EVALUATION RESEARCH

PRACTICE MANUAL

frank maidman associates

1989

1.WHAT IS EVALUATION RESEARCH?

"Evaluation research is the systematic application of of social research prtocedures ti assess the conceptualization and design, implementation, and utility of social intervention programs" (Rossi and Freeman,Pg 20).

(a)Fundamental evaluation questions

Evaluation research assists policy-makers, funders, planners and program staff answer the following questions:

.What is the nature and scope of the problem requiring actions?

.What interventions may be undertaken to ameliorate the problem significantly?

.What is the appropriate target population for the intervention?

.Is the intervention reaching the target population?

.Is the intervention being implemented in ways envisioned?

.Is it effective?

.How much does it cost?

.What are it's costs relative to it's effectiveness and benefits?

(b)DECISIONS, INFORMATION NEEDS AND PROGRAM EVALUATION

Program evaluation is applied research which helps provide different types of information for managers, program staff, funders, Boards of Directors, and other decision makers.

When deciding upon the kind of evaluation research to use in any given project, and the kind of design and scope, it is useful to ask oneself (a) what kinds of decisions or actions will be taken with the research results? and (b) what information will best aid such decisions or actions?

Evaluation research can help to ...

1.Make decisions. Three modal alternatives are (Rossi and Freeman)...

.the go/no-go decision

.developing a rationale for action: the conceptual use of evaluation

.legitimation and accountability

2.Reduce uncertainty about the real impact of a program or what specific things to do to make it better (Austin, Pgs.11-12),

3.Develop policy and administration

Policy evaluation (i) have potential impacts on large segments of the population or (ii) result in major organizational changes in the structure and activities
of groups delivering interventions (iii) are critical to the allocation to monetary, staff and other resources.

The purpose and eventual use of evaluation results may depend on the perceptions of key actors in the process. Also, the uses to which results are put may change as program organizational or community conditions change (eg. O.T.F.S.).

Several kind of decisions are possible:

1.The decision to end or continue a program

2.The decision to modify a program so that it better reflects initial planning, or that it has a better chance of reaching it's stated objectives.

2.PROGRAM DEVELOPMENT AND PROGRAM EVALUATION

The broad evaluation questions in the above section questions reflect three different foci for evaluation:

1.Program conceptualization and design

2.Monitoring and accountability of program implementation

3.Assessment of program utility (impact, efficiency)

Each type of evaluation is appropriately done at different stages of program development: innovative ones, those needing modification or fine-tuning, and established programs.

(a)How evaluation can contribute to an innovative program

Innovative programs are (i) those still in an emerging or research and development phase (ii) those in which the

delivery system or parts of it have not been tested (iii) the targets of the program are new or expanded, (iv) those in which a program originally undertaken in response to one goal is continued or expanded because of it's impact on another objective.

To assist the development of an innovative program, evaluators can...

1.Help clarify goals and objectives

2.Facilitate program design and development

3.Help design the delivery system

Each of these three functions will be discussed in the following sections.

(i)Methods for setting goals and identifying objectives

To assist the early stages of program development, evaluators can help management and staff be clear and specific about program goals and objectives.

1.The decision-theoretic approach (summarized by Rossi and Freeman,1982; detailed by Edwards, 1975, see Rossi). Involves data-gathering on preferred objectives and ranking by diverse groups, feedback of results, explication of reasons and reordering, until decision is reached as groups take into account their diverse views. A "delphi" process.

2.Evaluability assessment (see later sections)

3.Goal-attainment scaling (summarized by Rossi, 1982; detailed by Kiresuk, 1973)

4.Other techniques , particularly good for program management and staff can be found in Morris and Fitz-Gibbon, 1978.

Research for fine-tuning is done (i) to increase the efficacy of their efficiency, i.e. to increase the magnitude of their impact or decrease costs per unit of impact (ii) to provide equitable service delivery (iii) to reduce drop-outs from the target population

Established programs require evaluation usually to demonstrate
their effectiveness and efficiency.

(ii)Facilitating program design and development

1. Help design an "impact model" (sometimes called a "program model", "service model", etc.)

To undertake a successful evaluation, some statement of agreed-upon objectives, and how they are to be achieved, is necessary. The functions of an impact model are...

.to help understand why a program worked or did not work.

.to help reproduce it's effects on a broader scale and in other locations

.see also draft service model to N.C.F.S.T., 2(b), "Why is a service model important? " for other functions.

Impact models are formalized statements or hypotheses which guide the actions of a program, actions which strive to regulate, modify, or control community conditions. Included are ...

.expected relationships between a program and it's goal;
.strategies for closing the gaps between the goals and conditions or behaviors

.causal hypotheses

.hypotheses about interventions

.an action hypothesis.

(2) Contents of an impact model

(a)Causal hypothesis

A hypothesis about the influence of one or more processes or determinants on the behavior or condition that the program seeks to modify. The hypotheses should be stated in such a way that permits testing or measurement. Such operational measures are not the only statements consistent with the causal hypothesis.

(b)Intervention hypothesis

A statement specifying the relationship between what will be done in the program and and the process or determinant affecting the behavior or condition to be changed, as specified in the causal hypothesis. Other intervention hypotheses may be consistent with the causal one.

