EVALUATIONS OF INTERVENTION PROGRAMS IN CANADIAN CORRECTIONAL INSTITUTIONS: A METHODOLOGICAL ASSESSMENT

by

Laurel C. Cropley

B.A., York University 1976

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS (CRIMINOLOGY)

in the Department

of

Criminology

SIMON FRASER UNIVERSITY

August, 1982

APPROVAL

Name: Laurel C. Cropley

Degree: MASTER OF ARTS (CRIMINOLOGY)

Title of thesis: EVALUATIONS OF INTERVENTION PROGRAMS IN CANADIAN CORRECTIONAL INSTITUTIONS: A METHODOLOGICAL ASSESSMENT

Examining Committee:

Chairperson: John W. Ekstedt

Douglas F. Cousineau
Senior Supervisor

Ronald M. Roesch

Vincent F. Sacco
External Examiner

Date Approved: August 9, 1982

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of Thesis/Project/Extended Essay

Evaluations of Intervention Programs in Canadian Correctional Institutions: A Methodological Assessment

Author:

(signature)

Laurel Cropley

(name)

(date)

ABSTRACT

This thesis is a response to James Hackler's assertion that rigorous evaluation in Canada is an unproductive endeavour. Hackler claims that reviews of evaluation programs in the United States, most notably the review produced by Lipton, Martinson and Wilks, have indicated that evaluations are rarely conducted with a methodologically sound research design and, when they are, the results are likely to be negative. Therefore, he concludes, we should not make the same mistakes in Canada and we should not conduct evaluation research.

Academic interest has recently evolved in the area of 'meta-evaluation' research, that is, the appraisal of evaluation research. A review of the literature has produced several assessments of evaluation studies. A major focus of criticism and discussion in these assessments centres around the lack of methodological attention paid to the research designs employed, and the outcomes of the studies. This review also produced several generally agreed upon requirements for a properly conducted research design.

On the basis of the literature, 15 categories or criteria were established to evaluate the methodological adequacy of evaluation studies. Conclusions drawn by evaluators were also analyzed to see if they could be supported by the data. All Canadian Journals were reviewed and any study found dealing with an evaluation of an intervention program conducted in a correctional institution was assessed. The search produced

23 published studies conducted between the years 1960 and 1980.

The results of this study indicate that the quality of evaluation research in Canada is poor according to the rules of social experimentation. However, outcome discussions indicated that evaluators appear to be aware of the limitations of their studies.

Because of the frequency that the Lipton, et al. report is cited on treatment effectiveness, 10 published studies of intervention programs in prisons were selected from their review and analyzed according to the criteria employed for the 23 Canadian studies. The two sets of findings were then compared. The comparison indicated that the quality of research design was not significantly different between the American and Canadian studies. However, there was an indication that the Canadian studies were improving over time (from 1960 to 1980) in methodological rigor.

It is not possible to arrive at conclusive comparisons and statements of the state of evaluation research in Canada because of the small number of studies evaluated. However, the critical response to evaluation expressed by Hackler and Lipton, et al. is not supported by this evaluation. At present, the only statements we can make regarding program evaluation is that conclusions remain, at best, conditional and tenuous.

ACKNOWLEDGEMENT

I would like to thank Dr. Doug Cousineau for his many suggestions, and his support, Aileen Sams for her much needed assistance is putting this thesis together, and Dubbie and Dawn for their comments.

DEDICATION

Bob

Who understands and respects

what this represents

TABLE OF CONTENTS

TABLES

I. INTRODUCTION

This thesis is a response to James Hackler's (1978) assertion that rigorous evaluation research in Canada is an unproductive endeavour. Hackler, in The Prevention of Youthful Crime: The Great Stumble Forward (1978:23), has addressed the issue of evaluation research in the Canadian context, and asserts that if we were able to evaluate accurately, cheaply, and without the possibility of negative side effects, he would favour the evaluation of intervention programs in Canada. However, he is skeptical about conducting evaluation studies in Canada. He outlined three propositions to support his position.

In the first proposition, he identified three factors in support of his argument that it is unwise to establish evaluation activity in Canada. 1) He contends it is difficult to conduct an objective evaluation, 2) any attempts to do so may defeat other important purposes, and 3) external pressures (including political) to evaluate can aggravate this situation. He maintains (1978:28) that, "we do not have to repeat the same mistakes in Canada" that were made in the United States.

Third, he reviews several evaluation programs (three American and one Canadian) according to the formal requirements presented by Charles Logan (1972) to test the effectiveness of 'treatment' programs. These requirements include: 1) a clear set of program procedures, 2) some division of subjects into treatment and control groups (preferably random), 3) before and

after measures of the behaviour to be changed, 4) a definable measure of success, and 5) follow-up measures of the outcome variable in the community. Hackler indicates (1978:25) that, "tests of correctional or preventive effectiveness which met the required standards stated above are rare in Canada".

Finally, Hackler criticizes the use of an experimental design and contends (1978:28) that, "the 'soft' studies are almost universally successful. The 'hard' studies are almost universally unsuccessful or reflect no change". In addition, he indicates (1978:68) that, "sophisticated data analysis can be a barrier to communication between researchers and policy-makers". Therefore, he recommends an alternative strategy for evaluation. He suggests (1978:68) that we concentrate on the masses of data that are gathered by official agencies which can be analyzed from different perspectives. He suggests (1978:66) that, "careful record keeping may provide basic data for understanding, if not evaluating, some of the activities connected with the prevention of or supposed correction of delinquency".

However, despite the criticisms by individuals such as Hackler, the systematic evaluation of treatment or intervention programs has been established in the United States and is presently surfacing in Canada. To date there has been only one attempt to review Canadian intervention programs (Ross & Gendreau, 1980). We have therefore taken on part of that task in this thesis. Before beginning, there are several

counter-propositions which can be put forth in opposition to Hackler's position.

Beginning with his final proposition, that we should discontinue experimental research and concentrate on record keeping for evaluation data, we argue that manipulation of data should only be conducted once the reliability of the data is established through proper research procedures and found to be adequate. In addition, experimental evaluation might help to develop new techniques of intervention, and modify those that do not work. Evaluations should not be used simply to justify cancellation of programs.

Hackler's third assertion, that rigorous evaluations are rare in Canada, has not been supported in the literature. At the time of Hackler's writing there did not appear to be a systematic review of intervention programs conducted in Canada.

Finally, Hackler argues that it is unwise to evaluate existing or new programs in Canada, judging from the failures in the United States. However, despite his contention that evaluation is an unrealistic goal, researchers in the United States have concentrated on attempting to produce more rigorous evaluations. Henshel (1976:99), in discussing the virtues of evaluation, recognized the trap that many policy-makers and researchers, including Hackler, fall into. He indicated:

> In spite of the apparent obvious value of the systematic study of intervention, it is at once disturbing and fascinating to know that until very recently this type of study was rarely done, and even more rarely done in a manner which would permit meaningful conclusions to be drawn. In part this failure has been a reflection of a

> widespread lack of understanding of the essential elements of experimentation, and lack of recognition that the experimental approach could be applied to the appraisal of social reforms.

Hackler has based his claims primarily on research conducted in the United States. One of the sources of information utilized by Hackler in defence of his position is a report by Lipton, et al. (1975). This report has had some impact in the area of evaluation research, therefore it should be examined.

In 1975, Lipton, Martinson and Wilks produced one of the most extensive studies on the effectiveness of correctional treatment to be found in the literature. The purpose of their study was to accumulate information regarding effectiveness of rehabilitation techniques[1] as they recognized (1975:3):

> It is only through continuing evaluative research that it can be learned whether specific rehabilitative techniques are effective in dealing with various types of offenders. At the present time, there is no systematic program of evaluation of the offender treatment system.

They compiled 231 evaluation studies, from 1945 to 1967, covering eleven treatment categories, including probation, counselling, milieu therapy, and medical methods. Inclusion of studies in their analysis was based on five selection criteria: 1) The study must be an evaluation of a treatment method applied to criminal offenders, both adult and/or juvenile, 2) It must have been completed after January 1, 1945, 3) It must include empirical data obtained from an experimental or

quasi-experimental design, and must include some form of control or comparison group(s), which could include comparison with the general inmate population, matched control subjects, base expectancy rates or itself through comparison of pre and post measurements, 4) The data must be measures of performance improvement on some dependent variable(s), such as recidivism, attitude change or cost benefits, and 5) Clinical speculations and descriptive case studies were specifically excluded from the analysis. Other exclusion conditions were also employed. (See Lipton, et al., 1975:6).

Once the 231 studies which met the above five selection criteria had been collected, they were summarized and allocated to one of three categories.[2] The first category, "A" studies, were acceptable for their survey with minimal research shortcomings. "B" studies formed the second category, and included acceptable studies with research shortcomings that made interpretation of findings less clear. The third category was referred to as "Other Studies", and is comprised of articles excluded from further analysis for a variety of reasons, such as insufficient data being presented, extraneous variables found to confound the results, methods and variables inadequately defined or inadequate procedures used, or too small a sample size. They also assessed the studies regarding the type of research design employed. For this part of their evaluation they assigned a number from one to eighteen to each study, where one represented the best type of design possible and eighteen represented the

worst type of design. (See Lipton, et al., 1975:15-16).

The findings of the Lipton et al. study are far too extensive to discuss except in a very general manner. They came to a three-fold conclusion regarding the 'Inadequacies of Correctional Treatment and Research' (1975:627-628). First, they concluded that not only was more research needed, but a better grade of research will have to become a standard procedure before it can be determined whether treatment programs are effective or not. Their second and most widely quoted finding was that the narrow range of treatment techniques that are being employed in corrections are simply not effective. Finally, they recognized that the correctional system is very complex, and that perhaps the treatment model may not be the appropriate method to deal with the problem of recidivism.

Lipton, et al. have taken a meta-evaluation approach to the study of evaluation research. Academic interest has recently evolved in the area of 'meta-evaluation' research, that is, the appraisal of evaluation research. This particular type of research has been referred to as 'evaluations of evaluations' (Bernstein & Freeman, 1975:xii), and 'metaevaluative research' (Cook & Gruder, 1978:6). We will refer to it throughout this thesis as 'metaevaluation research'. A review of the literature has produced several assessments of evaluation studies. The major focus of criticism and discussion in meta-evaluation research has centred on two issues. The first concerns the lack of methodological attention paid to the research designs

employed. The second examines outcomes of the studies, i.e. whether the findings are positive or negative and the conclusions evaluators draw on the basis of them. Positive and negative findings refer to whether the intervention strategy has been effective in achieving the desired behaviour change. Our review of the literature also produced several generally agreed upon requirements for a properly conducted research design.

On the basis of the literature we reviewed on meta-evaluation research we established 15 catagories or criteria to evaluate the methodological adequacy of evaluation studies. Conclusions drawn by evaluators were also analyzed to see if they could be supported by their data. All Canadian Journals in Criminology, Sociology, Psychology and the Behavioural Sciences were reviewed and any studies found dealing with an evaluation of an intervention program conducted in a correctional institution were assessed. Our search produced 23 published studies conducted between the years 1960 and 1980. Published studies were chosen to be included in this assessment, as opposed to unpublished, on the assumption that peer review would be required before a piece of research would be accepted by a journal. This would implicitly incorporate, hopefully, a better quality of research than might appear if there was no external review.

The results of this study indicate that the quality of evaluation research in Canada is poor according to the rules of social experimentation. However, discussions of the outcomes of

their studies indicated that evaluators appear to be aware of the limitations of their studies.

Because of the frequency that the Lipton, et al. (1975) report is cited on treatment effectiveness (Hackler, 1978; Ross & McKay, 1978; Annis, 1979; Ross & Gendreau, 1980; Palmer, 1978), 10 published studies of intervention programs in prisons were selected from their review and analyzed according to the criteria employed for the 23 Canadian studies. The two sets of findings were then compared. The comparison indicated that the quality of research design was not significantly different between the American and Canadian studies. However, there was an indication that the Canadian studies were improving over time (from 1960 to 1980) in methodological rigor.

EVALUATION RESEARCH: REVIEW OF THE LITERATURE

Before we begin with our assessment, we shall briefly examine the area of research with which this study is concerned as a background to the following discussion.

The literature on evaluation research has frequently focused on the development of applied and basic social science research, and the differences between them (See Rossi, et al., 1978; Bernstein and Freeman, 1975; Rossi, et al., 1977; Polivka and Steg, 1978). Much of the discussion in this body of literature focuses on differences in funding arrangements, the audience for whom the research is conducted and research schedules.

A variety of definitions have been employed in discussions of evaluation research (Weiss, 1972:4; Riecken, 1972:86; Weinstein, 1975:134; Rossi and Wright, 1977: 5; Suchman, 1972:53; Bernstein and Freeman, 1975:1; Wholey, et al., 1970:23). There are several distinctive features of evaluation that can be identified from the definitions employed in the literature.

First, there is usually an assumption that there is an objective or goal of intervention that is desirable. Second, a planned program of intervention is designed to achieve the desired goal and third, a method is developed for determining the degree to which the desired goal is attained as a result of the planned program. This third feature assumes that change is measurable. Finally, the 'tools of science' are used to study the effects of the program. A variety of research designs have been and can be employed for assessing the intervention (Rossi and Wright, 1977:5; Suchman, 1972:65-66; Riecken and Boruch, 1974:3; Weiss, 1972:18), however, the general preference in the literature rests with a 'true classical experimental design'.

For the purposes of this thesis we define evaluation as the attempted assessment, using the 'principles of research design', of an intervention program conducted in a federal, provincial or juvenile institution in achieving its objectives, and which is published in a Canadian journal. This definition includes several specific characteristics: An assessment of the effectiveness of a program includes an examination of the stated

objectives and compliance of the program to those objectives. The principles of research design, as identified by Wholey, et al. (1970:23) refer to the use of a classical experimental design where possible to a) measure the effects of the program, b) to allow for comparison between competing programs, and c) to provide the causal connection between program and effects while controlling for extraneous influences or alternate explanations. This element of evaluation will be discussed in more detail in Chapter II, as well as a justification for the utilization of the proposed classical design.

We are also interested in evaluation research as it adds to the body of knowledge that has developed in this area of research and, in turn, generates further knowledge.

There have been a number of textbooks and journal articles dedicated to an analysis of the development of, and obstacles to, evaluation research (Rossi and Wright, 1977; Riecken and Boruch, 1974; Weiss, 1972; Bernstein and Freeman, 1975; Wortman, 1975; Weinstein, 1975; Hackler, 1978; Gottfredson, 1979; Polivka and Steg, 1978; Chelimsky, 1977; Freeman and Sherwood, 1970; Wholey, et al., 1970; Szabo and Rizkalla, 1978; Rossi, 1972; Scriven, 1972; Cavior and Cohen, 1975; Glaser, 1974; Rossi and Wright, 1978; Rossi, Wright and Wright, 1978; Cook and Campbell, 1976; Rossi and McLaughlin, 1979).

Most discussions focus on the aims or purpose of evaluation. The primary aims appear to be; 1) to assist policy-makers and administrators in their decision-making

functions (Suchman, 1972:55), and 2) to determine the cost/benefit ratios of competing programs (Bernstein and Freeman, 1975:4; Rossi and Wright, 1977:6). Bernstein and Freeman (1975:1) have also noted that, ideally, an additional purpose or goal of evaluation research is to add to the existing knowledge in social programming.

Discussions of the obstacles to conducting evaluations have surfaced in many articles. The major obstacles, according to the literature, include; incompetence of researchers, problems in the interaction of researchers and program staff, inadequate funding (Bernstein and Freeman, 1975:5-7), vague statements regarding program goals made by administrators and, in general administrative considerations (Weiss, 1972:7; Cavior and Cohen, 1975:238; Rossi and Wright, 1977:6). Rossi (1972:227) for example has noted: "One of the major obstacles to evaluation research is the interests in the maintenance of a program held by its administrators."

The major consequence of many of the above obstacles has been the reduction of technical quality in evaluation research (Rossi and Wright, 1977:9-11; Rossi, 1972:233; Cook and Campbell, 1976:300). Meta-evaluators have turned considerable attention to this phenomenon. One of the major sources of criticism of evaluation has focused on methodological issues, specifically, the type of research design employed to evaluate the effectiveness of intervention programs. This is the only aspect of evaluation research with which we are concerned in

this thesis. Therefore, before discussing meta-evaluation and the methodological issues surrounding evaluation research we shall examine the types and potential of research designs to assess the possible impact of intervention techniques.

