PSY 502

Program Evaluation

Program Evaluation

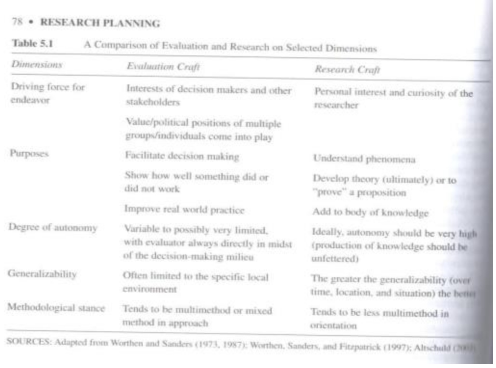
- Evaluation Research
- A process or set of procedures for providing information for judging effects of interventions designed to influence behavior or for determining merit and worth of process, product, or program
- Implemented in field to guide and facilitate decision making
- Most widely used type of applied research
- Can be quantitative or qualitative

History

- Grew in late 60s to provide “hard data” on many newly created govt. social programs
- Decline in 80s due to hostility to social programs and desire to cut spending

- Primary audiences are decision makers such as government administrators, legislators, school boards, and company executives.

Evaluation compared to Research



78 • RESEARCH PLANNING

Table 5.1 A Comparison of Evaluation and Research on Selected Dimensions

Dimensions	*Evaluation Craft*	*Research Craft*
Driving force for endeavor	Interests of decision makers and other stakeholders	Personal interest and curiosity of the researcher
	Value/political positions of multiple groups/individuals come into play	
Purposes	Facilitate decision making	Understand phenomena
	Show how well something did or did not work	Develop theory (ultimately) or to "prove" a proposition
	Improve real world practice	Add to body of knowledge
Degree of autonomy	Variable to possibly very limited, with evaluator always directly in midst of the decision-making milieu	Ideally, autonomy should be very high (production of knowledge should be unfettered)
Generalizability	Often limited to the specific local environment	The greater the generalizability (over time, location, and situation) the better
Methodological stance	Tends to be multimethod or mixed method in approach	Tends to be less multimethod in orientation

SOURCES: Adapted from Worthen and Sanders (1973, 1987); Worthen, Sanders, and Fitzpatrick (1997); Altschuld ([illegible])

Types of Evaluation

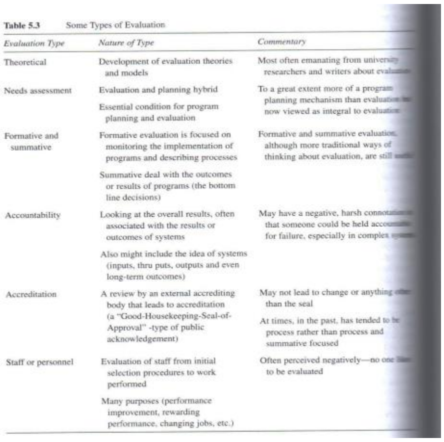


Table 5.3 Some Types of Evaluation

Evaluation Type	*Nature of Type*	*Commentary*
Theoretical	Development of evaluation theories and models	Most often emanating from university researchers and writers about evaluation
Needs assessment	Evaluation and planning hybrid Essential condition for program planning and evaluation	To a great extent more of a program planning mechanism than evaluation now viewed as integral to evaluation
Formative and summative	Formative evaluation is focused on monitoring the implementation of programs and describing processes Summative deal with the outcomes or results of programs (the bottom line decisions)	Formative and summative evaluation, although more traditional ways of thinking about evaluation, are still
Accountability	Looking at the overall results, often associated with the results or outcomes of systems Also might include the idea of systems (inputs, thru puts, outputs and even long-term outcomes)	May have a negative, harsh connotation that someone could be held accountable for failure, especially in complex systems
Accreditation	A review by an external accrediting body that leads to accreditation (a "Good-Housekeeping-Seal-of-Approval" -type of public acknowledgement)	May not lead to change or anything other than the seal At times, in the past, has tended to be process rather than process and summative focused
Staff or personnel	Evaluation of staff from initial selection procedures to work performed Many purposes (performance improvement, rewarding performance, changing jobs, etc.)	Often perceived negatively—no one likes to be evaluated

