Roesch

Raymond R. Corrado

Follow this and additional works at: https://scholarlycommons.law.northwestern.edu/jclc

 Part of the Criminal Law Commons, Criminology Commons, and the Criminology and Criminal Justice Commons

Recommended Citation

Ronald Roesch, Raymond R. Corrado, The Policy Implications of Evaluation Research: Some Issues Raised by the Fishman Study of Rehabilitation and Diversion Services, 70 J. Crim. L. & Criminology 530 (1979)

This Criminology is brought to you for free and open access by Northwestern University School of Law Scholarly Commons. It has been accepted for inclusion in Journal of Criminal Law and Criminology by an authorized editor of Northwestern University School of Law Scholarly Commons.

9901-4169/79/7004 0530$02.00/0
THE JOURNAL OF CRIMINAL LAW & CRIMINOLOGY
Copyright © 1979 by Northwestern University School of Law
Vol. 70, No. 4
Printed in U.S.A.

CRIMINOLOGY

THE POLICY IMPLICATIONS OF EVALUATION RESEARCH: SOME ISSUES RAISED BY THE FISHMAN STUDY OF REHABILITATION AND DIVERSION SERVICES*

RONALD ROESCH** AND RAYMOND R. CORRADO***

The Fishman study of rehabilitation and diversion services in New York City[1] was a major attempt to evaluate an intervention that had gained national attention and implementation.[2] Despite this widespread support, however, there has been virtually no methodologically adequate evaluation of diversion and other similar forms of intervention.[3] Since diversion projects spend large sums of money, and thus divert resources from other possible interventions, it is imperative that their effectiveness be studied closely. Fishman purports to have conducted such an evaluation, thus his study merits a close inspection.

The Fishman study, based on data from eighteen rehabilitation and diversion projects in New York City, concluded on the basis of an analysis of recidivism rates that the projects failed to rehabilitate their clients. While acknowledging the problems in generalizing from this finding to diversion projects nationally, Fishman suggested that "unless there is better comparable data with different results the findings from these eighteen projects are indeed the best estimate of the outcome to be predicted for a universe of projects."[4]

A critical review of the methodology used in the Fishman study, however, suggests that this negative assessment of the effectiveness of rehabilitation and diversion projects in New York City and the assertion that one can generalize about these types of projects constitute a highly questionable conclusion and generalization. Most of the basic or standard criteria for program evaluation were either glossed over or ignored completely. There were specific problems with the research design, the sampling procedures, the reliability and validity of treatment and response measures, and the causal inferences made. It is not contended that Fishman was totally unaware of the limitations of his study and that most such evaluation studies can employ classical experimental criteria in the face of insurmountable practical problems. This article intends only to demonstrate the limited nature of the conclusions Fishman could properly draw about the effectiveness of rehabilitation and diversion projects in New York City or in general. It is important, therefore, that definitive policy recommendations, such as Fishman's position that these programs should be abandoned, must be avoided when practical or other exigencies prevent the employment of fundamental program evaluation criteria.

Section one of this article will review the methodology used in the Fishman study in order to demonstrate that the data do not justify the conclusions and generalizations reached by Fishman. Section two will then consider the broader implications of the relationship between criminology research and policy.

I. A CRITIQUE OF THE FISHMAN STUDY

A. THE LIMITED NATURE OF THE SAMPLE

The Fishman study was sponsored by the Criminal Justice Coordinating Council of New York

* This article is a co-authored critical response to Fishman, *An Evaluation of Criminal Recidivism in Projects Providing Rehabilitation and Diversion Services in New York City*, 68 J. CRIM. L. & C. 283 (1977).

** Assistant Professor, Departments of Criminology and Psychology, and Director, Criminology Research Center, Simon Fraser University.

*** Assistant Professor, Department of Criminology, Simon Fraser University.

[1] *See* Fishman, *An Evaluation of Criminal Recidivism in Projects Providing Rehabilitation and Diversion Services in New York City*, 68 J. CRIM. L. & C. 283 (1977).

[2] *See, e.g.*, AMERICAN BAR ASS'N COMM'N ON CORRECTIONAL FACILITIES & SERVICES, DIVERSION FROM THE CRIMINAL JUSTICE SYSTEM (1972); NATIONAL PRETRIAL INTERVENTION SERVICE CENTER, PRETRIAL CRIMINAL JUSTICE INTERVENTION TECHNIQUES AND ACTION PROGRAMS (1974).

[3] *See* J. MULLEN, PRE-TRIAL SERVICES: AN EVALUATION OF POLICY RELATED RESEARCH (1974); R. ROVNER-PIECZENIK, PRETRIAL INTERVENTION STRATEGIES: AN EVALUATION OF POLICY-RELATED RESEARCH AND POLICYMAKER GUIDELINES (1974); Roesch, *Does Adult Diversion Work? The Failure of Research in Criminal Justice*, 24 CRIME & DELINQUENCY 72 (1978).

[4] Fishman, *supra* note 1, at 303.

City. The council was funding fifty-three rehabilitation projects which were directed primarily at job-training, remedial education, or mental health counseling. Some of the projects (Fishman did not specify the number) also diverted clients out of the criminal justice system. The council was interested in obtaining feedback on the effectiveness of these projects in order to "determine whether its rehabilitation programs were having individual and collective effect on the criminal behavior of their clients."[5]

The first step in the evaluation of the projects was the development of a Standard Intake Form designed to obtain individual client data from each project. Four of the fifty-three projects submitted no forms to the evaluation, while the remaining forty-nine projects sent in 27,733 forms. Data from thirty-one of these projects were not used, according to Fishman, either because "1) there was not enough time to process the records; 2) the projects did not contain enough clients to permit analysis; or 3) the clients were female."[6] Fishman explained that many of the projects did not have a sufficient number of male clients to satisfy statistical requirements for analysis. He added that many of the projects were new and either had too few clients or were not in existence long enough to allow a sufficient period of time (twelve months) to obtain followup data.[7] This process left eighteen projects, with a total of 20,924 Standard Intake Forms.