(c)Action hypothesis

Specifies the relationships between the natural (i.e. non-program induced) processes that may change the problem behavior, and the problem behavior.

(3)Sources of hypotheses

.experimental studies that permit testing of causal hypotheses

.well-developed theories

.clinical impressions

.statistical association studies

.results of other action programs

(4)Manipulability and feasibility

.intervention models must specify intervention variables that can effect targets directly or indirectly; eg. variables operating in the past (eg. previous socialization) are not good candidates.

.avoid selecting intervention variables with low feasibility, due to: lack of program acceptance by various stakeholders, ideological factors, community conditions, risked side effects, lack of technical knowledge, political factors, etc. (implementation framework?)

(5)Target populations

The impact model must lead into a specification of direct and indirect target groups.

.Direct target groups are those immediately subjected to the intervention

.Indirect target groups are those who benefit eventually as a result of those receiving service

The impact model should clarify the assumptions concerning how services to the indirect group eventually will affect the indirect target.

Models should also clarify the acceptance of the program by the target groups, as well as the conditions supporting or

inhibiting involvement in the program. Wrong assumptions may require dramatically shifting the program.

(iii)Facilitating Design of Delivery System

To document and assess a program, elements of the delivery system must be explicated and criteria of performance developed and measured. Elements include...

.identification of the target problem and population

.procedures and services provided

.qualifications and competencies of staff

.mechanisms for recruiting and obtaining cooperation of targets

.means for optimizing access, including location, physical characteristics

.referral and follow-up efforts

THE USE OF FORMATIVE EVALUATION TO AID PROGRAM DESIGN AND DEVELOPMENT

Formative evaluation during the program design and development can contribute in the following ways:

.developing the delivery system

.selecting targets

.structuring the intervention

.mini-impact studies

.to test evaluation procedures and instruments

THE USE OF SIMULATIONS TO AID PROGRAM DESIGN AND DEVELOPMENT

Simulations may be less costly and time-consuming than formative evaluation.

.Some are highly quantitative and formal

.Computer-based modelling

.Qualitative scenarios concerning certain courses of action

.Simulation of results from studies of similar programs

.Dummy tables showing estimated utilization and impact estimates; these can help identify inappropriate evaluation questions or suggest other indicators of impact.

(Rossi and Freeman, 1982)

EVALUATING ESTABLISHED PROGRAMS

Several different kinds of evaluation studies are typically conducted in the evaluation of established programs:

1.Evaluability assessment studies (to be inserted)

2.Accountability studies

(a)Impact accountability

(b)Coverage accountability

(c)Service delivery accountability: implementation

(d)Efficiency accountability: impact in relation to program costs

(e)Fiscal accountability: account for use of funds, costs per client, per service, locations, etc.

(f)Legal accountability: eg. informed consent, protection of privacy, community representation on Boards, equity in service provision, cost-sharing, compliance with legal requirements, etc.

Accountability studies are provided...

(a)On a continuous (Management Information System) versus cross-sectional basis (individual periodic studies)

(b) By insiders or outside consultants

FINE-TUNING ESTABLISHED PROGRAMS

Program modification which impacts markedly on intervention efforts.

Three sets of activiuties are usually involved:

(a)Reappraising objectives and outcomes

Awareness that program has failed to meet communiuty concerns requires modification of objective and outcome criteria. The need to redefine objectives comes from...

.after program implementation, dialogue during administrative and day-to-day activities

.special studies for consultants

.internal management information services

(b)Reputability assessments

Systematic efforts to obtain from stakeholders (including targets, service providers, and others) opinions and experiential data on which to judge the program's success in meeting it's objectives. Through special study or on-going monitoring effort. Questions may concern such things as waiting lines, fees, relations with practitioners,, desired services , unmet needs, etc.

(c)Programming Replanning and Redesign

Implementin fine-tuning refinements may require...

.redefinition and description of the problem

.operationalization of objectives

.revised impact model

.redefinition of the target poulation

.delivery system be redesigned

EVALUATION RESEARCH AS DIAGNOSTIC PROCEDURE

Evaluators'contributions to the identification and ranking of human and social deficiencies, and to the innovation and refinement of programs, has been rare. Such functions are typically fulfilled by politicians, advocacy groups, mass media and charismatic personalities.

However, evaluators can contribute systematic approaches to (i) identification of communal problems, and (ii) clarification of their scope.

Diagnostic research helps prevent problems in implementing programs. That is, preliminary research helps avoid problems in which ...

.the program is not delivered

.the target population did not exist, or was incorrectly identified

.target population did not seek the program

.target poulation made demands that the progranm was incapable of meeting

Diagnostic research activities are used to undertake ...

.needs assessments

.conceptualization of program targets

Techniques to accomplish such activities include ...

.key informant approach

.the community forum

.the rates under treatment approach

.social indicators

.surveys and censuses

(Rossi and Freeman, 1982, pgs.111-121)

(a)NEEDS ASSESSMENTS (Rossi and Freeman, 1982, Pgs.93-103; McKillop, Need Analysis, 1987.)

This refers to the process of verifying and mapping out the extent and location of a problem and the target population. Used to describe the program-relevant characteristics of targets.

(i)What is a target?

.define the unit of analysis clearly

.differentiate between direct and indirect targets

.specify the assumed relationship between direct and indirect targets

Targets can be specified by..

