ESSENTIALS OF PROGRAM EVALUATION

A WORKBOOK FOR SERVICE PROVIDERS

by

Ronald Jay Polland, Ph.D.

© 1989 by Ronald Jay Polland, Ph.D.
all rights reserved
LESSON ONE

WHAT IS PROGRAM EVALUATION?

- What is program evaluation?
- Purpose of evaluations
- Conceptualizing program intent and program design
- Program monitoring
- The need for good program data
- The evaluation system as an information system
- Program accountability
- Program impact
- Alternative models of evaluation
- The pragmatic paradigm -- or "Action research"
- Values in Evaluation
- Ethical considerations in evaluation
- Why evaluation results are often not implemented
- Sources of program failure
- When official program goals don't reflect actual operation

LESSON TWO

A SYSTEMS APPROACH TO MEASURING PROGRAM EFFECTIVENESS;

- The Systems Approach to Measuring Effectiveness (SAME)
- Brief explanation of the SAME

LESSON THREE

USING NEEDS ASSESSMENT TO IDENTIFY A PROBLEM AND RELATE IT TO PROGRAM GOALS AND OBJECTIVES;

- Interpreting a program's mission, goals, objectives and activities
- The problem with the word, "problem."
- Three ways (at least) of defining the "need"
- The role of needs assessment
- Needs assessment techniques - direct methods
- Needs assessment techniques - indirect methods
- Types of needs assessment
- A "cooking analogy" to defining the problem

LESSON FOUR

SURVEY METHODS FOR NEEDS ASSESSMENT AND OUTCOME ASSESSMENT

- What do surveys measure?
- Uses of surveys
- Choosing the types of data to collect
- Question types

LESSON FIVE

ASSESSING RELIABILITY AND VALIDITY OF EVALUATION DATA;

- Reliability and sources of variation
- The importance of assessing reliability and validity
<table>
<thead>
<tr>
<th>Statistical Test</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Analysis</td>
<td>68</td>
</tr>
<tr>
<td>Cochran's Q Test</td>
<td>69</td>
</tr>
<tr>
<td>Coefficients of Concordance</td>
<td>69</td>
</tr>
<tr>
<td>Concordant</td>
<td>69</td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td>69</td>
</tr>
<tr>
<td>Contingency Table</td>
<td>69</td>
</tr>
<tr>
<td>Continuous Variable</td>
<td>69</td>
</tr>
<tr>
<td>Covariance</td>
<td>69</td>
</tr>
<tr>
<td>Covariate</td>
<td>70</td>
</tr>
<tr>
<td>Cramer's V</td>
<td>70</td>
</tr>
<tr>
<td>Cronbach's Alpha</td>
<td>70</td>
</tr>
<tr>
<td>Crosstabulation</td>
<td>70</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>70</td>
</tr>
<tr>
<td>Differences</td>
<td>70</td>
</tr>
<tr>
<td>Discordant</td>
<td>70</td>
</tr>
<tr>
<td>Discrete Variable</td>
<td>71</td>
</tr>
<tr>
<td>Discriminant Analysis</td>
<td>71</td>
</tr>
<tr>
<td>Effect Size</td>
<td>71</td>
</tr>
<tr>
<td>ETA</td>
<td>71</td>
</tr>
<tr>
<td>ETA Squared</td>
<td>71</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>71</td>
</tr>
<tr>
<td>Friedman Two-Way ANOVA</td>
<td>71</td>
</tr>
<tr>
<td>Gamma</td>
<td>72</td>
</tr>
<tr>
<td>Goodman and Kruskal's Lambda</td>
<td>72</td>
</tr>
<tr>
<td>Hilologlinear Analysis</td>
<td>72</td>
</tr>
<tr>
<td>Homogeneity of Variances</td>
<td>72</td>
</tr>
<tr>
<td>Independent Samples</td>
<td>72</td>
</tr>
<tr>
<td>Independent T-Test</td>
<td>72</td>
</tr>
<tr>
<td>Interrupted Time Series</td>
<td>72</td>
</tr>
<tr>
<td>Kendall's Tau</td>
<td>73</td>
</tr>
<tr>
<td>Kolmogorov - Smirnov</td>
<td>73</td>
</tr>
<tr>
<td>Kolmogorov - Smirnov 2-Sample</td>
<td>73</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>74</td>
</tr>
<tr>
<td>Kruskal-Wallis One-Way Analysis of Variance</td>
<td>74</td>
</tr>
<tr>
<td>Lambda Coefficient</td>
<td>74</td>
</tr>
<tr>
<td>Main Effects</td>
<td>74</td>
</tr>
<tr>
<td>Mann-Whitney U - Wilcoxon Rank Sum W Test</td>
<td>74</td>
</tr>
<tr>
<td>McNemar Test</td>
<td>74</td>
</tr>
<tr>
<td>Mean</td>
<td>75</td>
</tr>
<tr>
<td>Median</td>
<td>75</td>
</tr>
<tr>
<td>Median Test</td>
<td>75</td>
</tr>
<tr>
<td>Minimum Expected Cell Frequency</td>
<td>75</td>
</tr>
<tr>
<td>Mode</td>
<td>75</td>
</tr>
<tr>
<td>Moses Test of Extreme Reaction</td>
<td>75</td>
</tr>
<tr>
<td>Multiple Analysis of Variance (MANOVA)</td>
<td>76</td>
</tr>
<tr>
<td>Multiple Range Test</td>
<td>76</td>
</tr>
<tr>
<td>Null Hypothesis</td>
<td>76</td>
</tr>
<tr>
<td>Observed Significance Level</td>
<td>76</td>
</tr>
<tr>
<td>One Sample T-Test</td>
<td>76</td>
</tr>
<tr>
<td>Paired Samples T-Test</td>
<td>76</td>
</tr>
<tr>
<td>Pearson's R</td>
<td>76</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td>77</td>
</tr>
<tr>
<td>Power</td>
<td>77</td>
</tr>
<tr>
<td>Principal Components Analysis</td>
<td>77</td>
</tr>
<tr>
<td>Proportional Reduction in Error (PRE)</td>
<td>77</td>
</tr>
<tr>
<td>Range</td>
<td>77</td>
</tr>
<tr>
<td>Regression Analysis</td>
<td>77</td>
</tr>
<tr>
<td>Regression Coefficient</td>
<td>77</td>
</tr>
<tr>
<td>R Squared</td>
<td>78</td>
</tr>
</tbody>
</table>
LESSON ONE

What is Program Evaluation?

OBJECTIVES

- Define "evaluation", "evaluation project", "proposal"
- Identify and define the major parts of a evaluation proposal
- Compare and contrast different types of proposals

Keywords

evaluation, evaluation project, proposal, funding agency, applied evaluation, theoretical evaluation, grants, dissertations, journal articles.

What is program evaluation?

Evaluation involves a process of systematically gathering, examining and reorganizing information to make informed decisions and to increase one's understanding of specific phenomena. The key word is "systematically" -- meaning "methodical, orderly, planned and reproducible." Evaluation makes use of observations made by various sources -- from "primary" observations (those made by you) and "secondary" observations (those made by others).

The purpose of a program evaluation is to determine if a program is achieving its intended outcomes. Program evaluation is a system for identifying relevant input and output measures that provide information about program effects and program implementation. Program evaluation is a systematic assessment of the process or product of deliberate and planned interventions. Evaluation employs processes that measure actual performance against set performance standards. Having valid performance standards that are based on measurable internal and external criteria is critical to the design of an evaluation.

Program evaluations are basically of two general varieties: a process or administrative evaluation that analyzes program inputs and outputs such as the numbers of people served or activities performed and an outcome or impact evaluation that examines the effects of a program's output on program participants. Both types of evaluation require that desirable objectives or goals have been set, a planned program of deliberate intervention occurs and a method for determining if achievement of desired objectives and effects are a result of the planned program.
While the intent of evaluation is to look for expected outcomes, the evaluator shouldn't ignore unanticipated results. In the real world, few social and educational programs achieve only large, positive effects. Most produce small effects or ones that are both positive and negative. Conversely, increased awareness of a social problem may be an unanticipated, positive outcome of a program. The evaluator should consider looking for unintended effects during the planning of a program evaluation.

**Purpose of evaluations**

The design of an evaluation is a plan for the efficient allocation of resources: to produce highly useful information within a specified budget. Evaluations may be done for management or administrative purposes, to meet accountability requirements for funding, to find ways of improving program effectiveness, for planning and policy improvements, to test ideas, to solve problems, to determine whether to expand or curtail programs or to compare one program versus another.

Evaluation may be used to fine-tune programs. Evaluation may involve existing programs or be conducted on new and innovative programs. Evaluations influence the actions and activities of individuals and groups who have an opportunity to change their actions and activities on the basis of the evaluation.

Evaluations may have different functions. One may be a theory-building function -- to clarify, validate, disprove or modify the body of theory on which the basis of the program's intent and goals were derived. Another may be an accounting function - to inform the funding agency of their value received for dollars spent. The third may be a dissemination function -- to make the results of the evaluation available to others. A fourth function present in some monitoring evaluations is a feedback function -- to refine and improve the program by a continuous process of feeding data back into the program planning process.

As previously noted, evaluations may be broadly distinguished as either assessing process or impact. However, a better classification is to examine the four main purposes of evaluation: to conceptualize program intent and program design, to monitor and improve program operation, to determine program accountability and to measure program impact.
Conceptualizing program intent and program design

Evaluability assessment

An evaluability assessment examines the feasibility of evaluating a program. An evaluability assessment is therefore a pre-evaluation. To determine evaluation feasibility and to gain a thorough understanding of the objectives, implementation and management of a program, an evaluator must collect and review relevant program data. Relevant program data include a program's legislative history, regulation and guidelines, budget justification, monitoring reports and reports of program accomplishments. An evaluator would then interview key policy-makers, managers and stakeholders on their assumptions and expectations about program resources, activities and expected outcomes. Based upon these data, the evaluator can construct a tentative program evaluation plan to present to its stakeholders.

In the process of doing site visits, interviews and document reviews, the evaluator obtains intimate knowledge of a program that will help in the design of the evaluation. Additionally, the evaluator will be able to determine if the social and political environment within which the program operates is favorable to conducting an evaluation.
To do a proper evaluation, the relevant program data has to be obtainable, measurable and useable. An evaluation may not be possible in instances where the data is not available or its in a form that’s not useable. Another consideration is whether the evaluation will have any real utility to the stakeholders, policy-makers and decision-makers. If their actions and decisions won't be affected by the outcome of an evaluation, then doing one might be a waste of time and resources.

Program monitoring

The following analogy will serve to illustrate the relationship between program monitoring and program evaluation. If your goal is to travel from Tallahassee to Orlando by car, program evaluation would tell you if you actually arrived in Orlando and how you actually got there. Program monitoring, on the other hand, would ensure that you can reach Orlando from Tallahassee by staying on the right roads and following the correct directions. Program monitoring is like navigation: it involves monitoring progress at many points in time to see if you're on course and recommends appropriate changes if you're not. Program monitoring only ensures that you stay on the course you've chosen to take. In contrast, you'll need to do a program evaluation to determine whether you've chosen the best possible course available.

Program monitoring accomplishes two things: whether programs are reaching their target population and whether delivery of services is consistent with program design specifications and objectives. It provides program managers with operation and performance information on a daily basis. It provides information for accountability purposes and as a necessary complement to impact assessment since program failure often results from faulty or incomplete implementation.

Monitoring information may be for the sole purpose of judging program impact or a supplement to utility assessment. Monitoring as part of outcome evaluation determines how a program was carried out and helps to link program inputs to outputs. Monitoring for management and accountability purposes is aimed at maximizing productivity and organizational effectiveness.

Today monitoring for management and monitoring for outcome assessment are more similar than they were ten years ago. One of the factors that has influenced this convergence is the widespread use of computerized, management information systems (MIS's). MIS's provide continuous, systematic data on what is happening within a program; e.g., how many persons are served by the program and what are their characteristics? What services are being delivered? How are funds being spent? How long do persons remain in the program?

The purpose of a well-designed MIS is to provide program mangers and funders with detailed, periodic reports on how well the program is functioning and to alert them to delivery problems so that they may be handled when they arise. At the same time, they can provide the necessary information to monitor program interventions and to assess impact. Thus, an evaluation system is system that makes use of the informational and analytical capabilities present in an MIS to accomplish the data requirements of ongoing monitoring and evaluations.
An information system performs two functions: one, stores, retrieves and reports information in convenient formats, and two, quantifies data by condensing and analyzing it into a few, relevant features that illustrate the meaning of the data. It's essential to first conceptualize all the data requirements from all sources that a program will require for monitoring and evaluation before creating an MIS. Also critical is for program staff who are charged with entering and retrieving data to be fully aware of the system's capabilities and limitations as well as their roles and responsibilities in its operation.

The need for good program data

Program monitoring and program evaluation depend upon a ready supply of reliable and valid program data. The methods typically used for acquiring program monitoring data include (1) observations and interviews, (2) records and reports, and (3) surveys. Those who are responsible for monitoring the operation of their programs will need to observe program functions at different times in a given day, on different days and at different points in the monthly or annual calendar of operations to ensure that have been observed. Information collection, classification, storage, access, analysis and reporting are critical elements of any system used for program monitoring.

A program that provides services to clients must maintain client records and administrative files. Client records will have biographic information, types of services rendered, dates of service and other service information. Administrative records would be kept on numbers of clients served, costs of service, staff turnover and any day-to-day changes in program operation.

Surveys, either by survey or interview, are useful in estimating the level of services actually received by clients as well as people's attitudes about programs generally or the quality of service obtained. They are also useful in assessing community support for programs and determining the size of potential clientele. Periodic surveys can be an essential part of program monitoring and evaluation and can provide reliable estimates of trends in the community in which a program functions as well as specific outcomes of the program itself.

The evaluation system as an information system

Evaluation is a process for providing information necessary to planning and decision-making. In the daily routines of life, evaluation is actually a common activity. For example, to maintain their safety, drivers regularly evaluate the operating condition of their cars. Each part of the car is expected to operate at some acceptable level of performance. Checks are made to ensure that the tires have sufficient tread and pressure, the engine and transmission have sufficient lubricant, the brake linings are thick enough to stop the car, and so forth. For a governmental program to function or work properly, it also needs regular check-ups.
The key concept here is "regular check-ups." If evaluation is to be a continual, on-going process, then this process must be part of a system. A system is comprised of functionally related processes. To assist drivers in monitoring and evaluating their cars, car companies have developed evaluation systems that provide for both daily monitoring (gauges, dip sticks) and periodic evaluations (warranty service check-ups). These evaluation systems provide information to the driver who must make regular decisions regarding the continued operation of the car, its safety and its performance.

Like cars, social programs have many functionally related processes that require on-going to ensure proper operation. Unlike cars, however, governmental social programs are not typically designed with built-in evaluation systems.

Therefore, a need exists for an evaluation system that can provide information on the performance level and administrative operation of social programs to all persons who are responsible for program operations and to stakeholders. An evaluation system must also be able to provide information to decision-makers and policy-makers on program operation, program outcome and program improvement.

**Program accountability**

Program managers are required to conduct their programs as efficiently as possible. Sponsors and stakeholders require evidence of program implementation such as "Is program reaching the specified population?" and "Is the intervention being implemented as specified?" There are several types of program accountability that an evaluation may identify:

**Accountability for Efficiency**

Efficiency of a program is a way to measure its impact in relation to program costs. Efficiency accountability is important in judging relative benefits and effectiveness against costs of different program elements.