It is important to note that this figure is only about 7,000 less than the original total, so despite dropping thirty-one projects from the analysis, Fishman was left with the majority of forms. But the process of eliminating subjects from the sample was not yet completed. Of the 20,924 forms, only 2,860 were used for analysis. Fishman acknowledged that "it could be contended that the 2,860 clients in the evaluation may not be a representative sample of the total population of either the criminal justice system or the 53 projects since representativeness was not demonstrated statistically."[8] Nevertheless, Fishman stated that it was his view that the clients in the sample were representative of the eighteen projects, the thirty-five projects not evaluated, *and* of the criminal justice system as a whole. This statement requires some analysis.

Given that the 2,860 clients represent a less than 10 percent nonrandom sample of all possible clients, it may be that they are completely unrepresentative of the eighteen projects, the fifty-three projects, and especially the criminal justice system as a whole. It may be that a disproportionate number of difficult offenders and those who were most likely to commit new crimes ended up in the final sample. This could be due to the methods of recordkeeping by projects if more attention and documentation was given to the more troublesome offender. At a minimum, the question of representativeness should have been answered empirically by comparing such factors as the demographic data, socioeconomic status, and criminal histories of the final sample to the eighteen projects.[9]

Fishman suggested that even if the sample is

[5] *Id.* at 284.

[6] *Id.* at 287.

[7] Since many of these projects were newly developed, it is appropriate to ask the following questions. What were the goals of these newer programs; to what populations were the projects directed; and did they differ from older, more established projects? These questions are important concerns since they relate directly to the representativeness of the 18 projects included in the final analysis.

[8] Fishman, *supra* note 1, at 289.

[9] It is particularly troublesome that there was a high error rate associated with the Standard Intake Forms. Fishman acknowledged that this was due to a concern for confidentiality and the resulting unwillingness of some project staff to provide any data for the evaluation, as well as a "lack of competence in record keeping." *Id.* at 287.

Despite extensive training programs designed to correct this problem, a high error rate continued. In addition, some information needed for police record retrieval was unavailable. The final result was that 18,064 of the 20,924 had to be excluded from the analysis. The only reason given for a nonrandom retention of approximately 10 percent of the available forms was that "the unused balance of the 20,924 forms were analogous to unreturned, incomplete or incorrect forms in an election-type survey." *Id.* at 289.

It simply is imperative to demonstrate that a systematic bias did not exist in reaching the final drastically reduced evaluation sample. Neither random sampling nor representative sampling techniques were employed. Thus, even though certain characteristics of the evaluation sample, such as ethnic group configuration, were between six to nine percentage points from the percentage of this characteristic in the relevant New York City sample population, a systematic bias still may have been in effect in reaching the evaluation sample. While acknowledging the potential limits of a nonrepresentative sample, Fishman maintained that a sample cannot be considered unrepresentative unless the following three questions are answered: "How was the method of selection biased? What kind of unrepresentativeness does this cause? How could the possible unrepresentativeness have resulted in the findings and conclusions presented?" *Id.* at 291. Usually, it is the responsibility of the author of such a report to answer these questions, especially when a strong possibility of an unrepresentative sample exists. Without intimate access to the project, particularly the actual sampling process and the entire data base, it is virtually impossible to answer the above questions.

unrepresentative "it is unlikely that they are more recidivistic than a comparable population in the criminal justice system,"[10] since the courts try to keep out defendants with more severe criminal histories. This is an example of an inappropriate inference. The concern is not whether they are more or less recidivistic than some population which is certainly *not* comparable (since their criminal histories will be a significant difference), but whether they are similar to the participants in the eighteen projects. In addition, the fact that defendants with more severe criminal histories are less likely to be diverted does not necessarily mean that only the better risk offenders are diverted. Forgers, petty thieves, and drug users may be diverted because their crimes are less serious, but the likelihood that they will be rearrested may be much greater than for murderers and rapists. Finally, one would need to determine how representative these eighteen projects are of the fifty-three projects in New York City and of diversion projects in general before questions of generalizability of results can be addressed.

Fishman stated that:

> [I]t appears far more useful and accurate to ask how the projects affect *specific* types of clients, by assessing each type separately, than to get one recidivism rate for a "representative sample" of *all* clients which "representative sample" may be composed of different proportions of males and females of different types and criminal histories.[11]

This statement misses a critical point. The proper question should have been whether the final sample was representative of the specific types of clients in each project. The sample may contain the same proportion of clients in each age group with similar criminal histories, socioeconomic status, and race, or, because of unknown bias in the selection of the final sample, it may be overrepresented (or underpresented) on any of these variables. If the sample was unrepresentative, then conclusions need to be limited only to the sample and not generalized to comments about the overall effectiveness of the projects.

B. PROBLEMS WITH THE OUTCOME MEASURE

The single outcome measure was a type of recidivism measured by the number and type of arrests during the twelve months after project entry. Fishman reviewed three possible measures of the incidence of crime—complaints, convictions, and incarcerations—but found weaknesses in each of the three that rendered them unusable. Yet in selecting arrests as the measure to be used, he failed to discuss the problems inherent in this measure of recidivism. For example, in the population served by the projects there were many clients who were well known to the police and thus might have been picked up and charged for a variety of reasons unrelated to criminal activity. Furthermore, defendants are often overcharged at arrest so that prosecutors will be in a better bargaining position.[12] Fishman should have pointed out these and other problems with the measure he selected.

It is also important to justify the use of recidivism as a response measure along lines such as Waldo and Chiricos[13] used in their study of work release. They posited five different theories concerning the way work release programs would be related to ex-offenders' behavior. Boruch and Gomez[14] point out that "[f]or confirmatory field research, and especially for policy-related evaluations, however, we believe there ought to be both sound theory and data to support the contention that the response variable proposed is indeed relevant to the treatment variable."[15] The Fishman study does not demonstrate this link for the arrest indicator he uses.

Moreover, while recidivism is clearly an important variable to be considered by any criminal justice intervention program, there are other variables which are just as important and which were directly addressed by some of the projects. For example, some projects dealt exclusively with addicts; thus a reduction in heroin use would be one measure of success of these programs. Other projects attempted to provide job counseling and referral; increases in employment would be a measure of success for these programs. Thus, while recidivism should always be included as an outcome measure, it should certainly not be considered the *only* measure of effectiveness.