.establishing appropriate boundaries, i.e. rules of inclusion or exclusion, which are neither so broad as to include everyone or so restrictive as to exclude most people.

.establishing definitions which are feasible, i.e. ones which lead to an observable and manageable number of criteria.

.including the various possible stakeholder perspectives on needs, hopefully leading to reconceptualizing the problem or intervention, or perhaps abandonning the program.

(ii)Conceptualizing program targets

Necessary to distinguish between target population and non-target units during program implementation. This can be done by...

.identifying the population at risk in probabilistic terms

.identifying a population at need, i.e. who currently manifest a given condition, as indicated by a specific criterion; the concept "need" should be distinguished from "demand".

.deciding whether it is more important to design programs based on incidence (new cases in a given time period) or prevalence (existing cases in a time period)

.knowing the rate of a particular problem i.e. the percentage of a particular unit number (eg. by sex and age group) based on demographics, or particular problems related to program implementation (eg. accessibility to program)

(iii)Selecting program targets

.overincluding the people not needing the program is (a) cost inefficient, or (b) unlikely to show an impact in evaluation

.underinclusion (a) possibly denies participation to those in need (b) may requie costly screening devises (c) which label certain people (c) possibly generates attagonism from those not include with those in need

(b)Needs assessment research techniques

(i)Key informant approaches

See Rossi, pg.111-112; and McKillop, Pg.81; and F.M.A. file

(ii)Community forum

See Rossi, Pgs. 113-114.

(iii)Rates under Treatment Approach

The estimation of target populations based on the utilization of services of similar people in similar communities. Likely a conservative estimate because of less than full coverage.

(iv) Social indicators studies

Estimates program need and target characteristics based on existing statistics, preferably time-series data, which give demographic, problem and other data known to correlate with problems. Much of these data are only available for large geographic areas, thus requiring reconstruction for analysis for smaller areas within cities.

(v)Surveys and censuses

3.PROGRAM MONITORING AND ACCOUNTABILITY

Evaluation research is also used to monitor program implementation. Implementation monitoring refers to a systematic attempt to measure (a) program coverage and (b) program process: extent to which service matches what was intended to be delivered. Together, these are also referred to as "outputs", as opposed to outcomes.

(a)The uses of monitoring

(i)Aids the development process by ...

.identifying problems of implementation

.documenting unexpected results and unwanted side effects

.aiding program diffusion by providing program descriptive data, critical points in implementation, potential managerial problems and solutions, administrative manuals, service delivery, personnel qualifications, etc.

(ii)Aids programs beyond the development stage by...

.providing information on coverage, process, and cost, leading to fine-tuning

.can be done through use of management information systems

(iii)Contributes to program accountability: coverage, service delivery, fiscal accountability, legal accountability.

(b)Foci and procedures for monitoring coverage

(i)Assessment of how programs arrange for, and facilitate the motivation and access of potential clients.

Two important concepts are:

."Coverage" refers to the extent to which a program obtains target population participation, as specified in program design.

."Bias" refers to the degree to which sub-groups of the target population participate differentially. Affects validity of impact evaluations.

When doing an assessment of coverage, look for ...

.factors affecting self-selection

.the effects of program actions and personnel

.location

.the actual impacts, negative or positive, of under- or over-coverage (see Rossi, pg.130) in terms of program goals and costs.

(ii)How to measure coverage

.coverage formula: (see Rossi, pg.131) for a formula combining over- and under-coverage.

.use of existing data from the organization's management information system records.

.surveys of program participants: periodically done in lieu of regular information-keeping; sampling basis.

.community surveys: particularly if the target group is community-wide, even with sub-groups as primary target groups.

.analysis of program drop-outs : including qualitative data on the experience of the program; also can use existing records, community surveys (non-participants)

(c)Monitoring Delivery of Services: Implementation and Process

(i)Why do implementation studies?

.decisions re program continuation or expansion

.are intended program specifications met?

.level of staff performance

.program failure is frequently failure to deliver program as planned

(ii)Why do programs fail?

.non-program or dilutted treatment: services are either not being delivered at all, or else are very limited in scope.

.wrong treatment: either (a) mode of delivery negates treatment (poor attitudes, poor training, resource failure) or (b) overly sophisticated delivery system

.unstandardized treatment in which different sites vary in the quality of their treatment as a result of local variations in support, staff and othert characteristics.

(iii)What delivery system features can receive attention in implementation evaluation?

.access:the structural and organizational arrangements that provide opportunities for and operate to facilitate program participation (eg. opening an office, outreach campaigns); evaluation can focus on

(i) consistency of access operations with program design
(ii)level of client participation and drop-out

(iii)accessibilty of potential targets to appropriate services

(iv)variation in utilization across various sub-groups

(v)participant satisfaction

.specificity of services: operational specification the actual services provided; program elements (units of service) can be defined in terms of time, costs, procedures, product.(See Rossi, pgs. 148-150 for principles on the necessary details for describing program elements.)

(iv) Methods for collecting information for evaluation of program implementation (Rossi and Freeman, 1982, pgs. 150-160; see als my notes on implementation evaluation literature, in later parts of this manual)

.observational data

.service record data

.service provider data

.program participant data

(v)Analysis of monitoring data

Analysis should address at least three issues: project description, comparison between sites, and program conformity.