**Accountability for Coverage**

Coverage implies the extent to which the appropriate persons are being served by a program. Almost all programs are required to keep records on persons in the target population served. The main concern is that the data is accurate and reliable, that appropriate forms and instruments are being used and that staff are properly trained in their use.
Accountability for Service Delivery

In assessing how the actual operation of a program conforms to program plans, the evaluator looks at service delivery; i.e., to see if those delivering the service to the target population are properly qualified and trained. Monitoring delivery of services is important for documenting not only the existence of treatments but also problems in the program treatments themselves.

Accountability for Legal Responsibilities

All programs require commitments to meet legal responsibilities including informed consent, protection of privacy, community representation on decision-making boards, equity in provision of services and cost-sharing. The consequences of not meeting legal responsibilities may be severe enough to lead to the cancellation of a program.

Accountability for Financial Responsibilities

Programs have to account for the use of all funds in their financial reports. Additionally, a range of other cost questions may be relevant including cost per client and cost per service, incremental and marginal costs. Costs may vary as a function of program site, time of year and competing programs.

Program impact

The prerequisites for assessing impact are clearly stated objectives in measurable terms and confirmation that an intervention has been properly and thoroughly delivered. The next step is the identification of one or more outcome measures that represent the objectives of the program. It's important to distinguish between gross and net outcomes. Net outcomes are only those impacts that can reasonably be attributed to the intervention free and clear of changes to the target population due to other factors.

The hardest programs to evaluate are ones that predict immediate, positive outcomes from the treatment or services they offer. One problem is determining a level of impact that is "acceptable." Another problem is the timing of program effects. Some programs will be expected to have immediate results while others may not demonstrate any effects for several years afterward. The key for programs with delayed effects is to establish, in advance, when measurable effects will occur.
Typically, evaluations are done on existing programs but they also may involve innovative programs. When a program involves an intervention that's too new to have any prior evidence of impact, then the evaluator should identify the goals of the sponsor, assess the context and content in which the program will operate to achieve these goals and determine the general framework or strategy that the program administrators and managers will use to achieve these goals. Subsequent decisions about improvement depend upon a careful analysis of how effectively and efficiently a program is achieving established goals.

Intermediate objectives represent the desired outcomes of a specific program. Immediate objectives are the enabling objectives that allow a program to achieve its intermediate objectives. Just as the attainment of intermediate objectives requires the achievement of immediate objectives, evidence that a program has achieved its intermediate objectives can be generalized to the achievement of its established goals.

**Models of evaluation.**

Social research generally deals with multicausal models in which no effect has a single cause and a single treatment can have multiple impacts. Multiple and interrelated events also apply to the field of evaluation. It means that program A becomes only one of many possible actions or events which may bring about the desired impact B. Program A and impact B will have many other consequences. In explaining the effectiveness of program A to achieve impact B, the evaluator relies on an explicit methodology and research doctrine that takes into account the preconditions under which the program was started, the intervening effects that occur and the consequences that follow those effects.

In an attempt to formalize their philosophy and methodology, evaluation researchers have prescribed to five main models of evaluation. They are goal-attainment models, impact models, system models, goal-free models and theory-driven models.

**Goal attainment models**

This model defines program effectiveness as the extent to which a program has attained its goals. Intervention programs are created to achieve goals and it follows that evaluators should assess effectiveness in terms of these goals. One of the problems with this model is that goals are often vague and not operationally defined. A goal model requires clear, measurable goals to evaluate effectiveness. Another weakness of this model is that goals are often formed to obtain funding rather than to serve as a plan for program activities.
**Impact models**

The logic of the impact model views the problem of determining program effectiveness as identical to establishing that a program causes some specified effect. Thus, determining program success is simply a matter of establishing causality and whether the addition of the program leads to an effect above and beyond what would have occurred without the program.

**System models**

The system model views a social program as a set of interdependent subsystems. If any subsystem performs poorly, then performance of the whole system will be affected. To survive, a system must maintain its internal stability in coping with the external environment. This model is concerned with how a system operates, how it survives and grows, how it acquires and allocates its resources, deals with conflicts and adapts to change.

**Goal-free models**

A goal-free model requires the evaluator to focus not on the stated goals of the program, but rather to attend to the actual effects of a program. The argument here is that the evaluator's judgement would be limited by focusing only on goals.

**Theory-driven models**

A theory-driven model requires evaluators to specify two kinds of outcomes in any list of outcome variables. The program plausible goals and theory-driven plausible outcomes which include unintended outcomes as well as intended ones. This model has a multiple outcome perspective. Unlike traditional impact evaluation, theory-driven evaluation has one or both of the following characteristics: (1) when assessing impact of the treatment on the outcome, impact evaluation uses theory-guided strategies to generate a broad evidence base and (2) when specifying the outcome in the study, impact evaluation uses both the stakeholders view and the existing knowledge and theory related to the program to assess the important outcomes both intended and unintended.

Intervention in social programs is assumed to have an effect according to some social model. This doesn't mean that it's theory-dependent or that it has to be classified according to any one particular theory. Rather, program designers should at least specify the theoretical or pragmatic basis for a program. It's entirely possible to have an intervention that "works" without anyone knowing how it actually works. Providing a theoretical base may add to its understanding but not necessarily to its pragmatic value.

If there is a solid theoretical framework that we can use to understand a program's intent, then we shouldn't disregard it since our mission is to learn as much about a program as we can or as is desirable.
**Alternative models of evaluation**

Within the past twenty years, there has been a movement away from scientific inquiry models of evaluation based on logical positivism to naturalistic inquiry models based on social constructivism.

**Logical positivism** contends that an external world independent of human experience exists and that objective, scientific knowledge about this world can be obtained through direct sense experience (as interpreted within the framework of the theory-based, hypothesis-testing laboratory experiment.

**Social constructivism** contends that reality is conceptually constructed by individuals and social groups. "Facts" and "raw data" can be known only within a particular pre-established cultural, social, historical and linguistic context. Reality is thus a by-product of one's personal beliefs, culture and language.

The basic tenets of constructivism are that

- "Truth" is a matter of consensus.
- "Facts" have no meaning except within some value framework
- "Causes" and "Effects" do no exist except by definition.
- "Phenomena" can only be understood within the context in which they are studied. Neither problems nor solutions can be generalized from one setting to another.
- Interventions are not stable
- Change cannot be engineered: it's a non-linear process
- Evaluators are subjective partners with stakeholders in the literal creation of data
- Evaluation data has neither special status nor legitimation.

**The pragmatic paradigm -- or "Action research"**

This model focuses on action-oriented approaches from engineering and research and development. A conceptually coherent program is designed to address a significant social or psychological problem within a naturalistic, real-world setting in a manner that is feasible, effective and efficient. Quantification is used to develop indicators of system functioning. Then the system is monitored in terms of baselines and changes.

The pragmatic paradigm focuses on getting programs to "work" within a particular real-world setting. Evaluation occurs in four phases. First, the type of decision to be made is identified. Next the context of the decision and the culture of the relevant decision-makers are described. The evaluator constructs a conceptual model for understanding the nature of the decision to be made.
A quantitative data methodology is developed that is explicitly linked to the decisions set forth in the first phase. In phase three, the methodology is pilot-tested. When the pilot-test is successful, the methodology is implemented at full scale. When this is successful, the methodology is disseminated at full scale. The pragmatic model is used to meet the decision-makers informational needs rather than purporting to discover the real state of affairs.

Values in Evaluation

Responsiveness

Responsiveness is the recognition that results should be relevant and useful to the needs and concerns of not only decision-makers but also other stakeholders such as program managers, program staff and clients.

Objectivity

Objectivity means reliable, factual or confirmable. Objectivity is attained through intersubjective agreement; that is, when another evaluator reproducing the same evaluation study would get the same results. The evaluator shouldn't have a personal stake in the success or failure of a program.

Trustworthiness

Trustworthiness is an assurance that the evaluation can provide convincing evidence that can be trusted by stakeholders and other consumers of evaluation results. This is related to the concept of internal validity.

Generalizability

Generalizability refers to the extent to which evaluation results can be applied to future pertinent circumstances or problems in which stakeholders are interested.

Ethical considerations in evaluation

There are a number of ethical considerations that an evaluator may have to deal with in the design and administration of program evaluation. There will be different considerations depending upon the intended purpose of the evaluation. One concerns whether preliminary evaluation data should be used to modify an on-going program; i.e., be fed back into the planning and implementation process. While this seems like a reasonable, logical and ethical thing to do, it may disrupt a planned research design that assumes no change in program parameters from start to finish.
Another ethical dilemma arises when the evaluator has a stake in the outcome of the evaluation. If an evaluator has an interest in a competing program or wants to change the existing program in some specific way, then discovering program success would run counter to his desires. Conversely, an evaluator would be reluctant to discover program failure if, in doing so, it would jeopardize his present or future working relationship with the program managers, administrators or stakeholders.

One is the degree of tension between the evaluators and the program operators and managers. The second is what is imposed by disciplinary boundaries that separate the various social sciences from one another. The third constraint is the ethical necessity for continuous feedback of research findings back into the programs themselves. This may affect a research design developed at the beginning of the program under the assumption that the nature of the program would remain constant. The fourth is imposed by the time dimension. The goals and effects of social action programs are usually long-range in nature. However, evaluation is generally demanded in the short-term. The fifth constraint is the degree of openness of the community in which the program is implemented. The community is not a laboratory where all variances can be controlled.

**Why evaluation results are often not implemented**

Institutions don't often change their behavior in response to evaluations. They explain away the results sometimes criticizing the evaluator on his understanding of the program or organization or on the validity of his methodology. Conversely, evaluators most often complain that their findings are ignored.

Organizations respond to factors other than the attainment of their official goals. These factors include a need to perpetuate itself, individuals' need for status and self-esteem, fear of the unknown and of change, their image with the public, costs, political ideology and a host of others. Evidence of program outcome usually can't override primary organizational behaviors and needs. Evaluators should therefore pay attention to both the official and the operative goals of an organization.

Acceptance of evaluation is more likely when few interests are threatened, when minor modifications will be required or the costs will be minimal. Acceptance of evaluation is also more likely when the quality of the evaluation can't be brought under question.

Use of evaluation results might be increased if they included (1) description of theoretical basis underlying the program and the relationship of the evaluation to the basis, (2) specification of the process model that show the linkages between program inputs and program outputs; (3) analysis of the effectiveness of the program components.

Other ways of increasing use is the early identification of evaluation results users and selection of the issues important to them; involvement of administrators and program managers in the evaluation process; prompt completion of evaluation and the early release of results; effective presentation methods and dissemination of results.
Administrators and program managers who are involved in the evaluation gain both an insight into the process of solving problems, a new awareness of their programs and a feeling of responsibility for initiating positive steps to improve their programs. Group participation is important in facilitating attitude change by invoking group pressures. Group participation provide a broader input of ideas and the pooling of ideas.

**Sources of program failure**

Evaluators conduct their work in continually changing environments: political shifts, changing interests of the stakeholders and changes in responsibilities of agencies that sponsor programs.

There are two sources which may explain program failure: one, inability of program to influence causal variable; and, two, the lack of validity of the theory linking the causal variable to the program objective.

Some of the problems in treatments that may occur are (1) the wrong treatment, (2) incomplete treatment or diluted treatment, or (3) nonexistent treatments. As an example of #2, a program delivered by highly trained and motivated staff may fail when its implemented in the real world outside the classroom or laboratory. Another problem that may exist is too much variance from one program site to another.

The most serious cause of program failure is the imprecision with which program goals are stated and program activities are implemented. Evaluators should be cautious of accepting as sufficient the description of the program by the program administrators. Their description may not be borne out in the actual operation of the program nor support the theory on which the program is based. Also a problem is imprecision in program input.

**When official program goals don't reflect actual operation**

Program goals may not accurately reflect what a program is actually doing. The main discrepancies between official and operative goals may be deeply rooted in policy formulation and program implementation. Often program goals are deliberately vague to help build coalitions. By making program goals broad and general, all members of a coalition can see something in them that appears to satisfy their needs. The best examples (or worst offenders depending on your point of view) are programs proposed by political candidates vying for reelection. Their designed to attract support and avoid opposition by being vague and addressing noble causes.

In comparison, operative goals involving details of resource allocations or value trade-offs only serve to highlight differences among coalitions and enhance the conflicts between them. When funding social programs, decision-makers and funders have high hopes for them. However, they may not know exactly how to translate these hopes or goals into actions. The job of designing the intervention, defining the target population, screening the applicants, contacting the clients, allocating resources, deciding what delivery systems are needed often is not considered to be the responsibility of funding agencies or decision-makers.
These problems are left to program administrators or managers. The providers at street-level who actually work directly with clients and make decisions on day-to-day operations are the ones who ultimately shape the program. Thus, a gap may exist between decision-makers and the expectations of program administrators and managers. This gap is magnified when each have their own needs and incentives.
LESSON TWO

A Systems Approach to Measuring Program Effectiveness;

OBJECTIVES

- List the steps of the Systems Approach to Measuring Effectiveness (SAME).
- Review the example of the SAME as applied to the Case Study in Appendix A.
- Relate the information required by each step to its possible sources.
- Describe the purpose of each activity in the SAME.

Keywords

need, goals, context, stakeholders, decision-makers, policy-makers, systems approach

The Systems Approach to Measuring Effectiveness (SAME)

The previous lesson discussed the various approaches or models used in the evaluation of programs. By virtue of several shared goals, the different evaluation models are not mutually exclusive. It is possible (and often preferable) to combine the approaches to provide the best possible explanation of what activities a program actually does and what effects a program actually has.

Given the dynamic nature of programs and the changing priorities and needs of stakeholders, program evaluators shouldn't rely on any one model to evaluate program effectiveness. In general, the nature of the program and the goals of evaluation will determine the activities an evaluator will need to do. The Systems Approach to Measuring Effectiveness (SAME) presented here is a forty step, "soup-to-nuts," approach to planning, designing and implementing a full-scale program evaluation. It's application is flexible in that evaluators can combine steps or eliminate those unnecessary to their evaluation goals.

For example, an evaluator may only be required by a supporter to document a program need or to provide a simple measure of program impact such as determining customer satisfaction. Regardless of the actual tasks the evaluator has been asked to do, the SAME can help to conceptualize the "big picture" in determining program effectiveness from several perspectives. The following are the steps to the SAME:
1. The need or problem targeted by the program.
2. History, scope and nature of the need or problem.
3. Social, political and historical context of the program.
4. Purpose of the program and how it addresses the problem.
5. Underlying theory supporting the purpose of the program.
6. Mission, goals and objectives of the program.
7. Relation of program activities to mission, goals, and objectives.
8. Client population served by the program.
9. Program methods for identifying clients and levels of service.
10. The stakeholders and their interest in the program.
11. The decision-makers' and policy-makers' interest in the program.
13. Needs and goals of the stakeholders and program funders.
15. Needs and goals of the evaluators.
16. Purpose of the evaluation.
17. Decisions dependent on the evaluation.
18. Ethical issues surrounding the evaluation.
19. Audience(s) of the evaluation report(s).
20. Evaluable of the program.
21. Evaluation resources, constraints and limitations.
22. The benefits and costs of doing the evaluation.
23. Evaluation questions to be asked.
24. Expected outcomes and evidence of program achievement.
25. Possible unexpected outcomes of program implementation.
26. The proposed process or improvement evaluation design.
27. The proposed impact or effectiveness evaluation design.
28. Available sources of data to be used in the evaluation.
29. Data sources and collection instruments to be developed.
30. Reliability and validity assessment of data sources.
31. Pilot-testing of methods and materials.
32. The evaluation management plan.
33. The proposed sampling plan.
34. Data coding and collection plan.
35. Qualitative analyses to be done.
36. Quantitative analyses to be done.
37. Statistical decisions to be made.
38. Cost-benefits analyses to be done.
39. Cost-effectiveness analyses to be done.
40. Methods of reporting and presenting the results.