The conclusion to be drawn from this discussion is that a program should not be evaluated on the basis of only a single measure. It would have been

[10] *Id.* at 291.

[11] *Id.* at 289 n.30 (emphasis in original).

[12] *See, e.g.*, Alschuler, *Prosecutors Role in Plea Bargaining*, 36 U. CHI. L. REV. 50 (1968); Kipnis, *Criminal Justice and the Negotiated Plea*, 86 ETHICS 93 (1976).

[13] Waldo & Chiricos, *Work Release and Recidivism*, 1 EVALUATION Q. 87 (1977).

[14] Boruch & Gomez, *Sensitivity, Bias, and Theory in Impact Evaluations*, PROFESSIONAL PSYCH. 411 (1977). *See also* Maltz & McCleary, *The Mathematics of Behavior Change: Recidivism and Construct Validity*, 1 EVALUATION Q. 421 (1977), for a discussion of establishing the construct validity of recidivism.

[15] Boruch & Gomez, *supra* note 14, at 414.

a far better evaluation of the effects of any or all of the projects if the conclusion had been based on several measures of the incidence of crime as well as a variety of measures of other potential effects of each of the projects (*e.g.*, employment, attitude change).

An additional difficulty with Fishman's use of recidivism was that he did not employ a severity index which could have combined the seriousness of arrests with the frequency of arrests. Fishman examined the seriousness of arrests in terms of the Uniform Crime Report categories of violent crimes, property crimes, and nonindex crimes, and then cross-tabulated seriousness with the frequency of arrests.[16] The study did not employ a standardized procedure such as the Sellin-Wolfgang index that could characterize "the frequency of an individual's arrest rates with the nature of arrest charges."[17] Fishman's failure to employ a summary measure of severity prevented him from distinguishing a recidivism record consisting of, for example, five burglaries as opposed to a record involving a single homicide. One could hypothesize that certain levels of severity of recidivism are more susceptible to diversion and rehabilitation programs than other levels. Without a severity index, Fishman was unable to examine this possibility.

Fishman did acknowledge the need to make distinctions in the severity of arrest records in his attempt to employ such an index for "prior arrest history." It is worthwhile examining Fishman's attempt to create a severity index for prior arrest history since it might shed light on why it was not used for the recidivism variable or for the "arrest history after project entry" variable. Fishman acknowledged that the Sellin-Wolfgang index was an appropriate measure for characterizing severity, but that he was unable to employ it for reasons of time, money, and availability of data. On the recommendation of Wolfgang and Figlio, Fishman substituted a Mean Seriousness Scores (MSS) procedure which had been used in the former's Philadelphia "cohort" study.

Fishman attempted to establish the validity of MSS for his New York City sample. The predictive validy of the MSS rests on its relationship with recidivism. The amount of variance predicted was less than 15 percent, a finding which Fishman admitted could have been simply a function of the large number of degrees of freedom. This inability of the MSS to predict recidivism was, according to Fishman, "an outcome without a ready explanation."[18] Still, he claimed that the predictive ability of the MSS was established by three different tests of significance. Then, with a minimum of explanation, Fishman dropped the MSS as his severity measure of prior arrest history and turned to the "average number of prior arrests." Because the latter measure accounted for the same variance in recidivism as the MSS, Fishman claimed that concurrent validity had been established. Given that "average number of arrests" was easier to understand and cheaper to employ, he chose it over the MSS as his measure of prior arrest history.

Clearly both measures of prior arrest history were of dubious validity.[19] Furthermore, unlike the MSS, no seriousness dimension was included in the substitute measure. Consequently, it is conceivable that an individual with ten shoplifting arrests could be classified as having a more severe prior arrest history than an individual who had been arrested for one rape and one homicide.

C. PROBLEMS WITH THE EVALUATION DESIGN

The basic design of the Fishman study was a one group pretest-posttest only design. The premeasure was based on arrest records during the second year prior to project entry, while the post data were based on arrests during the year after project entry. This design is subject to a variety of confounding extraneous variables which threaten both the internal and external validity of the results.[20] Any conclusions or generalizations from the data resulting from a design of this type should have been severely limited.

Threats to the validity of an experiment are usually controlled by the use of a randomly assigned control group. In a diversion study this

[16] Magnitude of arrests was also measured by the ratio of the total number of arrests after entry to the total number of arrests of all clients, both recidivists and nonrecidivists. This measure does not figure prominently in Fishman's analysis.

[17] Fishman, *supra* note 1, at 291.

[18] *Id.*

[19] It is difficult to infer prediction with any confidence based on the rejection of the null hypothesis. Tests of significance do not indicate the strength of relationships or explained variance which normally constitute the accepted measure for prediction. Also, while it remains a subject of debate, inferences based on tests of significance should not be made from a nonrandom sample such as Fishman's evaluation sample. *See* D. MORRISON & R. HENKEL, THE TEST OF SIGNIFICANCE CONTROVERSY (1970).

[20] *See* D. CAMPBELL & J. STANLEY, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH (1963). *See also* Cook & Campbell, *The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings*, in HANDBOOK OF INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY (M. Dunnette ed. 1976).

would mean random assignment of all eligible participants to treatment and control conditions. Fishman attempted to provide such a control by using a "control" group (N = 105) which had previously been selected by one of the eighteen projects.[21] The recidivism rates of this "control" group were compared with the male clients of eight other projects.[22] From a methodological point of view, this is an unsound procedure. A valid control group for all eight projects cannot be selected by only one project. For example, the project may have been attempting to select a certain type of offender and thus the individuals who were placed in a control group may not be representative of the offenders in the other projects. The bias could be in either direction; there could be a higher concentration of either more serious or less serious offenders. As a result, it is possible that the control group and the groups studied differed on a number of key variables.[23]

The Fishman study attempted to answer the basic question of the effectiveness of various diversion and rehabilitation programs. Since these programs are viewed as alternative methods of dealing with certain defendants (as opposed to court processing, probation, or incarceration), a judgment of effectiveness implies a comparison. Thus one would need to determine what would have happened to those or similar defendants had they been processed through the usual criminal justice system. An attempt at such a determination would have required a more controlled design than the one used by Fishman. It has been suggested that diversion programs can be more critically evaluated through the use of a randomized control group design consisting of three groups: diversion participants who receive the complete range of services offered by the diversion program; diversion nonparticipants who are diverted but receive no services; and a control group processed as usual through the legal system.[24] This kind of design would permit a comparison of the effects of diversion, including a range of possible services, with the usual method of handling defendants. The Fishman design falls far short of this kind of control, and it is primarily for this reason that the conclusions reached by Fishman were not supported by the design employed in the study.