.Project description: coverage, bias, service types, service intensity, participant's reactions; can utilize both narrative description and sophisticated quantitative analysis.

.Comparison between sites: sources of program diversity; facilitates efforts to achieve standardization; comparative effectiveness.

.Program conformity: discrepancies leading to program respecification, move project closer to design, judge appropriateness of impact evaluation or formative.

2.TRAINING AGENCIES IN EVALUATION RESEARCH

(i)Reasons for doing evaluation research

Agency management, staff and Board must value the importance of evaluation research before estaablishing an internal capability.

Reference

.Austin and Associates, Evaluating your Agency's Programs Pgs. 10-11.

3.<u>What considerations are taken into account in designing an evaluation?</u>

(i)Utility

(ii)Types of decisions to be informed

(iii)Technical considerations: will the data provide a fully accurate answer to the evaluation question?

(iv)Available resources

(v)Developmental stage of program

A.<u>Arguments in favour of comparison group design</u>

1.Strengthens evaluator's position against attempts to discredit the results based on political or other vested interest considerations

-Fitz - Gibbons, pg.10-11

B.<u>Reasons for the lack or inadequate use of comparison group designs</u>

1.Funder perceptions of programs as one-shot effort

-Fitz-Gibbons, pg.12.

2.Evaluators called in too late for appropriate design

-Fitz-G., pg 12.

3.Difficulty of designing good evaluation because of ethical, legal, therapeutic or political reasons (eg. with-holding treatment from some, asking non-program sexual abuse victims to talk about their experiences)

4.Youth of social science

- good experimental design lacking (Fitz-G., 13)

5.Researchers cannot agree on appropriate design

B.Formative Evaluation

1.Comparisons between formative and summative evaluation

F-G, pg.11.

2.Major contributions to program development

(i)Encourage staff to scrutinize and re-think assumptions

(ii)Test alternative courses of program action

3.The use of comparison group design in formative evaluation

(i)Alternative versions of programs (Fitz-Gibbons,16)

(ii)Relaxing strict design requirements (F-G.,17)

(iii)Conduct short experiments to test different program alternatives (F-G,17)

4.Case examples of formative evaluation in education

(i)F-G., pgs. 15, 16, 18, 20, 21, 32,

5.Case examples of formative evaluation in business

(i)Marketing F-G.,17.

(ii)Reducing absenteeism F-G.19

(iii)Employee Assistance program evaluation F-G, 33.

6.Using comparison design in situations with special client populations

Some organizations serve people (eg. clients with certain disabilities) who by law or other ethical considerations should not be denied a program. Establishing a pure control group design is difficult in such circumstances. The following evaluation designs are recommended:

(i)Non-equivalent control group design (eg. another setting with no program or a different program)

-F-G. 21
(ii)Use a formative approach and evaluate program components

-F-G, 21

(iii)Compare diverse programs in terms of a common indicator

(iv)Compare program outcomes to pre-established criteria (eg. blood pressure range, ability to read road signs, etc.)

-F-G,22

(v)Conduct theory-based evaluations

-correspondence of program rationale to good theory, F-G,23

-implementation of program aaccording to theory
F-G, 23.

STRATEGIES FOR IMPACT ASSESSMENT

1.What is an impact assessment?

An impact assessment is designed to establish whether or not an intervention is producing it's intended effects. Estimates are not made with absolute certainty, but within limits of error, and with varying degrees of plausibility. Therefore systematic and rigorous research design is necessary.

Outcomes are assessed by comparing information about program participants before and after the intervention, or by less powerful designs. The aim is to reject alternative explanations of outcomes.

2.When is an impact assessment best conducted? And prerequisites?

(i)To test new or proposed programs, or changes in existing programs.

(ii)Review of existing or on-going programs.

(iii)The project has objectives defined in order to measure goals. Or evaluator establishes objectives.

(iv)Intervention should have been well-implemented so that critical elements of program have been delivered to targets.

3.The critical questions in impact assessment

(i)Does a particular program produce more of an effect than would have occurred without the program or with a different kind of program.

(ii)What are the net outcomes of the program, that is those outcomes produced by the program. This contrasts to gross outcomes which are those net outcomes plus confounding factors (events, experiences, etc.). (Rossi, 1982, pgs. 169-170)

4.What are the various confounding factors that impact design and analysis must try to rule out when trying to

estimate the true program effects?

(A)Extraneous factors

(i)Endogenous change. Changes that result from natural processes or events in the program participants' lives, causing the condition to change of own accord.

(ii)Secular drift. Long term trends in the community or country in question which enhance or mask the program effects.

(iii)Interfering events. Short-term masking or enhancing events.

(iv)Maturational trends. Developmental changes in people's lives.

(v)Self-selection. Those target population easiest to reach and involve in the program are also those most likely to change in the direction specified by the program. Also drop-outs are likely to be different from those whom stayed.

(B)Measurement error

(i)Stochastic effects. Chance or random fluctuations in a particular sample, given the true outcomes. Tests of statistical significance allow one to estimate how often a particular result (i.e. difference between control and experimental groups) would happen by chance, if there were no real differences.

(ii)Unreliabilty of measurement. This is lack of consistency in measures using the same instrument, when measures are taken at different points in time with the same or another group.