NOTES

1: We might question why Lipton, et al., did not acknowledge the previous evaluations conducted by Pawlicki (1970), Logan (1972), Fisher and Erickson (1973), Slaikeu (1973) and Davidson and Seidman (1974). Although evaluations of this nature are not consistently conducted on all intervention programs implemented in the 'offender treatment system', to ignore the attempts that have been made to establish a program of evaluation negates the import of these previous works for evaluation research.

2: Lipton, et al's., review of the articles contained nine features: 1) treatment method; 2) desired area of change, 3) setting for treatment, 4) nature and size of population, 5) research design, 6) time in treatment, 7) time in follow-up, 8) outcomes, and 9) research shortcomings. This final feature was critical in determining in which category the reviewed studies were included. Lipton, et al., (1975:6) define research shortcomings as, "those aspects of the research methodology that may call to question the results the researcher obtained." For further discussion of these categories, see Lipton, et al., 1975, pages 7-20.

II. RESEARCH DESIGN

INTRODUCTION

There is very little that can be said about research designs that has not been exhausted in a multitude of textbooks and journal articles on research methods. However, there are many different types of designs that have been employed in evaluation research, and each design varies in its methodological rigor, consequently differing in the usefulness of its findings. This thesis focuses on the methodological quality of evaluation studies; the central concern stemming from the nature of the research design. We will, therefore, examine briefly the various research designs that have been utilized in evaluation research and the advantages and problems connected with them. However, before we begin with our discussion of design, we should note how these designs are applied, i.e. how the results of evaluation studies are used or misused.

One of the central policy issues that concerns evaluation researchers is the applicability of their findings within the context of administrative goals and programming. A common error that has often been made by researchers is claiming more than the design will allow (Suchman, 1972:65; Szabo & Rizkalla, 1978:23; Freeman & Sherwood, 1970:110).[1] Roesch and Corrado (1979:536) stress that researchers must not only be aware of the

limitations of their studies, but must also acknowledge them when presenting their findings.

With this in mind, we shall now turn to a discussion of the functions of research and types of designs employed in evaluation studies.

FUNCTIONS OF RESEARCH

The primary purpose of research is to develop and evaluate practices, concepts, and theories of social relations and to develop and evaluate methodologies that test those practices, concepts and theories. (Selltiz, Wrightsman and Cook, 1976:7) There is fairly general agreement in the literature that evaluation research, while different from other forms of research, i.e. basic research, in its purpose, use and relationship to social and political institutions, is not different in its methods. (Selltiz, Wrightsman and Cook, 1981:83)

The purpose of evaluation research was aptly put by Chelimsky (1977:442) when she indicated that:

> While it is true that evaluation quality derives, at least in part, from the insight and objectivity of the evaluators, the incremental pay-off to that insight and objectivity is directly proportional to the empirical evidence available on which judgements can be made...if evaluative data are not available, then programs cannot be well assessed and effectiveness must remain a matter of conjecture.

Suchman (1972:64-65) identified three major methodological requirements of evaluation research; 1) description and analysis

of input which clearly identifies the active components in the program; 2) understanding the cause/ effect of the behaviour change underlying the desired objective; and 3) definition of the desired goal in terms of criteria which permit valid and reliable measurements of attainment.

Selltiz, et al. (1976:161) have noted that; "To be useful, the data-collection techniques and the rules for using the data must produce information that is not only relevant but correct." Correctness of information is coincident with research design, so we will now turn to a discussion of the experimental design, which we maintain should be utilized in evaluation research.

EXPERIMENTAL DESIGN

In recent years there has arisen a growing concern for useful and reliable methods of assessing the effectiveness of program results (Conner, 1977:195). There is fairly general agreement in the literature that a true experimental design is definitely preferable to a non-experimental or quasi-experimental design (Rossi, Wright & wright, 1978:179; Riecken & Boruch, 1974:8-9; Weiss, 1972:18; Szabo & Rizkalla, 1978:241; Wortman, 1975:562; Lundman & Scarpitti, 1978:219; Conner, 1977:195; Freeman & Sherwood, 1970:107; Powers & Alderman, 1979:89; Riecken, 1975:6; Rossi, 1972:225; Rossi & Wright, 1977:13). Riecken and Boruch (1974:xiii) claim that: "experimentally designed trials of interventions provide the least equivocal evidence possible regarding the effectiveness of

an intervention."

It has been recognized that evaluation research, as a form of social experimentation, is subject to the same rules as experiments in other fields. (Freeman & Sherwood, 1970:103) To elaborate, Selltiz, et al. (1981:83) note that evaluation research contains the same research design and measurement problems as basic research; it is also subject to the same threats to validity; the same issues that arise concerning measurement devices and indices in basic research also occur in evaluation research; and finally, the same problems are encountered in operationalizing definitions and procedures.

The research formula utilized by evaluation researchers, as identified by Weiss (1972:6), follows the format of a classical experimental design in that first, the goals of the program must be identified. They are then translated into measurable indicators, and data is collected on those indicators for both experimental and control groups. Once data collection is completed, the two groups are compared in terms of the goal criteria.

The essential features of the classical design, then, include: 1) some form of treatment or treatments; 2) recipients of the treatment preferably drawn at random from the same population; 3) measurements made on all individuals (of a baseline nature) preceding the intervention; and 4) random assignment of all individuals to experimental or control groups (Riecken & Boruch, 1974:44-48).

PROBLEMS WITH EXPERIMENTAL DESIGNS

Many problems have been identified regarding the plausibility of conducting true experimental designs. The ethics surrounding the random assignment of individuals to control and experimental groups has been identified as one problem (Freeman & Sherwood, 1970:106; Riecken & Boruch, 1974:250; Gordon and Morse, 1975:341). However, it is primarily the methodological problems that beset experimental assessment of treatment programs. Many of these problems stem from the ethical constraints, as well as from administrative needs and demands.

Rossi, Wright and Wright (1978:180) noted that many of the problems encountered when attempting to implement this design result from the actual quality of the design. They claim that an experimental program is better conducted than would occur if the program was to proceed without being evaluated, and the results may be excessively favourable or unfavourable. This could, they suggest, create unrealistic conclusions and expectations. They also noted that experiments are expensive, complex and time-consuming. Gordon and Morse (1975:342) suggest that for an experimental design to be employed, the program must remain stable for a lengthy period of time (until at least the evaluation has been completed), however, it is possible that administrators view their programs as flexible and continuously changing and progressing.

ADVANTAGES OF EXPERIMENTAL DESIGNS

Riecken and Boruch (1974:9) discuss the advantages of experimental design in considerable detail. They note that experimental designs are useful in ruling out 'causal displacement'. Causal displacement refers to the cause/effect relationship of treatment; i.e. that it is in fact the treatment that created the particular change in behaviour, and not some other alternative condition or intervening factor. By ruling out causal displacement, this allows for the comparison of two or more equally plausible kinds of treatments, by clearly identifying and measuring the components of each. This is possible because experimental designs, conducted properly, should force operationalization of definitions, and require that the treatment be explicitly defined and described.

As experimental designs provide a better estimate of the impact of a particular treatment than any other type of design, and the primary purpose of impact evaluation is to provide that information, there is little more that needs be said regarding its value. As Rossi, Wright and Wright (1978:180) have recognized, because of their superior inferential power, experiments are here to stay.

QUASI-EXPERIMENTAL DESIGNS

It should be noted at this time that although an experimental design is definitely preferable to most other methods and should be employed wherever possible, it is not the only method. Riecken (1975:6) suggested that in recent years quasi-experimental designs have become more and more accepted and acceptable, as techniques are being developed to increase their usefulness. He (1975:6) noted that:

> Quasi-experimental techniques generally yield more equivocal results than randomized experiments do, yet they are a promising device for identifying and minimizing the inferential equivocality which is impossible to escape outside of an experiment.

Quasi-experimental designs are being utilized more and more, and there is a variety of designs that can be applied in an evaluative context. We will, therefore, turn our attention to quasi-experimental designs and some of their advantages.

Quasi-experimental designs differ from 'true' experimental designs in that groups are assigned in non-random fashion in the former situation (Cook & Campbell, 1976:224). It has been noted throughout the literature that evaluation can occur on many levels and at different stages of the development of an action program, and that many different types of research designs, varying in their approximation to a classical design, can and should be used (Suchman, 1972:57; Rossi & Wright, 1977:15; Freeman & Sherwood, 1970:106).[2]

There are several advantages to the use of quasi-experimental designs which results from the problems encountered when attempting true experimental designs. First, they are easier to conduct.

Second, it is suggested that experiments make oversimplified assumptions about how programs affect individuals (Kennedy, 1979:662). Because treatments are hard to define and manipulate, and because there are many 'extenuating circumstances' and 'intervening influences' that must be considered, Kennedy suggests this design should be replaced by more illustrative techniques such as case studies[3] that can identify the more intricate details of the program.

Finally, situations where a true experimental design can be conducted are rare. Rather than have no assessment of program effectiveness, quasi-experiments are an alternative. However, there is a variety of designs being employed in evaluation research that differ in their approximation to a true design, and it is this mistaken belief, that any form of evaluation is better than none, that has led to the utilization of unreliable and invalid methods to assess the effectiveness of treatment programs.

CONCLUSION

Rossi (1972:233) has noted that we need powerful designs to detect results. He suggests that before we attempt to use 'soft techniques' perhaps we should determine just how 'good' or 'bad' they are.

Freeman and Sherwood (1970:107) also commented on the necessary balance between rigor and reality. They indicated that although we cannot ignore the constraints imposed on research, it is essential that evaluators examine the manifestation of these constraints on the type of design required to "ascertain whether the compromises necessary for its conduct are so severe that it would be inadvisable to pursue the research."

The pressures to evaluate are very apparent (Hackler, 1978:39; Freeman & Sherwood, 1970:107), however, the integrity of the research design must not be compromised because of those pressures. Perhaps we should be asking ourselves when should evaluations be conducted, rather than settling for less than adequate designs. Riecken (1972:100-101) has responded to this, saying:

> The question is an extremely important one, involving not only professional prestige, financial support for research, the advancement of technique and accumulation of knowledge, but, quite as important, responsibility to society, assistance to practitioners, the cost of failure in terms of wasted effort and unfulfilled expectations, and, not least, the probability that research results can have some effect on action.

Social scientists are now becoming more concerned with the quality of research. In recent years a new branch of evaluation

research has developed which is concerned with the evaluation of conducted evaluation research. We will now turn to Chapter III and a discussion of 'Meta-evaluation Research'.

NOTES

1: This becomes especially problematic when researchers compromise their methodological standards and settle for less than adequate designs (Rossi & Wright, 1977:13).

2: It has also been suggested that rather than dichotemize the use of 'soft' and 'hard' data, i.e. qualitative versus quantitative, researchers should attempt to coordinate the use of both to develop a more rounded picture of the entire operation (Szabo & Rizkalla, 1978:18).

3: Case studies, as employed in this paper refer to studies with an N=1.

III. META-EVALUATION RESEARCH

INTRODUCTION

In 1949, Merton contended that one responsibility of social science, which has been seriously neglected in the past, is to study the performance of professional social scientists (Bernstein, 1978:25). Despite recognition of this responsibility over thirty years ago, there appears to be relatively little concern before the last decade for the methods used by social scientists in evaluation research. Most of the research in this area has been conducted in the United States. As noted earlier, one of the most comprehensive reports on the effectiveness of correctional treatment was produced by Lipton, Martinson and Wilks (1975). Their conclusion argued that correctional treatment is ineffective. Prior to publication of the above report, Martinson, one of the authors, (1978:805) implied that 'nothing works' in correctional treatment. The reaction to Martinson's argument was immediate and varied, for example, Ross and McKay's (1978:279) cryptic observation of this report illuminates the attitude several researchers have taken: "Martinson, the funeral director, may have signed the death certificate for treatment through his critical review of the published research on treatment in corrections." On the other hand, as Annis (1979:5) recognized: "correctional treatment

programs have been subjected only infrequently to valid evaluation using an adequate experimental design."

As we noted in Chapter I, academic interest has evolved in the last ten or twelve years regarding the methods employed in, and impact of, evaluation research. Although it may be classified under different titles (i.e. evaluation of evaluations or meta-evaluation), the major goal of meta-evaluation remains the same; to improve the 'state of the art' in evaluation research. The need for evaluations of evaluations has been recognized throughout the literature (Bernstein & Freeman, 1975:xii; Riecken & Boruch, 1974:38-39; Roesch & Corrado, 1979:541). We shall therefore turn to a discussion of 'Meta-evaluation research'.

META-EVALUATION RESEARCH

Cook and Gruder (1978:6) employ the term 'meta-evaluation' as referring "only to the evaluation of empirical evaluations - studies where the data are collected directly from program participants within a systematic design framework.". The purpose of meta-evaluation is: "simply to help evaluators meet their goals by providing diagnostic feedback and helpful advice about what to do." Riecken and Boruch (1974:38-39) also indicate that evaluation studies be reevaluated or reanalyzed and replicated. They suggest (1974:39) that:

> ...because social experimentation is designed to shape public social policy, the public has a corresponding right to know whether the experiment has been well conducted and whether its results are dependable.

Cook and Gruder (1978:6) have identified what they call three research traditions that have particular import for metaevaluation:

1. The first is to acquire an evaluator's data and reanalyze it to answer either the same questions or new ones;
2. The second tradition is simply to see how technically competent studies are in general; and
3. Finally, 'research on research' forms the third tradition. Although not well defined in this article, the purpose of this final method appears to be related to determining ways of articulating and producing research. Questions are asked, such as, what are the consequences of writing requests for proposals in structured or unstructured ways?[1]

Cook and Gruder (1978:28) discuss metaevaluation conducted after the completion of the evaluation study using published reviews as the source of research data. However, they suggest that metaevaluation might be better conducted at different stages of the evaluation study itself, using different techniques of analysis. For example, they indicate that empirical reevaluation of a program evaluation may improve the quality of the results as raw data are being reassessed.

Several attempts have been made in the last ten or twelve years to increase methodological rigor in evaluation studies, primarily in the United States. (See: Emery & Marholin, 1977; Davidson & Seidman, 1974; Logan, 1972; Gordon & Morse, 1975; Bernstein, 1978; Slaikeu, 1973; Fisher & Erickson, 1973). The

attention of researchers in Canada has surfaced only recently, however, the need for socially responsible research has been established, and will hopefully continue to set a new trend in evaluation research (Russon, 1963; Pawlicki, 1970; Quinsey, 1973; Henshel, 1976; Gendreau & Ross, 1978; Ross & McKay, 1978; Roesch & Corrado, 1979; Normandeau & Hasenpusch, 1980). Before detailing the criteria for assessment that will be employed in this study, we shall briefly review the literature on metaevaluation research in order to elucidate the central issues which must be reexamined in order to elevate program evaluation into an acceptable body of research.

We found eleven meta-evaluators concerned with the assessment of evaluation studies in terms of methodological criteria (Pawlicki, 1970; Logan, 1972; Fisher & Erickson, 1973; Slaikeu, 1973; Davidson & Seidman, 1974; Gordon & Morse, 1975; Lipton, et al., 1975; Emery & Marholin, 1977; Ross & McKay, 1978; Bernstein, 1978;[2] Annis, 1979). In general, a review of the literature was conducted by each meta-evaluator and studies were selected and content-analyzed. They were then analyzed according to a specific set of methodological criteria.[3] Overall, our review of these meta-evaluation studies found approximately thirteen criteria to be important for the assessment of evaluation studies. Not all of the criteria were employed by each meta-evaluator. See Table I and Legend for an outline of methodological requirements that have been used by each meta-evaluator to assess evaluation studies.

Table I

Methodological Requirements Found in Evaluation Literature

Author	Date	Control Groups	Follow-up	Baseline Measures	Systematic Variation of Treatment	Adequate Definition of Program and Techniques	Adequate Definitions of Success and Failure	Unbiased Observers	Randomization	Routinization	Multi-Collinearity	Multiple Measures	Target Behaviours	Context of Group Treatment	Total
Pawlicki	1970	*	*	*	*	*		*		*					7
Logan	1972	*	*	*	*	*	*		*	*					8
Fisher & Erickson	1973	*		*	*					*	*				5
Slaikeu	1973	*	*	*		*	*				*			*	7
Davidson & Seidman	1974	*	*	*	*			*	*			*	*	*	9
Gordon & Morse	1975	*	*			*			*			*	*	*	7
Lipton et al.	1975	*	*	*		*					*			*	6
Emery & Marholin	1977		*										*		2
Ross & McKay	1978	*	*												2
Bernstein	1978	*							*			*			3
Annis	1979	*	*	*					*						4
Total		10	9	7	4	5	2	2	5	3	3	3	3	4	

Legend for Methodological Requirements
(for Table I)

1. Control Groups.

   Any group that is not receiving the specific treatment being evaluated. The Control group should be equivalent to the treatment group or at least provide a comparison.