Each of the steps of the SAME is both a possible evaluation activity ("What to do") and an information requirement ("How to do it"). Accomplishing each step requires the evaluator to obtain supporting information from one or more sources. When you read the steps of the SAME, do it in two ways. First, precede each step with the question, "What is." For example, for Step 1, ask yourself, "What is the need or problem targeted by the program?" Next, precede each step with the question, "How will I determine," as in "How will I determine the need or problem targeted by the program?"
Figure 1 is an example of how the SAME can be used to determine the information requirements of the evaluation. Each step has a corresponding source of information. This form of the SAME can be used as a planning guide to determining the data requirements and data sources for the evaluation.

The rest of this workbook is structured as a "How to" guide to accomplishing the major steps in the SAME. Appendix A is an example of how all of the steps in the SAME were used to design a comprehensive, large-scale program evaluation. As presented, the SAME also serves as an outline for a report on the evaluation. It is beyond the scope of this workbook to provide readers with everything they need to know to do a complex evaluation like the one described in Appendix A. However, using the SAME as a guide and the list of references supplied in the workshop, the reader should be able to locate additional sources of support in designing and conducting complex program evaluations.

**Brief explanation of the SAME**

1. The need or problem targeted by the program.

   This is the "intended" focus of a particular program activity defined in terms of the need(s) to be met or the problem to be solved. "Intended" is in quotes because programs sometimes are begun for reasons other than solving a real need or problem; e.g., a program may be funded simply because someone wants the program to be a source of jobs for program workers.

2. History, scope and nature of the need or problem.

   The origin, extent and significance of a problem or need. For example, most everyone knows about the problem of HIV. However, what many don't know is how extensive is the spread of HIV, how fast rates of infection are increasing, what factors are contributing to its spread among different subgroups and how difficult it is to accurately assess the need.

3. Social, political and historical context of the program.

   The context of a program to solve a problem or address a need refers to the social and political environment that led to the development of the program. The historical context includes previous attempts made to alleviate the problem by other means. The social and political context describes the forces at work to shape public opinion and public policy. In the example of HIV, public programs to eradicate HIV and to educate the public about HIV didn't receive much support as long as it remained a problem of homosexuals and drug users. Once it spread to the general populace, public outcries for HIV programs pressured political leaders into taking action.
4. Purpose of the program and how it addresses the problem.

What is the specific approach that the program takes to address the need or problem and what are the specific program activities related to that approach. For example, one program approach to the problem of HIV is teaching safe sex while another approach is teaching abstinence. Both approaches seek to eliminate the primary mode of HIV transmission: the exchange of bodily fluids.

5. Underlying theory supporting the purpose of the program.

Theory as defined here can also mean the value system on which a theory is based to incorporate the constructivist view of reality. For example, "social learning theory" is the basic, underlying theory supporting the teaching of safe sex or abstinence as a way to change behavior. However, specific interpretations of that theory result from different value systems. Consequently, those supporting an abstinence program may argue that teaching safe sex encourages promiscuity.

6. Mission, goals and objectives of the program.

These are what the program and the people who designed the program say it's supposed to do and what it's supposed to accomplish. The mission is a general, sometimes philosophical statement of purpose. Goals are statements of the desired, ultimate program outcomes while objectives are statements of immediate and intermediate-range program outcomes.

7. Relation of program activities to mission, goals, and objectives.

A program initiates activities leading to the achievement of a set of objectives which, in turn, lead to the achievement of a goal and ultimately, fulfilling the mission of the program. Part of the evaluator's job is to check the correspondence among these elements as well as matching them back to the problem the program is supposed to solve.

8. Client population served by the program.

Clients are the intended recipients of products or services provided by a program. Identification of the target population allows the evaluator to determine if the actual people served by the program are those for whom the program is intended.

9. Program methods for identifying clients and levels of service.

Except for instances where the client population is fully identified and never changes during the life of a program, there will be some program mechanism to identify who should receive services, what services they should receive and how much. If the "wrong" people are receiving services, most likely, the fault lies here.
10. The stakeholders and their interest in the program.

There are four basic groups of people who have a stake in the operation and outcome of a program: the people who receive service (clients), the people who provide the service and manage the program (program managers and staff), the people who fund the program (public and private agencies) and the people ultimately affected by the outcome of the program (society). Each group will have a different interest in the outcome of a program.

11. The decision-makers' and policy-makers' interest in the program.

These two groups are different from stakeholders in that their responsibilities aren't directly affected by the outcome of a program; i.e., they will continue to make decisions and set policy regardless of whether a program succeeds, fails or doesn't even exist. However, the kinds of decisions they make and the policy they set may be influenced by a program's outcome.


Decision-makers and policy-makers primary needs are for documentation of program operation and program outcomes and for timely and accurate information on which to base decisions. Their goals are to hold programs accountable for their legal responsibilities.

13. Needs and goals of the stakeholders and program funders.

Program funders goals are to get their monies worth from a program; i.e., efficiency and fiscal accountability. Their needs include receiving timely financial statements and evidence of cost benefits and cost effectiveness. Program managers have a goal of ensuring accountability of coverage and service delivery as well as operating a program as efficiently as possible. Their needs include daily information on service delivery and program performance.


The goals of clients can be generalized to those which deal with their quality of life and sense of self-worth. Their needs include getting the services they require in an equitable manner and being treated with respect. The goal of the community is to be free of the problems that led to the need for the program. Their needs also pertain to improving their quality of life and knowing that their support of a program has been worthwhile.

15. Needs and goals of the evaluators.

The evaluator goals include conducting the best possible analysis of a program given the available resources and constraints. Needs include cooperation among stakeholders and support from decision-makers, policy-makers and the community and recognition for a job well-done.
16. **Purpose of the evaluation.**

What the evaluation intends to discover or demonstrate about a program. The purpose also includes why the evaluation is being done at this time and in this manner.

17. **Decisions dependent on the evaluation.**

Decisions may range from deciding the ultimate fate of a program or simply what program activities need to be changed as a result of the evaluation.

18. **Ethical issues surrounding the evaluation.**

There will be different ethical issues depending upon the intended purpose of the evaluation. For example, one concerns whether preliminary evaluation data should be used to modify an on-going program; i.e., be fed back into the planning and implementation process. While this seems like a reasonable, logical and ethical thing to do, it may disrupt a planned research design that assumes no change in program parameters from start to finish. Another ethical dilemma arises when the evaluator has a stake in the outcome of the evaluation. If an evaluator has an interest in a competing program or wants to change the existing program in some specific way, then discovering program success would run counter to his desires. Conversely, an evaluator would be reluctant to discover program failure if, in doing so, it would jeopardize his present or future working relationship with the program managers, administrators or stakeholders.

19. **Audience(s) of the evaluation report(s).**

It's important to identify the information needs and levels of sophistication of the recipients of evaluation reports. The utility of an evaluation report depends upon the relevance of the information to the needs of the audience and its comprehensibility.

20. **Evaluability of the program.**

Before embarking on a program evaluation, one has to determine if an evaluation is doable; i.e., if there are sufficient resources and minimal constraints for doing evaluation. For example, if an evaluation requires raw client utilization data on a daily basis and the program only keep summary counts by quarter, evaluation of service delivery will be difficult.

21. **Evaluation resources, constraints and limitations.**

Identifying resources, constraints and limitations is a function of doing an evaluability assessment and in the design of an evaluation plan. The design of an evaluation has to take into account the resources that will facilitate evaluation and the constraints and limitations that will inhibit it.
22. The benefits and costs of doing the evaluation.

Also related to evaluability assessment, even if an evaluation is doable, it may not be worth the time, cost and effort if no one will make use of the results. Conversely, the potential benefits of an evaluation may be high but, so, too, are the costs to do it right. One should take into account both monetary costs and non-monetary costs (such as a loss of credibility).

23. Evaluation questions to be asked.

These are specific research questions pertaining to program operation and program impact to be answered by the analysis of the data collected; "How many people have gone through this program?" and "How has their behavior changed as a result of this program?"

24. Expected outcomes and evidence of program achievement

The expected outcomes should be given by the specific program objectives. From the objectives, the evaluator will need to decide if they already specify what the measures of program achievement are to be. If not, the evaluator will need to acquire or develop them.

25. Possible unexpected outcomes of program implementation.

Because the evaluator is often not part of the program and can objectively assess program objectives and activities, it's possible to anticipate when and where unexpected outcomes will occur. So, if an increase in the proportion of people reporting safe sex practices is an expected outcome, an unexpected outcome might be an increase (or decrease) in the reported frequency of sexual activity.

26. The proposed process or improvement evaluation design.

This is the methodology and plans to analyze program inputs and outputs such as the numbers of people served, activities performed and objectives achieved.

27. The proposed impact or effectiveness evaluation design.

This is the methodology and plans to analyze the effects of a program's output on program participants.
28. Available sources of data to be used in the evaluation.

The watchword here is "Don't reinvent the wheel." If demographic data can be obtained from census data, client records or some other source, evaluators shouldn't waste their time creating a survey instrument to capture that data. There are a lot of potentially valuable sources of data kept by the program and by other agencies that an evaluator can use.

29. Data sources and collection instruments to be developed.

For all the data requirements not met by existing data sources, the evaluator will have to develop them from scratch or adapt existing instruments.

30. Reliability and validity assessment of data sources.

Erroneous data lead to erroneous conclusions. The evaluator has to ensure that the data measured is what it's supposed to be measured and that it's accurate and precise.

31. Pilot-testing of methods and materials.

Pilot-testing on a small scale before implementing at full-scale can save the evaluator (and the sponsors) lots of time, money and aggravation. Make sure your methods are viable, your materials are analyzable and your data is reliable and valid.

32. The evaluation management plan.

This is a strategic plan for the management of all evaluation activities. It will typically include a time and work schedule highlighting when and how the major tasks of the evaluation project will be done.

33. The proposed sampling plan.

Unless your evaluation budget and plan allows you to analyze an entire population, there will be a need to select a representative sample from that population. The sampling plan describes the size needed, the characteristics of the sample required and the methods to be used to obtain the sample.

34. Data coding and collection plan.

This is plan defining the types and sources of data to be collected, the methods used for data collection and the process used to code the data for analysis.
35. Qualitative analyses to be done.

As part of the evaluation design, the evaluator will list the specific types of qualitative analyses (ethnographic surveys, participant observations, etc) that will be done.

36. Quantitative analyses to be done.

Closely following a description of the quantitative analyses to be done, the evaluator will also specify the types of quantitative analyses (T-test, regression, etc) that will be done.

37. Statistical decisions to be made.

Statistical decisions pertain to two areas: (1) the assumptions about the data set required by the analyses to be done and (2) the decision rules that will be applied if hypothesis testing is to be done.

38. Cost-benefits analyses to be done

These are the analyses that relate benefits to program costs in terms of money saved. Cost benefits are stated in dollars.

39. Cost-effectiveness analyses to be done

These are the analyses that compare one program with another on the basis of the costs to provide a certain level of outcome. Cost effectiveness is expressed in units of program outcome.

40. Methods of reporting and presenting the results

What presentation techniques will be used to convey the results to your audience? Tables? Graphs? Will only summary statistics be presented or will more detailed information be shown? Will the evaluation report be in printed form or in electronic form?
LESSON THREE

Using Needs Assessment to Identify a Problem
and Relate it to Program Goals and Objectives;

OBJECTIVES

- Identify a program's intent from its mission statement, goals and objectives.
- Match a program's goals and objectives with its intent and activities.
- Classify objectives as measurable and non-measurable.
- Revise non-measurable objectives to make them measurable.
- Restate a problem as a "need" that identifies a gap between the way things are now and the way they should be.
- Define theory" and "theoretical framework" and its relationship to the problem and program intervention.

Keywords

research problem, observed problem, need, phenomena, theoretical framework, mission statement.

Interpreting a program's mission, goals, objectives and activities

If you were to create a hierarchy of program action statements from most general to most specific, it would be in the following order: mission, goal, objective and activity. A program mission is usually a global appraisal of a need with a noble-sounding premise such as "To end world hunger in our lifetime" or "To provide affordable, quality health care to low-income workers." As mentioned above, building coalitions among stakeholder necessitates having a mission that all can agree upon (even though it may have contradictory concepts such as "quality health care " and "affordable"). Nothing contained in a program mission statement is directly verifiable nor measurable.

Program goals are generally long-range statements of specific outcomes. They should directly address a stated need and indicate what change will occur by program's end or by some specified future date. A program to increase seat belt use might have as its goal "To increase the national average for seat belt use to 90 percent by the year 2000." Program goals that have a stated unit of measure will be much easier to evaluate than goals without a specific measure of change or impact.
Program objectives specify expected program outcomes. To be complete, program objectives should identify what is being measured in observable terms, the conditions under which it occurs and how it will be measured. Clearly written objectives typically have a single purpose and a single outcome expected to occur within a given time frame. As an example, a cholesterol screening program set up by a hospital might have as an objective "By March 31st, the program will be able to draw and analyze blood samples from 25 to 100 patients per day, five days a week."

Program activities are the day-to-day tasks that program staff, management and participants perform to achieve their objectives. Implicit in the hierarchy of mission, goals, objectives and activities is the dependency of one on the other. Taken together, the performance of daily activities are necessary to achieve specific short-term objectives. Achievement of these objectives should lead to the achievement of general long-term goals which, in turn result in ultimately fulfilling the program's mission.

Part of the evaluator's job is to check the correspondence among these elements as well as matching them back to the problem the program is supposed to target. Since the achievement of one leads to the achievement of the other, evaluators can check the correspondence by generating their own set of objectives (from the actual activities of the program) and their own set of goals (from their set of generated objectives). The evaluator would compare the generated set against the program's to see if the two reasonably agree.

If a program has no formally stated objectives or goals, this method can be used to create them. In this situation, the evaluator should ask program managers what they perceive their programs objectives or goals to be and then compare their responses to the evaluator's set of goals and objectives. A lack of correspondence among program activities, objectives and goals is one of the reasons why programs fail to fulfill their mission or to solve the problems they were designed to solve.

The problem with the word, "problem."

Before embarking on an evaluation to determine if a program has addressed a problem, the evaluator should be very specific in the use of the term, "problem." The evaluator should distinguish among an "observed problem" as typically contained in a program's mission statement, a "research problem" and a "need." A "need" is a discrepancy or "gap" between what is known and not known, between the ways things are and the way they should be or between two or more potentially related bodies of knowledge. An "observed problem" is an observation of the consequences resulting from a need. Whereas an observed problem focuses on the observable consequences of a need without ever defining the need, a "research problem" is an assumption that defines both the need and its consequences.

The following are observed problems:

1. The proportion of black felons sentenced as "habitual offenders" by judges is twice that of white felons similarly sentenced.

2. Leon County has the highest rate of gonorrhea in the state.
3. Thirty-seven million Americans are without health insurance.

4. Men on low cholesterol diets are more likely to die as a result of violence than men who eat normally.

5. Over 25,000 people died last year in alcohol-related traffic accidents.

While each of the above are observations about a specific area of concern, they don't define the need and are therefore not researchable problems. Here are the above observations restated as research problems:

1. Judges who are supposed to be impartial to race may be basing their sentencing decisions on skin color.

2. Instead of practicing safe sex, Leon County residents may be more likely to engage in behaviors that lead to the spread of sexually transmitted diseases.