(C)Intervention-related obscuring factors

(i)Hawthorne effect

The effect of being in the experiment, typically mediated by the meaning that an experiment has for the participants (eg. "I'm special")

(ii)Delivery system contaminants

Physical plant, personnel, rules and regulations, labelling of targets, etc. alll have an effect on program participants.

Note: The confounding effect of measurement error and delivery factors are most seriously consid. when the program impact is weak.

5.What are the design options for impact assessments?

(i)Randomized controls

(ii)Constructed controls

(iii)Statistical controls

(iv)Reflexive controls

(v)Generic controls

(vi)Shadow controls

These options are not mutually exclusive. For example, statistical controls can be used in random assignment designs.

6.Factors affecting the choice of impact assessment design

(i)Degree of coverage of target population.

In total coverage situations, it is not possible to find people who have not received the program for control group purposes. Reflexive controls and before-after data collection and comparisons are necessay. Eighty percent or over should be considered full coverage, because small proportios are likely to be significanly different. Natural variation (eg site differences) in program delivery may be used as the basis for group comparisons.

C.THE ELEMENTS OF EVALUATION DESIGN

1.Who is the group receiving the program or treatment?

.individuals? departments? schools? whole organizations?

2.Who is the group acting as the control group?

.is it a true control group or a comparison group?

3.What methods are used to establish the control group?

(i)Randomization as the best method for establishing a true control group; randomization accomplishes the following:

.helps to rule out alternative explanations concerning program effects, resulting from the characteristics of the people in the experimental and the control groups.

.F-G, pg.27

.Steps in the randomization process

(ii)Use another new program in lieu of testing people with no exposure to an experimental program

-F-G,28

(iii)Use people in the control group who are less in need of the program (borderline control group strategy)

(iv)Taking turns strategy

.first one group then another is given the program

(v)Delayed program strategy

.delay entry of a randomly selected control group, who will eventually receive the program; use the delayed group as a control group until they go into the program

(vi)Delayed evaluation strategy

.if too late to form a control group, use less credible design and lay groundwork for stronger design

4.Methods for establishing a non-randomized control group (comparison group)

Sometimes it is necessary to find a comparison group wwhich although not a true control group, is as similar as possible to the experimental group. Often such non-equivalent control groups are selected from other locations where similarities can be proven or strongly assumed. Here are some principles for using non-eqivalent groups:

(i)If the experimental group is selected according to some pre-test procedure (eg. math scores) select the non-randomized comparison group, using similar or same procedures.

(ii)Give both groups the same tests throughout the evaluation, so the testing itself does not make a difference to the achievement outcomes through focused attention, practice on tests, affecting motivation, etc.

(iii)Document all similarities and differences between the two groups, in order to answer skeptics' questioning. It is likely useful to know the suspicions of the critics beforehand so that data-based responses can be made.

5.What program should the control group get?

Keep in mind that program evaluation provides information to assist in decision-making. Optional decisions could include "no program" "another program", a "modified version of the experimental program", or whatever.

(i)Best solution. The control group should receive a representative of the alternative couse of action, whether it be no program, or some other program.

(ii)Next best solutions

A program with similar aims and objectives.

.The old program

.A program never used, and never intended.

(iii)Weakest solution. The use of no program at all even though this option is not being considered. This option helps...

.To decide whether the program is better than no program at all

.To asses whether the program does harm.

.To establish some base-line data, against which the experimental group results can be compared. Controls for maturation, the possible change in program impact due to the growth of the participants.

(iv)When comparing programs of similar aims...

.Make sure that comparisons are done on those factors that clearly distinguish the programs, i.e. the experimental program should be able to produce better results in areas representing it's own special emphasis

.Report three sets of results: one set for each programs major distinct goals, and one set for those held in common.

6.Appropriate implementation of the experimental program

.F-G, 35.

7.The timing of measurements

(i)Post-tests

1.If determined by the timing of the report, leave yourself twice as much time as you think you need to write the report.

2.Be alert to those special occasions in the school or work life of the program participants, occasions which could affect program results.(eg. school breaks, fiscal year dead-lines, holidays , seasonal changes in company activities, etc).

(ii)Pre-tests

Pre-test measures can be used to...

.select people

.check assumptions in program planning

.check comparability of groups

.establishing basis for checking program gains

.obtain sensitive test of program effects

(iii)Pre-tests to select people for the program

1.The problem of regression or regression towards the mean. Beware of using pre-test measures as program outcome tests, particularly if extremely high or low scoring people were selected. The group scores may regress towards the mean on subsequent scoring, meaning that extreme scoring people may change regardless of program effect.

2.What to do about the problem of regression in the selection process?

.Have a control group.

.If a non-equivalent control group, assure that control and experimental group members come from similar home, school, or work backgrounds (F-G,38)

.When selections are made from low or high scores, test again, because second tests are more likely to be normally distributed.

(iv)Pre-tests to check program assumptions

.i.e. proper implementation of the program (eg. characteristics of participants as specified?)

(v)Pre-tests to assure group comparability

This assures that two groups are similar to begin with, particularly if randomization has not been done (pre-test is essential here) or if done with small numbers (less than 15 per group). Also, pre-tests can be used to assist the randomization process ("Blocking").
(vi)Pre-tests to check gains made during program.