Stage 1: Goal Definition

- Goals must be defined conceptually and operationally
- Begins with needs assessment
  - The process of determining that a problem exists and of designing a solution for the problem
  - Theory underlying program should be explicated
  - Should include goals defined by stakeholders
    - The people who are affected by the program and its success or failure

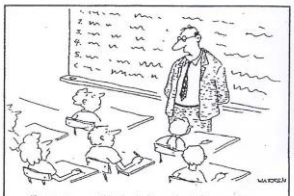
Common Stakeholders in a Program

- Policy makers: decide whether a program should be started, continued, or changed
- Program sponsors: fund the program
- Program designers: decide on the content of the program
- Program staff: deliver services to clients
- Program clients: use the services
- Opinion leaders: Community members who can influence public attitudes toward the program

Priorities

- Course Evaluation
- Student
- Teacher
- Colleagues
- Administrator
- Employer
- Board of Trustees

Can you help me, Mrs. Martin? This wasn't covered in any of my education courses.



'I expect you all to be independent, innovative, critical thinkers who will do exactly as I say'

Possible Sources for Identifying Problems

- Social indicators: Continuing measures of various social phenomena
  - Examples: Crime rates, unemployment data, admissions to psychiatric care facilities
  - An increase in negative social indicators suggests a need to address the problem
- Surveys conducted to assess people's perceptions of needed community services
- Policy maker's intuition that something is wrong

Developing Solutions

- Intervention program should identify
  - the client population: Who will be served by the intervention?
  - the content of the program: What services will be provided?
  - possible causes of the problem, ideally based on theory and research

Evaluability Assessment

- The process of examining a program to determine the information necessary to conduct an evaluation
- Need to know
  - the goals of the program
  - the expected consequences of achieving those goals
  - the expected impact of the program
  - the theory of the program that links the elements of the program to the expected consequences
  - the implementation guidelines of the program
    - Includes an explanation of how theory will be put into practice
  - the resources the program needs to achieve its goals effectively

Goal Specification

- To evaluate a program's success, need to measure goal attainment
  - Measurable goals are clear, specific, and concrete
  - Broadly stated goals can have many different meanings
  - ***Mission Statement****. Oakland University is a preeminent metropolitan university that is recognized as a student–centered, doctoral research institution with a global perspective. We engage students in distinctive educational experiences that connect to the unique and diverse opportunities within and beyond our region.*
  - Researcher must focus on meaning to be addressed in the proposed evaluation

Goal: Improve Education

Possible Points of View		
School board	School administrators	Teachers
Parents	Students	Legislators
Possible Meanings of “Improve Education”		
Students enjoy classes more	Students show more interest in school	Students can apply what they have learned
Possible Behavioral Indicators of Increased Interest		
Students participate more in class discussion	Students ask more questions	Students do outside reading

Goal Specification

- Each goal must be specified in terms of specific consequences that can be operationally defined
- Consequences that should result from the program fall into three categories
  1. What the client should *know*
  2. What the client should *believe* (changing beliefs ≠changing behavior)
  3. What the client should *do*

Goal Specification

- Measures must be developed for each specified program outcome
- Specification process must also lay out the expected impact the program will have on each measure, including
  - the expected timing of effects
  - the magnitude of effects
  - the durability of effects

Goal Specification

- Researcher needs to determine the relative importance of the goals and their consequences
- Necessary for two reasons
  1. The available resources might not support assessing all goals or all consequences of a goal
  2. Success of the program is best evaluated by considering the relative importance of the obtained goals versus unattained goals

Program Outcomes

- It is useful to distinguish between
  - proximal outcomes: Direct effects of the program
    - Are expected to occur while the clients are taking part in the program
  - distal outcomes: Indirect effects of the program, occurring
    - after the client has completed the program
    - in environments not controlled by the program
    - at a higher level of analysis than is addressed by the program
    - Social impact outcomes

Role of Program Theory

- The theory underlying the program specifies the
  - kinds of treatments clients should receive
  - expected outcomes of the treatment
  - moderating variables that can limit treatment effectiveness
  - mediating variables that come between the treatment and the outcome

Advantages of Theory-Based Evaluations

- Allow a test of the theory's validity
- Can determine which components of a program are necessary to accomplish its goals and which are not
- Helps specify the conditions necessary for the program to have its intended effects
- If program is not effective, can point to reasons why

Stage 2: Program Monitoring

- The on-going assessment of how well the program is being implemented
  - Also called *process* or *formative* evaluation because it evaluates the process or form of the program