2. Follow-up.

   A delayed measure taken in the community, preferably for both the experimental and control groups.

3. Baseline Measures.

   Measures taken of the treated behaviour before the intervention is introduced. Measures should be taken of both the experimental and control groups.

4. Systematic Variation of Treatment.

   A breakdown in the elements of the treatment that may be responsible for the observed change in behaviour. This includes evidence that the treatment group is receiving the treatment and the control group is not.

5. Adequate Definition of Program and Techniques.

   The aims of the program and intervention techniques employed should be sufficiently operational to determine if the treatment is actually being received by experimental subjects.

6. Definitions of Success and Failure.

   Outcomes of the program should be operationalized so that reliable measurements of the subjects' performance can be made.

7. Unbiased Observers.

   Also referred to as 'blind analysis'. Individuals taking the measurements are not aware of whether the subjects are in the experimental or the control group.

8. Randomization.

   The assignment of subjects to experimental or control conditions in a manner determined by chance.

Legend (continued)

9. Routinization.

   The ability or potential for particular elements of a program, the techniques employed and subsequent results to be extended and applied to other behaviours, environments and administrative styles.

10. Multicollinearity.

   The intercorrelation of two or more variables which, when statistically measured could produce deceptive results of the weight of the variables.

11. Multiple Measures.

   The use of more than one measurement device to test for behaviour changes.

12. Target Behaviours.

   Behaviours to which the intervention is directly applied and assessed throughout the evaluation.

13. Context of Group Treatment.

   Situational factors in the environment which may affect an individual's behaviour.

The thirteen criteria include: control groups, i.e. any group, equivalent to the experimental group, which is not receiving the specific intervention being evaluated; follow-up, that is, a delayed measure taken upon completion of the intervention; baseline measures, which are measures taken of the behaviour to be "treated", for both the experimental and control groups, before the intervention is introduced; systematic variation of treatment, i.e. a breakdown in the elements of the intervention techniques that may be responsible for the observed change in behaviour; adequate definition of the aims of the program and the intervention techniques employed, which means that these terms should be operational and clearly expressed; definitions of success and failure which should be operationalized so that reliable measurements of the subjects' performance can be made; unbiased observers, which is also referred to as 'blind analysis', i.e. those taking the measurements are not aware of whether the subjects are in the experimental group or the control group; randomization, which is the assignment of subjects to experimental or control conditions in a manner determined by chance; multicollinearity, which is concerned with the intercorrelation of two or more variables which, when statistically measured could produce deceptive results of the weight of the variables; multiple measures, i.e. the use of more than one measurement device to assess the possibility of behaviour changes; target behaviours, i.e. behaviours to which the intervention is directly applied and

assessed throughout the evaluation; and context of group treatment, which are situational factors in the environment, such as the prison setting, which may affect an individual's behaviour.

It appears from the meta-evaluation studies reviewed that methodological rigor is lacking in many evaluation studies. With a few exceptions, every criterion assessed by the above meta-evaluators was met by less than half of the studies they reviewed.

The above reviews were concerned primarily with evaluation studies conducted in the United States. With one exception (Ross & Gendreau, 1980), there has not been an assessment of Canadian research attempts. Therefore, in Chapter V we shall employ the same techniques of assessment that were used by Logan and several other meta-evaluators to examine the 'state of the art' in Canada. Before doing so, however, in Chapter IV we shall examine in greater detail the criteria we will be employing in our assessment.

NOTES

1: Bernstein (1978) also discusses this third kind of research tradition. She concentrates on the social factors that either limit or enhance achievement in applied social science. Within this general framework, she examines (1978:26) one aspect of achievement, "conformity of a set of methodological norms, and how a variety of social factors affect such conformity".

2: Bernstein (1978) attempted a different approach to the analysis of evaluation studies reviewed. In her study she attempted to construct a scale by which to weight the assessment criteria and then measure adherence to them. Unfortunately, the procedure she used for establishing the scale was unclear in her study and we cannot attempt any further analysis of it. This approach, however, once articulated and perfected may provide invaluable assistance to meta-evaluators in the future by establishing a priority system for the assessment of evaluation studies.

3: One problem we found with several meta-evaluators' studies was that they did not provide sufficient information regarding inclusion of studies into their analyses, what kinds of programs

the studies were evaluating, or how the meta-evaluators determined whether the studies adhered to methodological criteria.

IV. METHODOLOGICAL CRITERIA

INTRODUCTION

In attempting to establish quidelines by which evaluation studies can be effectively assessed, there are two concerns that become apparent. The first concern refers to the actual method of evaluation; the research design. The options discussed earlier indicate that there is a wide discrepancy among evaluators regarding methodological considerations, however, there is general agreement in the literature that a 'true' experimental design is the preferred method. In accordance with this view, some of our criteria assess directly the degree to which evaluations comprise the elements of a true experimental design.

The second concern focuses on the actual data that is collected regarding the effectiveness of a program, and the types of inferences that can be drawn from them. The import of this aspect of evaluation should not be minimized, as this is the central question examined by impact evaluations; does the program work? It also provides the fuel for 'counter-evaluations', i.e. those wishing to challenge the effectiveness of a program. Finally, the examination of the results of the data has become a target of evaluators in recent years and has led to the conclusion by some that intervention

strategies are ineffective (Lipton, et al.,1975).

The gravity of the above considerations is not inconsequential for the continuation of programmatic activity in the field of corrections, and must therefore be stressed in any analysis.

Before we begin, however, it must be stressed that, "methodological rules and criteria are guides to action rather than laws of nature." (Hirschi and Selvin, 1973:7), and that adherence to one set of criteria may result in conflicts with other sets. Therefore, whatever the criteria decided upon, they must be realistic, i.e. achievable in the 'real world', and clear and relevant to the issue at hand, i.e. the adequacy of evaluative studies in reliably assessing the effectiveness of various intervention programs.

With this in mind we shall now turn to a discussion of the criteria for assessing the methodological adequacy of evaluations on corrective intervention of incarcerated offenders.

CRITERIA

ADEQUATE DEFINITION OF PROGRAM AND TECHNIQUES [1]

Rossi and Wright (1977) have noted that well designed research needs careful conceptual and operational specifications of the major relevant variables, identified as the goals of the program. Quay (1977) has echoed this contention, indicating that intervention integrity requires the unambiguous specification of

conceptualizations, or what exactly the intervention is composed of, in order to determine the accuracy with which the independent variable can be described and measured.

There must be an adequate definition[2] of the program and techniques being tested. Russon (1963:236) noted that in discussions of correctional programs, the words 'technique' and 'treatment' are frequently employed, however, he also notes that, "these must reflect an extensive scientific background of clinical research, study, classification, and methodology, to be meaningful." Logan (1972:378) also maintains that the definition of the program should be operational and succinct. In addition, the terms and theoretical propositions underlying the program must be clear and unambiguous. Vague notions such as 'the psychotherapeutic state' should be avoided. In addition, Russon (1963:237) asserts that vagueness and variability of meaning tend to be a result of uncertainty and the convenience of remaining uncommitted to a binding standard.

There are two elements of a definition of the intervention that should be considered. (a) The aims of the program should be clearly stated. This would include specification of the phenomenon that the program is attempting to change, or what has been referred to in the literature as 'target behaviour'. (b) Russon (1963:239) has questioned the use of the term 'treatment' and 'treatment program' in correctional settings. He argues (1963:239) that the term 'treatment':

> ...conveys the meaning of remedy, repair, or cure. Its use immediately becomes suspect if that which is to be

remedied or cured is not identified with sufficient clarity or if the word treatment appears to have a variety of meanings each suited to the transient purpose of the user.

These definitions, therefore, should also include an adequate description of the specific program and techniques of 'treatment' being used.[3]

ROUTINIZATION

The program, or techniques used should be capable of routinization. Logan (1972:379) suggested, and we agree, that there are three subcategories for this criterion: Routinization means that (a) the technique should be designed such that it can be used in any setting with different types of individuals. In other words, it must not be specific to any particular institution, environment, or persons. (b) There must be adequate information in the article to train others to do the program. References to and citations of the sources of such information in the article would be sufficient. However, vague descriptions, such as 'creating a permissive atmosphere', do not provide sufficient material by which the effectiveness of the technique can be measured, nor permit the training of other persons to carry out the "treatment" in some other setting. Indeed, this type of description could lead to the conclusion that it was some particular characteristic of the individual(s) who is doing the treatment that resulted in the success or increase of positive performance of his client, rather than the actual program. (c) There must be adequate information in the article to replicate the intervention strategy in a research sense,

using 'intervention' alone as the independent variable.

Because of the many difficulties with which evaluation researchers are faced when attempting to apply a dependable research design, this criterion becomes extremely important and valuable for the generalizability of the findings, and for correctional policy implications. Replication of intervention and concurrence of findings is one method of providing policymakers with dependable information regarding a particular program.

CONTROL GROUPS

This criterion, along with the next criterion, randomization, satisfies the basic requirements of the 'true' experimental design, as discussed in Chapter II.[4] Riecken and Boruch (1974:5) noted that:

> If an effect can be demonstrated in a group of units (persons, places, or institutions) chosen at random and subjected to a specified intervention while a similar group that is not treated does not show the effect, one can be reasonably confident that the intervention produced the effect.

Control group is defined as any equivalent group that is not receiving the specified intervention or a group that is receiving alternate intervention.[5] This concept has often been misrepresented to mean only those receiving 'no intervention', however, this narrowly defines the potential of this criterion as a feasible methodological requirement. In dealing with institutional environments especially, it is often difficult, if not impossible, to avoid some form of intervention or environmental condition aimed at altering behaviour. Therefore,

there must be some provision of a control group. If it is not feasible to set up a control group, there should at least be a comparison group.[6]

When we are dealing with 'intervention', there must be some evidence that the designated experimental group is in fact receiving the intervention and that the control group is not. The intervention may be performed on a fairly regular, or on a one shot basis, but unless there is direct evidence to the contrary, it should also be assumed that there is a possibility that the control group is receiving elements of the intervention program as well, or that it may be indirectly influenced by the intervention effect. Logan (1972:379) noted that this criterion is rarely considered in evaluations. It is, however, extremely important in institutional settings, where subjects are often in close proximity to each other.

Therefore, since the purpose of studies is to assess intervention effectiveness, there must be direct indications that it is, in fact, the 'intervention' that is causing the observed change and not some other extraneous variable.

RANDOMIZATION

Randomization is perhaps one of the most crucial requirements for both the external and internal validity of any study (Boruch et al., 1978). The use of randomization has been recognized as one method of obtaining unbiased estimates of the effect of a program. Boruch et al. (1978:657) indicate that:

> field experiments are usually designed to estimate size of program effects in natural (or nearly natural)

> settings. The practical as well as statistical significance of differences in multiple outcomes-costs, benefits, negative effects-are important. Unbiased estimation of the program effects in the setting at hand is usually more important than generalization; an important secondary objective is understanding generalizability of results to settings which vary in similarity to the one at hand.

At times the purpose of the study is to examine individuals on particular characteristics and to compare the intervention-no-intervention groups on those characteristics. At other times it may not be feasible to randomly assign individuals or groups to experimental and control situations. In these situations it is not advisable to attempt to randomize the placement of subjects into groups, but rather to match them on the specific traits or characteristics that have been targeted for the program.[7] As Conner (1977:241) noted:

> Although it is the best procedure available to produce similar groups, randomization does not guarantee equivalency between groups. This is particularly true for small groups, where it may be necessary to first block clients in important characteristics, then randomly assign similar kinds of clients to both the intervention and control groups.

Due to the requirements of specific programs, and the difficulties[8] encountered in randomly assigning individuals to control or experimental groups, studies should be analyzed for both matching or randomization of subjects into groups.

While we may match subjects so they are comparable to each other on certain traits, the placement of each matched pair must be decided upon by some randomization technique. As noted

earlier, this criterion has been recognized in most textbooks on research methods as a basic requirement of a true experimental design (Selltiz, Wrightsman & Cook, 1976;137 Kerlinger, 1973:123; Hirschi & Selvin, 1973:40). In addition to this requirement, randomization assures group equivalency before intervention, which can be useful in determining causation. It is useful in solving other problems related to equivalency of groups, such as favouratism in the selection of individuals for groups, and other biases resulting from using the most needy or making choices on a first-come-first-serve basis (Conner, 1977).

Another of the possible consequences of experimental research affecting both the internal and external validity of the study is the possibility of alternative explanations (Kerlinger, 1973:388-390) or spuriousness (Hirschi & Selvin, 1973:73-89; Selltiz et al., 1976:490-495). Randomization would provide some control against such an occurrence. Hirschi and Selvin (1973:40) note that:

> It is always possible, however, that the process of randomization does not completely remove the association between the extraneous variable and the independent variable... with the techniques of statistical inference, it is possible to calculate the probability of such an occurrence.

BEFORE AND AFTER MEASUREMENTS

Measurements[9] of the behaviour to be "treated" should be taken before and after the intervention has been introduced. This should be carried out: (a) to check for the 'randomness' of the sample, i.e. to be sure that the individuals chosen for the

study are as similar as possible on specific characteristics; and (b) to provide a baseline upon which we can compare post-intervention data.

Before-after measurements should be taken for both experimental and control groups. The comparisons then can be made of the two measures for each group and, if possible, between the two groups on each measurement.

SUCCESS AND FAILURE

Studies assessing the effectiveness of an intervention program must include operational definitions of success and failure. Although many studies say that the interventions are successful,[10] the reader is often unaware of what the success relates to. Success and failure can be, and should be, examined in at least two ways. First, the target behaviour outcomes should be stated and measured, i.e. those behaviours which are specific to the particular program. If, for example, the object of the program was to change the inmates' attitudes towards institutional rules, then success must be defined in those terms.

The second way to view definitions of success and failure is in terms of 'conventional outcomes' or criminality (Logan, 1972:379). It is important that these definitions refer to criminal behaviour rather than personal adjustment alone as it must be assumed that one purpose of correctional intervention programs is to prepare the individual to return to society.[11]

FOLLOW-UP

The primary purpose of including follow-up data in an evaluation is to determine if there was an effect, and whether it was able to continue over time or was only short-term (Gibbons, 1976:321).[12]

Wholey, et al. (1970:96) have noted that the relationship between short-term and long-term objectives is often unknown, however, an important issue that should be examined when assessing program impact is the duration of the effects. They state that not all program evaluations require follow-ups of the same length, but, they suggest that anything less than one or two years is simply insufficient.

Evaluators have lamented the fact that policy-makers or administrators often require information regarding results of a study before a sufficient amount of time has passed to adequately test the effectiveness of the program. Gordon and Morse (1975:342) raise several issues that must be entertained when considering follow-up data. The first concerns 'lag time', or the amount of time that must pass before measurable effects can be expected. The second issue concerns the determination of effects over time. We must be aware of and question how long changes in behaviour can be expected to last.[13] Wholey, et al. (1970:97) have suggested that follow-up measures continue to be taken even after the initial findings have been delivered to administrators. They cite an example of the importance of continuing follow-up: A five-year follow-up of a MDTA (Manpower

Development and Training) program suggested that the MDTA had greater effects than were apparent after a one or two year follow-up.

Lengthy follow-ups, however, are also beset with problems. McCord (1978) conducted a thirty-year follow-up assessment of the Cambridge Somerville Youth Project. Her assessment included a comparison of intervention and control groups, using official records and personal contacts, to obtain information on the long-term effects of the project on marriage, children, occupations, drinking, health, attitudes, and how the experimental group felt the intervention had helped them.