1.Not necessary in case of random assignment to experimental and control groups, unless concern of non-similarity due to small numbers or high variabilty on important variables.

2.Necessary when using a non-equivalent control group.

3.Necessary when criterion-referenced tests are used (pre-determined criteria referencing mastery of specific subject matters, are used to examine specific content changes). (F-G,40)

4.Inappropriate to use pre-tests where standardized tests are used. (F-G,40)

(vii)Pre-tests to increase the sensitivity of program test

1.Use pre-tests when program impacts may not be that evident and more powerful methods are needed to detect small differences.

2.In this case, the researcher gathers information to help explain the differences observed, particularly when the changes could be strongly affected by initial level of the quality being addressed by the program. (eg. initial knowledge or attitudes). Deviations in outcomes from what was expected by the initial score may be attributable to the program.

(viii)Pre-tests should not be used when...

1.A true control group is part of the design, since randomization typically assures similarity in the two groups.

2.Pre-testing might alter the participants in some way. For example, attitude tests lead people to focus on certain ideas, or to guess what the program evaluation is trying to find out. To measure initial attitudes, particularly in large groups (eg. 30 or more), take a random 1/2 of those receiving the program.

3.When using one would be meaningless, as in any situation where the participants could not possibly have knowledge or an attitude towards something (eg. program for teaching French to non-speaking people).

4.When the program is already in progress. Unless you can assume that the program progress has not affected certain factors, retroactive pretesting may usefully describe the relative standing of the two groups before the program started.

5.It is costly in time and money.

Mid-tests

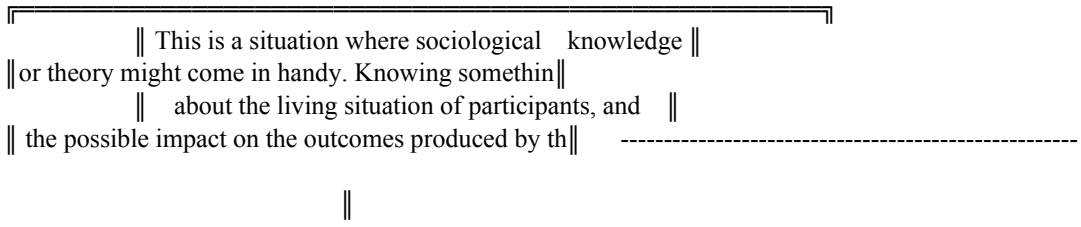
Can be used for...

1.Showing the impact of the program across time; some programs don't have impacts until some time has elapsed; Of particular interest might be the question of when
program effects take place, and what things are going on in the program.

2.Situations where a control is lacking

3.Tracking the impact on sub-groups of persons exposed to the program

4.Situations where the impact of the program might lessen after the program ends (eg. summer vacation). Another test showing results after the program has stopped gives added conviction tro the strength of the program.



5.Time series tests are paticularly useful when a control group is missing, creating the need for trend data. Taking several measures before the program begins, gives you an idea of what changes to expect without program intervention. Taking another set of data after the program has been implemented allows comparisons with program intervention.

(F-G, 46-48)

TYPES OF DESIGNS

1.True control group: pre-test, post-test design

2.True control group: posttest only design

3.Non-equivalent control group: pretest, posttest design

4.Single group time series design

5.Time series with a non-equivalent control group

6.Before and after design, experiment group only.

(These designs are briefly summarized in F-G, pgs. 55-62)

MAJOR THREATS TO THE IMPLEMENTATION OF EVALUATION DESIGNS

A. Experimental or quasi-experimental designs

1. Differences between experimental and control group participants in amount (and quality-F.M>) of time spent in the program.

2. Attrition

3. Confounds are things happening extraneous to the program which affects one, but not the other, group. Large numbers of experimental participants help to remove this problem, through the "averaging out" phenomenon.

4. Contamination is the problem of the control group somehow being affected by the experimental group (eg. idea-sharing between therapists). This is minimized when the two groups are located in different sites.

(These implementation problems are discussed by F-G on pages 62-63)

Rules for selecting appropriate evaluation designs

1.Possible to select more than one design for the same evaluation or experiment, depending on what is being measured, availability of control groups for all parts of program.
(F-G,49).

2.If the evaluation can be planned before the program starts, try for the strongest possible design, the true control group. If not possible, work through the next strongest: non-equivalent control group, time-series design, and (last resort) pre-post design.(F-G,49)

3.If planning evaluation after program has started, check how participants were selected. If truly random, you may be able to conduct a true control group design, by randomly creating a control group, and comparing test scores. (F-G,49)

4.Review the Fit.-Gib summary of choices based on two design elements: participants and timing of measures.(pgs. 50-53)

PROGRAM EVALUATION DESIGNS: STEPS, ANALYSIS AND CAUTIONS

1.True control group: pretest-posttest design

(A)Steps in implementing design

(i)Identify people or groups some of which could get the program

(ii)Pretest all people

(iii)Randomly assign some people or groups to the experimental group and some to the control group

(iv)Experimental group gets the program being evaluated; control group gets an alternative program or none at all.

(v)Posttest both groups with the same instrument under identical conditions.