A final problem with the basic design concerns the length of participation required by a project. Fishman stated: "[T]he arrest recidivism was measured over the period of twelve months after project entry, even though a client may not have remained in the project during the entire period."[25] Two questions are critical: How many of the 2,860 clients successfully completed the project and what was the length of participation? Many diversion projects in the United States are designed to allow for relatively brief periods of participation for clients.[26] Most projects require only three to six months participation, with only a small percentage of the projects requiring participation for a year or more. Length of participation is an important variable to consider in evaluating these projects because if successful completion meant three months involvement in a program and the client dropped out after only one week or one month, this might alter the effect the project would have on the client's behavior. It is clear that these two variables, length of participation and number of dropouts, could result in a significant bias in the results if there were a substantial number of dropouts or short-term participants.

[21] The procedures for selecting this control group were not described by Fishman and, without access to a project report on this procedure, it is impossible for the present authors to ascertain whether this group was an actual control group. A question arises as to whether these subjects were randomly assigned from a group of all eligible diversion participants (who had agreed to participate in the diversion program) and, if so, whether they were processed as usual through the criminal justice system. The exact selection procedures must be known before it can be determined whether this was a valid control group, at least for the project which selected it.

[22] This allowed, according to Fishman, a comparison of projects with male clients age 21 and older. It must be noted, however, that six of these projects also contained clients who were under 21 years of age.

[23] Obviously, Fishman considered the use of some type of comparison group to be an important design factor, a conclusion with which the present authors are in complete agreement. But the use of a comparison group which is not in fact "comparable" may only serve to mislead rather than enlighten. A decision as important as the one Fishman would like to reach as a result of the evaluation calls for at least an attempt to create some methodologically sound comparison groups, even if this must be done within the constraints of a post hoc analysis. One possibility would be to use a procedure like the multivariate matching technique developed by Sherwood, Morris & Sherwood, *A Multivariate, Nonrandomized Matching Technique for Studying the Impact of Social Interventions*, in 1 HANDBOOK OF EVALUATION RESEARCH (E. Struening & M. Guttentag eds. 1975).

[24] *See* note 3 *supra*. *See also* Roesch, *Pretrial Interventions in the Criminal Justice System*, in CHALLENGES TO THE CRIMINAL JUSTICE SYSTEM: THE PERSPECTIVES OF COMMUNITY PSYCHOLOGY (T. Sarbin ed. 1979).

[25] Fishman, *supra* note 1, at 294.

[26] *See* Roesch, note 24 *supra*.

D. THE USEFULNESS OF THE RESULTS

Given the problems in the design of the study, the results are necessarily of limited utility and generalizability. Nevertheless, some of the findings are useful and perhaps open to interpretations other than those offered by Fishman.

Recidivism at the twelve-month followup ranged from 24 to 51 percent, with a clear tapering off of criminal activity with age. In fact, by collapsing Fishman's data into two age categories, one finds that the mean recidivism rate for those age twenty and under is about 47 percent, compared with a mean of approximately 35 percent for the twenty-one to seventy-one year-old clients. Within levels of severity, those clients who had the least serious history (at least in terms of number of arrests) and were eighteen years old or less had considerably lower recidivism rates when compared with other levels in their age category. This may suggest the possibility, limited by the data, that diversion programs should concentrate on defendants who are either twenty-one or older, or under twenty-one with little prior contact with the law.

Fishman also looked at violent crime recidivism, as defined by the Uniform Crime Report classification—homicide, forcible rape, robbery, and aggravated assault—and found that such offenses accounted for 29 percent of the total arrests. Since Fishman concluded that this figure was "the main reason for the conclusion that the human costs of recidivism are too high,"[27] these data deserve a closer inspection.

It is interesting that nearly 60 percent of the violent crimes were robbery. When compared with the other three offenses in the category, robbery may be the least serious of the crimes classified as violent. As has been previously pointed out, a seriousness index for each crime and occurrence would have been more appropriate than classifying them together. Furthermore, the violent crimes were committed by only 16 percent of the clients, which is a result of the fact that many clients were arrested for more than one violent crime. A closer look at the personal and criminal histories of these individuals might have revealed information which could have been used to avoid selecting these defendants for diversion. Fishman did this for two variables, age at project entry and arrests for violent crime before project entry, and was unable to find any strong relationships, but many more variables could be added to strengthen the prediction.[28]

Fishman concluded that rehabilitation was a failure based on three major findings: 1) an overall 41 percent recidivism rate; 2) no significant differences between the "control" group and the eight projects with which it was compared; and 3) 29 percent of the total crimes committed one year after project entry were violent. Forgetting, for the moment, the differences which were found between age groups, one can ask whether this 41 percent recidivism rate was significant. Fishman suggested that the "control" group comparisons provide an indication of what would have happened without participation in the rehabilitation program. But if this group is not a true control group then the issue only becomes more clouded. Furthermore, a comparison of this "control" group with the project participants who made it into the final sample may present a distorted picture. Suppose that the sample participants were unrepresentative of the total population of participants and were more likely to recidivate. If so, a comparison showing no differences between the two groups would be possible.