She found that although the program seemed successful in obtaining the short-term objectives, which was to establish a confidential relationship between the social workers and teenagers, none of the measures indicated that the longterm goal, the general improvement of the lives of the intervention group, was attained. In fact, as she suggested, the results presented a disturbing picture. Not only did the intervention fail to prevent clients from committing crimes, but it may have produced negative side effects as well. She concluded (1978:289) that the message was quite clear from her follow-up data: "Intervention programs risk damaging the individuals they are designed to assist."[14]

Sobel (1978:290) criticized McCord's conclusions, however, she also recognized (1978:291) that while McCord's speculations may be inaccurate, her study was important because it

demonstrated the possibility of conducting a carefully designed intervention with constant evaluation and a longitudinal follow-up.

In addition to Sobel's (1978:290-291) observations, another question raised by McCord's (1978) study is whether we can realistically expect intervention to have such a long-term effect? There are several questions[15] that remain to be answered about the appropriate length of follow-up, however, that does not negate the importance of this criterion in a methodological assessment.

Follow-up data should ideally be gathered once the individual has left the institutional environment as well as simply post-intervention (Logan, 1972). Quinsey (1973:351), discussing intervention with child molesters, observed that: "it's difficult to know when the goals of therapy have been achieved in an institutional setting because the target behaviours cannot occur."

UNBIASED OBSERVERS

The final criterion requires that design and measurement throughout the implementation of intervention, and in the follow-up, be taken by unbiased observers. This criterion involves two necessary conditions: the first concerns who designs the evaluation of the program; the second focuses on who takes the measurement.

In discussing the two elements of this criterion we will first examine the evaluator's relationship with program

administration and staff. There has been considerable attention given to this issue in the literature. Debate has arisen as to whether it is more advantageous to have evaluators situated within (in-house) or outside (independent) the actual program (Riecken, 1972:99; Szabo & Rizkalla, 1978:22-23; Suchman, 1972:78-79; Riecken & Boruch, 1974:35; Conner, 1977:223-226).

In-house evaluations have several advantages over evaluations conducted by independent researchers. It is contended that much of the resistence to the research by program staff is minimized, as the researchers would have greater familiarity with subject matter (Riecken, 1972:99), a closer acquaintance with the subjects of the program (Riecken, 1972:99), a more detailed knowledge of the organization and all of its programs (Szabo & Rizkalla, 1978:22), as well as of the demands placed on administrators (Conner, 1977:224). In addition, valuable time is saved by in-house evaluations by not needing to acquaint the researcher with the program and staff.

The result of the above advantages should allow for easier access to data (Szabo & Rizkalla, 1978:22) and greater cooperation from staff (Suchman, 1972:79). In addition, the evaluator is in a better position to help interpret findings and detect unanticipated results of the evaluation (Riecken, 1972:99) and subtle changes in the program (Suchman, 1972:79).

On the other hand, it is argued that the greatest advantage of independent evaluation is the ability of the researcher to maintain objectivity (Szabo & Rizkalla, 1978:22; Riecken,

1972:99; Suchman, 1972:79; Riecken & Boruch, 1974:35). Objectivity is maintained because the evaluator is not identified with the program and therefore receives less pressure from colleagues and interest groups (Riecken, 1972:99). He maintains greater freedom of movement and ambiguity of status (Riecken, 1972:99; Szabo & Rizkalla, 1978:22).

In getting acquainted with the program, the independent evaluator may be in a better position to recognize additional research ideas and alternatives that those close to the program may not see (Riecken, 1972). In addition, independence may allow the researcher to include evaluative criteria that may question some of the organizational premises (Szabo & Rizkalla, 1978).

Finally, it is believed that the independence of evaluators creates less resistence to the findings of the study, as there is a lesser degree of committment to the program itself (Riecken, 1972). As well, where internal conflict arises, the independent evaluator may be able to act as a mediator (Szabo & Rizkalla, 1978), thus assisting in the maintenance of a smooth-running and consistent program.

Bernstein (1978:32-42) was interested in determining the influence of affiliation of the researcher on conformity to methodological norms. She examined (1978:32-38) the types of organizations with which researchers were associated, i.e. non-profit, profit or university. Bernstein (1978:37) developed two models of researcher affiliation. The first model, she termed 'academic'. Researchers in this model received grants

from research-oriented agencies. They tended to come from educational institutions and defined their audience as 'cosmopolitan', and worked as 'insiders', i.e. decisions were made together with the program staff. The second model was referred to as 'entrepreneurial'. Research in this model received funding through contracts from service-oriented agencies, they identified their audience locally, and they worked 'outside' the program, making research decisions independent of program administration. She concluded that, "the academic model is more conducive to the likelihood that there will be adherence to technical norms". Gordon and Morse (1975:348-349) also examined the issue of affiliated versus non-affiliated researchers with a program. They found that in general, researchers who were affiliated with the program tended to produce less rigorous studies with more positive results while non-affiliated researchers produced more rigorous studies with more negative results. They suggested (1975:348) that, "The pattern of findings, we think to some extent, can be explained by unconscious bias and experimenter effect". Gordon and Morse (1975:349) suggest, however, that the number of cases they examined was insufficient to fully support their conclusion and, in general, this issue has not received adequate attention in the literature to warrant any conclusive interpretation, but requires further investigation.

The second component of this criterion focuses on both who takes the measurements, and who provides information regarding

the various indices of the intervention program. Unbiased observers in this context would include not only analysts with no interest in the research problem, but would also exclude people involved in the program by virtue of the nature of their jobs.

'Blind analysis' means that those taking the measurements should have no knowledge of the nature of the research project (i.e. the aims of the intervention) nor of the division of the subjects into experimental and control groups. Although blind analysis has not received an abundance of interest in the literature, the value of this procedure has been recognized (Suchman, 1972:65). The use of a double-blind procedure, where both the experimenter or data gatherers and subject are unaware of the intervention conditions, has also received some attention (See: Wortman, 1975:571; Gendreau & Ross, 1979:477).

One of the goals of evaluation research is to provide reliable information about the impact of a program. Conformity to these goals, i.e. objectivity and greater reliability of research findings, is probably facilitated when evaluators are independent of the organization and program.

PROBLEMS OF CATEGORIZATION

Before continuing with the analysis of evaluation studies in Canada, it would be apropos to comment first on some general problems that face us when we attempt to categorize the studies according to their adherence to a specific set of criteria.

One problem Slaikeu (1973:95) has recognized in evaluating research done in penal environments is that because intervention does not take place in a vacuum, we should look for indications of the effects of the penal environment on intervention, such as consideration of the 'intervention-custody conflict', or group intervention in prison as opposed to group intervention outside the prison. Reppucci and Clingempeel (1978) also address this issue as one problem that confronts psychologists who do correctional research.[16] They have noted (1978:732) the "empirical neglect of situational factors as they affect an individuals' behaviour". This has led, they contend, to a lack of attention paid to the interaction effects of environment and situational factors in research designs, which ultimately leads to limited predictive power and generalizability of results.

A second problem that arises in analyzing published studies is determining how much weight to put on claims being made, i.e. to say that one randomly assigns individuals to groups does not guarantee that the method of randomization was appropriate - indeed that they were actually randomized at all. As Conner (1977:200) noted:

> discussions with project staff were essential to obtain the necessary information about randomization: project reports very rarely present anything but brief, superficial discussions of the process.

Quay (1977:342) refers to this as 'program integrity'. He indicates that, "we need to be as equally concerned with the 'what' of evaluation as with the 'how'". He also suggests that before we can make any kind of conclusive statement about

whether a correctional intervention has not worked there is much we need to know that exceeds the experimental design and outcome criteria. Program integrity includes conceptualization of the intervention or a determination of what exactly the 'intervention' is composed of, which is determined by the accuracy with which we are able to describe and measure the independent variable. This includes a determination of whether what is happening to the individuals in the program actually meets the specifications of the intervention. It also involves an examination of the expertise of those who are delivering the service. It is not uncommon for 'intervention' studies in corrections to omit information regarding program integrity (Quay, 1977:342-346). In conclusion, he argues that without due attention to the integrity of the intervention, critical errors will be made with serious policy and practical consequences, most notably the recommendation of program discontinuation based on an insufficient demonstration of progressive intervention effects (Quay, 1977:353).

NOTES

1: The methodological criteria are discussed in general terms in Chapter IV. For a breakdown of criteria into units of analysis, see Legend for Table IV in Chapter V.

2:In discussing the improper use of technical terminology, Russon (1963) focused on several terms which have been consistently misused. The word 'definition' was one term he identified; as he says (1963:238):

> Definition is frequently confused with the words 'description' and 'discussion', that is, when instructed to define, many people tend to describe or discuss, so that while what is said is quite true and correct in itself, it does not actually define; nevertheless having been used, it is assumed that a definition has been given.

3: We have avoided the use of the word 'treatment' in this thesis because of the varying types of programs being assessed which stretch the definition. In place of treatment, intervention strategy has been employed as this term indicates only the interference into the lives of institutional subjects, and does not necessarily imply 'cure'.

4: See Conner (1977), Rossi (1972), Powers and Alderman (1979) and Freeman and Sherwood (1970) for a discussion of techniques that can be employed to facilitate the use of control groups in evaluation research. Powers and Alderman (1979:89), for example, maintain that the constraints that have been commonly accepted as obstacles to implementing true experimental designs can be made to work in the researcher's favour.

5: Control and control groups have been used interchangeably throughout this paper. There are several meanings that could be applied to the term 'control', such as statistical control, or experimenter control, however, the term 'control' throughout this paper refers only to control groups.

6: The use of comparison groups to replace control groups has been regarded as an acceptable alternative (Powers & Alderman, 1979; Riecken & Boruch, 1974; Cook & Campbell, 1976), however, the dependability of results using comparison groups has also been questioned (See Rossi, 1972:233).

7: Tufte (1974:22) has indicated that matching helps provide information to the reader on what is going on with the data. In addition, it helps control for extraneous variables that could possibly affect the outcome of the intervention. He notes, however, (1974:22-24) that there are several limitations to the use of a matching procedure.

First, contrary to Conner's (1977:241) assertion, Tufte (1974:22) contends that in complex situations it is difficult to do a good job of matching without a large number of cases. This results in a high degree of inaccuracy. As we attempt to control for more variables, we end up with more combinations. This results in a level of complexity that makes it difficult for the reader to understand.

8: Freeman and Sherwood (1970:104-106) provide a discussion of the difficulties of randomization. They noted that it is difficult to implement in most cases because of its impracticality. Ethical responses to the denial of intervention also raise questions regarding the possibility of employing this technique. Finally, administrators believe that they have an overriding responsibility when it comes to assigning cases to either intervention or control. Conner (1977:221) also recognized the problems that arise when individuals other than the researcher, which includes administrators as well as program staff, have control over the assignment of individuals to groups.

9: Measurements should include both the dependent and independent variables. The dependent variable, or target behaviour, should be measured in order to determine if and how much change occurred during the course of the intervention and upon its completion. The independent variable, or intervention

strategy, should be measured to ensure consistency of the program, i.e. that the components of the intervention have not altered over the course of implementation.

10: It is possible for a study to be considered successful, or implied successful, although the actual intervention may have failed, depending on the goals of the program as indicated by administrators or program directors. Ambiguity of goal definition has been noted earlier in this paper, and recognized as an obstacle to empirical assessment of program effectiveness and utilization of evaluation results.

11: The question of what constitutes an adequate measure of success with correctional subjects has centered on the concept of recidivism. Fisher and Erickson (1973:181) maintain that the measurement problem related to recidivism is one of three major limitations in evaluation studies that prevents them from demonstrating a genuine intervention effect. One of the principle sources of error is the use of cumulative arrest rates. This creates two problems according to Fisher and Erickson; the first problem results from the inability to determine the distribution of arrests, as a cumulative record often does not indicate when they occurred. Second, cumulative rates tend to diminish the importance of recidivism rates, with the more intervals that are included. Fisher and Erickson suggest that perhaps arrest data should be measured

independently for different time intervals.

12: Gibbons (1976) has remarked that social programming should be cumulative and add to the existing knowledge. Follow-up data not only provides information regarding the effectiveness of a program, but also extends our knowledge about that program over time, and can provide insight into side effects, areas of program change, and adequacy of intervention techniques.

13: This becomes especially important when dealing with intervention methods such as behaviour modification and psychotherapy, where the duration of change and the extent to which change is reversible are questioned (Friedman, 1975:741).

14: For further information of McCord's findings, see McCord, 1978, pages 284-289.

15: Future research on the length of follow-up should include an examination of a) how long persistence of 'treated' behaviour can realistically be expected to last, b) what are the effects of the institutional environment on behaviour within and outside the institution, and c) what external factors and pressures influence the individuals' performance in the community, and how much of an impact these pressures exert.

16: Reppucci and Clingempeel (1978:727) recognize that the

problems they discuss, specifically the influence of the environment on research, are not unique to correctional populations, but rather, occur whenever 'clinical' populations are studied, i.e. any group subjected to institutional 'intervention', including mental patients as well as inmates of correctional institutions.

V. METHODOLOGICAL ASSESSMENT

INTRODUCTION

The central issues concerning evaluation research and critical reviews of previous evaluation attempts have been discussed in this thesis. It was determined that methodological issues merit considerable attention and concern, as the quality of research design is directly related to the usefulness of the research findings.

Hackler (1978:23) argues that rigorous evaluation is a potentially destructive endeavour. He notes that various researchers in the United States have attempted to establish evaluation as an integral part of program activity. However, when critical reviews of their methods have been conducted the results have been disheartening to those who search for effective intervention strategies. Therefore, as Hackler (1978:28) contends, we should learn from the mistakes made in the United States, and cease this fruitless endeavour in Canada.

This thesis was prompted in part by Hackler's (1978:27-28) assertion that researchers in Canada can do no better than their American counterparts. We will therefore turn to a methodological assessment of Canadian evaluations of institutional intervention programs.

All studies in this assessment were taken from relevant Canadian Journals. We examined all the journals and any study which assessed the impact/effectiveness of an intervention program conducted in a provincial, federal or juvenile institution was assessed. The studies assessed in this evaluation spanned twenty years; from 1960 to 1980 (1960:2; 1961:3; 1963:1; 1964:2; 1966:1; 1967:1; 1968:1; 1971:1; 1972:2; 1973:1; 1974:1; 1975:1; 1976:1; 1977:1; 1978:1; 1979:1; 1980:2). The majority (19 studies) were published in the Canadian Journal of Criminology. Two studies were published in the Canadian Psychologist, one in Crime and/et Justice, and one in the Canadian Journal of Behavioural Science. As we mentioned earlier, only published evaluations were assessed in this study, as it is assumed that, due to peer review for journal publication, the studies found in journals would comprise a sample of the "most" methodologically rigorous designs available. A content analysis was conducted of the 23 studies included in this analysis to determine information regarding: 1) the date of the evaluation; 2) authors; 3) the problem stated, i.e. the aims of the program; 4) the principle methods used; 5) research sites; 6) research data; 7) independent variables; 8) dependent variables; 9) control groups; 10) follow-up; and 11) research findings. (See Appendix I for a description of each category.)

In addition, an examination of the bibliographies of the 23 studies was conducted to determine whether researchers cited

and/or utilized information regarding research methods and critical approaches to evaluation research (i.e. meta-evaluation research).

Information from the content analysis was then used to determine:

a. The degree to which each study met the assessment criteria set out in Chapter IV;
b. Whether improvements in research design have become apparent over the last 20 years; and
c. Whether certain Journals in Canada maintain stricter controls over the quality of research published than other Canadian journals.

CONTENT ANALYSIS

The primary information extracted from the 23 studies to aid us in our assessment of methodological rigor can be readily seen in Table II. The data was broken down into five categories. Definition of Treatment was obtained by examining the problem, research data, and independent variables. Methods of evaluation were obtained from an examination of the principle methods used, including measures as well as premeasures of independent and dependent variables. Population and research sites are self explanatory, however, target behaviour was identified through an examination of the dependent variables and research findings.