(B)Content of report:other than data

(i)Program implementation: minimally, did the experimental group receive the program?; see material on implementation for other possible material on implementation; explai gaps in implementation and any possible implications for outcome results.

(ii)Contamination: document any contamination, and possible implications.

(iii)Confounds:examine any confounds which may have affected the differential results.

(iv)Attrition: include table summarizing attrition resons for both groups; explain any differences in attrition rates for the two groups, interview drop-outs from both groups if possible.

(C)Analysis, reporting and discussion of results

Depending on the type of data, the following data analysis methods are used:

(i)analysis of differences between mean scores of experimental and control groups on pretest and posttest scores; tabular presentation summarizes data, including both sets of scores, standard deviations, and t-tests, showing statistical differences. (F-G,69-70)

(ii)analysis of variance is a more powerful statistical technique, to be tried if above analysis shows no diff.

(ii)graphic presentations are quickly understood and interpreted (F-G,71)

(iii)if analysis of pretest scores shows signifant differences several options are possible:

.reassign people or groups to programs,
perhaps using blocking, matching, or stratification techniques.

.drop any possible extremely low or high scores from the analysis

.recheck the randomization procedures and correct any deviations from good randomization.

.treat design as non-equivalent control group.

(iv)when analyzing posttest scores:

.do not test difference between gains on scores, because the higher group is usually penalized.

.gain added strength by presenting confidence limits, ie. how often the differences would happen over 100 tests.

(v)analysis of results other than means

.where appropriate, use other summary statistics
such as percentage differences; adapt table or graph presentations to any single number expression of differences.

.testing on several program objectives is difficult and not practical to summarize in single summary measures; options are (74-77)...

.use a single mean or proportion that seems to adequately represent group performance

.augment single number analyses with detailed presentation of program results

.use graphs (bar and line graphs) and tables to summarize findings per objective (75)

(D)Describing the practical significance of program effects

Because statistically significant differences do not necessarily communicate clear practical implications (a question of value),
steps must be taken to address these:

(i)use staff to examine the questions in the instruments and identify particularly meaningful ones in terms of their practical judgements and expectations.

(ii)compare program results on different types of participants (eg. high gain in skills or self-esteem among those who have been out of work)

(iii)compare score differences to national norms

(iv)do item analysis to ascertain how the two groups faired on different types of items, and what is the perceived significance of the patterning.

(Items i-iv are reviewed in F-G,79-80)

(v)compare quantitative analysis to parallel qualitative data which gets at various indicators of value. (F.M.)

2.TRUE CONTROL GROUP: POSTTEST ONLY DESIGN

(A)Steps in implementing design

(i)Identify people or groups, some of which would get the experimental program

(ii)Randomly assign some people, groups or organizations to one or other of experimental or control group

(iii)Experimental group gets new program, control group gets none, or an alternative program

(iv)Posttest both groups with the same test under identical conditions

(B)Content of report other than data

(i)Randomization procedure: Particularly important to warrant absence of pretest.

(ii)Program implementation

(iii)Contamination

(iv)Confounds

(v)Attrition

(C)Analysis, reporting and discussion of results

(i)Use tables or graphs to display data (F-G,83, 84)

(ii)If simple or blocked randomization, use t-test for unmatched groups ("non-correlated t-test")

(iii)If matched randomization, use t-test for matched groups.

(iv)If significant differences do not appear on global analysis of differences between means, do an internal analysis of data (eg. analysis of variance) to determine if program worked better for some individuals. Use tests of significance to guard against drawing conclusions based on small sub-samples. Also, have a rational for internal analysis.

(D)Practical significance of program effects [see 1 (D)]

3.NON-EQUIVALENT CONTROL GROUP: PRETEST, POSTTEST DESIGN

(A)Steps in design

(i)Identify people who will be getting experimental program

(ii)Identify people who will not be getting program, but who are otherwise as similar as possible to experimental group.

(iii)Collect information about the ways the two groups are alike and different

(iv)Pretest both groups

(v)Experimental group gets new program, control group does not.

(vi)Posttest both groups

Variations on the basic design

Such "quasi-experimental" variations have been much discussed in the literature because of the basic practicality of the design, in comparison to the true randomization model:

(i)Regression projection model: (see Horst, Tallmadge and Wood, 1975)

(ii)Regression-discontiuity model: regression models, and uses participants in control groups who clearly "outperform" the experimental group members on the pretest. (Cook and Campbell, 1976)

(iii)Variation in the control group's programs and the the types of post-tests.

(iv)Methods to control for the effects of cognitive development in children

(B)Content of report other than data

(i)Detailed information on the similarities of both groups, particularly in relevant characteristics thought to possibly affect the outcomes. Also similaritry on whatever the program intends to change.

(ii)As in other designs, report on implementation, possible contamination, confounds, attrition.

(C)Analysis, reporting and discussion of data

(i)Tables and graphs; analysis of differences between mean scores on pretest and posttest; t-test or confidence limits, using t-test for unmatched groups (F-G, 90,91).

(ii)If pretest analysis shows difference between the two groups, several things can be done. Some of these analyses have been criticized as useless by other researchers.

.Analysis of covariance: adjusts postest scores up or down according to pretest performance; requires assumption of strong similarity between groups on other variables.

.Post-hoc matching: in which people from both groups are matched after the fact on a person by person basis; t-tests are then calculated for these new sub-samples.