Target Population

- The particular group of people the program is designed to effect
- Establishing a program does not guarantee it will reach its target population
  - Programs should include an advertising component to reach as many people as possible
  - And out-reach to attract target group

Target Population

- Bias is introduced if clients are selectively enrolled in the program
  - E.g., if selection is based on clients who
    - are most likely to benefit from the program
    - the program staff feel most comfortable with
- Consequences of bias include
  - some people who need services do not receive them
  - internal validity of the evaluation is threatened
  - generalizability of the program is questionable

Target Population

- Evaluation team should check to ensure the program is reaching the people it is designed to serve
- Should be checked against
  - program goals
  - surveys of potential clients who are not enrolled

Program Implementation

- Is the program being carried out as it was designed?
- If clients' experience does not match what was intended for them, program is not being implemented properly
- Can have blind observer attempt to categorize different treatments (e.g. leadership style)

Sources of Implementation Failure

- A lack of specific criteria and procedures for program implementation
  - Program should have procedures manual that describes procedures, techniques, and activities in detail
  - Program criteria can also state the basis for determining how well the program is being implemented

Sources of Implementation Failure

- Staff is insufficiently trained
  - Effective programs thoroughly train and test staff on their ability to carry out program
- Inadequate staff supervision
  - Provides opportunity for treatments to "drift away" from intended course
- Staff does not believe in program effectiveness
  - Can lead to resistance or even sabotage

Assessing Implementation

- Programs are complex and absolute criteria are rarely available
- Must establish criteria for whether program has been properly implemented
  - Should be done in consultation with stakeholders
- Should also evaluate potential for unintended effects

Client Resistance

- Potential clients may be resistant to program due to
  - distrust of goals, program sponsors, etc.
  - uncertainty about program effects
  - anxiety due to unfamiliarity with new program

Client Resistance

- Clients may be reluctant to use services due to
  - inaccessibility, such as lack of transportation, location of services
  - threats to client dignity, such as unnecessarily intrusive questioning or rude treatment
  - failure to consider clients' culture, such as special treatment needs or language difficulties
  - provision of unusable services, such as printed material at too high a reading level

Reducing Client Resistance

- Be sure to include all stakeholder groups at the program design stage
- Conduct informational programs that address clients' uncertainties
  - Stakeholder focus groups can provide information about program goals and structure
  - Clients can provide feedback and modifications can be made to address concerns

Stage 3: Impact Assessment

- Addresses the question of how much effect the program had on its clients in terms of achieving its goals
- Data are collected to make a summative evaluation of the program's effectiveness
  - Address the overall effectiveness of a program relative to its goals

Criteria For Evaluating Impact

- What degree of change was brought about relative to each of the program's goals and desired outcomes?
  - Change experienced by each client can be averaged to provide overall change index
  - Can also compute an effect size as indicator of program's effect
  - However, group average can obscure individual differences in response to treatment

Criteria for Evaluating Impact

- Importance of the change
  - Statistical analyses do not address practical importance of effect
  - Can be assessed in terms of the percentage of clients who meet the program's goals
  - Can be operationally defined in terms of meeting some preset criterion of improvement
  - number of program goals achieved
    - No reduction in achievement gap but improvement for disadvantaged children (Sesame St. program)
  - the durability of treatment outcomes as assessed by follow-up data

Criteria for Evaluating Impact

- Costs of the program, including
  - monetary costs of program administration
  - method of providing treatment (i.e., group vs. individual administration of treatment)
  - required level of professional qualifications of staff
  - costs to staff, such as increased stress or burnout

Criteria for Evaluating Impact

- Costs of the program for clients, including
  - monetary costs such as transportation, child care costs
  - psychological costs such as lowered self-esteem, disruptions to daily life

Criteria for Evaluating Impact

- Acceptability of program
  - If potential clients do not like the program they will not use it
  - If program creates unpleasant working environment, recruiting and retaining staff will be difficult
- Effective programs that are not utilized are failures

Evaluation Researchers Can Ask Four Questions About a Program

1. Is the program effective?
   - Addressed by program package design
     - Program is compared to no-treatment or waiting-list control group
     - Design acknowledges but ignores that programs are complex and that several aspects could be investigated separately
   - Addressed by comparative outcome design
     - Answers the question of which program is more effective by comparing two different programs that have same goals, may also include control