Table II

Content Analysis of Canadian Evaluation Studies by Author

		Type of Treatment	Methods of Evaluation	Population	Sites	Target Behaviour
Flint	1960	Milieu therapy, group therapy individual therapy	observation interviews	'girls' 18-40	reformatory	self-esteem anti-social att.
Frechette	1960	Group psycho-therapy	partic.observ. recidivism rates	adult males juvenile males	penitentiary	verbalize conflicts 'rehab'
Hartman	1961	Group psycho-therapy (on paedophile)	dir. observ. clinical assess.	adult males (20-35)	psychiatric hospital (Tor.)	self-awareness introspection
Turner	1961	group psycho-theray (on sex deviations)	observation	adult males	psychiatric hospital (Toronto)	modification of underlying psychopathology
Philip	1961	Brief group psychotherapy	projective tech. (TAT) stats. tests	boys (14-16)	training school	hostility, guilt, insecurity, bland-ness, dependence
Campbell	1963	group counselling	partic. obser-vation	adult males and their wives	John Howard Soc. office (using Haney Correct-ional Instit. inmates)	recognition of individual strivings in relation to authority, peers, and selves marital adjustment adjust. to community after release
Achille	1964	therapeutic milieu re-education	observation	boys (17-20)	Boscoville Insitution	verbalization recog. and expression ofinternal problems and feelings

Table II (continued)

	Type of Treatment	Methods of Evaluation	Population	Sites	Target Behaviour
Coutts 1964	group work approach includ. group counselling and good conduct ratings	pay ratings observation	adults females	Oakalla Prison Farm	rule infractions perform. in work assignments and planned group activities relationships with others and staff.
Klonoff 1966	Sodium amytal (with sexual deviants)	psychometric tests, clinical assessments	adult males	B.C. Penitentiary	disinhibition
Landreville 1967	re-education	official records (police, instns.)	boys ($\bar{x}$16)	Boscoville Institution	recidivism
Grygier, et al. 1968	informal, small-unit residential atmosphere	sociometric tests interviews	boys (under 12)	White Oaks Village	interaction with each other and staff
Parlett & Ayers 1971	programmed instruction	psychometric tests	adult males	William Head and Matsqui Institutions.	'socialization' recidivism
Coons 1972	psychotherapy stressing inter-personal inter-action	psychometric test	adults males and females	psychiatric hospital	improvements of post-therapy 'protocol'
Coons 1972	interpersonal interaction and formal group therapy	psychometric tests	adults males and females	psychiatric hospital	patients 'ajustment' and group cohesiveness
Ross & Doody 1973	behaviour modification using continuous, partial and intermittent punishments	psychopathic Deviate Scale psychometric tests	girls (12-17	Grand View School	'correct responses'

Table II (continued)

	Type of Treatment	Methods of Evaluation	Population	Sites	Target Behaviour
Andrews & Young 1974	Short-term structural multiple grp. approach	psychometric tests observation	males (16-21)	Provincial Min. Security Inst.	prison adjustment attitudes, behav. ratings, misconduct reports
Quinsey & Sarbit 1975	Behaviour modification token economy	psychometric tests	adult males	max. security psychiatric hospital	point earning behaviour
Quinsey et al. 1976	aversion therapy (shock therapy)	apparatus	adult males	mental health centre	skin conductance penile circum.
Reker & Meissner 1977	Life skills program	psychometric tests	adult males	Federal Penitentiary	self-worth, positive attitudes towards life, change in personality traits
Davidson et al. 1978	regular correctional educ. prog.	Peabody Indiv. Achievement Test	males($\bar{x}$17.5)	adult training centre	rate of progression in terms of grade equivalents
Annis 1979	intensive group therapy	interviews official records psychometric tests self-reports	adult males	Mental Health Centre	personality and behaviour change prison adjustment post-release adj. recidivism.
Daigle-Zinn & Andrews 1980	role-playing and didactic discuss. approaches	behaviour rating scales self-report scales psychometric tests	young adult males	minimum security institution	attitudes towards self and others self-esteem interpersonal adjust.
Gendreau et al. 1980	inmate volunteer program	attitudinal scales work rating sheets	adult males	med. sec.inmates at Rideau Regional Centre & Brockville Pysch. Hospital	attitude change hospital staff ratings success on job staff ratings

One of the aims of this chapter was to determine whether certain Canadian journals maintain stricter controls over the quality of research being published than other Canadian journals, however, there is not enough variance between the journals, due to the large number of studies found in the Canadian Journal of Criminology, to make any definitive conclusion regarding this question.

The number of references made to research methods[1] and evaluation research[2] was also examined. See Table III for inclusion of relevant references in the bibliographies of the Canadian evaluation studies assessed in this thesis. Three observations can be made from Table III. The first and most obvious one is that very few studies made reference to the relevant literature; four studies mentioned one reference; one study indicated between two and five references; and only one study employed more than five references. All of these studies, with the exception of one, were conducted after 1972.

Table III

Reference to Methodology (Research Methods)

Literature in Bibliographies of 23 Canadian Studies

Author	Year	No Reference	No Reference Applicable	1 Reference	2 - 5 References	Over 5 References
Frechette	1960	*				
Flint	1960	*				
Philip	1961	*				
Turner	1961	*				
Hartman	1961		*			
Campbell	1963	*				
Coutts	1964	*				
Achille	1964		*			
Klonoff	1966		*			
Landreville	1967				*	
Grygier, et al.	1968		*			
Parlett, et al.	1971		*			
Coons	1972		*			
Ross, et al.	1973			*		
Andrews, et al.	1974		*			
Quinsey, et al.	1975		*			
Quinsey, et al.	1976			*		
Reker, et al.	1977		*			
Davidson, et al.	1978			*		
Annis	1979					*
Daigle-Zinn	1980		*			
Gendreau, et al.	1980			*		
TOTAL		6	10	4	1	1

The second and third observations concerning references to the relevant literature on evaluation research and/or research methods, are closely connected. First, six (26%) of the 23 studies did not include bibliographies. It should be noted that all six studies were conducted prior to 1965, and it is possible that the exclusion of bibliographies could be due to early publication standards maintained by journals. Ten studies (43%) contained bibliographies, but made no reference to the relevant literature, i.e. research methods and evaluation research literature. Finally, it is apparent from table III that the trend towards employing references, specifically references to research methods, is moving in a positive direction from the 1960's to the 1980's. This may be due, in part, to increased expectation of referencing demanded by journals, and, in part, to an increased awareness of the need for more rigorous research designs.

The content analysis indicated that a wide variety of intervention programs are researched in Canadian institutions. Group psychotherapy appears to be a predominant mode of intervention researched in Canada. Five studies (21%) assessed the effects of group psychotherapy, three (13%) with adult males, one (4%) with adult males and females, and one (4%) with juvenile males. Group therapy was conducted in three (13%) studies: adult and juvenile 'girls'; adult males; and adult males and females. Individual therapy was also evaluated in one (4%) study, with adult and juvenile females. Educational

programs (four (17%) studies: three (13%) with juvenile males; and one (4%) with adult males) formed the next largest group of evaluations. Behaviour modification programs were also of interest to evaluators; two (9%) studies evaluated aversion therapy with adult male sex offenders, and two (9%) studies evaluated operant conditioning techniques. Group counselling (two (9%) studies; one (4%) with adult females and one (4%) with adult males and females) were also evaluated. One (4%) study assessed an institutional program with young juvenile males which provided an informal, homelike atmosphere; one (4%) study evaluated a volunteer program conducted in the community with adult males; one (4%) study dealt with role-playing, and didactic discussion with young male adults was also evaluated; one (4%) study was concerned with a multi-group approach with young male adults; and one (4%) life skills program with adult males was also assessed. There is a greater number of intervention techniques than studies evaluated, as some studies employed multiple techniques or compared a variety of programs.

Measurements taken of the intervention program relied primarily on psychometric tests (nine studies-39%).[3] Observation, either direct or participant, also provided a large number of assessments (eight studies-35%). Interviews were conducted in three (13%) of the studies, and "clinical assessments" counts for two (9%). Official records, that is, police files and/or prison records were employed in two (9%) studies, self-reports in two (9%) and recidivism rates were

identified in only one (4%) study. Sociometric tests provided measures in one (4%) study and projective techniques were used in one (4%). Other types of tests such as behaviour rating scales or attitudinal scales were employed in three (13%) studies. Pay or work ratings were used in two (9%) studies, and 'apparatus' was employed in a study evaluating shock therapy (4%). More than one technique for measuring the effectiveness of the intervention was employed in eleven (50%) of the studies.

Finally, target behaviours were primarily identified in reference to non-criminal behaviours. Four (17%) studies identified individual adjustment as the desired behaviour change, three (13%) aimed at institutional adjustment, four (17%) at community adjustment (or socialization, but not related to criminal behaviour) and one (4%) examined marital adjustment. Ten (43%) studies focused on specific personality characteristics: three (13%) attempted to alter self-esteem, one (4%) focused on self-awareness; four (17%) were concerned with anti-social attitudes; and two (9%) aimed at improving verbalization of inner problems and feelings. Ten (43%) studies dealt with interpersonal relationships: four (17%) focused on relationships with peers; three (13%) examined relationships with authority figures; and three (13%) assessed relationship with self. Education or academic improvement was assessed in one (4%) study, 'protocol'[4] was assessed in another (4%), and physiological and psychological changes were evaluated in two (9%) studies employing aversive therapy techniques.[5] Only three

(13%) studies assessed recidivism.

The information amassed here will now be utilized to assess the degree to which each study meets the eight assessment criteria set out in Chapter IV.

GENERAL EVALUATION

The findings of this study can readily be seen in Table IV The table sets out the methodological criteria and demonstrates the degree to which evaluation studies published in Canadian Journals meet these requirements. See the Legend to Table IV for a brief description of each methodological criterion. Each criterion has been classified as fully met (indicated by by an asterisk), meaning that the study gave evidence of the specific requirement, or partially met (indicated by a plus sign), meaning that the specific requirement was indicated but could not be verified due to a lack of information in the article. If a space was left blank the criterion was either not employed in the study or was unknown. In order to provide evidence of accordance, either a measurable indicator is present, or, as in most cases, we must rely on what the evaluator tells us. Observations were made regarding whether the criteria were met in any way, either fully or partially. Then the degree to which they met the criteria was examined. The following observations of the 23 studies can be drawn from Table IV:

Table IV

Analysis of 23 Studies on the Effectiveness of Intervention Programs in Canadian Correctional Institutions

		I		II			III		IV		V	VI		VII	VIII		No. of Criteria met by each study			
		IA	IB	IIA	IIB	IIC	IIIA	IIIB	IVA	IVB		V1A	VIB	VII	VIIIA	VIIIB	Total (fully met)	Total (partially met)	Total	Outcome
Frechette 1960	N=74	*	+	+		*							+	+			2	4	(6)	Partial -/+
Flint 1960	@ N=10	+	*	+	+	+											1	4	(5)	Positive +
Philip 1961	N=86	+			+		*	+		*	*		+		+		3	5	(8)	Positive +
Turner 1961	?											+		+			0	2	(2)	Partial -/+
Hartman 1961	?	+	+								+						0	3	(3)	Positive +
Campbell 1962	N=8 N=10	*	+	+									on-going				1	2	(3)	Positive +
Achille 1964	N=15	+	+	+								+			+		0	5	(5)	Partial +/-
Coutts 1964	@ N=10	+	+	+		+						+					0	5	(5)	Positive +
Klonoff 1966	N=13	*	+		*	*					+						3	2	(5)	Partial +/-
Landreville 1967	N=214	*	+						+		+		+	*	+		2	5	(7)	Partial +/-

Table IV (continued)

	N=	I IA	IB	II IIA	IIB	IIC	III IIIA	IIIB	IV IVA	IVB	V	VI V1A	VIB	VII	VIII VIIIA	VIIIB	No. of Criteria met by each study Total (fully met)	Total (partially met)	Total	Outcome
Grygier, Guarino, Nease, and Sakowice 1968	N=40	+	+	+								+					0	4	(4)	Positive +
Parlett and Ayers 1971	N=65	+		*			*	+		*	*	+					4	3	(7)	Positive +
Coons I 1972	N=66		+	+			*		+		*	+				*	3	4	(7)	Positive +
Coons II 1972	N=56	+	+	*	*	*	+	+	+		+	+					3	7	(10)	Partial -/+
Ross and Doody 1973	N=36	*	*	*	*	*	+	+	+		+	+				+	5	6	(11)	Positive +
Andrews and Young 1974	N=47	*	+	*	*	+	*	+	+			+		+		*	5	6	(11)	Partial +/-
Quinsey and Sarbit 1975	N=12	+	*	+	+	+					*	+					2	5	(7)	Partial +/-
Quinsey, Bergersen and Steinman 1976	N=10	*	*	+	*	*					*						5	1	(6)	Positive +
Reker and Meissner 1977	N=48 attrition N=38	*	*		+	*	*	+	+		*	+			+		5	5	(10)	Positive +
Davidson, Willis and Cole 1978	N=36		*	*	+	*	+				*	*					5	2	(7)	Partial +/-
Annis 1979	N=150	+	+			+	*	+	+		*	+	+	*			3	7	(10)	Negative -
Daigle-Zinn and Andrews 1980	N=43 Original (loss of cases)	*	*	+	*	*	*		+		*	*			+		7	3	(10)	Partial +/-

Table IV (continued)

Name and Year		N=	I IA	I IB	II IIA	II IIB	II IIC	III IIIA	III IIIB	IV IVA	IV IVB	V	VI V1A	VI VIB	VII	VIII VIIIA	VIII VIIIB	No. of Criteria met by each study Total (fully met)	Total (partially met)	Total	Outcome
Gendreau, Burke and Grant	1980	N=19 Both Groups	+		+		+	*	*			+	*			+		3	5	(8)	Positive +
Total (fully met)		Number of Stud. meeting each criterion	9	7	5	6	8	8	1	0	2	9	3	0	2	0	2				
Total (partially met)			11	12	11	5	6	3	7	8	0	6	12	4	3	6	1				
Total			(20)	(19)	(16)	(11)	(14)	(11)	(8)	(8)	(2)	(15	(15)	(4)	(5)	(6)	(3)				

Legend for Methodological Assessment
(for Tables IV & VI)

Methodological Criteria

1. Adequate Definition of Program

(A) Operational definitions of the terms, theoretical propositions and aims of the program, including specification of the phenomenon the program is attempting to change (target behaviour).

(B) Adequate description of the specific elements of the program and techniques employed.

2. Routinization

(A) The technique must be designed such that it can be used in any other institution with different types of individuals.

(B) There must be adequate information in the article to train others to do the program.

(C) There must be adequate information in the article to replicate the therapy in a research sense, using 'intervention' alone as the independent variable.

3. Control or Comparison

(A) There must be some provision of a control or comparison group, either not receiving the intervention, or receiving an alternative intervention.

(B) There must be some evidence that the treatment group is in fact receiving the 'intervention' and that the control group is not.

4. Randomization

There must be some evidence that individuals are drawn at random from the institutinal population and are then:

(A) randomly assigned to the treatment and control groups; or

(B) where it is not feasible to randomly assign subjects, to match individuals in both groups on pertinent characteristics.

Legend (continued)

5. Before and After Measures

Measures should be taken of the target behaviour prior to commencement of the intervention to provide a baseline upon which comparisons can be made with the post intervention data, for both experimental and control groups.

6. Operational Definitions of Success and Failure

Expectations of measurable outcomes should be clearly stated for:

(A) target behaviours, i.e. those behaviours specific to the particular program; and

(B) 'conventional' outcomes, i.e. criminality.

7. Follow-up in the Community

Measurements should be taken of the individual's behaviour upon release from the institution.

8. Provision of Unbiased Observers

(A) Independent researchers, i.e. those not associated with the institution or the treatment program, should conduct the evaluation of treatment effectiveness.

(B) measurement throughout the study and in the follow-up should be taken by 'unbiased observers', including analysts who have no interest in the program, and excluding individuals involved in the program by virtue of the nature of their jobs.

Legend (continued)

Program Outcomes*

1. Positive: (+)

   Subjects 'cured' or 'improved'
   Recommendation of future Research
   Recommendation of continuance of program

2. Negative: (-)

   Failure of program or specific techniques.
   Recommendation of other alternatives

2. Partial: Both positive and negative results
   (∓) Further research before any conclusions drawn.
   (± Despite partial failure program should be continued.

*based on Authors' conclusions

1. At first glance it would appear that a good number of the studies attempted to proceed under some form of experimental design. Thirteen studies (57%) appeared to have met 7 or more of the criteria - to some degree. Seven (30%) studies met between 4 and 6 (approximately one-third) of the criteria. Three (13%) studies were seriously inadequate.

2. However, on closer examination it can be seen that of the thirteen studies that met half (7) of the criteria, only one (7%) study fully satisfied the 7. Six (46%) studies fully satisfied between 4 and 6 criteria. When the studies were examined for "partial satisfaction", sixteen (69%) studies were left grossly inadequate.