Fishman's strongest argument against continuing the programs focused on the costs to the victims of these recidivists, in terms of cost of theft, property damage, injury, and death.[29] This implies that if these programs were unavailable, the crimes would have been prevented, thus "the decision to continue the programs also continues the high recidivism rates and the consequent high rate of crime."[30] This is an erroneous conclusion in which Fishman failed to acknowledge a fact about the criminal justice system which he previously mentioned in his article in a different context. If, as Fishman asserted earlier, incarceration is the weakest measure of the incidence of crime because relatively few offenders are sent to prison, then those offenders who would be eligible for diversion would probably be even less likely to be sent to prison than the general population of offenders since the selection criteria usually excludes serious offend-

[27] *Fishman, supra* note 1, at 299.

[28] Based on Fishman's own analyses of the predictive utility of a limited number of variables, and on a number of studies which show that dangerousness is vastly overpredicted, it would appear that, even with increasing the number of potential predictor variables, the identification of the individuals likely to commit violent crimes will be an extremely difficult task. *See* H. STEADMAN & J. COCOZZA, CAREERS OF THE CRIMINALLY INSANE (1974); Wenk, Robison & Smith, *Can Violence be Predicted?*, 18 CRIME & DELINQUENCY 393 (1972).

[29] Fishman, *supra* note 1, at 299.

[30] *Id.* at 303.

ers.[31] The fact that these individuals were diverted did not cause the subsequent crimes as it is likely that they would not otherwise have been imprisoned. Thus abolishing these programs may not decrease the number of crimes committed by convicted criminals.

E. SUMMARY

Fishman suggested that the results of the study provide a sufficient basis for making decisions about discontinuing rehabilitation and diversion services. He argued that given the high recidivism and incidence of violent crime, the continuance of the programs "may be difficult to justify on academic grounds alone."[32] The methodological problems raised in this article about the Fishman study suggest that this conclusion was both premature and unwarranted. The study did not lend itself to confident conclusions about the eighteen projects and certainly not about projects in existence in New York City or elsewhere. The present critique suggests that due to a variety of flaws in the design and data collection, the conclusions which can be drawn are largely limited to the group of participants who survived the process of reducing the initial sample from which they were drawn. The group selected for the study represented a less than 10 percent nonrandom sample of possible subjects for the study. One commentator has argued that "most evaluations, if done at all, are not well designed and usually lead to estimates of program effects which are so confounded with competing influences that making unequivocable statements about size and direction of impact is typically impossible."[33] The Fishman study is a prime example of such an evaluation.

II. Policy Implications of Evaluation Research

A. POLICY ISSUES

The limitations of the Fishman study in both research and policy terms are evident in many evaluation studies.[34] While ignorance of appropriate evaluation techniques might explain some of the reasons for this imperfect state of evaluation research, there is little doubt that financial, time, and ethical constraints are critical. Fishman clearly worked under financial and time constraints, and the major criticism of the study is that it did not qualify the policy inferences in the light of the many limitations of the research. Evaluation research has evolved to the state where criteria exist that provide the basis for qualifying policy inferences, but it is imperative that researchers acknowledge the limitations of their studies in terms of certain procedures and concepts of evaluation research. This article will propose a guideline for such evaluations based upon a modified version of a model initially discussed by Wortman.[35] Wortman established six major components of an evaluation: construct validity, process or formative evaluation, internal validity, summative or outcome evaluation, statistical conclusion validity, and external validity. For the purposes of this article, process and outcome are so heavily interrelated that they will be discussed in the same section. Internal and statistical conclusion validity also will be discussed together. Each of these components will be discussed in the context of developing program evaluation and a research base in criminology which could potentially be used to make policy decisions.

B. A MODEL FOR EVALUATION RESEARCH

1. Construct Validity

Various forms of validity constitute the crucial criteria for assessing evaluation research. Since Campbell's classic works on the concept of validity in evaluation research,[36] there has been an extensive concern in the evaluation literature with this concept.[37] Complete agreement on either the definitions of the various forms of validity or their order in the evaluation process is not evident. For the purposes of this article, however, sufficient agreement exists to provide a checklist of criteria against which evaluation researchers and policymakers can compare a particular study.

[31] The question of what would have happened to diversion participants if diversion were not available of course could be answered empirically through the use of the control group design discussed earlier in this article.

[32] Fishman, *supra* note 1, at 303.

[33] Boruch, *On Common Contentions About Randomized Field Experiments*, in Experimental Testing of Public Policy 139 (R. Boruch & H. Riecken eds. 1975).

[34] *See* I. Bernstein & H. Freeman, Academic and Entrepreneurial Research (1975).

[35] *See* Wortman, *Evaluation Research: A Psychological Perspective*, 30 Am. Psych. 562 (1975).

[36] *See* D. Campbell & J. Stanley, Experimental and Quasi-Experimental Designs for Research (1969); Campbell, *Reforms as Experiments*, 24 Am. Psych. 409 (1969).

[37] *See, e.g.*, Validity Issues in Evaluative Research (I. Bernstein ed. 1976); Readings in Evaluation Research (F. Caro ed. 1971); P. Rossi & W. Williams, Evaluating Social Problems: Theory, Practice, and Politics (1972); C. Weiss, Evaluation Research: Methods of Assessing Program Effectiveness (1972).

Construct validity involves, in part, the determination of whether causal inferences concerning independent and dependent variables have been derived properly from theory.[38] Construct validity is particularly difficult to determine in criminology because theories are often not elaborated in a manner that facilitates the direct derivation of variables and their relationships. Many theories in criminology are at a high level of generality in that they encompass such varied phenomena as deviancy and dangerousness. Consequently, many important constructs are not easily translated into operational measures. One example of this is the extensive effort needed to operationalize the construct of severity of crime.[39] Perhaps because of these problems, criminologists often do not attempt to establish a theoretical base.[40] In addition to this, there is the problem of the complexity of levels of analysis in which variables such as personality, nationality, age, culture, and history are central to the development of a construct or a theory. These variables are often quite difficult to operationalize and measure and thus lead to substantial controversy in criminology.

Construct validity is, therefore, especially critical when a theoretically controversial variable, such as "reduction of crime," is central to a research evaluation. At a minimum, multiple measures of key constructs should exist and if possible, an additional attempt made to establish the intersubjective validity of these constructs. More elaborate procedures involving multitrait and multimethod forms of measurement, treatment, and analysis[41] are often costly and difficult to employ. It is occasionally possible, however, to mitigate these restraints by engaging in substudies with restricted samples, while still employing multitrait and multimethod procedures.