3. All Americans require access to basic medical care. Since health insurance is intended to provide access, the 37 million people in this country without health insurance may be denied access to care.

4. Lowering cholesterol levels is supposed to increase longevity by reducing the risk of contracting certain fatal diseases; however, it doesn't appear to reduce the overall mortality rate.

5. Even though the media repeatedly warns the public about the dangers of drinking while driving, people who were involved in a traffic fatality may not have believed that alcohol could adversely affect their driving.

Each of the research problems above focuses on the problem as a discrepancy between the way things are and the way things should be. Each research problem also identifies the "likely" consequences of that discrepancy. The purpose of the research will be to systematically gather evidence that will allow evaluators to measure that likelihood.

Three ways (at least) of defining the "need"

- Need may depend entirely upon what consumers or potential consumers of services say they want. This is a "consumer-oriented" approach to defining need.
- Need may be estimated by analyzing the people who now use services. This is called the "rates-under-treatment" approach.
- Need may be defined by using experts to specify the criteria for service type, location and eligibility. This is called the "expert judgement" approach.
The role of needs assessment

One of the difficulties in conducting an objective evaluation of an existing program is that it may have been initiated in response to subjective assessments of the original problem. Those who are too close to a problem tend to exaggerate it whereas those who are unfamiliar with a problem will minimize its importance. Different groups of stakeholders may have different perspectives of the nature and scope of the problem as well as having different interests in the success of the program.

A needs assessment is a special type of evaluation that serves to document a program's existence. Needs assessment is a systematic and objective process that identifies the problem a program should address irrespective of any preconceived notions of what the program's purpose should be. Needs assessments make use of a systematic problem-solving model to identify a problem and to generate a series of intervention strategies to solve the problem. Stakeholders will then rank the list of possible strategies on the basis of preference and arrive at a choice of strategies presumably by consensus. The selected strategy will embody the rationale used to define a program's intent as well as the method and means for program design and implementation.

Since a needs assessment provides the rationale and justification for a program intervention, it should logically precede its implementation. However, very few programs are preceded by a well-conceived needs assessment that identifies the problem and proposes solution alternatives. Too often, stakeholders choose a solution first, develop a program to implement it and then request an evaluation long after a program has been in operation.

Needs assessment techniques - direct methods

key informant surveys

These are interviews and surveys of individuals who have direct knowledge of the extent of a need by virtue of being in contact with the source of a problem. An example would be case workers who have a better appraisal of the incidence of child abuse than program planners would.

community forums

These are gatherings of concerned community leaders and citizens much like a typical "town meeting" except that its purpose is to elicit responses to structured questions about needs.

field surveys

Field surveys, like ethnographic or epidemiological surveys, are in-depth interviews or observations conducted in a community or subpopulation to identify areas of need.
Needs assessment techniques - indirect methods

Social indicator comparisons

This type of needs assessment is done by using statistical models to indirectly estimate need for services. The models use social data such as number of households in poverty, divorce rate, suicide rate, single person households, etc. The models can range from simple, ranking procedures to complex regression analyses.

Analysis of rates-under-treatment

This type of needs assessment is done by applying observed treatment rates found in demographic subgroups to similar subgroups living in the area where the needs assessment is to be done. Once the numbers of cases are estimated in all subgroups, they are added to obtain a total figure of need for that area.

One of the problems in conducting needs assessments of an ongoing service program is separating out estimates of who "receives" services versus those who "need" services. Unmet need can lead to political consequences such as public reaction or even lawsuits over inadequate or inequitable funding. On the other hand, identifying unmet need may lead to program expansion and improvements in the allocation of resources.

Determining the types, numbers and geographic distributions of persons needing services of various kinds is the principle goal of social services needs assessment. Other goals of needs assessment include deriving estimates of how much each kind of service is necessary to meet each type of problem and making estimates of the duration and resultant cost of the service incidents required.

Types of needs assessment

Needs based on demonstration of achievement

After a program has been in operation for a period of time, this needs assessment seeks to identify the gaps between desired outcomes and the actual immediate, intermediate and ultimate outcomes that a program produces. For example, if all participants in a weight reduction program have lost an expected number of pounds, then, for this particular group, result and moment in time, there is no need for program change. This type of needs assessment attempts to identify the level of achievement in relation to predetermined criteria or standards.
Needs based on the desired level of achievement

This is like the previous type except that a person's opinion of the desired level of program achievement or deficiency is used instead of actual measured outcomes. This needs assessment compares what people expect a program can do with what they would like that program to do.

Needs based on the importance of achievement

This type of needs assessment rank-orders desired outcomes according to their perceived importance to a person.

Needs based on the satisfaction with program outcomes

This needs assessment measures customer satisfaction or a person's opinion of the quality and effectiveness of service provided by a program.

Needs based on requests for service

This type differs from the previous ones in that it implies a personal dissatisfaction with a program -- this dissatisfaction forms the basis of the need for program change.

A "cooking analogy" to defining the problem

Imagine that your boss asks you to prepare a covered dish for an office party that will feed at least eight of your co-workers. The problem is that you don't have anything already made you can bring to the party and you don't want to bring a store-prepared item. Here's how you would define the problem.

Arising from the problem is the purpose for this activity; i.e., what you will do to solve the problem. Before you can decide on a solution, you need to consider what that solution will look like and what it will accomplish. These are the criteria for your problem solution:

Affordable - does not exceed your budget for food

Easy to prepare - can be completed with minimal additional expertise or training

Produces desired and expected effect - The finished product will what the recipe says it is. It will look and taste like it is supposed to.

Reliable - If the recipe calls for cooking 30 minutes in a 350 degree over, you expect no variance from 30 minutes or 350 degrees.
Conforms to requirements of hosts - the hosts don't want 30 dishes of baked beans or cole slaw nor do they want something like lobster newburg that only feeds one, so they will ensure that each participant bring something according to a set of specifications; i.e., salad, dessert, side dish, feeds 8 people.

Satisfies personal needs - you have a need to produce something of value either because its expected or required of you and/or because it will bring you recognition (assuming people like it).

Satisfies needs of hosts - the hosts do not have the resources to prepare a multitude of dishes to feed a large group: that is why they have others doing it.

People will benefit from it - Hopefully, it will satisfy their hunger and perhaps acquaint them with new ways of preparing food which they may use in the future.

After determining the solution requirements, the next step in the process is locating the sources of information you will use to support your solution to the problem. In this example, these would be cookbooks, newspapers, magazines, television shows and personal cooking experience.

A number of resources will be required to accomplish your solution - resources required might be a stove, an oven, a beater, mixing bowls, measuring devices, baking dish, utensils, time keeper and the ingredients for the meal. Some of these resources will already be in your possession; some may be borrowed and the rest will need to be purchased from outside sources. The feasibility of your solution will greatly depend upon the quality of the resources you can assemble.

The solution requirements imposed by a funding agency would be analogous to this cooking example in that your solution must fit within an allotted budget, be doable given your resources and reliably produce the desired outcome. Just as the accumulation and allocation of resources is critical to achieving your cooking goal, so, too, is the accumulation and allocation of resources critical to achieving your evaluation goal.

The point of this exercise is that careful planning is necessary to achieve success in any project involving many related activities and interdependent resources. The research skills and experience of the evaluator determine the overall quality of the evaluation.
LESSON FOUR

Survey Methods for Needs Assessment and Outcome Assessment

OBJECTIVES

- Define and give examples of four levels of measurement
- Define and give examples of five measurement domains
- Define and give examples of eight types of scales
- Contrast the five types of measurement scales
- Compare and contrast interval versus ordinal scales
- Give examples of simple and weighted composite scales

Keywords

nominal, ordinal, interval, ratio, attitudes, opinions, beliefs, behavior, attributes, Likert, Semantic differential, Thorndike, Rating, Open-ended, Partially Closed, Paired Alternatives, Matching/Attribute Sampling,

What do surveys measure?

The most accurate and objective way to assess the behavior of people is to observe them repeatedly and directly over time. If we want to know how often people go to food stores, we could follow them every time they leave their house. If we wanted to know what they eat for breakfast, we could sit at their table in the morning and record what they ate. As you may suspect, we rarely have and opportunity to directly observe people and their behavior. Also, there are aspects of human behavior that are not directly observable -- such as attitudes, beliefs and opinions. What are the alternatives to direct observation?

One of the alternatives is to find a measure that represents the attitudes, behaviors, opinions and beliefs of people. The alternative is the survey: a systematic way of asking people to volunteer information about their attitudes, behaviors, opinions and beliefs by responding to questions. The success of this technique depends upon how closely the responses match reality or how well we are able to repeatedly measure human characteristics and behavior through the process of reported answers to surveys.

The first issue that a researcher needs to address is whether the survey method will produce the information that is needed. Is this survey necessary? Is the purpose of the research to generate hypotheses, to test hypotheses, to generate projections, to evaluate people or programs? Can the data be obtained by other means? What level of detail is required?
The second issue is how accurate will this method be. How available are reliable and valid indicators? Can the same data be obtained on several occasions and from different settings? What is the range and scope of data available? How generalizable will the results be? How doable, given one's resources is this method?

Uses of surveys

**Descriptive research**

The goal is to obtain a precise measurement of certain phenomena such as political party preference. The purpose is to conceptualize a phenomenon and summarize it.

**Causal explanation**

To establish a causal connection, three conditions must be satisfied. First the assumed cause and effect must be associated with each other. Second, the cause must precede the effect. Third, all other possible explanations of the effect have been ruled out.

**Prediction**

Using data from survey to forecast results. Accuracy depends upon how stable the measures are over time.

The specific applications of surveys range from highly practical public opinion polls and market research studies to highly theoretical analyses of social influence. Planners and administrators in many countries have used surveys as a rapid and effective means of gathering base-line information for policy decisions. Social scientists use surveys to measure voter behavior, psychological influences on the spending and saving behavior of consumers, attitudes, values and beliefs related to economic growth and the correlates of mental health and illness. Economists rely on regular consumer surveys for information on family financial conditions and surveys of business establishments to measure recent investment outlays.

The survey is an appropriate means of gathering information under three conditions: when the goals of the research call for quantitative data, when the information sought is specific and familiar to the respondents and the researcher has prior knowledge of the responses likely to emerge.
Choosing the types of data to collect

The method of analysis that you choose and the type of research design will impose some limits on the types of data you can use. Choosing a survey design requires the researcher to decide exactly what will be measured and how it will be measured. When something is measured, it is assigned a unique position along a dimension or numerical scale. When numbers or words are used to identify or categorize things, the scale is called nominal or categorical. Nominal variables would be measures like sex and race. When the numbers represent an ordering of things, the scale is called ordinal. The ranking of football teams is done using an ordinal scale. The third type of scale is called an interval scale. An interval scale has units with equal intervals and an arbitrary zero point. The fourth type of scale is called a ratio scale. A ratio scale has equal units and an absolute or true zero point.

Instruments or measuring scales in the social sciences are the counterpart of the weight scales, measuring tape, thermometer and various other type of physical devices used for scientific research. Unlike the physical sciences, the social sciences are faced with the problem of measuring properties that are constantly changing. When researchers measure the characteristics of people, they look at people's behaviors, attitudes, beliefs and opinions. With the exception of behaviors, the other aspects must be inferred from surrogate measures.

The kinds of information sought

- attitudes -- what people say they want
- opinions -- what people think might be true
- beliefs -- what people know is true
- behavior -- what people actually do
- attributes -- what people are (demographic characteristics)
- preferences -- what people would choose

It's critical to distinguish among the different types of information. Questions of each type tend to pose different writing problems. Often the true aim of a question is not what it appears to be (face validity). Some questions are particularly susceptible to wording. In which case, it is wise to ask several questions on the same subject.

Question types

- Open-ended
- Partially Closed
- Closed

Rating scales
- Semantic differential
- Likert
- Thorndike

Checklists

Categorical attribute items

Ranking scales
- Rank ordered alternatives
Open-ended questions are used primarily in two distinctly different situations: one, where respondents can express themselves freely and two, where a precise response is needed and listing all of the possible responses would be unwieldy. Open-ended questions are used when the researcher cannot anticipate what the possible responses will be. They are used when no limit is placed on the length and depth of a response. They are most useful for exploratory studies.

Open-ended questions of this type exhibit some problems. These type of questions tend to be very demanding. Often the answers will be too short, uninterpretable, or irrelevant. Another disadvantage is the difficulty in quantifying them for statistical purposes. However, analyzing their content will provide for meaningful statistics and description. Unlike the first type of open-ended question, the second type doesn't encourage free expression; rather, it seeks a specific fact. One possible disadvantage is that the respondent may not recall all possible answers (if that is required).

Close-ended questions with ordered answer choices presents each choice as a gradation of a single dimension of some concept or construct. Questions with ordered choices tend to be quite specific, restricting respondents to thinking about a very limited aspect of life in a very limited way. To answer them, respondents must identify the response dimension that underlies the answer choices and place themselves at the most appropriate point on a scale that is implied by the answer choices. This type of question uses the responses to determine the extent to which each respondent differs from every other one.

Close-ended questions with unordered answer choices have each choice as an independent alternative. To answer, it is necessary to evaluate each alternative separately.

Partially closed questions provide an "Other" category under which a respondent can provide additional alternate information. This is a compromise between close-ended and open-ended forms.

Semantic differential scales make use of a series of seven-point rating scales with each end point anchored by an adjective or phrase. These adjectives, called bipolar adjectives, are direct opposites. The scales have three recurring factors: an Evaluative factor covering such dimensions as good-bad, pleasant-unpleasant and positive-negative; a Potency factor representing the dimensions of strong-weak, hard-soft and heavy-light, and an Activity factor with such scales as fast-slow, active-passive and excitable-calm.

Likert scales are steps on a continuum like the semantic differential except that the scale may cover any range of values and may have more or less than seven steps. Like the semantic differential, the steps are identified by adjectives or phrases -- however, unlike the semantic differential, all of the steps may have an adjective or phrase associated with it. Typical of this type of Likert scale is one that measures agreement using a five point scale with steps labelled "Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree."
In a Thorndike scale, there are a series of statements that cover a range of possible responses. Thorndike scales are created by first asking people to respond to an open-ended question. Then, the most commonly mentioned responses comprise the question alternatives.

Checklists are lists of items that respondents can check off. Usually they are questions about attributes or facts and the respondent is asked to check as many items as they feel applies to them. Categorical attribute items are like checklists except that respondents select one and only one alternative (e.g., questions of race, sex, SES, etc.).

Ranking scales require the respondent to "sort" items by assigning a numerical ranking (e.g., 1st, 2nd, 3rd...etc.) to each response item. The main difference between ranking scales and rating scales is that rating scales assign an absolute value to an item without regard to the other items; ranking scales assign a relative value to an item in relation to the items before and after it. Paired alternatives is a special type of ranking scale that compares every response item with each other. Paired alternative scales provide a more accurate ranking of items than can be done on a simple, rank-order scale.
LESSON FIVE

Assessing Reliability and Validity
of Evaluation Data;

OBJECTIVES

- Define "random variance", "true variance," "measurement error"
- Describe the rationale for domain sampling
- Define "reliability", "validity", "precision", "accuracy"

Keywords

Face validity, content validity, concurrent validity, discriminant validity, construct validity, test-retest reliability, internal consistency, stability.