3. One of the objectives of this study is to determine whether or not there have been improvements in research methods over time. It is quite apparent, upon examining Table IV, that the criteria are better represented by a greater number of the latter studies. From 1960 to 1970, two studies out of eleven (18%) met at least 7 of the criteria in some way (either partially or fully), and only 2 criteria from each of these two studies were fully met or satisfied. On the other hand, eleven studies out of twelve (92%) conducted between 1971 and 1980 met over 7 criteria (either partially or fully). At least 3 criteria were fully met by four (33%) studies, and as many as 7 criteria were met by one (8%) study. The rest fell somewhere between with the exception of one study (Quinsey and Sarbit, 1975), which met only 2 criteria.

4. It is apparent from our review of the evaluation literature that there are certain criteria that are more clearly recognized as essential components of evaluation research. Logan (1972:380) identified what he considered to be four of the most crucial criteria as: 1) adequate definition of the program and techniques, that is, the intervention strategy; 2) the presence of control groups; 3) adequate definition of success and failure; and 4) follow-up assessment. Our assessment of the literature confirms recognition by meta-evaluators of the need for control groups, follow-up and operationalization of program and treatment techniques as important criteria. Success and failure appear to be of less concern in the literature reviewed. The importance of this criterion for assessing the impact of a program, however, has formed the basis for many critiques of evaluation research (Hackler, 1978:41), and therefore merits recognition as a crucial element of any study. In addition to these four criteria identified by Logan (1972:380), premeasurement or baseline measures was also discussed widely in the literature, and should also be included in this list. This brings the number of 'crucial criteria' to five. Because of the large number of programs deemed to be inadequately evaluated, we shall re-examine the 23 studies on these 'crucial' components of evaluation research to determine if they, at least, are being adhered to. These five criteria are utilized to determine whether or not the results of a study can reliably be expected to be of any value or whether the entire study must be rejected

as an inadequate assessment of the intervention program. These five criteria, then, form the base requirements for acceptance of a study as meriting further assessment. Upon further examination of Table IV, it can be seen that only one (4%) study met all 5, however, only 3 of the criteria were fully met.

5. In the discussion of points 1 to 3 above, we examined the number of studies which met each criterion. Reading down the table we can see the number of criteria which were met by each study. It can be seen that 8 (53%) of the criteria were met in some way by at least half of the studies (eleven or more studies-50% or more). An additional 2 (13%) criteria were met by a third of the studies (eight or more-35% or more). The remaining 5 (33%) criteria were seriously neglected. However, when we removed studies that only partially adhered to the criteria, not one criterion was adequately represented. Only 4 (27%) criteria - adequate definition of the aims of the program; information to replicate; provision of control; before and after measures - were fully met by at least one third of the studies.

6. Although number of subjects was not included as a criterion for our study, it is important to note that most of the studies employed designs with a minimum of ten subjects and a maximum of 214. Only two (9%) studies conducted in the early 1960's (Turner, 1961 and Hartman, 1961) did not indicate the number of subjects.

SPECIFIC EVALUATION OF CRITERIA

In general, the results of this study have been both disheartening and encouraging. We shall now turn to a discussion of each criterion to determine where evaluators' greatest weaknesses lie.

1. Adequate definition of program. This criterion was satisfied most often, including both subcategories, (a) aims of the program (twenty studies-87%) and (b) intervention technique (nineteen studies-83%). Most of the studies were very clear about the purpose for conducting the research. For example, Reker and Maissner, (1977:293) identified the purpose of their project as an evaluation of the effects of Life Skills training on inmate participants by comparing their progress with inmates receiving an attention placebo and those receiving the regular institutional program (each technique being fully described). The target behaviours were then clearly set out and objective measures to evaluate them were described, such as a Purpose-in-life test, a Life Areas Survey, the Edwards Personality Inventory and a Life Skills Evaluation Questionnaire.

Other studies, although apparently providing a clear indication of purpose, or aims of the program, failed to indicate the techniques employed or a measurable criterion of the target behaviour. Quinsey and Sarbit (1975:178), for example, intended to 'add to the Behaviour Modification literature' by 1) examining behaviour changes associated with

increasing the value of points earned, and 2) demonstrating the efficacy of a token economy in a maximum security setting. The techniques of treatment were adequately discussed and objectively measurable (i.e. the number of points earned in a week), however, target behaviour was simply presented as 'point earning behaviour'.

Less clear program goals were also presented in several studies. Turner (1961:485), for example, identified the problem; to examine the effects of group psychotherapy on sexual deviations, however, treatment was vaguely defined in terms of 'permissiveness', 'support', and 'stimulation', and the target behaviour was referred to ambiguously as 'symptoms of deviancy' and 'underlying psychopathology'.

Although twenty studies met the first subcategory of specifying the aims of the program and nineteen studies met the second (definition of intervention techniques), when the degree to which the criteria were considered to be fully satisfied rather than partially met was examined, only nine (39%) and seven (30%) respectively were fully satisfied.

2. Routinization. All three of the subcategories: (a) the intervention is not specific to an institution/ population; (b) there is adequate information to train others to perform treatment; and (c) there is adequate information to replicate the research, were met by over half the studies; sixteen (69%), eleven (50%) and fourteen (61%) respectively. However, partial agreement removed, the numbers were reduced to five (21%), six

(26%) and eight (35%).

3. Control or Comparison Groups. Eleven (50%) studies indicated that a control or comparison was included in their assessments, however, only eight (35%) of the studies provided sufficient information of their provision. Parlett and Ayers (1971), for example, examined four groups on programmed instruction, two groups from one institution were to be compared with two groups from a second institution. These groups were matched as closely as possible for length of sentence, type of crime and age. Each institution had one group attend school full time and one group performing regular institutional duties.

There was some evidence in the 23 studies reviewed that the experimental groups were receiving the intervention, and the control groups were not, although the regularity and quality of intervention are unknown. Table IV indicates that, seven of the eight studies (88%) which provided evidence of intervention, were not sufficiently clear to determine the reliability of the claim, i.e. they did not clearly explain what they actually did. Only one study (Gendreau, Burke and Grant, 1980) provided some evidence. In this study of an inmate volunteer program, the volunteers performed a job outside of an institution in two settings: 1) One group worked at the Rideau Regional Centre for the retarded; and 2) the other group worked at Brockville Psychiatric Hospital, on the geriatric ward. They were compared with each other and with a non-volunteer group who remained at the Rideau Correctional Centre. In this study it can be assumed

that the distance between the three groups disallowed the intervention to flow over from one group to another. However, in this case inmates returned to the institution in the evenings, and it is possible that experiences could be shared.

4. Randomization. Randomization is perhaps one of the most essential requirements of a good experimental design, yet not one study fully conformed to this criterion, and only eight (35%) even partially met this criterion. Subjects were matched in only two (9%) studies.

5. Before and After Measures. Fifteen (65%) of the studies indicated premeasurement, however only nine (39%) succeeded in providing evidence of its existence. Davidson, Willis and Cole (1978) measured subjects on their arrival at an institution, using the Peabody Individual Achievement Test to score and establish an educational grade equivalent value.[6] The subjects then entered the correctional education program. Retests, using the same measure, were conducted within one week of release from the institution. Pretests were indicated in other studies. For example, Hartman (1961:493-494) indicated that psychological assessments were made by a clinical team before intervention, however, the tests were not specified and, although clinical assessments may be helpful in providing additional information regarding subjects, they must be accompanied by objective measures, or at least they must tell us what 'information' they gathered.

6. Operational Definition of Success and Failure. Both subcategories of success and failure were severely neglected in the studies. Success, as it related to target behaviours, was indicated in fifteen studies (65%), however, only three studies (13%) operationalized and measured the criteria for success and failure. Gendreau, Burke and Grant (1980:68) for example, defined failure as removal from the program for rule violation or poor work performance. In addition to specific criteria that determine success or failure, tests specifically designed to determine success/failure responses also constitute an objective and clearly defined procedure. Thus Daigle-Zinn and Andrews (1980:323), in assessing the effectiveness of role-playing and didactic-discussion approaches, tested for acceptance of self as one of their criteria for success. Success was measured using Berger's Acceptance of Self and Others Scales.[7] In most of the studies, however, a nebulous notion of success was more the rule, including both the definition of the concept and how it was measured. Such vague notions as "expectations or "recognition and expression of internal problems" (Achille, 1964:71), or a change in patterns of behaviour and values (Grygier, et al., 1968:254), formed the basis for concluding the effectiveness of the intervention.

Conventional definitions (i.e. criminality) of success and failure were almost entirely neglected; only four studies (17%) attempted to determine success in terms of criminality, and none succeeded in providing objective measures. Recidivism was the

criterion employed in all three studies. Frechette (1960) and Landreville (1967) both employed recidivism rates as measures of success, but failed to explicate the criteria for recidivism. Annis (1979:11) used two measures of recidivism: 1) number of convictions; and 2) number of days served in prison. She did not, however, indicate whether return to an institution resulted from new offences (serious or minor) or parole violations. Nor did she indicate whether days served in prison included parole suspension as well as revocation.

7. Follow-up in the Community. Table V indicates the length of follow-up conducted in the 23 Canadian studies. Table V clearly demonstrates the lack of attention evaluators pay to this criterion. Out of 23 studies, thirteen (57%) did not include any follow-up, three (13%) provided post-intervention information for up to six months, two (9%) continued for up to two years and only one (4%) clearly indicated a follow-up of more than two years. Two (9%) studies claimed on-going follow-up but did not clearly define the types of measures, or where the measures were taken. Landreville (1967) followed his subjects for two years, measuring recidivism, and Annis (1979) provided a twelve month post-release analysis, using two methods: 1) the first involved employment records, involvement with the law (i.e. contact with the police), and use of drugs or alcohol; and 2) the second method involved the use of recidivism (discussed above).

Table V

Length of Follow-up in the Community of 23 Canadian Studies

Author	Year	None	less than 6 mths.	6 mths. to 1 yr.	1 - 2 years	2 years or more
Flint	1960	*				
Frechette	1960	unknown				
Hartman	1961				*	
Turner	1961					on-going
Philip	1961	*				
Campbell	1963					on-going
Achille	1964	*				
Coutts	1964	*				
Klonoff	1966	*				
Landreville	1967					*
Grygier	1968		*			
Parlett	1971	*				
Coons	1972	*				
Coons	1972	*				
Ross	1973	*				
Andrews	1974		*			
Quinsey	1975	*				
Quinsey et al.	1976	*				
Reker & Meissner	1977	unknown				
Davidson et al.	1978	*				
Annis	1979				*	
Daigle-Zinn & Andrews	1980		*			
Gendreau et al.	1980	*				

8. Provision of Unbiased Observers: Independent Evaluators. This final requirement, like many of the others, was insufficiently attended to by the majority of researchers. Six studies (26%) indicated that perhaps the research was conducted by independent evaluators. This assumption is made only on the knowledge of affiliations of the researchers; i.e. Philip (1961), Achille (1964), Landreville (1967), Reker and Meissner (1977), Daigle-Zinn and Andrews (1980), and Gendreau, Burke and Grant (1980) were all (or some members if more than one author) associated with universities. None of the studies, however, indicated that this was, in fact, the case.

Provision of Unbiased Observers: Blind Analysis. Blind analysis was recognized in only three studies (13%) as a necessary element for objectivity. Ross and Doody (1973:297), examining the application of behaviour modification techniques, provided two unlabelled lists so that the experimentor would not know the category of subject he was assessing during the study. Blind analysis was also utilized by Andrews and Young (1974:9). One of the measures, attitude scales, was administered by the institution's psychometrist who was not a part of the program. Other measures were taken by on-line staff who were not aware of the experiment. Coon's (1972) study, assessing psychotherapy, indicated complete accordance to this criterion. The 'judge' was unaware of the identity of the subjects, the treatment and which protocol was pre or post therapy.

Although there are several criteria that are grossly underrepresented in evaluation studies (conventional definitions of success and failure, follow-up in the community, unbiased observers, and randomization), many studies have made some attempt to employ more rigorous components in their designs, with more criteria being met in the latter studies. According to Table IV there is an observable difference in the quality of studies conducted in the early 1960's compared with the 1970's and early 1980's. This would appear to coincide with the emergence of reassessment reviews in the literature, beginning around 1970.

LIPTON, MARTINSON AND WILKS: A REASSESSMENT

Hackler (1978) claimed that in Canada, we should learn from the mistakes made in the United States. This was in part a response to the findings of studies on intervention effectiveness which appear to be more negative as methodological rigor increases. The Lipton, et al. (1975) report provides the primary source of these conclusions, therefore a reassessment of the relevant studies conducted by Lipton, et al., according to our criteria might prove enlightening to future researchers in Canada. In addition, a comparison of some of the alledgedly 'best' published studies available in correctional intervention in the United States with published studies available in Canada may also indicate where research in Canada stands, and where it should be going.

A review of the 231 studies collected by Lipton, et al., (1975) produced a sample of 10 published articles dealing with institutionalized offenders. Table VI presents the results of our reassessment of these 10 studies according to our criteria set out in Chapter IV. The following observations can be drawn from the table:

1. Three (30%) of the studies met seven or more of the criteria, six (60%) met between 4 and 6 criteria, leaving only one (10%) study as seriously inadequate.

2. When partial accordance was subtracted from the total, none of the studies fully satisfied half of the criteria. Two (20%) studies fully satisfied between 4 and 6 criteria. The remaining eight (80%) studies were seriously inadequate.

3. Of the five criteria identified in Point 4 of General Evaluation, as most crucial to evaluation research, (i.e. adequate definition of program; control groups; adequate definition of success and failure; follow-up; and baseline), only one (10%) study (Persons, 1966) met all five, however, only two of the criteria were fully met. Four (40%) more studies met 4 of the 5 criteria (Freeman & Weeks, 1956; Schnur, 1948; Cabeen, 1961; and Blake, 1965) to some degree.

4. Reading down the table, it can be seen that 7 (47%) of the criteria were met by at least half of the studies (five or more-50% or more). Three (20%) more criteria were met by a third of the studies (three or four-30 to 40%). The remaining criteria were grossly under-represented.

Table VI

Methodological Re-Assessment of Lipton, et al. Using Criteria Employed for Canadian Assessment

Name and Year	N=	I		II			III		IV		V	VI			VII	VIII		No. of Criteria met by each study Total (fully met)	Total (parially met)	Total	Outcome
		IA	IB	IIA	IIB	IIC	IIIA	IIIB	IVA	IVB		V1A	VIB			VIIIA	VIIIB				
Schnur 1948	N=1,762	+					+					+	*		*	+		2	4	(6)	Positive +
Gerstenlauer 1950	N=44						*			+	*							2	1	(3)	Positive +
Freeman and Weeks 1956	N=237	+					+			+			*		+			1	4	(5)	Partial +/-
Benson 1059	N=262	*				+	+				*		+					2	3	(5)	Negative -
Cabeen 1961	N=120		+				+	+		+		*	+		*			2	5	(7)	Positive +
Gillooly 1965	N=96		*		*		*					*	*			+		6	1	(7)	Positive +
Blake 1965	N=97	*	*				*					+	*		*			5	1	(6)	Partial +/-
Persons 1966	N=82	+	+					*		+	*	*		+	+	+		3	6	(9)	Positive +
Jurjevich 1966	N=97	*	+					*					+					2	2	(4)	Partial +/-
Levinson 1966	N=300	*					+	*					+					2	2	(4)	Partial +/-
Total (fully met)		4	2	1	1		2	4	0	0	1	5	2	2	3	0	0				
Total (Partially met)		3	3	0	0		3	4	0	2	2	1	4	2	2	3	0				
Total		(7)	(5)	(1)	(1)		(5)	(8)	(0)	(2)	(3)	(6)	(6)	(4)	(5)	(3)	(0)				

5. All studies reassessed contained large samples; the smallest included 44 subjects, the largest comprised 1,762.

There does not appear to be any continuity in improvements over time in these studies. This may be due, in part, to the fact that the studies reassessed from Lipton, et al. spanned the years 1948 to 1966 (1948:1; 1950:1; 1956:1; 1959:1; 1961:1; 1965:2; 1966:3). As mentioned above, there was little concern for more rigorous research methods prior to the early 1970's.