Fishman acknowledged some of the shortcomings of the single rearrest measure of the construct "reduction in crime," yet made no argument other than cost and convenience for using this single measure. Cost, time, and convenience clearly play a paramount role, but nonetheless, Fishman could have done further analyses of even the partial data which appeared available from evaluations previously completed by the diversion projects. These earlier evaluations apparently contained other measures of the reduction in crime.[42]

If, as often happens, additional resources are not available to offset threats to construct validity, then the policy personnel should be alerted to the potential problems in determining the success or failure of a program. In the Fishman study, there was a strong threat to construct validity; the absence of the hypothesized cause and effect relationship between diversion programs and reduction in crime could have been the result of the improper translation of certain independent variables and the dependent variable from theory. But Fishman seemed unaware of this threat to construct validity since he maintained that the diversion programs caused an increase in crime.

Determining the existence of a cause and effect relationship in a research setting requires an as-

[38] As Cook and Campbell suggested, the primary area of concern for construct validity is that of arriving at operational definitions which can be interpreted in terms of more than one construct. They refer to this as "confounding."

> Confounding means that what one investigator interprets as a causal relationship between A and B another investigator might interpret as a causal relationship between A and Y or between X and B or even between X and Y, and later experiments might support one or the other of these reinterpretations.

Cook & Campbell, *supra* note 20, at 238. Cronbach and Meehl have also discussed construct validity, but largely focused upon its relationship to measurement development. This aspect is particularly relevant for criminologists attempting to define and measure phenomena which are not operationally defined. Cronbach & Meehl, *Construct Validity in Psychological Tests*, 52 PSYCH. BULL. 281 (1955).

[39] *See* T. SELLIN & M. WOLFGANG, THE MEASUREMENT OF DELINQUENCY (1964).

[40] Wortman has discussed this problem in the broader context of research dealing with social problems. His comments are especially true for much of the criminology research:

> The problem is not so much one of establishing the causal relationships that produce the phenomenon, but those that will eliminate it. Unfortunately, the emphasis is all too often on the immediate solution, and is accompanied by precipitous action to institute the remedy. The more time-consuming creation of explanatory theories accounting for social problems is an important, and often overlooked, activity in deducing likely solutions.

Wortman, *supra* note 35, at 562.

[41] *See* Campbell & Fiske, *Convergent and Discriminant Validation by the Multi-trait—Multi-method Matrix*, 56 PSYCH. BULL. 81 (1959).

[42] Fishman claimed that these evaluations varied considerably in quality and that since the evaluators were responsible directly to the projects, the objectivity of at least some of the evaluations was questionable. While we do not want to engage in another debate about this latter statement, it appears to us that it may have been useful to directly examine the nature and quality of the data to determine their appropriateness for the kinds of analyses we are suggesting here.

sessment of internal validity, yet frequently evaluation research will start with internal validity and ignore construct validity.[43] Given the paucity of theories that are structured in the form of nomological nets, evaluation research in criminology ought to begin with construct validity. It then would be possible for hypotheses to follow a hierarchial deductive order with cause and effect variables clearly delineated. The deductive process also is inherent in determining construct validity. This process forces the researcher to think through theories systematically before and after a study is in effect. Internal validity, however, is usually the primary concern in evaluation research since the central question often is whether or not one can change a dependent phenomenon. The construct validity concern of properly labeling the constructs involved in the causal relationship may be secondary to the internal validity concern of establishing that the desired causal change took place.

2. *Internal Validity*

Internal validity involves the determination of whether a causal relationship exists between treatment and outcome variables. This determination must include an attempt to establish whether the causal agent has been applied properly to the treatment unit. Formative or process evaluation is therefore critical to establishing internal validity, as will be discussed shortly. Internal validity should also include an attempt to eliminate competing or alternative explanations of the empirical relationship between treatment and outcome variables. The alternative explanations can be defined as threats to internal validity, and commentators have catalogued a wide variety of these threats.[44] Some of the more serious threats include selection, maturation, mortality, regression to the mean, instrument decay, and the interaction of selection with some of the remaining threats, *e.g.*, selection-maturation.[45]

Evaluation research in criminology generally appears to be faced with the need to establish the internal validity of findings that indicate either no causal relationships or very weak ones.[46] Few studies in criminology can afford, for a variety of cost and ethical reasons, to undertake the random assignment of individuals to control and treatment groups and, in addition, to employ the various research designs that are normally needed to offset most threats to internal validity. Evaluation researchers in criminology therefore have to be extremely cautious about the internal validity of their findings. There are methodologies available, which have been used with some success in practice, that can at least mitigate threats to internal validity without unreasonable political or other costs.[47]

The selection threat can be of special importance to evaluation research in criminology because of the difficulty of providing appropriate control or comparison groups. Alwin and Sullivan state that "[t]here is general consensus among policy researchers that evaluation of social programs is ineffectual in the absence of comparison groups.... Because evaluation research is often characterized by designs involving comparisons among non-equivalent treatment and control groups, pre-intervention differences must be taken into account."[48] Although Fishman felt that the study had provided for a comparison group, it is difficult to have confidence in a comparison or control group that was formed by selecting individuals from one treatment group who were not subjected to the entire range of treatment that members of other treatment groups experienced. If the comparison group had to be selected in this fashion, then comparison groups should have been selected from each treatment group; thus Fishman should have had eighteen comparison groups. An attempt

[43] *See* Cook & Campbell, note 20 *supra*.

[44] *See* SOCIAL EXPERIMENTATION (H. Reicken & R. Boruch eds. 1975); Alwin & Sullivan, *Issues of Design and Analysis in Evaluation Research*, in I. BERNSTEIN & H. FREEMAN, note 34 *supra*; Cook & Campbell, note 20 *supra*.