Reliability and sources of variation

Measurement theory maintains that the variance of a score can be partitioned into two parts: true score variance (due to random fluctuations) and error variance (due to measurement error). Reliability is actually a ratio of true score variance to total variance—the smaller the contribution of measurement error to the total variance, the more reliable (and accurate) the score. Random variation is something that occurs naturally in this world in all things. Error variance is more complex because it is derived from many sources. These sources are primarily the questions on the instrument, the persons who administer the instrument, and the people who respond to the instrument.

The importance of assessing reliability and validity

In order for information to be useful, it has to be consistent, dependable, accurate and, most of all, true. Too often, we are presented with information that fails on one or more of these criteria. In research, these criteria are represented by the concepts of reliability and validity.

When we say that information is reliable, we mean that we can expect to obtain the same information time after time. The concept of reliability can be applied to sampling. If we repeatedly draw random samples of equal size from a population, we can expect to get the same sample mean and standard deviation each time (plus or minus a certain amount due to sampling error). If a person can add a series of two-digit numbers, it shouldn't matter which two digit numbers we use to test that ability. We can repeatedly sample addition problems and expect that the person will be able to solve all of them.
When we say that information is valid, we mean that it is presented or used in the way for which it was intended. An IQ test is valid only if it is used to measure intelligence -- it is not valid if it used to assign individuals to groups. A psychological test that is a valid measure of anxiety is not a valid measure of stress.

Types of validity.

Face validity is when the obtained information appears to be what was expected. A question that asked "Do you smoke?" would appear to have face validity as a measure of smoking behavior.

Content validity is when the measurement adequately reflect the major factors of an underlying behavior. Content validity is established by having experts who are knowledgeable with the data that the test seeks to measure evaluate the relevance of the test items to that data. The question on smoking by itself would not have content validity for smoking behavior since it does not adequately represent the dimensions of smoking behavior.

Concurrent validity is when an assessment device is comparable to another assessment device that validly measures the same content or construct. Concurrent validity is established by correlating one test with another that has previously been validated. An example might be correlating the results of a new I.Q. test with the Stanford-Binet I.Q. test (a well-validated intelligence scale).

A variant of concurrent validity is discriminant validity -- the assessment device is associated with behaviors that discriminate one group from another. Achievement tests that are used to determine success in college would have discriminant validity if the students who graduate score higher on the test than students who leave before graduation.

Another variant of concurrent and discriminant validity is predictive validity which holds that an assessment device can be used to predict behavior. Predictive validity is to concurrent validity what linear regression is to correlation. The intent of the GRE (Graduate Record Examination) is to determine how successful students are likely to be in graduate school. When graduate GPA is regressed on GRE scores, the resulting R-squared is quite large. By virtue of its ability to predict GPA in graduate school, the GRE has "predictive validity" with respect to its measurement intent.

The last type of validity is called construct validity. A construct is a theoretical dimension (construct) upon which an assessment device is based. Construct validity is established by relating the correlational structure of a set of questions to a defined construct. Factor analysis is the most common mathematical tool used to evaluate construct validity. An example of construct validity might be if answers to the following questions are highly correlated, then one might assume an underlying construct of "unsafe driving" is what is actually being measured:

Q: I usually drive faster than the speed limit. YES NO
Q: I try to "get even" with drivers who cut me off. YES NO
Q: I usually don't have time to signal lane changes. YES  NO

An example of a familiar construct in psychology is self-esteem. Self-esteem does not exist by itself but is represented by behavioral responses. In this example, construct validity measures the extent to which these behavioral responses can be labelled as self-esteem. Construct validity is determined by having experts on a particular behavior rate the adequacy of an assessment device in measuring that behavior.

**Types of reliability**

**Test-retest reliability** is obtained by administering the same test on two or more successive occasions and then correlating the scores.

**Internal consistency** is obtained by correlating the scores on several questions that pertain to the same content to the sum total of the scores. The average item-total correlation is a measure of how consistently people respond to related items on a test. For example: if a math test had several sets of items that required people to multiply two and three digit numbers, you would expect persons who could not correctly multiply two digit numbers to be unable to multiply three digit numbers. Likewise, you would expect persons who can multiply three digit numbers to be able to multiply two-digit numbers. These expectations should be "consistent" from person to person and would be borne out by high item-total correlations of the items.

**Stability** refers to how much a person's score can be expected to change from one administration to the next. A perfectly stable measure will produce exactly the same scores time after time. This concept is similar to test-retest except that in test-retest situations there is no assumption that the absolute value of each person's test score will stay the same.
LESSON SIX

Quantitative and Qualitative Evaluation Designs

OBJECTIVES

- Classify types of evaluation designs
- List advantages and disadvantages of experimental and quasi-experimental designs
- Give examples of evaluation designs and identify the advantages and disadvantages of each
- Given an evaluation design, select an appropriate statistical treatment
- Given an evaluation problem, decide on necessary resources

Keywords

The importance of the evaluation design

The term, "evaluation design," may refer to the type of evaluation procedures used or may refer to an experimental design. In the most general sense, the evaluation design is a set of rules and systematic procedures for collecting, analyzing and interpreting data. One other characteristic of evaluation design is that the most commonly used types have well-known labels (case study, experiment, survey, interview).

Evaluation designs may be of two major types: qualitative and quantitative. In qualitative designs, the main purpose is to collect and describe information about unknown phenomena that will be useful for more in-depth analysis. Qualitative designs do not make any suppositions about the collected data other than it is valid, reliable and representative. Quantitative studies also require the collection of valid, reliable and representative data; however, the data collected must be sufficient enough to make inferences about the uncollected data and about unknown phenomena. Researchers who conduct qualitative studies limit their judgements about a data set to whatever the data set has to show. There are no theories to test nor predictions to make. In contrast, quantitative studies require evaluators to demonstrate that their evaluation findings relate to specific theories and that they represent true events and not chance occurrences.

The choice of evaluation design depends in large measure on the nature of evaluation problem and the purpose of the evaluation. Qualitative evaluation primarily involves qualitative designs and descriptive methods while quantitative evaluation employs quantitative designs and inferential methods. Regardless of the type of problem or evaluation design, there should always be descriptive data provided. The types of descriptive data will, of course, vary depending on the evaluation design.
Qualitative methods like participant interviewing, field work, case study, involves having the evaluator interact with the people being studied in their natural setting. The purpose or objective of qualitative evaluation is to describe social phenomena from the perspective of the people being studied and not the evaluators. Evaluators use three general methods for making observations in a systematic way: (1) a set scheme for classifying events; (2) a data guide containing a set of short-answer questions, or (3) a structured rating scheme. Systematic observations include observable physical actions as well as expressed emotions, attitudes and beliefs.

Unlike qualitative designs, quantitative designs do not require any interaction between evaluator and participants (unless the project is designed that way) nor do they require people to be studied in their natural setting.

**Types of quantitative designs**

**One-shot case study (or posttest only design)**

This method involves studying a single group once following its exposure to some condition or treatment. This is actually a qualitative design rather than quantitative, experimental design. An example of this type of design is a post-class survey given to people who attend a driver improvement class. Students are asked to indicate if the class has changed their attitudes about driving. There is no pre-class survey (pretest) to compare with the post-class survey (posttest) and no control group for comparison. There is no way to determine what effect the class had on attitudes. Since none of the extraneous variables are controlled, one can't make any reliable statements about the effect of the treatment (the class) on the outcome (the survey of attitudes).

**One group pretest-posttest design**

This design overcomes the obvious limitations of the posttest only design by assessing a group before a treatment and then following treatment. The main problem with designs of this nature is the time between the administration of the pretest and the posttest. If the span of time is too long, changes in the group's responses on the posttest may be due to unknown events occurring in the interim. The best pretest-posttest design is one with a very short interim time (a week or less). The other problems with this design are (1) test effect where the group has been sensitized to the pretest (has learned the pretest) and does "better" on the posttest and (2) the stability of what the tests measures changes over time.
**Static group comparison design**

In this design, a control group is used to compare against a treatment group. The control group is a "no-treatment" group. Ideally, people are randomly assigning to both groups -- however, this design is often used with an intact treatment group and a randomly selected or cohort-matched control group. The major weakness of this design is justifying that the treatment was valid and not a Hawthorne effect. Since the treatment group received attention and the no-contact control group did not, one should always suspect whether the contact itself (rather than the treatment) led to the effects measured. Another weakness of this design is the assumption that both the treatment and control groups were identical at the beginning of the study. If both the treatment and control groups consist of randomly assigned participants, then this assumption is probably met.

**Pretest-posttest control group design**

This design overcomes the weakness of justifying the treatment in the static group comparison design by administering a pretest to both the treatment and control groups. Again, random assignment to groups is the best way to ensure that treatment and control groups are equivalent. The control group is still a no-contact group -- unless people could normally experience the experimental treatment outside of the experiment, then a no-contact control does not really contribute much beyond showing that an effect did occur to the treatment group. For example, if the treatment is a class in quantum physics, it is unlikely that people who do not take the class will pick up that knowledge on their own.

A better design would be to have a control group that receives a treatment equal in time and substance to the treatment but differs in focus or intent. Using the example of the driver improvement class, having a control group take a class equal in length and content amount but dealing with a subject other than safe driving -- like automobile safety devices.

**Time series and repeated measures designs**

The basic concept underlying this design is that the treatment group also serves as the control group by taking many measurements of the treatment group across time. In time-series designs many measurements are taken before, during, and after the treatment has occurred. Measurements taken before the treatment serve to describe a baseline period (where the behavior of the measurement is consistent and steady). Measurements taken during the treatment can then be compared back to the baseline to measure changes in level or trend. Finally, measurements taken after the treatment can then be compared back to the treatment and baseline to measure the overall impact of the treatment. One of the key elements of a time-series design is that all measurements must be made within equally spaced time intervals.
Repeated measures designs are similar to time-series design in that multiple measurements are made -- however, that is where the similarity ends. In repeated measures designs, there may be many measurements of one treatment or one measurement of many treatments. The key element in a repeated measure design is that all measurements are related or associated. If the same treatment is given many times, then, by virtue of it being the same treatment, all measurements will be related. The main strength of this design is that multiple measurements will provide stronger evidence that a treatment effect has occurred.

**Confounding factors that threaten the validity of the evaluation**

Confounding factors are extraneous factors that explain in whole or in part changes in the target population. Since the evaluator's role is to demonstrate that only program impact was responsible for changes in the target population, control of extraneous factors is an essential element in effective evaluation designs. There are several broad categories of confounding factors:

**Testing.** The effect of taking a pretest on subsequent posttests.

**Endogenous change.** Social programs operate in environments where ordinary and natural sequences of events influences outcomes.

**Secular drift.** Long-term trends may produce changes in outcomes that enhance or mask the true effects of a program.

**Interfering events.** Short-term events may produce changes in outcomes that enhance or mask the true effects of a program. An example would be a natural disaster, a social upheaval or a sudden economic downturn.

**Maturational trends.** Over time maturational processes may produce changes that mimic or mask program effects.

**Selection.** This is the processes that are not under control of the evaluator that lead to some types of targets being more likely to participate in a program under evaluation. Self-selection is a good example. Selection bias: people who volunteer for job training programs are more likely to look for jobs.

**Stochastic effects.** Whether effects are due to chance effects or true differences. Statistical regression to the mean.

**Instrumentation.** Occur when changes in the calibration of an instrument or how an instruments is used from one time to the next.

**Reactive effects of testing.** Occurs when the participants react to the testing.

**Reactive effects of innovation.** Also known as the Hawthorne effect. Participants may perform better simply because they're excited about taking part in an innovative program.
Multiple program interference. When the impact is a consequence of two or more program effects interacting.

Contaminants. Contaminants means that program effects rarely occur in isolation. The site where a program is provided, the delivery system, the physical plant and personnel involved can all effect the outcome of a program.
Qualitative Designs

Qualitative designs are characterized by the methodology used to assess a subgroup or population. The main methods used in qualitative designs are surveys, interviews, observation and field studies. (NOTE: surveys were covered the preceding lessons).

Qualitative interviews

There are four types of qualitative interviews that researchers currently use. The first is called an "in-depth, unstructured interview." Respondents are encouraged to talk on a subject selected by the researcher but the direction the discussion takes is guided by the respondents and not the researcher. The second type is called an "in-depth, structured interview." This method uses open-ended questions to structure the content of the interview; however, there is no particular order to the interview topics. The third is called the "focused interview." This type is also a structured in-depth interview -- the difference between this one and the former is that the focused interview concentrates on the reactions of respondents to specific experiences. This method is often employed in focus groups which typically consist of eight to ten people who share a common background or orientation. The fourth type is the "ethnographic interview." This type of interview is unstructured and non-directive. It seeks to uncover data about people's behaviors and beliefs as perceived within their social and cultural context.

All qualitative interviewing techniques allow the interviewer flexibility in the choice of topics and questions. They use open-ended questions rather than closed or fixed-choice. Information is provided from the respondent's point of view and care is taken to avoid biasing or influencing the responses. All require and encourage rapport between the interviewer and respondent.

Participant observation

This is a technique commonly used in anthropological studies. It involves taking an active role in the culture one is studying to understand it from the perspective of its members. The researcher becomes an unobtrusive observer by blending in with the group being observed. An educational example of participant observation would be having the researcher attend a class as a student to observe teacher-student interactions. An example from social work or mental health would be having the researcher spend time with a family to observe its functioning and interpersonal dynamics.

Field studies

Field studies are a form of observational data collection. In field studies, researchers keep notes of every contact they have with others and of every activity that occurs during their observations. Later, these notes are compared for content validity and assessing interrater reliability.
LESSON SEVEN

Developing a Sampling Plan

OBJECTIVES

- Distinguish between various types of sampling giving examples of when they would be applicable.
- Determine sample size given desired precision
- Perform various types of sampling
- Use random number table for random sampling and random assignment
- Determine size of sample based upon various study factors
- Contrast matched pairs and cohort studies with random sample
- Perform test of randomness and representativeness
- Identify methods for reducing initial sample differences
- Identify common sampling errors and how to correct them.
- Event sampling and construct sampling versus people sampling.
- Identify biased and unbiased sampling estimators

Keywords

sample, random selection, random assignment, stratified sample, cluster sample, systematic sample, simple random, systematic, multistage random, stratified, cluster, stratified cluster, multiple/sequential, judgement/quota.

Identifying the target population

From the target population will come the people to be studied or served by the program. Populations are groups of people assigned on the basis of common characteristics. The population is delimited by its characteristics. Thus, all males in the Tallahassee area would be one population; all black males, single, 20 to 25, currently unemployed, would be another. Obviously, the size of the population decreases as the number of characteristics increases.

The concept of population in evaluation has often been a source of confusion. The reason is because in statistics, population also refers to a group of independently measured quantities: IQ scores, age, sex, race, blood pressure, number of arrests, and so on. In quantitative designs, evaluators make judgements about the mathematical nature of the measurements. In qualitative designs, evaluators make judgements about the people who are the source of the measurements. Quantitative -- populations of numbers; qualitative -- populations of people.

This is not to say that quantitative designs ignore populations of people for the sake of measurement. Numbers would be meaningless if they were not related back to the people who produced them. That is why quantitative designs pay particular attention to selecting samples of people that are large enough to be representative and typical of their population.
Sampling theory and sample selection

One big advantage of qualitative studies over quantitative is that you don't have to worry about sample size. Probably the most often question asked of statistical consultants is, "How large a sample should I take?" The stock answer to that question is, "As many as you can afford!" The size of the sample is a function of your research design and your desire for precision. What is of most value are the methods for determining the minimum sample size needed for a particular type of analysis and a particular degree of precision.