Evaluation Researchers Can Ask Four Questions About a Program

2. What aspects of the program are important?
   - The extent to which a program is effective may be due to some components more so than others
   - Can be assessed by dismantling design
     - Tests the necessity of including a component in the program
     - Compares client group that experiences that component to client group that does not

Evaluation Researchers Can Ask Four Questions About a Program

- Program aspects can also be compared with client and program variation design
  - Can address
    - whether program is equally effective for all client groups
    - whether program is equally effective if implemented in different ways
    - E.g., high and low impairment alcoholics benefit from programs run by professional and non-professional staff (interaction)

Evaluation Researchers Can Ask Four Questions About a Program

3. How can the program be improved?

- Can be evaluated with constructive design
  - A component is added to successful program
  - Whether the addition improves goal attainment is assessed
- Can also be evaluated with parametric design
  - Looks at the degree to which clients experience a component of the program
  - E.g., participating 3 times vs. once a week

Evaluation Researchers Can Ask Four Questions About a Program

4. How valid is the theory of the program?
   - Can be evaluated by any of the designs used to answer the previous questions
   - Key is to look at which design includes the theoretically relevant variables
   - Program effectiveness and validity of the theory can be tested simultaneously

Research Designs

- Minimum requirement for evaluation research study is 2 x 3 factorial design
  - 2 (treatment and control group)
  - 3 (pretreatment, end of treatment, follow-up)
  - Pretest ensures that treatment and control group are equivalent at starting point
  - Follow-up assesses extent to which effects of program last over time

True Experiments

- In principle, ideal strategy for evaluation research
  - Can test whether the program caused changes in clients
- In practice, infrequently used
  - Difficult to assign participants to conditions
  - Controlling situational variables often impractical

True Experiments

- Even if control of situational factors is difficult, random assignment is possible when program resources don't allow all clients to be served
  - Random selection of who can enroll may be fairest way of allocating resources
  - Can also assign clients to program on first-come, first-served basis

True Experiments

- Quasi-random assignment also possible when
  - patients admitted to facility are assigned to wards or other units
    - Different versions of the program can be offered in different units
  - clients have no preference among alternative versions of the program
    - So, are okay with being randomly assigned to version of the program

Quasi-Experiments

- Nonequivalent control group design commonly used in evaluation research because
  - natural units are often studied (e.g., classrooms, industrial settings)
  - evaluations are often designed and conducted after a program has been instituted

Quasi-Experiments

- Control groups can be formed after the fact to rule out alternative explanations
  - Recall these are called *patched-up quasi-experiments*

Threats to Internal Validity

- *Control group contamination*
  - Treatment diffusion: Members of the control group learn about the treatment from members of the treatment group
    - Try to apply the treatment to themselves
  - Staff might attempt to compensate the control group for being deprived of the benefits the treatment group receives

Threats to Internal Validity

- Control group contamination can also be a result when
  - people who are aware they are part of a control group feel rivalry with the treatment group
    - Work to outdo them on the posttest
    - Compensatory rivalry
  - members of the control group resent being deprived of a program that could benefit them
    - Might reduce efforts to solve their problems themselves
    - Resentful Demoralization
    - Compensatory Equalization

Threats to Internal Validity

- Local history: An event outside the research that affects the dependent variable
  - Affects *only* the treatment group or *only* the control group (so is local to it)
  - Threat is strongest when treatment and control group are geographically separated
  - Threat can be reduced if several groups are used, each in a different location

Pre-experimental Designs

- Pretest-posttest design without a control group
- Pre-experimental because cannot assess the impact of internal validity threats
- Design is used in situations where
  - it is impossible to find a no treatment control group
  - researcher is conducting pilot test to determine if full evaluation is necessary

Meta-Analysis

- Can be used to determine the average effect of a treatment across a number of programs that use the same treatment
- Can also test for possible moderators of a program's effectiveness
  - Addresses the question "Under what conditions is a program more or less effective?"
  - Helps practitioners decide whether a particular intervention is appropriate for a particular client group

Interpreting Null Results

- Possible sources of null results
  - Program failure: A true null result
    - The program does not bring about desired effect
  - Type II errors, including
    - failure to implement program as designed
    - low statistical power
    - poor research design, including poor content validity