[45] Cook & Campbell, note 20 *supra*, have added the following threats which we believe should be the concern of criminologists: 1) diffusion or imitation of the treatment, in which control and experimental groups communicate with each other, allowing controls to also receive the treatment; 2) compensating equalization of treatment, in which administrators attempt to compensate for perceived inequalities between experimental and control groups; 3) compensatory rivalry, which occurs when assignment to groups is made public and members of the control group are motivated to reduce or reverse the expected effect; and 4) resentful demoralization of respondents receiving less desirable treatments.

[46] The Kansas City patrol experiment, reported in G. KELLING, T. PATE, D. DIECKMAN & C. BROWN, THE KANSAS CITY PREVENTIVE PATROL EXPERIMENT (1974), is a good example of a study in which the authors argued that the intervention was not effective. THE INTERNATIONAL ASSOCIATION OF CHIEFS OF POLICE, POLICE CHIEF 16 (1975), and Larson, *What Happened to Patrol Operations in Kansas City?*, in T. COOK, EVALUATION STUDIES REVIEW ANNUAL (1978), challenged the validity of the results, conclusions, and policy implications. *See also* Waldo & Chiricos, note 13 *supra*.

[47] *See, e.g.*, Waldo & Chiricos, note 13 *supra*.

[48] Alwin & Sullivan, *supra* note 44, at 103.

must be made to assess the selection differences between the comparison and treatment groups. It can not be assumed that no selection differences exist between the comparison and treatment groups unless the individuals were randomly assigned to treatment and control groups.

The maturation and the mortality threats to internal validity have to be major concerns in criminology-based evaluations. The maturation threat exists because age appears related to distinctive patterns of behavior in terms of the type and frequency of crimes committed.[49] Mortality is important since it can be extremely difficult to maintain the cooperation of individuals who are involved in areas of concern for criminology such as criminal proceedings and corrections. An attempt has to be made to account for high mortality rates when they occur, because these rates can affect the determination of whether the causal relationship exists.

Without a true experiment, threats to internal validity exist for any research, but they can be sometimes mitigated by properly utilizing certain quasi-experimental designs.[50] The nonequivalent control group design and the cohort design are particularly accessible to evaluation research in criminology. Developing extensive, reliable, and valid pretest and posttest measures is absolutely critical when the nonequivalent control group design and the cohort design are employed. The weaker versions of these designs, where the appropriate pretest and posttest measures are missing, are generally uninterpretable in terms of assessing internal validity. The design in the Fishman study is an example of an uninterpretable design where threats of internal validity such as selection and maturation remain unaccounted for.

Statistical conclusion validity represents another dimension of internal validity. The determination of whether relationships exist usually involves statistical inferences, a process which requires the use of caution.[51] An important statistical conclusion validity concern for evaluation research in criminology is the presence of weak relationships between treatment and outcome variables. The internal validity of such findings can be related partly to the selection threat and its interaction with other threats. If potentially important variables are not identified and statistically controlled, then variability becomes uncontrolled, and error variance increases to the extent where true differences among treatment and control groups are obscured. It is extremely difficult for criminological evaluation research to control all of the complex variables, in terms of generality and levels of analysis, in actual field settings. When these conditions prevail, one must exercise caution in arriving at no-difference conclusions. The danger of accepting the null hypothesis or weak relationships as valid is compounded "when sample sizes are small, significance is set low, one-sided hypothesis are incorrectly chosen and tested and most kinds of distribution-free statistics are used for hypothesis testing."[52] This is also true for measurement reliability and the reliability of treatment implementation; if measures are not highly reliable and treatments are not standardized, then error variance is inflated and true differences are further obscured.

3. *Process and Outcome*

Program evaluators have found it useful to distinguish between two types of evaluation: process and outcome. Process evaluation involves determining whether a program is actually implementing what it has intended to implement and has also been referred to as formative evaluation. Outcome evaluation, also referred to as summative evaluation, is basically concerned with whether a program has succeeded in reaching its goals in terms of some measurable and relevant criteria such as a change in the participants' behavior.

One concern with Fishman's approach to evaluation was that it was designed by an outsider who was unable to negotiate and discuss the matter with the individual program coordinators. Wholey discussed several important steps in defining and evaluating a program from two perspectives, that of the user and that of the evaluator.[53] These steps included determining the objectives and goals of a program; relating activities and programs to goals, intended impacts, and assumed causal links; and agreeing on a set of measures that will evaluate the presumed relationship between program activities and program outcomes. Determining the nature of these criteria is necessary prior to an outcome evaluation. There is often little correlation between

[49] *See* G. NETTLER, EXPLAINING CRIME (1978); M. WOLFGANG, R. FIGLIO & T. SELLIN, DELINQUENCY IN A BIRTH COHORT (1972).

[50] *See* Cook & Campbell, note 20 *supra.*

[51] *See, e.g.*, Berk & Brewer, *Feet of Clay in Hobnail Boots: An Assessment of Statistical Inference in Applied Research*, in T. COOK, EVALUATION STUDIES REVIEW ANNUAL (1978).

[52] Cook & Campbell, *supra* note 20, at 232.

[53] *See* Wholey, *Evaluability Assessment*, in EVALUATION RESEARCH METHODS: A BASIC GUIDE (L. Rutman ed. 1977).

stated program goals and activities, and the manner in which a program actually operates. Rutman pointed out: "The major purpose for monitoring the program's operation is to determine whether there are uniform activities that are implemented in a systematic manner."[54] If one is unaware of what a program is doing, and what its goals are, the results of the evaluation will not be useful. The Fishman study may have missed important aspects of each of the programs because of the failure to adequately consult and negotiate with each of the program coordinators and because of the study's focus on an outcome evaluation measured solely in terms of recidivism.