In survey research, the primary decision the researcher has to make is "How close to the true response do I want the obtained response to be?" In experimental research, the primary decision the researcher has to make is "How small is the phenomena I'm searching for and what level of risk am I willing to take to find it?" In both instances, the end result is determining the level of confidence that the results the researcher obtains are true and a function of the observations and not due to external factors (such as pure chance).

The basis for sampling theory is grounded on measurement theory. The theory holds that any measurement is comprised of two parts: a "true" score and measurement error. The smaller the measurement error, the more precise (closer to the true score) is the measurement. How does measurement error arise? It's primarily comprised of two sources: one arising from sampling, and the other from the instability of the measured phenomena. Sampling theory focuses on why sampling error occurs and how to minimize it.

If you flip a coin, the chances of it being heads or tails is one out of two (1/2) or 50 percent. Does that mean you will alternately get a head followed by a tail (or vice versa)? Of course not. You're more likely to get several heads or tails in a row. If the coin is unbiased (meaning each side is the same weight and shape), after many, many flips, the number of heads and tails should even out. For any number of coin flips, the expected proportion of heads (or tails) will always be 50 percent. However, in practice the actual proportion obtained will be more or less than 50 percent due to random variation.

What is random variation? In the coin flip example, each coin flip is independent of every other coin flip. The likelihood of it coming up heads or tails on each successive coin flip has nothing to do with the outcome of the previous flip. This is what is meant by random: every coin flip is equally likely to produce a head or a tail and each head or tail flipped is independent of every other head or tail flip.

One of the most useful and common statistics that we experience in our daily lives is the arithmetic average of a group of numbers -- also known as the mean. If you are a baseball fan, you are well aware of batting averages. If you are a football fan, you can probably recite the average yards per game of your favorite team. When you flip a coin 100 times, the average number of heads you expect to get will be 50. The average serves two purposes: one, it describes the value that is most typical of a group of numbers and two, it is the most likely value you'd get if you were to randomly select a number from a group of numbers.
When you randomly select a number from a group of numbers, that value may not equal the average but will be more or less or equal to the average. This variation from the average is a function of sampling. The more numbers you sample, the closer will the average of those numbers be to the average of the main group of numbers. The average of the sample is used to estimate the average of the population from which its drawn. The larger the sample, the more alike will the numbers (scores) be to those of the population, the less will be the variation, and the sample mean will be closer to the population mean.

The other source of variation mentioned above was measurement error. If we use a coin with one side heavier than the other, the results of the coin flips will be biased toward that heavier end and the outcome will not be as expected. This is the variation due to the nature of the measurement. The main purpose of statistical analysis is to identify and separate the true value from measurement error and random variance.

**Target estimation**

A variety of techniques can be used to estimate the target population. One is the key informant approach -- identifying surveying and collecting information from knowledgeable leaders. The next is community forum approach which is like holding a town meeting. The third is the rates under treatment approach which is like benchmarking the services used by a similar community for the same target population. The fourth is indicators approach or statistical data from federal, state and community offices. The fifth way is through sample surveys and censuses.

**Census or sample?**

A sample is, by definition, a part of a larger population. When the entire population is available for study, the result is a census. Since statistics describe the characteristics of populations based upon sample characteristics, there are no statistics to be collected from a census. This may be a bit disconcerting to the novice researcher who plans on doing a statistical treatment of the data. This author once worked with a graduate student who was studying an entire population. I advised him that no statistical analysis need be done since he was working with a population. When he told his doctoral committee about what I had said, they recommended that he take a sample so that he could do statistical analysis!
When speaking of population characteristics, researchers refer to them as parameters. When speaking of sample characteristics that are used to estimate population parameters, researchers refer to them as statistics. Since statisticians deal with populations of measurement and not people, one can correctly reason that every population of numbers (or observations) represents a subpopulation or sample drawn from a much larger population of numbers (or observations). There is conceptually no such thing as one and only one measurement. A measurement (observation) is a specific event that occurs at a specific point in time. That event represents one of thousands of other similar events that could occur at different points in time. As you will learn, true scientists never trust a single measurement made at a single point in time.

The concept that measurements we make are actually samples of the total population of measurements that could be made at this or any other time is the basis for reliability theory. When we say that a method yields a reliable measurement, we mean that the method will produce the same measurement over and over again regardless of when or how it is made. Reliability of measurement is an important issue given that one of the goals of research is to produce reproducible results.

**Obtaining an ample sample**

When researchers are unable to obtain or analyze an entire population of people or numbers, they will need to obtain a subset or sample of the population. Below is a list of the objectives and advantages of sampling and a discussion of each:

1. to obtain a manageable collection of objects to study
2. to provide a qualitative representation of population characteristics
3. to provide quantitative estimates of population characteristics (parameters)
4. to control for factors that are extraneous to the design and focus of the research

**Obtaining a manageable collection of objects to study**

Ever wonder why political researchers sample about 1,000 people for their polls? The main reason is that it's a lot easier (and cheaper) than surveying all 260 million people in this country! While this example is perhaps a bit extreme, it illustrates the problem faced by all researchers: "How large of a sample can I afford to take?" The counterpart to that question is "How small a sample can I afford to take?" The difference between the two problems is how one defines, "afford."
Larger samples require more resources to manage them. As with all projects, there are upper limits to the amount of resources a researcher can dedicate to the sample being served or analyzed. Resources cost time and money and researchers on a budget must know fairly clearly how much they can afford to spend. At the other end of the spectrum, the costs of taking a sample that is too small may be less tangible but more serious. The danger of taking too small a sample are that (1) due to attrition, the study cannot be completed as designed, (2) the phenomena under study are too subtle to be measured in a small sample, (3) the analytical procedures used will be invalidated or unreliable and (4) the results of the project will not be generalizable to real-world applications.

The first steps in the sample selection process are to identify the goals of the sampling, define the limitations and constraints of the project budget, sampling methodology, and target population and determine the minimum sample size required for a reliable, valid and relevant analysis.

Providing a qualitative representation of population characteristics

A sample can be thought of as a portrait of a population. The sample is the population in a microcosm. All that can be known about the population must be inferred from the sample. The goal therefore is to draw a sample that adequately represents the population on all relevant and important characteristics. For characteristics to be relevant they must adequately define the target population and relate to the design and purpose of the study; to be important they must be ones which influence the outcome of the study. An ideal sample is one that differs from its population only in size.
Providing quantitative estimates of population characteristics

Here, too, the goal is to draw a sample that adequately represents the population on all relevant and important characteristics. The difference between qualitative and quantitative characteristics of a sample is that the qualitative characteristics of the population are known while the quantitative characteristics typically are not. When some of the quantitative characteristics are known, then, these can be used to assess how representative the sample is of the population in much the same way as was done with qualitative characteristics.

Generally, the quantitative characteristics of a sample (statistics) are used to estimate the unknown quantitative characteristics of a population (parameters). The larger the size of the sample, the more precise the population estimates. Statistical analysis is based on the mathematical relationships between sample statistics and population parameters. By observing the outcomes of sampling, statisticians have developed mathematical probability models that can be used to determine the likelihood of each of those outcomes. For example, using a model of probability called the binomial expansion, statisticians can determine the likelihood of getting 75 heads in 100 coin tosses (given that the expected number is 50).

What the likelihood estimate is really indicating is whether that particular sample was drawn from a population with known parameters. The "population with known parameters" may be either a real population or a theoretical population of numbers created mathematically by repeated sampling of numbers similar in nature to that of the sample (as in the case of the binomial population of coin flips). If the mean of a population is known (for the binomial population or distribution, the mean is .50), then, it is a simple mathematical procedure to determine the likelihood that the mean of the sample represents (is similar to) the mean of the population.

Controlling for extraneous factors

In experimental research, one of the key goals is to have a well-controlled study. Control in this instance refers to reducing or eliminating the effects from sources other than what is being directly studies or manipulated by the researcher. Recall that data consists of two sources of error: measurement error and random error. Since random error is strictly a function of sampling, proper random sampling methods will ensure that the error from sampling will not exceed a known maximum. Measurement error is another story since error can arise from many sources in the process of measurement. A typical problem in social research are population characteristics that are interrelated such as race and income. If a researcher is interested in looking at the effect of race on a particular outcome, the design of the research needs to account for (or control for) the possible effect that income may have on that outcome given its strong association with race.
A special form of sampling could be used to reduce the likelihood of income being a factor. The technique is called stratified sampling. In this method, the population is divided into groups or strata based upon differing levels of a population characteristic. In the case of income, the technique might be to let five income ranges represent the strata and then to draw an equal number of samples from each strata. If, on the other hand, you were interested in measuring the impact of income on an outcome, you would want to ensure that your sample has the same levels of income in the same relative amounts (proportions) as the population. To accomplish this would also require stratified sampling with one difference: the number of values sampled in each level or strata would be according to the same proportions as the population.

This how a sample of 300 would be drawn:

<table>
<thead>
<tr>
<th>Income Range</th>
<th>Percent of Population</th>
<th>Number in sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>less than 10,000</td>
<td>35</td>
<td>105</td>
</tr>
<tr>
<td>10,000 - 19,999</td>
<td>25</td>
<td>75</td>
</tr>
<tr>
<td>20,000 - 39,999</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>40,000 - 59,999</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>above 60,000</td>
<td>5</td>
<td>15</td>
</tr>
</tbody>
</table>

Here is a list of the types of sampling procedures commonly used in research:

- **simple random** - assign numbers to objects and then use a table of random numbers to select objects
- **systematic** - first object selected at random, then every other object selected at equal intervals.
- **multistage random**
- **stratified** - sampling done within levels of a variable
- **cluster** - sampling done within clusters of like objects
- **stratified cluster** - combination of cluster and stratified
- **multiple/sequential** - sampling done in stages according to progressively specific criteria
- **judgement/quota** - set number needed for sample selected on basis of physical characteristics
- **convenience sample** - you take whatever you can get
**Intact groups**

In many areas of social research, many groups receiving services, treatment, or analysis were already in existence before the advent of research and therefore were not created by any sampling procedure. A classroom of students or traffic violators would be such a group. When people are assigned to a group on the basis of some action they take to be in that group, then they comprise a self-selected sample. An example would be people who volunteer for drug testing. The problem with intact groups or self-selected groups is that there is no control over their representation: they may be totally unlike the population from where they reside. Additionally, in self-selected groups, the reason for the existence of the group depends upon factors that are common among group members but not common among population members. In other words, any analysis of this group would be biased towards whatever factors led to its creation.

One of the methods used to control for extraneous factors in intact groups is to find or assemble another group with similar characteristics. This research method is known as a cohort study or matched pairs design. The logic behind this approach is to strive for a situation where the only difference between two groups of people is group membership and not individual characteristics.

**Methods for determining sample size**

In statistics there are some general rules of thumb regarding what is a small and what is a large sample. The number that divides small from large is 30. This doesn't mean that you will only need 30 people or values in any situation; nor does it mean that you will need that many. What it means is that for most samples of thirty, the characteristics of the sample begin to approach those of the population. The kinds of analyses that you intend on doing will dictate the size of the sample you will need. In every case, though, the size will be a function of how precise you wish your measurements to be.

**Sample size in survey research**

Determining a sample size for estimating a single parameter (like the average number of heads in 100 coin tosses) is easy and straight-forward. Determining the appropriate sample size for a survey that may ask 100 questions (and each question could be considered a parameter), is not so easy and requires some compromises. The compromise occurs in making an assumption about the data that's collected. If each item on a survey will be analyzed apart from every other item, the researcher will assume that each item was obtained independently of every other item. Alternately, the researcher will assume that each item was answered independently of every other item? Why is this important? Because the mathematics of random sampling require that the drawing of an object from a population is independent of every other drawing. Note: it is the act of selecting the object that is independent -- not the object itself.
By treating each survey measurement as an independently sampled event, the researcher can apply the rules for determining precision on one measurement to all measurements taken. Survey researchers speak of "statistical confidence" and "error rate." For example, political surveys typically will quote the percentage of people responding "Yes" and "No" to a question along with a "margin of error" that is so many percentage points greater or lesser than the quoted amount. In a sample of 800 to 1,000 persons, that margin of error (error rate) will be plus or minus 3.5 percent. It's 3.5 percent regardless of the question asked since the error is a function of the size of the sample. A sample size of 400 would have an error rate of plus or minus 5 percent.

Since the survey is only performed one time, researchers would like to know how confident they can be in the results they obtain. If they were to repeat the sampling 100 times, how many times would they get a sample that yields similar results. When political researchers quote error rates, they should also quote their confidence in that error rate. The concept of confidence also relates to sample size. If you have a very large sample, you need not repeat the sampling process many times in order to capture the true values of the population characteristics. However, with large populations and small samples, you would need to do a very large number of repeated samplings in order to be sure of capturing the true values. If you were to repeat the sampling process 100 times and 90 of the 100 samples taken would produce the same percentages and error rates, you would be confident that you've obtained the true value 90 out of 100 times. This is what is meant by a 90 percent level of confidence.
When you see or hear the results of a political poll, you can rest assured that the researchers are working under a 90 percent level of confidence. Can you have higher levels of confidence, say 100 percent? Yes, but, it will require a larger sample size for the same margin of error. If we think of error rate as relating to the precision of a sample, then the level of confidence would be its accuracy.

**Sample size in hypothesis testing**

The essence of hypothesis testing is to determine whether a sample is drawn from a population with known characteristics or from a population with unknown characteristics. If we know how all values in a population are distributed, then we can tell if a sample is from that distribution by looking at the value of its characteristics. The characteristics of samples that are most important are the measurements that describe the sample as a whole: its mean and standard deviation. In doing repeated sampling, the value of the sample mean will vary above and below an average value: the average being the mean of the sample means. How much it varies is its error rate. The larger the sample, the less it varies and the smaller the error rate. Using a similar procedure for determining sample size for categorical responses (such as "Yes" and "No") we can determine what sample is required given a particular error rate desired.

Since we can estimate all values of a distribution given its mean and standard deviation, we can determine the likelihood of any value being part of that distribution. That is what is done in hypothesis testing. If we have a distribution of sample means and the standard deviation of that distribution, we can determine if the sample we have was drawn from that population by comparing the mean of the sample to the distribution of sample means. If the sample mean value does not fall anywhere within the range of values in the distribution, then we can say that the sample was drawn from a "different" population.

Fortunately for researchers, there are published tables that indicate the required sample size given the degree of precision one desires and the type of analyses performed.
LESSON EIGHT

Devising a Data Collection Plan

OBJECTIVES

- Operationally define the variables of interest in the study
- Demonstrate procedures for developing a survey
- Define the major methods of data collection in quantitative studies and qualitative studies
- Demonstrate how to determine if responses are random and representative
- Given non-quantified information, develop a coding scheme that reduces the data to quantifiable terms.

Keywords

variable, coding scheme, tabulation,

Sources of data

Sources of data are categorized into two groups: primary sources and secondary sources. Primary sources of data are obtained directly from project participants and are made directly by the project evaluators. Secondary sources of data are obtained from sources and individuals not directly connected with your project. In an evaluation proposal, you would normally review secondary sources of data in the related literature section of your project proposal. Do not however, confuse secondary sources with secondary analysis.

Secondary analysis refers to a reanalysis of a previously studied data set with a perspective different from prior evaluators or for a purpose different from prior evaluation objectives. The main advantage of secondary analysis is that collecting and reanalyzing existing data is much cheaper and faster than having to plan, collect and analyze new data. Given that data collection represents a major cost of evaluation, secondary analysis can accomplish more with existing data since the focus of allocations will be towards analysis.