COMPARISON

When the results of the Canadian studies are compared to those assessed by Lipton, et al. several interesting observations can be made:

1. There was no difference between the two sets of studies in the number of criteria met by the studies. The assessment of the Canadian studies found that 8 criteria (53%) were met by half the studies; 2 criteria (13%) by a third, leaving 5 (33%) criteria virtually unattended to. In comparison, the Lipton, et al., (1975) assessment found 7 (47%), 3 (20%), and 5 (33%) respectively.

2. Only one study in both the Canadian (Annis, 1979) and American (Persons, 1966) assessment met all five of the most important criteria, (i.e. adequate definition of program; control groups; adequate definition of success and failure; follow-up; and baseline).

3. Reading across Tables IV and VI, the results pointed in the same direction for criteria being fully satisfied by the studies, although the Canadian studies appear to be meeting more of the criteria; one (4.3%) met 7 or more, six (26.1%) met between 4 and 6, and sixteen (69.6%) met less than 3 criteria. This is compared with the American studies, zero (0%), two (20%), and eight (80%).

4. However, when studies which are simply attempting to employ more rigorous designs in evaluating institutional intervention programs (whether fully or partially) are compared the Canadian studies fare much better than the American studies. Whereas thirteen (56.5%) of the 23 Canadian studies met 7 or more criteria, only three (30%) of the 10 American studies did; seven (30.4%) Canadian studies met between 4 and 6 criteria as opposed to six (60%) American studies. Three (13%) Canadian studies and one (10%) American study met 3 or less criteria.

5. Finally, the types of interventions were compared to determine if in fact the techniques of intervention have changed, or if different types of intervention are being evaluated. See Table VII for a comparison of intervention strategies employed in tha Canadian and American studies. The results of the comparison indicate that there appears to be no difference between the American and Canadian studies regarding types of intervention strategies.

Table VII

Comparison of Intervention Strategies between American (from Table VI) and Canadian (from Table IV) Studies Based on Methodological Criteria

	CANADIAN STUDIES				AMERICAN STUDIES			
	N Studies	For more Criteria	4 - 6	Less than 4	N Studies	For more Criteria	4 - 6	Less than 4
group therapy	N=3	2	1		N=2	1	1	
group psychotherapy	N=5	2	1	2	N=1	1		
individ. therapy	N=1		1					
group counsell.	N=2		1	1	N=1		1	
individ. counsell.					N=1		1	
milieu therapy	N=2		2		N=1		1	
education programs	N=4	3	1		N=2	1	1	
skill development					N=2	1	1	
aversion therapy	N=2		2		N=1		1	
instnl. program	N=1							
behaviour modif.	N=2	2						
multi-group appch.	N=1	1						
life skills	N=1	1						
role playing & didactic discsn.	N=1	1						
inmate volun. prog.	N=1	1						
imprisonment	N=1		1		N=2		2	
Total	27	13	10	3	13	4	9	

FINDINGS

Researchers appear to be acknowledging the methodological limitations of their studies and the implications of these limitations for their conclusions, findings and recommendations. For example, although Landreville (1967:343) reported unequivocal success on four measures of recidivism, he also recognized and discussed the methodological weaknesses of his study, specifically noting that without a control group, "Les taux de succes que nous avons trouves ne peuvent etre contestes, mais il nous est impossible d'etablir si les taux de succes des jeunes confies a Boscoville auraient ete superieurs, ou inferieurs, si ceux-ci n'avaient ete soumis a aucun traitement ou a un traitement different." (The success rate we found could not be contested, however it is impossible to establish whether the success rate of the youth confined in Boscoville would have been better or worse, unless xome were submitted to no treatment or a different treatment). In addition, he recognized that, although the purpose of the intervention program was the internalization of self-control and feelings of self-worth, the only objective measure used in the study was not for the target behaviour, but rather in terms of recidivism. Therefore, although it appeared that the results of this study were positive, an interpretation of those results by the author, taking into account the methodological flaws in design, indicated only partial success, with future 'experimental'

research recommended to validate the results.

Of the 23 studies analyzed, only one (4.3%) indicated negative findings (i.e. the intervention strategy was not effective or there was no difference), twelve (52.2%) presented positive findings (i.e. the intervention produced beneficial results), and ten (43.5%) found equivocal results. Despite equivocal findings, i.e. partial failure in results, seven (30%) recommended further research or analysis of the intervention program. Although the numbers in this study are few, this does not appear to support the claims in the literature that evaluations with negative findings necessarily result in condemnation and ultimate discontinuation of programmatic activity, as claimed by Hackler (1978:58).

Outcome data presented in the American studies was comparable to the findings of Canadian studies. Five studies (50%) produced positive findings; one study (10%) indicated negative findings; and the remaining four (40%) found equivocal results. Lipton, et al's. analysis of the findings was not significantly different from ours. They found five studies (50%) with positive results, two (20%) with negative results (or no difference) and three (30%) with equivocal results.

CONCLUSION

The number of studies included in both sets of assessments is small, therefore conclusive comparisons are not possible. However, one tentative conclusion that can be suggested by this comparison is that the Canadian studies are not of poorer quality than the American studies. In fact, there is indication that the Canadian studies are improving in methodological rigor. It should be noted that the improvement of methodological rigor appears to be coincident with the appearance of meta-evaluation research in the literature. As both the Canadian and American studies were published, this can not be the sole reason for more rigorous methodology. It would appear that time may be a more important factor. The literature in recent years has produced an increased number of critical reviews and discussions in the field of evaluation research. This appears to be reflected in research being conducted and published.

In addition, the claims that have been made that methodologically conscious studies will only produce negative results, have not been supported by this study. It is apparent that, although evaluation studies in Canada require greater attention to research methods, evaluators are becoming aware of the weaknesses in their designs, and are, consequently, qualifying their conclusions accordingly.

NOTES

1: Research methods include conceptual analyses and appropriate designs for evaluation research.

2: There is a large body of literature on evaluation research. Many issues have been raised regarding types of research questions and obstacles encountered, as well as appropriate designs. As evaluation research develops and becomes established, the accumulation of knowledge in the field expands. It would seem that reference to this body of knowledge and expertise would also accompany this trend.

3: Psychometric tests employed in the Canadian studies assessed in this evaluation include MMPI, Wechsler-Bellevue Intelligence Scale, Cattel's 16 Personality Factor Test, Ellsworth MACC Scale, Internal-External Locus of Control Scale, Edwards Personality Inventory, and Thematic Apperception Test. An examination of the studies employing these methods of measurement found that the majority of the studies did not provide an indication of reliability or validity measures on the tests employed.

4: Protocol in the study conducted by Coons (1972) refers to two psychotherapeutic approaches to mental hospital patients where interpersonal interaction is stressed in the absence of insightful content and where insight was stressed with very little interaction.

5: Physiological changes induced by shock therapy conducted by Quinsey (1976) involved a comparison of skin conductance and penile circumference during a constant number of classical aversion conditioning sessions. Psychological changes associated with a study conducted by Klonoff (1966) involving the use of the drug sodium amytal included cognitive abilities and personality changes.

6: A grade equivalent value represents the correlation between the academic achievement of institutional students as indicated by the Peabody Individual Achievement Test and the average student's achievement in a regular school. For example, Davidson, et al. (1978:56) found that the mean score for 36 institutional students was 7.9, which indicated that the average student in this group was academically equivalent to a regular school student in the ninth month of the seventh grade.

7: We noted that, unlike many scales utilized in these studies, Daigle-Zinn and Andrews (1980) have indicated that the tests used in their study had reliability tests conducted on them.

VI. CONCLUSION

A review of the literature on evaluation, specifically program evaluation, indicates that members of the social sciences are becoming concerned with the quality of research conducted in the field. Policy implications have created an awareness that the role of the social scientist has the potential of being a very powerful one. This responsibility has led to more attention to procedure in methodology and research design.

It is apparent, in view of the literature on evaluation research, that more methodologically rigorous and detailed evaluation is essential in assessing the effectiveness of different intervention programs. Aside from the policy implications of the research, the addition of this research to a more general body of knowledge should also be considered.

The results of our review of program evaluations has been on the one hand disheartening, while on the other very encouraging. The findings indicated generally that approximately half of the Canadian studies attempted to employ some sort of experimental design. However, when these studies were more closely examined it was found that only one study clearly indicated adherence to the important methodological requirements. There was a notable difference in methodological rigor between studies conducted from 1960 to 1970 and from 1971 to 1980, where the latter period produced much more rigorous

designs. This trend coincides with the emergence of evaluative evaluation research beginning around the 1970's.

The studies reassessed from the Lipton, et al. review produced fairly similar results. Although the numbers are small, a comparison of the two analyses, that is, the Canadian studies compared to those from the Lipton, et al. alledged assessment, indicated little difference between the two in the number of criteria met in general, and by specific types of interventions (i.e. similar interventions were being assessed, and adherence to methodological norms was comparable for both sets). However, in terms of the number of studies attempting to employ more rigorous designs, the Canadian studies appear to be advancing ahead of the American studies. It is important to note, however, that the difference between the years in which the studies were conducted (in Canada between 1960 and 1980, in the United States between 1948 and 1966) makes it very difficult to compare them too closely, especially in view of the recency of academic attention in this area.

RESEARCH SHORTCOMINGS

One of the themes of this thesis focuses on researchers recognizing the flaws, or research shortcomings of their own work. We should, therefore, also adhere to those requirements and be critical of some of the methods we employed in our analysis and the assumptions upon which they were made.

1. Procedure of Analysis. The method employed for analyzing adherence to the methodological requirements was based on a subjective assessment of the studies. Although we felt that at this time in the development of methodological meta-evaluation it is the 'best' way to approach an assessment, we suggest that for more reliable results this method can be improved through the use of multi-analysts.

2. Published Studies. Only published studies were analyzed in this thesis, that is published studies found in Canadian journals. There are two considerations we must address regarding our selection of studies. The first point we should recognize is that the sole use of published studies excludes the majority of research conducted, either by in-house organizations or private firms.

A second concern addresses the quality of research maintained in Canadian journals as opposed to American, British or International journals. There is a possibility that there is a much larger number of better quality Canadian research being published in journals outside of Canada. It has been suggested that by only including Canadian journals the results may be unnaturally skewed in an unfavourable direction.

This study was a preliminary attempt to determine the quality of research produced in Canada. The impetus for our decision to take this approach stemmed from the small amount of information contained in Canadian journals. For the student of social science research in Canada this is a disturbing reality

which can no longer go unnoticed. However, future research in this area could well benefit from a comparative analysis of international journals regarding containment of Canadian research. Such an analysis would provide a wider and more representative sample of Canadian research. It would also, if published in a Canadian journal, provide an information base for students in the field.

3. Intervention in the Institutional Environment. Another two-pronged criticism of the population we chose to study concerns: (a) the specificity of the population, that is, a correctional institutional population; and (b) the sole use of 'treatment' intervention as the targeted intervention in studies to be assessed.

It was decided upon beginning our analysis that the limits of the study should be determined beforehand. The specifications we decided upon included only the populations of federal, provincial and juvenile institutions. There are two directions this study or a future study in this area could take: (1) either the study could become more specific and examine only one type of institution, if there is a sufficient number of studies available to do so; or (2) perhaps a more general approach would include an assessment of institutional as well as non-institutional populations.

The primary concern in the majority of literature on program evaluation has focused on the concept of 'treatment' in corrections. This background was instrumental in the development

of this thesis. There are, however, many other important concerns in corrections which are ignored by concentrating soley on 'treatment' or ameliorative types of programs. It is probable that many of these concerns are more immediate and more important at the administrative level than the impact of 'treatment' programs, such as programs designed to maintain institutional control and order. It is possible that inattention of these issues in this thesis is a result of non-publication of evaluation studies, perhaps due to the unpopularity of programs focusing on 'control' rather than 'treatment'.

On another level, although for policy purposes, evaluation of the 'control' type of programs may have more impact, for the student of program evaluation and research methods in general, published evaluations will provide the basis for her studies and the subject matter of the evaluation is not the issue but rather, how it was conducted.

4. Weighted Criteria. There was no actual attempt in this thesis to weight the relative importance of the criteria, that is, to determine which criteria were more important. The selection in Chapter V of five crucial criteria was a beginning, however the method employed was unsophisticated and not a valid measure. As we note in 'Recommendations for Future Research', this is in particular a direction research should explore.

We do not feel that acknowledgement of the shortcomings of this research undermines the value of this study. It is, however, recognition that the conclusions we arrive at must be

read in light of the above considerations.

RECOMMENDATIONS FOR FUTURE RESEARCH

Several interesting features of evaluation research were identified in this paper. Many of them suggest further study is required before any definitive statements can be made regarding the 'state of the art'. Indeed, perhaps some further questions must be asked and answered before evaluation can develop and mature as a valid source of information. There are several issues of particular relevance to this thesis.

1) One of the first features that is very apparent in program evaluation is the relationship between evaluator and the program, program staff and administration. This feature has been discussed extensively in the literature, and there appears to be two focal perspectives that are being maintained. The debate centres on whether evaluation should be conducted by researchers associated with the program or external to it.

2) Second, the majority of methodological reviews, including this one, have examined general requirements that "should" guide researchers in implementing adequate research designs. As evaluative evaluation research develops, new ways of examining the issues and identifying the pertinent features and qualities should be explored, such as focusing on establishing specific guidelines for conducting this unique type of evaluation. Bernstein (1978) attempted to apply a weighted scale to the analysis of evaluations, that is, she attempted to assess the

relative importance of several of the criteria used in evaluating evaluation research. Further exploration of weighting procedures may assist in the refinement of assessment techniques employed in evaluative evaluation research.

3. Third, much of the debate regarding program effectiveness has focused on the concept of recidivism (Hackler, 1978; Lipton, et al., 1975; Logan, 1972). It is interesting to note that although recidivism, or the discontinuance of criminal behaviour in general form the basis for this debate, we only found three studies which dealt with criminality as the behaviour to be changed. Further study into the goals of intervention strategies with institutional populations may be needed to address the debate regarding effectiveness of these programs in reducing recidivism.

4. A fourth issue we discovered is closely related to the goals of intervention strategies. We found in our review that there is a wide variety of intervention programs being assessed. This diversity of programs may possibly have deleterious effects on the development of cumulative knowledge in the area of corrections, in that there is insufficient information produced regarding any particular program, it's goals, techniques of intervention and ultimately it's effectiveness.

5. We noted earlier in Chapter Five that there was often insufficient information in the articles we reviewed. In order to assess the studies we were often left with the choice of leaving the criterion blank or of trusting the researchers

allusions to adherence. This phenomenon poses an interesting question for further examination. As we were only using articles published in journals, it may be pertinent to publication to determine the effects of journal policies regarding the amount of information or reduction of information different journals require for acceptance. It may be that much of the information we found missing was a result of journal exclusion policies.

In conclusion, it is not possible, at this time, to say treatment is not effective as Lipton, et al. (1975:627) would contend, for the studies cited in this paper indicate that indeed their methods do incur positive results, or at least merit further investigation. Hackler's (1978:23) assertion that it is not possible to conduct rigorous evaluations in Canada is also not fully supported by the evaluation studies reassessed in this paper. Although the studies continue to employ faulty designs, clearly there are attempts at improving these techniques in recent research. These indications reflect changing attitudes towards evaluation research, and require further exploration and attention.

At this point in the evolution of program evaluation it is not possible to draw definitive statements and conclusions. Rather than saying 'nothing works', or declaring that evaluation is destructive and unproductive, we should say that, at present, our knowledge in this area remains superficial, and that any conclusions we can produce remain conditional and tenuous.

Appendix I

Content Analysis Categories Employed in Canadian Assessment

1.	Date:	Year article was published Years covered in the evaluation
2.	Authors:	Authors of article Where they were hired from, i.e. university vs. institution. References to research methods literature Journal in which article was published.
3.	Problem:	Definition of Aims of Program Purpose of Evaluation.
4.	Principal Methods:	Number of Subjects methods of Data Collection, i.e. observation Pre-measurement
5.	Research Sites:	Where researach conducted Where subjects from i.e. institution
6.	Research Data:	Demographic Data on Subjects Selection criteria and procedures.
7.	Independent Variables:	Techniques of Treatment Instruments employed.
8.	Dependent Variables:	Behaviour attempting to change (i.e. target behaviour) Instruments employed.
9.	Control:	Provision of Control or comparison group or statistical comparison Selection procedure, i.e. random.
10.	Follow-up:	Length of follow-up Where measurements taken, i.e. post-release. How measurements taken.
11.	Findings:	Results of Assessment Definitions of Success and Failure Conclusions, i.e. positive.