One must also question whether the activities which are designed to lead to certain goals make any theoretical sense. Suppose, for example, that the stated goal of a program is to increase the employability of participants. If the primary activity of this hypothetical program was individual therapy designed to enhance self-esteem there is a question as to whether this activity logically can be assumed to lead to increased employability. Consider the following example. Kassebaum, Ward, and Wilner conducted a large scale and controlled study of group therapy in prison and found that the therapy had no effect on recidivism.[55] This study thus challenged the theoretical link between prison therapy and the goal of decreased recidivism. This study points to a strong need for a closer examination of the values underlying our approaches to both intervention and research.[56] Repucci and Clingempeel point out that much of corrections research focuses on examining offender deficits, a fact which "constricts the research questions asked and the methods used."[57] Another way of looking at this, drawing upon the terminology used by Caplan and Nelson,[58] is that much of our efforts are directed at person-centered interventions, while many of the problems faced by offenders may be more directly relevant to system-centered interventions, including such activities as the creation of new settings and programs which will increase the power, autonomy, and self-control of disenfranchised groups.[59]

The examination of values underlying a program's approach to a problem clearly needs to occur before one can develop causal links between program activities and objectives. In the case of diversion, there is a definite need for an examination of the program itself as well as its relationship to other systems (*e.g.*, courts, police) and the community (*e.g.*, employment opportunities, training facilities). The alleged failure of diversion, as suggested by Fishman, may not actually be a failure of diversion per se, but a lack of appropriate and necessary services and interventions.

4. *External Validity*

External validity is concerned with the extent to which results can be generalized to other populations, settings, treatment variables, and measurement variables. External and internal validity are highly related to each other since it would not be possible to generalize from an internally invalid experiment. However, the presence of internal validity does not by itself ensure external validity. For example, an evaluator might ask ten diversion programs to participate in a research program, and have only five of them agree. The evaluator could then randomly assign potential divertees to treatment and control conditions. Assuming that other internal validity questions were properly addressed, the evaluator would have an internally valid experiment, but would not be justified in generalizing the results to the five nonparticipating programs. Since the nonparticipating programs may represent a biased sample, the evaluator would have to be quite cautious in any generalization from the results.

Although there are many methods for increasing the external validity of experiments,[60] it may be best for criminology research to focus its efforts on establishing internal validity. Refining external validity will often involve trade-offs which adversely affect internal validity. An evaluation which lacks internal validity is of little use, whereas internally valid research can be replicated. This will result in the availability of a larger data base which, over time, might establish external validity.

[54] Rutman, *Formative Research and Progam Evaluability*, in EVALUATION RESEARCH METHODS: A BASIC GUIDE 62 (L. Rutman ed. 1977).

[55] *See* G. KASSEBAUM, D. WARD & D. WILNER, PRISON TREATMENT AND PAROLE SURVIVAL (1971).

[56] *See* Seidman, *Justice, Values and Social Science: Unexamined Premises*, in RESEARCH IN LAW AND SOCIOLOGY (R. Simon ed. 1978).

[57] Repucci & Clingempeel, *Methodological Issues in Research With Correctional Populations*, 46 J. CONSULTING & CLINICAL PSYCH. 727, 729 (1978).

[58] *See* Caplan & Nelson, *On Being Useful: The Nature and Consequences of Psychological Research on Social Problems*, 28 AM. PSYCH. 199 (1973).

[59] *See* J. RAPPAPORT, COMMUNITY PSYCHOLOGY: VALUES, RESEARCH, AND ACTION (1977).

[60] *See* Cook & Campbell, note 20 *supra*.

C. CONCLUDING COMMENTS

One commentator has argued for an experimenting society "in which we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify or discard them on the basis of apparent effectiveness of the multiple imperfect criteria available."[61] If our society is to continue such social experimentation, we must be able to design evaluations which will provide policymakers with sufficient information from which to make decisions about program changes and general policy orientation. In part, this will require the establishment of "metaevaluation research," or evaluation of evaluations. Cook and Gruder suggest several methods for evaluating an evaluation, including independent reviews of proposals and subsequent reports, reanalysis of data to verify conclusions or examine new questions, the use of consultants, and, of course, multiple and independent evaluations.[62] This latter method is especially relevant to Fishman's study, for even if the design had a higher degree of internal validity, it still would have been necessary to demonstrate that the results were applicable in other settings, such as smaller cities or rural communities. The metaevaluation process would also provide additional checks on unwarranted or inappropriate conclusions, since the process would involve an examination of threats to internal and external validity.

The greatest use of program evaluation at present is not in determining whether a program is successful in a general sense, but rather in providing specific information about certain aspects of a program. Evaluations can help distinguish between those clients for whom the program was successful, and those for whom it was a failure and, in addition, help determine if there are particular staff/client matches which are effective. Information such as this can provide valuable feedback to a program and can lead to changes which might ultimately effect overall outcome.[63] The Fishman study might have been better used in providing feedback of this type to the programs that it studied.

Criminological researchers must accept the existence of design limitations and methodological difficulties and begin building a significant research base. Recognizing that this base often does not exist, it would be wise to proceed with extreme caution when attempting to develop policy based on any single study. Cook and Campbell have noted that:

> Improvements in design need to be made, can be made, and should be made in order to facilitate better causal inferences. But we would delude ourselves if we believed that a single experiment, or even a research program of several years' duration, would definitely answer the major questions associated with confidently inferring a causal relationship, naming its parts, and specifying its generalizability.[64]

Of course, it is not possible to design or execute a perfect experiment.[65] Given the state of the art, criminologists should limit their goals along the lines suggested by Mahoney. "Our goals, then, should be to strive toward conducting the least fallible inquiries, to cautiously interpret our experiments in accordance with their logical warrant, and to guard against the paralysis of complacency regarding the adequacy of current research methods."[66]

[61] Campbell, *supra* note 36, at 409.

[62] *See* Cook & Gruder, *Metaevaluation Research*, 2 EVALUATION Q. 5 (1978).

[63] *See* Nicholas, *Evaluation Research in Organizational Change Interventions: Considerations and Some Suggestions*, 15 J. APPLIED BEHAVIOR ANALYSIS 23 (1979); Walker, *The Ninth Panacea: Program Evaluation*, 1 EVALUATION 45 (1972).

[64] Cook & Campbell, note 20 *supra*, at 227.

[65] *See* W. WEIMER, NOTES ON METHODOLOGY (1977); Mahoney, *Experimental Methods and Outcome Evaluation*, 46 J. CONSULTING & CLINICAL PSYCH. 660 (1978).

[66] Mahoney, *supra* note 65, at 671.