The biggest problem for the evaluator who intends on doing secondary analysis is locating data sources that are relevant to the evaluation problem. Although there are thousands of studies published in the literature, evaluators only release summaries of the data and not the actual raw data collected. The reasons for not making the original data set available are that (1) the original data has been lost or destroyed, (2) the data set is considered proprietary information, (3) evaluators may be reluctant to have their work scrutinized, (4) evaluators may not have nor wish to expend the resources required to produce the data set and (5) the data may not be in a format applicable to the secondary analyst's needs.
The collection of data for secondary analysis involves locating catalogs, guides, and directories of data lists from library sources, choosing evaluation studies that are related to what you intend on doing and ordering the codebooks (a codebook assigns a range of numbers for each variable corresponding to the various response categories). The evaluator then either borrows or purchases the data sets and copies them.

My philosophy and strategy regarding sources of data is to locate and tabulate any existing sources of data that will satisfy your evaluation needs and the needs of your evaluation objectives. The counterpart of "reinventing the wheel" in evaluation is the collection of data that has already exists. The likelihood that the data you need already exists is greater than the likelihood it doesn't. However, given the reasons listed above, the likelihood that the original data set is not readily available nor in a useable format is also high. It is up to the evaluator to do a diligent search and review of all existing data sources before setting out to collect one's own.

The fine art of coding

Coding is the process of converting information into a quantifiable format (usually numerical) so that a systematic analysis of the information can be done. The coding process is often made easier by precoding responses and enabling circled or checked numbers to be entered directly into a data set. The codebook is like a foreign language dictionary in that it translates English into numerical values.

Since nearly all analyses of large data sets will occur on a computer, the coding system used should model or be easily convertible to the format required by the computer's software. If the information is nominal or categorical in nature, then, the choice of whether to represent a data value as a number or a label depends upon the analyses that will be done. If you are only going to measure frequency counts and percentages, then either a number or label is acceptable. If you are going to do any statistical analyses, then numbers are mandatory. Labels cannot be manipulated mathematically. If you are collecting qualitative data and intend on doing content analysis, then your data might be entered as words and phrases instead of numbers.

When data can't be collected on one or more variables from one or more sources, then this creates the problem of missing values. Statistical computer programs deal with missing values in different ways. Generally, they give you the option of identifying what values are to be considered, "Missing," or allowing the computer to assume that zeroes or blank spaces represent missing values. Should you assign a number to represent missing values? Yes, if you intend on either analyzing the missing values or replacing them with other values. However, you must inform the software what your missing values are. What about blanks?
When data is entered into a computer, it will generally be organized in a spreadsheet orientation where the rows represent different cases and the columns represent different variables. If you leave spaces blank instead of typing in a value, you may not know if the space represents a missing value or a missing data entry. By entering values in every space provided, you will know exactly what data has or has not collected.

**Data entry and quality control**

Data entry is the process of transferring or transcribing the information from coding sheets, field notes, surveys and surveys onto the computer. Sometimes the process of data entry is facilitated by use of machine-readable sheets or by image scanners and optical character recognition software -- in which case, manual data entry is not required. However, whether data entry is accomplished by person or by computer, there is a need to check the accuracy of the data (especially in the latter case). Since your analyses hinge on the accuracy of the data, good quality control procedures are needed for error checking and entry consistency.

There are computer programs that facilitate the quality control function of data entry directly by checking for acceptable values at the time they are entered. For example, if you recording ages of adults in a study, values lower than 18 would be errors. Manual checking of the data may or may not spot the error. What the data entry program would do is immediately reject any entered value that does not fall within the accepted range or is not of the proper form. If you don't have this program, then you might use the capability of your statistical analysis program to identify incorrect values and then to recode them.

**Missing data handling**

Prepare for nonresponse as treating it as one of the categories (example, a "No" response). Other situations, you might want to assign the average value to retain the rest of the data for analyses that require it. If you're doing individual item analysis, then you'd drop the missing data -- adding constants don't change the overall results. For statistical purposes, you'd need to keep as much of the original sample size as possible.

Regression can be used to estimate missing data using answers to other responses. The problem may be avoided by having more than one question dealing with a topic. Not everything you ask will be of equal importance -- some items you may afford to lose data. The main problem with nonresponse is that you don't know if someone skipped a question or decided not to respond. Including a non-response category will ensure that people respond. Remind them both at the beginning of the survey and at the end to answer ALL times. Don't give people a way not to respond -- give them the opportunity to respond. You can identify respondents who essentially are non-participants: people who mark the middle response for all items. Only slightly better than nonresponders. Do you keep them in the data bank? No, since their responses will skew the data.
LESSON NINE

Pilot Testing

OBJECTIVES

- Determine if pilot testing of project will be necessary
- Determine sample size needed
- Demonstrate techniques for debriefing participants
- List types of information to be collected during pilot test

Keywords

debriefing, formative evaluation, reliability, content validity, construct validity, concurrent validity, predictive validity, face validity

Why do a pilot study?

The pilot study is useful for demonstrating instrument reliability, the practicality of procedures, the availability of volunteers, the variability of observed events as a basis for power tests, participants' capabilities or the investigator's skills. The pilot study is a good way to determine the necessary sample size needed for experimental designs. From the findings of the pilot study, the evaluator can estimate the expected group mean differences as well as the error variance. Even a modest pilot study conducted informally can reveal flaws in the evaluation design or methodology beforehand.

Any surveys that have not been used in the past or have been modified in any way should always be pilot-tested. Any procedures that require complex instructions should be pilot-tested. Any methodology requiring time estimates should be pilot-tested.

Selecting the pilot study sample

The sample for the pilot study should be as close as possible to the actual sample that will be drawn for the main project. When this is not possible, then you should strive to get a sample with approximately similar characteristics. Depending upon the availability of participants, you may need to save as many participants for the main project as you can -- in which case, you don't want to include them in a pilot study.
Some evaluators often will do a pilot study on a subset of their sample and then include them as part of the main sample. That is tantamount to mixing apples and oranges. If you make any change whatsoever to your study as a consequence of the pilot, then the participants in the pilot will have experienced something different from those in the main study. Additionally, one of the purposes of doing a pilot study is to debrief the participants after the study by asking questions about the methods, instruments, and procedures.

Information to be collected

The pilot study should be run exactly as if it were the actual study. The exception here is that you will be collecting data on how long procedures take, what actions facilitate or inhibit the operation of the study, whether instructions are understood and if the data you obtain is in the form expected.

It may be necessary to have more than one pilot study -- especially in the situation where instructional materials or methods have been developed. In the case of instructional materials or methods, you would do a formative evaluation of the materials and methods. Unlike a pilot study where the evaluator may not interact with participants, you would be asking questions of the participants as they read, watch, or listen to the instruction and when they are quizzed on what they have learned.

Participant debriefing

If your study involves surveys or interviews with people, you should have a debriefing session at the completion of the pilot study. Ask the participants if they understood all of the instructions, if they had any particular problem with any of the questions asked, if they understood the intent of the study, and if they had any recommendations how to improve the study.
LESSON TEN

Data Analysis
Descriptive Statistics to Hypothesis Testing;

OBJECTIVES

- Give examples of descriptive statistics
- Explain how graphical techniques can be used to analyze data.
- Given a data set, create a bar graph, pie chart and trend line.
- Given sets of research questions and data, select the most appropriate statistical analysis
- Describe four parametric tests of significance and their nonparametric equivalents.
- Construct a chi-square table
- Describe when nonparametric statistics should be used

Keywords

content analysis, frequency distribution, mean, median, standard deviation, histogram, pie chart, T-test, ANOVA, Mann-Whitney, Friedman, chi-square, association, Phi, Rho, Tau, Lambda, PRE, alpha, beta, power, significance test.

Descriptive analysis

Descriptive analysis involves describing the common underlying characteristics of data. In quantitative evaluation, descriptive analysis involves arranging the data into a frequency distribution which groups each value into categories from low to high. If it is a normal distribution, then most of the values will fall towards the center of the distribution and decrease in frequency further out from the center. The two most important descriptive statistics of a normal distribution are the mean and the standard deviation. The mean is a measure of central tendency (in addition to the median and mode) and the standard deviation is a measure of dispersion (in addition to the range variance).

If you are analyzing nominal or categorical variables, then you would want to know how many and what percent of each value fell within each category. If you are analyzing ordinal variables, in addition to what you would show for categorical variables, you might want to know the average ranking of a variable. The average rank is calculated in the same manner as the mean except that ranks are added, not values. Ranking is a process of ordering values from low to high and then assigning a rank from 1 to n (where n is the number of values). In the Division 1-A College Football Poll, judges assign a rank from 1 to 25 to 25 of the 216 colleges in that division. The average of the judges ranks then represent the actual ranking of a team in the poll.
An extension of constructing frequency distributions for one variable is to construct a combination of two related frequency distributions. The process is called cross-tabulation and the product is called a contingency table. The contingency table shows one variable's categories broken down by another. The contingency table is one of the ways that an evaluator can assess the relationship between categorical variables.

To assess the strength and statistical significance of the relationship between variables, evaluators have a number of procedures to choose from. These procedures come under the heading of inferential statistics.

**Inferential statistics**

Inferential statistics are used to make judgements about the relationship between two or more variables and to determine the likelihood that samples are drawn from the same or different populations. Measures of association allow an evaluator to determine the amount of change in one variable that is a function of change in another. Common measures of association for categorical variables are chi-square, phi coefficient, contingency coefficient; for ordinal variables are Kendall's Tau, Somer D, and Gamma, for interval variables, Pearson's correlation coefficient.

Inferential statistics that are used to determine if two or more samples are drawn from the same or different population are called measures of difference. These are the statistics commonly used in significance testing and hypothesis testing. Significance testing involves determining the probability that two or more samples are drawn from the same population. Hypothesis testing involves making decisions on whether to reject or not to reject a test of significance showing that two or more samples are drawn from the same population.

One of the most common tests of significance used to determine if two samples are drawn from the same population is the t-test. When there are more than two groups to be compared, a technique called analysis of variance (ANOVA) is used. When the goal of the analysis is to assess the relationship between one variable and several variables, a technique called multiple regression is used. These techniques are used when the sample data is drawn from normally distributed populations. Since the function of these statistics involves estimating parameters from normally distributed populations, they are known as parametric statistics. When the type or shape of the population distribution is not normal or is unknown, then a series of techniques called nonparametrics procedures must be used. The nonparametric tests of significance that have parametric equivalents are the Mann-Whitney (equivalent to the t-test) and the Friedman (equivalent to analysis of variance or ANOVA).

The choice of statistical analysis depends upon whether the evaluator is interested in testing hypotheses or in testing the significance of the results. Both will assist in providing answers to evaluation questions. The evaluator should be aware of the intent of each and not confuse their purposes. Also, an evaluation study is not made "more scientific" by the addition of hypotheses to test -- hypothesis testing is only relevant for experimental evaluation designs.
A BRIEF STATISTICS PRIMER

1-TAILED PROBABILITY

Probability of obtaining results as extreme as the one observed, and in the same direction, when the null hypothesis is true.

2-TAILED PROBABILITY

The probability of obtaining results as extreme as the one observed, in either direction, when the null hypothesis is true.

95% CONFIDENCE INTERVAL FOR MEAN

A range of values that 95% of the time includes the population (true) value of the mean. It is approximately equal to the mean +/- (plus or minus) two times the standard deviation.

ALTERNATIVE HYPOTHESIS

A statement about a situation against which the null hypothesis will be tested. The null hypothesis is always tested against an alternative. (See NULL HYPOTHESIS).

ALPHA LEVEL

The probability of falsely rejecting the null hypothesis; i.e., rejecting the null hypothesis when it is true. The critical level of alpha is the smallest significance level at which the null hypothesis would be rejected.

ANALYSIS OF VARIANCE (ANOVA)

A procedure that divides the total variation in the dependent variable into components (effects) and produces tests for the statistical significance of each component. ANOVA is often used to generate an F-test for differences among three or more independent samples or groups.

BETA COEFFICIENT

Beta coefficients are the regression coefficients when all independent variables are expressed in standardized (Z-score) form. Transforming the independent variables to standardized form makes the coefficients more comparable since differences in the units of measurement are eliminated. However, the beta coefficients do not in any absolute sense reflect the importance of the various independent variables since they depend on the other variables in the equation.
**BETWEEN GROUPS**

The part of total variability in the dependent variable that can be accounted for by differences in group means.

**BETWEEN MEASURES**

The portion of the "within people" (within cases) variation of a scale that can be attributed to differences between the items.

**BETWEEN PEOPLE**

The portion of the total variation in a scale that can be attributed to differences among the cases.

**BINOMIAL TEST**

A test of whether a sample comes from a binomial distribution with the specified probability of success (p).

**BOX-PLOTS**

Plots of the distributions of a dependent variable for each level of an independent variable. The upper and lower boundaries of the boxes are the upper and lower quartiles. The box length is the interquartile distance so the box contains the middle 50 percent of values in a group. The asterisk (*) inside the box identifies the group median. The larger the box, the greater the spread of the observations. The lines emanating from each box (the whiskers) extend to the smallest and largest observations (marked with X) in a group that are less than one inter-quartile range from the end of the box. O (outlier) marks points outside this range but less than 1.5 interquartile distances away. E marks points more than 1.5 interquartile ranges from the end of the box.

**CHI-SQUARE**

Statistic used to test the hypothesis that the row and column variables are independent. It measures the discrepancy between observed frequencies and expected frequencies.

**CLUSTER ANALYSIS**

A statistical procedure that identifies homogeneous groups or clusters of cases based on their values for a set of variables.
COCHRAN Q TEST

A test of the null hypothesis that several related dichotomous variables have the same mean. The variables are measured on the same individual or on matched individuals. This is an extension of the McNemar test to the k-sample situation.

COCHRAN'S C

Test that groups come from populations with the same variance. It is based on the ratio of the largest group variance to the sum of all the group variances. For sufficiently large sample sizes, a nonsignificant P value means there is insufficient evidence that the variances differ. The test is quite sensitive to departures from normality.

COEFFICIENT OF CONCORDANCE

A measure of agreement of raters. Each case is a judge or rater and each variable is an item or person being judged. For each variable, the sum of ranks is computed. Kendall's W ranges between 0 (no agreement) and 1 (complete agreement).

CONCORDANT

Two cases for which the values of both variables for one case are higher (or are both lower) than the corresponding values for the other case.

CONTINGENCY COEFFICIENT

A measure of association based on chi-square. This coefficient is always between 0 and 1, but it is not generally possible for it to attain the value of 1. The maximum value possible depends on the number of rows and columns in a table.

CONTINGENCY TABLE

A table containing the joint frequency distribution of two or more variables that have been classified into mutually exclusive categories.

CONTINUOUS VARIABLE

A variable that does not have a fixed number of values. For example, the variable INCOME, measured in dollars, can take on many different values.

COVARIANCE

An unstandardized measure of association between two variables.
COVARIATE

A concomitant variable that is measured in addition to the dependent variable in analysis of variance. It represents a source of variation in the dependent variable that has not been controlled for in the experiment. For example, in an experiment on reading comprehension the covariate might be the subject's IQ.

CRAMER'S V

A measure of association based on chi-square. Cramer's V is always between 0 and 1 and can attain a value of 1 for tables of any dimension.