Bibliography

Achille, Pier Angelo. "La Psychotherapie avec des groupes de jeunes delinquants en internat, de reeducation: les difficultes du demanage." Canadian Journal of Corrections vol. 6 (1964):65-73.

Andrews, D.A. and J.G. Young. "Short-Term Structural Group Counselllng and Prison Adjustment." Canadian Journal of Criminology and Corrections vol. 16 (1974):5-13.

Annis, Helen M. "Group Treatment of Incarcerated Offenders with Alcohol and Drug Problems: A Controlled Evaluation." Canadian Journal of Criminology vol. 21 (1979):3-15.

Bayley, David H. "Perspectives on Criminal Justice Research." Journal of Criminal Justice vol. 6 (1978):287-289.

Benson, Sir George. "Prediction Methods and Young Prisoners." British Journal of Delinquency vol. 9 no. 3 (1959):192-199.

Bernstein, Ilene Nagel. "Social Control in Applied Social Science: A Study of Evaluative Researchers' Conformity to Technical Norms." Social Science Research vol. 7 (1978):24-47.

Bernstein, Ilene Nagel and Howard E. Freeman. Academic and Entrepreneurial Research. New York:Russell Sage Foundation, 1975.

Blake, B.G. "The Application of Behaviour Therapy to the Treatment of Alcoholism." Behaviour Research and Therapy vol. 3 no. 2 (1965):75-85.

Boruch, R.F., A.J. McSweeny, and E.J. Soderstrom. "Randomized Field Experiments for Program Planning, Development and Evaluation: An Illustrative Bibliography." Evaluation Quarterly vol. 2 no. 4 (November, 1978):655-695.

Cakeen, Charles W. "Group Therapy with Sex Offenders: Description and evaluation of a Group Therapy Program in an Institutional Setting." Journal of Clinical Psychology vol. 17 no. 2 (1961):122-129.

Campbell, Jay. "An Experience in Group Counselling." Canadian Journal of Corrections vol. 5 (1963):90-97.

Cavior, Helene E. and Stanley H. Cohen. "Evaluative Research: Perspectives from a Corrections Setting" Criminal Justice and Behaviour vol. 2 no. 3 (September 1975):237-257.

Chelimsky, Eleanor. "The Need for Better Data to Support Crime Control Policy." Evaluation Quarterly vol. 1 no. 3 (August 1977):439-474.

Conner, Ross F. "Selecting a Control Group: An Analysis of the Randomization Process in Twelve Reform Programs." Evaluation Quarterly vol. 1 no. 2 (May 1977):195-244.

Cook, Thomas D. and Donald T. Campbell. "The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings." Handbook of Industrial and Organizational Psychology Marvin D. Dunnette (ed.) 1976:223-326.

Cook, Thomas D. and Charles L. Gruder. "Metaevaluation Research." Evaluation Quarterly vol. 2 no. 1 (February 1978):5-51.

Coons, W.H. "Psychotherapy and Verbal Conditioning in Behaviour Modifications." The Canadian Psychologist vol 13. no. 1 (January 1972):3-29.

Coutts, Dorothy. "The Oakalla Prison Farm Program for Treatment of Narcotic Addicts." Canadian Journal of Corrections vol. 6 (1964):14-21.

Daigle-Zinn, Wendy J., and D.A. Andrews. "Interpersonal Skill Training for Young Adult Prisoners." Canadian Journal of Criminology vol. 22 no. 3 (July 1980):320-327.

Davidson II., William S. and Edward Seidman. "Studies of Behaviour Modification and Juvenile Delinquency: A Review, Methodological Critique, and Social Perspective." Psychological Bulletin vol. 81 no. 12 (1974):998-1011.

Davidson, Paul, J. Willis and R.J. Cole. "A Practical Model for Evaluating Correctional Education Programs." Crime and/et Justice vol. 6 no. 1 (1978):53-58.

Emery, Robert E. and D. Marholen II. "An Applied Behaviour Analysis of Delinquency: The Irrelevancy of Relevant Behaviour." American Psychologist (October 1977):860-873.

Fisher, Gene A. and M.L. Erickson. "On Assessing the Effects of Official Reactions to Juvenile Delinquency." Journal of Research in Crime and Deliquency vol. 10 (1973):177-194.

Flint, Maurice. "An Experiment in the Rehabilitiation of Women Offenders." Canadian Journal of Corrections vol. 2 (1960):240-254.

Frechette, Marcel. "La therapie de groupe au penitencier." Canadian Journal of Corrections vol. 2 (1960):255-263.

Freeman, Howard E. and C. C. Sherwood. Social Research and Social Policy. New Jersey: Prentice-Hall, Inc., 1970.

Freeman, Howard F. and A.H. Weeks. "Analysis of a Program of Treatment of Delinquent Boys." American Journal of Sociology vol. 62 no. 1 (1956):56-61.

Friedman, Paul R. "Legal Regulation of Behaviour Modification." In Justice and Corrections, pp. 730-744. Edited by N.Johnston and L.D. Savitz. New York: John Wiley & Sons, Inc., 1978.

Gendreau, Paul, D. Burke and B.A. Grant. "A Second Evaluation of the Rideau Inmate Volunteer Program." Canadian Journal of Criminology vol. 22 No. 1 (January 1980):66-77.

-------- and B. Ross. "Effective Correctional Treatment: Biblio-Therapy for Cynics." Crime and Delinquency (October 1979):463-489.

Gerstenlauer, C. "Group Therapy with Institutionalized Juvenile Delinquents." American Psychologist vol. 5 (1950):325.

Gibbons, Don C. "Comments on the Efficacy of Criminal Treatment." Canadian Journal of Corrections vol. 2 (1960):165-174.

-------- B.D. Lebowitz and G.F. Blake. "Program Evaluation in Correction." Crime and Delinquency vol. 22 (1976):309-321.

Gillooly, William B. "A Revolution in Reading Instruction at the Training School." American Journal of Correction vol. 27 no. 2 (1965):30-31.

Glaser, Daniel. "Remedies for the Key Deficiency in Criminal Justice Evaluation Research." Journal of Research in Crime and Delinquency (July 1974):144-154.

Gordon, Gerald and E.V. Morse. "Evaluation Research." Annual Review of Sociology vol. 1 (1975):339-361.

Gottfredson, M.R. "Treatment Destruction Techniques." Journal of Research in Crime and Delinquency (January 1979):39-54.

Grygier, Tadeusz, M. Guarino, B. Nease, and L. Sakowicz. "Social Interaction in Small Units: New Methods of Treatment and its Evaluation." Canadian Journal of Corrections vol. 10 (1968):252-260.

Hackler, James C. "Invitation to Error: The Dangers of Evaluation and Some Alternatives." Canadian Journal of Criminology vol. 20-21 (1978-79):39-51.

--------.The Prevention of Youthful Crime: The Great Stumble Forward. Toronto:Metheun Publications, 1978.

Hartman, V. "Some Observations of Group Psychotherapy with Paedophiles." Canadian Journal of Corrections vol. 3 (1961):492-499.

Henshel Richard L. Reacting to Social Problems. Don Mills: Longman Ltd., 1976.

Hirschi, T. and H. Selvin. Principles of Survey Analysis. New York:The Free Press, 1973.

Jurjevic, Ratibor M. "Personality Changes Concomitant with Institutional Training of Delinquent Girls." Journal of General Psychology vol. 74 no. 2 (1966):207-215.

Kennedy, Mary M. "Generalizing from Single Case Studies." Evaluation Quarterly vol. 3 no. 4 (1979):661-678.

Kerlinger, Fred N., ed. Foundations of Behavioral Research. 2d. ed. New York:Holt, Rinehart & Winston, Inc., 1973.

Klonoff, H. "Drug Induced Psychological Changes in Sex Offenders." Canadian Journal of Corrections vol. 8 (1966):81-89.

Landreville, Pierre. "Boscoville Centre de Reeducation: Etude 'Follow-Up'." Canadian Journal of Corrections vol. 9 (1967):337-345.

Levinson, Robert B. and H.L. Kitchener. "Treatment of Delinquents: Comparison of Four Methods for Assigning Inmates to Counsellors." Journal of Consulting Psychology vol. 30 no. 4 (1966):364.

Lipton, Douglas, R. Martinson, and J. Wilks, The Effectiveness of Correctional Treatment: A Survey of Treatment Evaluation Studies. New York:Praeger Publishers, 1975.

Logan, Charles H. "Evaluation Research in Crime and Delinquency A Reappraisal." Journal of Criminal Law, Criminology and Police Science vol. 63 (1972):378-387.

Lundman, Richard J., and F.R. Scarpitti. "Delinquency Prevention: Recommendations for Future Projects." Crime and Delinquency (April 1978):207-220.

Martinson, R. "What Works? The Martinson Report." In Justice and Corrections, pp. 788-810. Edited by L.D. Savitz. New York:John Wiley & Sons, Inc., 1978.

McCord, Joan. "A Thirty-Year Follow-Up of Treatment Effects." American Psychologist (March 1978):284-289.

Normandeau, Andre and B. Hasenpusch. "Prevention Programs and Their Evaluation." Canadian Journal of Criminology vol. 22 no. 3 (July 1980):307-319.

Palmer, Ted. "A Critique of Martinson." In Justice in Corrections. Edited by N. Johnston and R.D. Savitz. New York:John Wiley & Sons, Inc., 1978.

Parlett, T.A.A. and A.D. Ayers. "The Modification of Criminal Personality Through Massed Learning by Programmed Instruction." Canadian Journal of Criminology and Corrections vol. 13 (1971):155-165.

Pawlicki, Robert. "Behaviour-Therapy Research with Children: A Critical Review." Canadian Journal of Behavioural Science vol. 2 no. 3 (1970):163-173.

Persons, Roy W. "Psychological and Behavioural Change in Delinquents following Psychotherapy." Journal of Clinical Psychology vol. 22 no. 3 (1966):337-340.

Philip, B.R. "An Objective Evaluation of Brief Group Psychotherapy of Delinquent Boys." Canadian Journal of Corrections vol. 3 (1961):463-466.

Polivka, Larry and Eric Steg. "Program Evaluation and Policy Development." Evaluation Quarterly vol. 2 no. 4 (November 1978):696-707.

Powers, Donald E. and D. L. Alderman. "Practical Techniques for Implementing True Experimental Designs." Evaluation Quarterly vol. 3 no. 1. (February 1979):89-96.

Quay, Herbert C. "The Three Faces of Evaluation: What Can be Expected to Work." Criminal Justice and Behaviour vol. 4 no. 4 (December 1977):341-354.

Quinsey, Vernon L. "Methodological Issues in Evaluating the Effectiveness of Aversion Therapies for Institutionalized Child Molesters." The Canadian Psychologist vol. 14 no. 4 (October 1973):350-361.

--------. L. S.G. Bergersen and C.M. Steinman. "Changes in Physiological and Verbal Responses of Child Molesters During Aversion Therapy." Canadian Journal of Behavioural Science vol. 8 no. 2 (1976):202-212.

--------. and B. Sarbit. "Behavioural Changes Associated with the Introduction of a Token Economy in a Maximum Security Psychiatric Institution." Canadian Journal of Criminology and Corrections vol. 17 (1975:177-182.

Reker, Gary T. and J.A. Meissner. "Life Skills in a Canadian Federal Penitentiary: An Experimental Evaluation." Canadian Journal of Criminology vol. 19 no. 3-4 (1977):292-302.

Reppucci, N.D. and W.G. Clingempeel. "Methodological Issues in Research with Correctional Populations." Journal of Consulting and Clinical Psychology vol. 46 no. 4 (1978):727-746.

Riecken, Henry W. "Memorandum on Program Evaluation." In Evaluating Action Programs. Edited by Carol Weiss. Boston:Allyn & Bacon, Inc., 1972:85-104.

--------. "Introduction: Experiments for Program Development and Evaluation." In Experimental Testing of Public Policy: The Proceedings of the 1974 Social Science Research Council Conference on Social Experiments, pp. 1-12. Editors R.F. Boruch and H.W. Riecken. Boulder:Westview Press, Inc., 1975.

--------. and R.F. Boruch, eds., Social Experimentation: A Method for Planning and Evaluating Social Intervention. New York:Academic Press, Inc., 1974.

Roesch, Ronald and R. Corrado. "The Policy Implications of Evaluation Research: Some Issues Raised by the Fishman Study of Rehabilitation and Diversion Services." The Journal of Criminal Law and Criminology vol. 70 No.4 (1979):530-541.

Ross, R.R. and K.F. Doody. "Persistance in the Psychopathic Personality." Canadian Journal of Criminology and Corrections vol. 15 (1973):292-305.

--------. and P. Gendreau, eds. Effective Correctional Treatment Scarborough:Butterworth & Company (Canada) Ltd., 1980.

--------. and H.B. McKay. "Behavioural Approaches to Treatment in Corrections: Requiem for a Panacea." Canadian Journal of Criminology vol. 20 (1978):279-295.

Rossi, Peter H. "Boobytraps and Pitfalls in the Evaluation of Social Action Programs." In Evaluating Action Programs. Edited by Carol Weiss. Boston:Allyn & Bacon, Inc., 1972.

--------. and S.R. Wright. "Evaluation Research: An Assessment of Theory, Practice, and Politics." Evaluation Quarterly vol. 1 no. 1 (February 1977):5-51.

--------. J.D. Wright and S.R. Wright. "The Theory and Practice of Applied Social Research." Evaluation Quarterly vol. 2 no. 2 (May 1978):171-191.

--------. and D.H. McLaughlin. "Establishing Evaluation Objectives." Evaluation Quarterly vol 3. no. 3 (August 1979):331-346.

Russon, G.W. "The Importance of Scientific Standards in Correctional Research." Canadian Journal of Corrections vol. 5 (1963):236-242.

Schnur, Alfred C. "The Educational Treatment of Prisoners and Recidivism." American Journal of Sociology vol. 54 no. 2 (1948):142-147.

Scriven Michael. "The Methodology of Evaluation." In Evaluating Action Programs. Edited by Carol Weiss, pp.123-136. Boston:Allyn & Bacon, Inc., 1972.

Selltiz, Claire, L.S. Wrightsman and S.W. Cook. Research Methods in Social Relations. 3rd ed. New York:Holt, Rinehart and Winston, 1976.

--------. L.S. Wrightsman and S.W. Cook. Research Methods in Social Relations. 4th. ed. New York:Holt, Rinehart and Winston, 1981.

Slaikeu, Karl A. "Evaluation Studies on Group Treatment of Juvenile and Adult Offenders in Correctional Institutions." Journal of Research in Crime and Delinquency (January 1973):87-100.

Sobel, Suzanne B. "Throwing the Baby Out with the Bathwater." American Psychologist (March 1978):290-291.

Suchman, Edward A. "Action for What? A Critique of Evaluative Research." In Evaluating Action Programs pp. 52-84. Edited by Carol Weiss. Boston:Allyn and Bacon, Inc., 1972.

Szabo, Denis and S. Rizkalla. "Conducting Evaluative Research and Implementing Its Results: Dilemmas for Both Administrators and Researchers." International Journal of Comparative and Applied Criminal Justice vol. 2 no. 1 (Spring, 1978):15-28.

Tufte, Edward R. Data Analysis for Politics and Policy. New Jersey:Prentice-Hall, Inc., 1974.

Turner, R.E. "The Group Treatment of Sexual Deviations." Canadian Journal of Corrections vol. 3 (1961):485-491.

Weinstein, Malcolm S. "Two, Four, Six, Eight - Everyone Evaluate: a Review of Evaluation of Behavioural Programs in Community, Residential and School Settings." Canadian Psychological Review vol. 16 no. 2 (April 1975):134-138.

Weiss, Carol, ed., Evaluating Action Programs. Boston:Allyn & Bacon, Inc., 1972.

--------. "Evaluating Educational and Social Action Programs: A Treeful of Owls." In Evaluating Action Programs pp. 3-27. Edited by Carol Weiss. Boston:Allyn & Bacon, Inc., 1972.

Wholey, Joseph S., J.W. Scanlon, H.G. Duffy, J. S. Fukumoto, and L.M. Vogt. Federal Evaluation Policy. Washington: The Urban Institute, 1970.

Wortman, Paul M. "Evaluation Research: A Psychological Perspective." American Psychologist (May 1975):562-575.