.Analysis of gain scores: in which the gain scores between pre- and posttest measures are calculated, and difference of mean analyses are computed on these new scores.

4.SINGLE GROUP TIME SERIES DESIGN

(A)Steps in design

(i)Prepare or select an outcome measure which can be used repeatedly.

(ii)Decide on composition of experimental group: same group measured repeatedly? successive groups of different people?

(iii)Collect at least three measurements prior to program, and made at regular intervals; methods of measurement must remain the same.

(iv)Check implementation of the experimental program.

(v)Continue to collect measures at same intervals even after program conlusions.

If possible seasonal variation in outcomes or correlates could affect, data so that test scores can be compared from the same time of the year. This, of course, may not be practical.

Possible variations in samples

(i) all members of experimental group measured

(ii)randomly selected members measured

(iii) successive groups, deemed representative of the program participants, are measured.

(B)Content of report, other than data

(i)implementation

(ii)confounds; one way of assessing confounds is to exanmine the speed of changes. Do external events seem to have an effect on the rate of change over time?

(iii)changes in methods of obtaining measurements

(iv)changes in composition; to counter, collect other data to examine and rule out other explanations through interrnal analysis of data.

(v) program introduced in response to crisis

(C)Data analysis and reporting

(i)Statistical analysis of time series data is complex, needing advanced technical expertise. However, simple examination of trend data can be done.

(ii)Plot summary statistics scores on a graph in which the horizontal axis represents time, and the vertical axis represents the outcome scores. Indicate clearly the time period during which program began and ended. (see F-G, pgs. 104-105)

(iii)Look for trend changes (ie. alteration in the rate of change)
or jump changes (sudden increase or decrease in scores) (F-G, Pg.104)

(iv)Use lines analysis to view and interpret the time series results, (see F-G, pgs. 106-110), which includes a method for visually extrapolating the preprogram trend beyond the program period.

Interpreting time series graphs

(i)If trend changes or jump changes are evident, suggest that it was the program that produced observed changes. However, to be cautious, other questions must be raised:

(ii)Could some other change at or about the same time as the program possibly have produced the change? (group composition? measurement method? other events? crisis?

(iii)If the results are too unstable to permit conclusions to be drawn from inspecting the plotted data, a statistical analysis may be in order.

(iv)To consider possible different affects on different types of people, plot scores of different groups (eg. age, sex, etc.)

5.TIME SERIES WITH A NON-EQUIVALENT CONTROL GROUP

(A)Steps in design

(i)Identify group which will get program

(ii)Prepare or select an outcome measure which can be collected repeatedly

(iii)Locate group similar to experimental group, and from which you can collect observations on the outcome measure.

(iv)Collect at least three measures from each group at the same time before the experimental program starts.

(v)Check the implementation of both programs.

(vi)Collect observations at regular intervals from both groups.

(vi)After program ends, continue to collect data at same regular intervals if possible.

(B)Data Analysis

See notes under #4.

6.THE BEFORE AND AFTER DESIGN

Note: This is the least adequate design because no comparison data are gathered. Can be used for implementation studies, but avoid when evaluation question has to do with results.

(A)Steps in design

(i)Pretest all people involved

(ii)Document the implementation of the program

(iii)Posttest all people.

The few strengths available from this approach are:

.Not necessary to monitor two groups

.Possible to take more measures, more info, delve deeply into activities

Because of these possibilities great pains should be taken to do a good description of the program, examining such things as materials, activities, and the relationship to theoretical base. (B)Analysis and description of outcomes

(i)Often used in comparison to other norm-referenced achievement scores. i.e. gives the means of interpreting the

pretest and posttest results relative to the performance of a (possibly national) standardized group. The validity of this approach depends on how close the skills in standardized group are to the program skills. These considerations should be made very clear to lay audiences, along with the meaning of standardized tests.

(ii)Other suggestions for lending credibility to the before-after design are...

.When planning to compare program results to standardized group scores, (a) use standardized test to terst program material (b) do measurements at the same time as the standardized population, and (c) provide detailed info on the characteristics of the norm group

.Look at different program emphases (eg. results of different program applications in different sites)

.Examine differential impact of the program on people with differenty charactertistics

.Develop and try out many instruments which might be sensitive measures of program goals

.If possible, focus evaluation on program objectives , thus (a) pointing out program strengths and weaknesses (b) identify important objectives

DATA ANALYSIS METHODS

1.Difference of means test

This test, typically done with a "t-test", is used to make comparisons when the program results are measured as interval data.
(Eg. math scores, self esteem measures, etc.).

Note that different statistical tests are used depending on whether the experimental and control groups were assigned on the basis of (a) randomization or (b) samples are matched

2.Difference of proportions test

Used when the program outcome data is based on nominal or ordinal data.

3.Analysis of Variance (ANOVA)

ANOVA is used when the program evaluation focuses on three or more programs, and/or is interested in whether the different programs have different sorts of effects on different individuals in the programs. In technical terms, ANOVA is used to test the interactive effects between program and other variables.

References

1.Fitz-Gibbon, C. and Morris, L., How to Design a Program Evaluation, Chapter 7.

2.Blalock, H., Social Statistics , Ch.16.

1.

3. EVALUATION OF PROGRAM IMPLEMENTATION