Interpreting Null Results

- Researchers typically compare the effectiveness of two treatments
- However, this is not the only possible question
  - For example, two treatments might be equally effective, but one might cost less or be more acceptable to clients
  - In this case, can conclude that one treatment better fits program goals (such as cost savings)

Stage 4: Efficiency Analysis

- Compares the outcomes produced by a program to the costs required to run the program
- Does *not* address the question of whether the outcomes are worth the cost
  - These value judgments are made by policy makers and administrators, not scientists

Cost-Benefit Analysis

- Compares the dollar cost of operating a program to the benefits in dollars provided by the program
- Assumes
  - all outcomes can be expressed in monetary terms
  - the only important outcomes can be expressed in monetary terms
- Can address both direct and indirect cost/benefits

Cost-Effectiveness Analysis

- Goals and outcomes can (and probably should) consider psychological factors, such as
  - reduction in psychological distress
  - increase in educational outcomes
- Particularly useful when programs with similar goals are compared
  - Helps determine which program has greatest effect for lowest cost

Stage 5: Information Utilization

- Instrumental utilization: Research results are directly used for making decisions or solving problems (ideal)
- Conceptual utilization: Information influences a policy maker's thinking about an issue (usual)
  - Even if does not have a direct influence on decisions about the issue
- Persuasive utilization: Evidence is used to convince others to support a political position or to defend a political position from attack (avoid)

Criteria for Research Utilization

- Relevance: Addresses the needs of all stakeholder groups
  - Does the program meet the clients' need for services?
  - Rooted in the inclusion of all stakeholder groups in all stages of the evaluation
    - Goal is for stakeholders to "own" the evaluation

Criteria for Research Utilization

- Truth: Addresses whether the research produces valid results
  - Research should adhere to principles of all types of validity
  - Stakeholders should also perceive that the research results are valid
    - Based in part on whether user has faith in the researcher and the research process

Criteria for Research Utilization

- Utility: Addresses the extent to which the research focuses on policy variables as IVs and on high-utility DVs
- Also address whether program evaluation provides information that aids program development
  - Answers questions such as why the program did (or did not) work

The Political Context

- Evaluation research is often conducted in a social environment pervaded by political motivations
  - People's incomes, power, and prestige can be negatively affected by the outcome
  - So, quality of research is not the only factor affecting whether research is used

The Political Context

- Ritual evaluations: Those that are carried out although no one has any intention of applying their results to the program
- For example, might
  - be required by law or regulations
  - seem like right thing to do at the time
  - be seen as useful for persuasive purposes

The Political Context

- Stakeholder interests: Vested interests of those involved with program can affect the acceptance of the results of an evaluation
  - May be proponents of information favorable to their interests
  - May suppress or downplay results seen as unfavorable

Measuring Change

- Difference scores: A person's score on a measure at a posttest minus that person's score on a pretest
  - Many statisticians believe these scores have low reliability
  - If so, implies low validity
  - However, more recent analysis of this question suggest a more optimistic view

Reliability of Difference Scores

- Reliability of difference score increases as the
  - reliability of the measure increases
  - correlation between the pretest scores and the difference scores increases
  - correlation between the pretest scores and the posttest scores approaches zero

Reliable Change Index (RCI)

- Is a person's difference score on a measure divided by the standard error of the difference for the measure
  - Computed from the standard error of measurement of the measure
  - Can only be used with standardized measures
- If value of the RCI is greater than 1.96, then the probability that the difference score reflects only random change is less than 0.05

Reliable Change Index (RCI)

- Can be used in program evaluation in two ways
  1. Researchers can compare the mean RCI scores of a treatment group and a control group
     - Addresses whether the average amount of reliable change was due to the treatment
  2. Can also compare the proportions of people in the treatment and control group who show reliable change

Necessary Knowledge & Skills

- Evaluators more eclectic
- May work with expert in subject matter
- King et al. (2001)
- Systematic inquiry
- Competent evaluation practice
- General skills for evaluation practice
- Evaluation professionalism
- Capacity building

After the Evaluation

- Provide detailed explanation of partial results
- Carefully organize and document statements made in the report with data
- Include multiple measures and data
- Communicate widely
  - Multiple formats

Critiques of Program Evaluation

- Less control?
- Not based on theoretical understanding?

Think about Possible Programs to Evaluate

- Reasons to conduct?
- Issues in external stakeholders?
- Defining worthy outcomes?
- Restrictions on reporting of results?