CRONBACH'S ALPHA

Reliability coefficient based on the internal consistency of items within a test. It ranges in value from 0 to 1. A negative value for alpha indicates that items on the scale are negatively correlated and the reliability model is inappropriate.

CROSSTABULATION

A cross-classification table showing a cell for every combination of values for two or more variables. Each cell shows the number of cases having a specific combination of values.

DESCRIPTIVE STATISTICS

Summary information about the distribution, variability, and central tendency of a variable.

DIFFERENCES

Differences classified by sign - Diffs The number of differences between two variables having negative signs. + Diffs The number of differences between two variables having positive signs.

DISCORDANT

Two cases in which the value of one variable for a case is larger than the corresponding value for the other case, and the direction is reversed for the second variable.
DISCRETE VARIABLE

A variable that has a limited number of values. For example, INCOME with values high, medium, and low is a discrete variable. Compare continuous variable.

DISCRIMINANT ANALYSIS

A technique similar to regression that estimates the linear relationship between a categorical dependent variable (group membership) and one or more independent variables. The linear relationship is used to predict or explain group membership.

EFFECT SIZE

The magnitude of the difference between samples expressed in standardized units of the distribution of the differences.

ETA

A measure of association that is appropriate for a dependent variable measured on an interval scale and an independent variable with a limited number of categories. Eta is asymmetric and does not assume a linear relationship between the variables. Eta squared can be interpreted as the proportion of variance in the dependent variable explained by differences among groups.

ETA SQUARED

Eta squared is interpreted as the proportion of the total variability in the dependent variable that is accounted for by variation in the independent variable. It is the ratio of the between groups sum of squares to the total sum of squares.

FACTOR ANALYSIS

A technique based on correlation used to create a weighted, linear combinations of variables. The linear combinations, called factors, represent the unique shared variance components of the variables.

FRIEDMAN TWO-WAY ANOVA

Tests the null hypothesis that two or more related variables come from the same population. For each case, the k variables are ranked from 1 to k. The test statistic is based on these ranks.
GAMMA

A measure of association between two variables measured on an ordinal level. It can be thought of as the probability that a random pair of observations is concordant minus the probability that the pair is discordant, assuming the absence of ties. Gamma is symmetric and ranges between 0 and 1. If a crosstabulation involves more than two variables, conditional gamma is calculated for each subtable.

GOODMAN AND KRUSKAL’S LAMBDA

PRE measure of association which reflects the reduction in error when values of the independent variable are used to predict values of the dependent variable. A value of 1 means that the independent variable perfectly predicts the dependent variable. A value of 0 means that the independent variable is of no help in predicting the dependent variable.

HILOGLINEAR ANALYSIS

A technique similar to analysis of variance used to examine the relationships among the variables in a multiway crosstabulation.

HOMOGENEITY OF VARIANCES

Tests that the groups defined by an independent grouping variable are taken from populations with the same variance. Analysis of variance techniques work reasonably well even when the assumptions of equal variances and normal distributions are not exactly met.

INDEPENDENT SAMPLES

Samples selected in such a way that there is no relationship between the members of the samples. There is no pairing of observations between the samples.

INDEPENDENT T-TEST

A statistical test of the null hypothesis that the means of two independent samples are drawn from the same population.

INTERRUPTED TIME SERIES

A time series whose pattern changes due to a change in some outside condition which occurs at a known time. For example, you might be observing a time series measuring automobile crash fatalities before and after the passage of a seatbelt law.
KENDALL'S TAU

Nonparametric measure of correlation for ordinal variables that takes ties into account. It has a value between +1 and -1. Only in square tables can it attain +1 or -1.

KOLMOGOROV - SMIRNOV

Used to test the hypothesis that a sample comes from a particular distribution (uniform, normal, or Poisson). The value of the Kolmogorov-Smirnov Z is based on the largest absolute difference between the observed and the theoretical cumulative distributions.

KOLMOGOROV - SMIRNOV 2-SAMPLE

A test of whether two samples (groups) come from the same distribution. It is sensitive to any type of difference in the two distributions -- median, dispersion, skewness, etc. The test is based on the largest difference between the 2 cumulative distributions.
KURTOSIS

A measure of the extent to which observations are clustered in the tails. For a normal distribution, the value of the kurtosis statistic is 0. If a variable has a negative kurtosis, its distribution has lighter tails than a normal distribution. If a variable has a positive kurtosis, a larger proportion of cases fall into the tails of the distribution than into those of a normal distribution. With the skewness statistic, kurtosis is used to assess if a variable is normally distributed.

KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE

Tests whether two or more independent samples are from the same population. Nonparametric equivalent of the one-way ANOVA.

LAMBDA COEFFICIENT

In a loglinear model, the log of the expected frequencies in a cell is expressed as a linear combination of lambda coefficients. There are lambda coefficients for all levels of factors and interactions. Positive values of lambda indicate that a factor level or interaction term is associated with an increased number of cases in a cell.

MAIN EFFECTS

A component of the total variation in the dependent variable that can be attributed to a single independent variable or factor. In general, the greater the differences between the group means of the factor and its overall mean, the greater the main effect of the variable. The usefulness of a main effect test of significance depends on whether it is involved in significant interaction effects with other variables. If there is no significant interaction, the main effects are of interest.

MANN-WHITNEY U - WILCOXON RANK SUM W TEST

A nonparametric test that two independent samples come from the same population. It is more powerful than the median test since it uses the ranks of the cases.

MCNEMAR TEST

For 2 related dichotomous variables, tests whether the unlike pairs of responses (0,1) and (1,0) are equally likely. The test is most useful in 'before and after' experimental designs, to detect switches from one category to another.
MEAN

The sum of the values of all observations divided by the number of observations. It is also called the arithmetic average.

MEDIAN

The value above which and below which half of the cases fall. For example, if there are 5 cases, the median is the third largest (or smallest) observation. When there is an even number of observations, the median is the average of the 2 "middle" observations.

MEDIAN TEST

A test to determine if two or more independent samples have been drawn from populations with the same median. The median for all groups combined is calculated. A contingency table with the number of cases in each group greater than or less than/equal to the overall median is computed.

MINIMUM EXPECTED CELL FREQUENCY

The smallest number of cases in a cell that would be expected to occur if the null hypothesis is true. If any cell has an expected frequency less than 1, or if more than 20% of the cells have expected frequencies less than 5, the significance level of the chi-square statistic may not be correct.

MODE

The most frequently occurring value (or values). A unimodal distribution has one mode; a bimodal distribution, two modes.

MOSES TEST OF EXTREME REACTION

Nonparametric test for comparing the range of a variable for two groups. The group with the lower identifying value is labeled the "control" group and the other group is labeled the experimental group. The procedure combines the two groups and ranks values from smallest to largest. The span of the control group is the difference between the ranks of its largest and smallest values plus 1. An exact 1-tailed probability is computed for the span and then recomputed after dropping a specified number (default = 5%) of control group members from each end of its span. It is a measure of how much extreme values affect the span of the control group. Recoding to reverse the "control" and "experimental" groups is possible.
MULTIPLE ANALYSIS OF VARIANCE (MANOVA)

A series of procedures that incorporates methods common to ANOVA and FACTOR ANALYSIS to create linear combinations of two or more dependent variables and to divide the total variation of these combinations into components (effects) for testing.

MULTIPLE RANGE TEST

A procedure for comparing all possible pairs of group means. There are many varieties of multiple comparison tests which differ only in how they adjust for the fact that many comparisons are being made.

NULL HYPOTHESIS

The hypothesis of no difference. Actually, the null hypothesis is an assumption about a situation we hold to be true until we challenge it with evidence showing that an alternative situation is more likely. Thus, the null hypothesis is the "incumbent" and the alternative hypothesis is the "challenger."

OBSERVED SIGNIFICANCE LEVEL

The probability that a statistical result as extreme as the one observed could have occurred if the null hypothesis were true. The observed significance level is compared with the critical alpha level selected for the hypothesis test. If the observed significance level is equal to or smaller than the critical alpha level, the null hypothesis can be rejected.

ONE SAMPLE T-TEST

A test of the null hypothesis that the mean of a sample is equal to the mean of a population. The purpose is to determine if a sample is drawn from a population with a known mean.

PAIRED SAMPLES T-TEST

A statistical test of the null hypothesis that two means are equal used when the observations for the two groups can be paired in some way. Pairing can be used to reduce extraneous influences on the variable being tested.

PEARSON'S R

A measure of linear association. The value of R ranges between -1 (a perfect negative relationship in which all points fall on a line with negative slope) and +1 (a perfect positive relationship in which all points fall on a line with positive slope). A value of 0 indicates no linear relationship. The correlation coefficient is a symmetric measure.
PHI COEFFICIENT

A chi-square based measure of association that involves dividing the chi-square statistic by the sample size and taking the square root of the result. For tables in which one dimension is greater than 2, phi need not be bounded by 0 and 1.

POWER

The probability of correctly rejecting the null hypothesis; i.e., rejecting when the null hypothesis is false. Power, effect size, sample size and alpha level are interrelated. Power can be increased by (1) increasing sample size, (2) increasing effect size and/or (3) increasing alpha.

PRINCIPAL COMPONENTS ANALYSIS

Used to form uncorrelated linear combinations of the observed variables. The first component has maximum variance. Successive components explain progressively smaller portions of the variance and are all uncorrelated with each other. Principal components analysis is used to obtain the initial factor solution. It can be used when a correlation matrix is singular.

PROPORTIONAL REDUCTION IN ERROR (PRE)

Measures of association that are essentially ratios of a measure of error in predicting the values of one variable based on knowledge of that variable alone and the same measure of error applied to predictions based on knowledge of an additional variable.

RANGE

The difference between the maximum value and the minimum value of a variable.

REGRESSION ANALYSIS

Estimation of the linear relationship between a continuous dependent variable and one or more independent variables or covariates.

REGRESSION COEFFICIENT

Estimate of the change in the dependent variable that can be attributed to a change of one unit in the independent variable. Sometimes B is called the unstandardized regression coefficient and, in multiple regression, the partial regression coefficient.
R SQUARED

A measure of the goodness of fit of a linear model. It is sometimes called the coefficient of determination. It is the square of the multiple R, the correlation of the observed and predicted values of the dependent variable. It is the proportion of the variation in the dependent variable explained by the regression model.

RELIABILITY COEFFICIENT

An index of how reliable a scale is as an estimate of a case's true score.

RELIABILITY ANALYSIS

Procedure for evaluating multiple-item additive scales. In general, the concept of reliability refers to how accurate, on the average, the estimate of the true score is in a population of objects to be measured.

REPEATED MEASURES

Variables consisting of two or more related measurements made on the same sample. Typically, the measurements are of the same type and are taken at successive time intervals under constant conditions and/or taken simultaneously under varying conditions.

RUNS

Any sequence of like observations in an ordered group of values. The likeness may be the same value, the same sign, or being from the same sample.

RUNS TEST

A one-sample nonparametric test for randomness in a dichotomous variable. A run is any sequence of cases having the same value. The total number of runs in a sample is a measure of randomness in the order of the cases in the sample. Too many or too few runs can suggest a non-random (dependent) ordering. The runs test is only appropriate when the order of cases is meaningful.

SIGN TEST

A nonparametric procedure used with two related samples to test the hypothesis that the distributions of two variables are the same. The sign test makes no assumptions about the shape of the distributions. The differences between the two variables for all cases are computed and classified as either positive, negative, or tied. If the two variables are similarly distributed, the numbers of positive and negative differences will not be significantly different.
SKEWNESS

An index of the degree to which a distribution is not symmetric or to which the tail of the distribution is skewed or extends to the left or right. In a normal distribution which is symmetrical, the skewness statistic is zero. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. With the kurtosis statistic, skewness is used to assess if a variable is normally distributed.

SOMERS' D

An asymmetric extension of gamma that differs only in the inclusion of the number of pairs not tied on the independent variable. Somers' D indicates the proportionate excess of concordant pairs over discordant among pairs not tied on the independent variable. SPSS/PC+ also calculates a symmetric version of this statistic.

SPEARMAN'S RHO

A measure of symmetrical association equivalent to Pearson's R in every respect except that the variables correlated are ordinal.

STANDARD DEVIATION

The square root of the variance. The standard deviation is a measure of dispersion that is expressed in the same units of measurement as the observations.

STANDARD ERROR OF THE ESTIMATE

An estimate of the standard deviation of the error terms in regression. It is an estimate of the standard deviation of the dependent variable for cases which have the same values of the independent variables.

STANDARD ERROR

The standard deviation of the sampling distribution for a statistic. For example, the standard error of the mean is the standard deviation of the sample mean.

T-TEST

Statistic used for testing the null hypothesis that two means are equal. In regression analysis, the t statistic is used to test the null hypothesis that there is no linear relationship between a dependent variable and an independent variable or, in other words, that a regression coefficient is equal to 0.
TIME SERIES

A variable whose values represent equally spaced observations of a phenomenon over time.

UNCERTAINTY COEFFICIENT

A measure of the proportional reduction in error based on entropy criteria. SPSS/PC+ calculates both symmetric and asymmetric versions. The closer the uncertainty coefficient comes to its upper bound of 1, the more information that is provided about the value of the second variable from knowledge of an observation's value on the first. Its lower bound is zero when no information about the value of the second variable is obtained from knowledge of an observation's value on the first.

VARIANCE

A measure of the dispersion of values about the mean. It is computed as the sum of the squared deviations from the mean divided by one less than the total number of valid observations.

WALD-WOLFOWITZ RUNS TEST

A nonparametric test of the hypothesis that 2 samples come from the same population. The values of the observations from both samples are combined and ranked from smallest to largest. Runs are sequences of values from the same group. If there are too few runs, it suggests that the 2 samples come from different distributions.

WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST

A nonparametric procedure used with two related samples to test the null hypothesis that the distributions of two variables are the same. It makes no assumptions about the shapes of the distribution of the two variables. The absolute values of the differences between the two variables are calculated for each case and ranked from smallest to largest. The test statistic is based on the sums of ranks for negative and positive differences.

WITHIN PEOPLE

Estimate of the variability of the responses of the same individual.

Z-SCORE

A transformation for standardizing the scale of a variable. Z-scores are computed by subtracting the original mean of the variable from the value of each case and dividing by the standard deviation. The Z-score transformation generates a new variable with a mean of 0 and a standard deviation of 1.
Z-VALUE

A test statistic which, for sufficiently large sample sizes, is approximately normally distributed. Often it is the ratio of an estimate to its standard error.
LESSON ELEVEN

Writing Conclusions and Recommendations

OBJECTIVES

- Given a set of results, write conclusions that refers back to the questions raised in the study.
- Given a set of data and data analysis, choose the most appropriate presentation medium and format.
- Describe the contents and format of the report.
- Given a set of conclusions and limitations, write a set of recommendations.

Keywords

conclusions, interpretation, theoretical framework

The results and only the results

While some reports may combine the results and conclusions section, it's best to develop them separately. The results section should present the results of the analyses with minimal interpretation since the following section on conclusions will relate the findings to the research questions. Any charts, tables, graphs should be presented or referenced here. Charts, tables, and graphs should speak for themselves; i.e., the reader should be able to interpret them without textual explanation. A rule of thumb in discussing their contents is not to repeat in text what appears in the tables, charts and graphs. For example, you wouldn't say, "50 percent of the respondents answered, 'Yes'" if your table reports that. Also, you wouldn't say "the data shows an upward trend" if your line graph clearly indicates one.

The language of the results section should be short and sweet: say and show what happened and let the conclusions section take care of the interpretation and inference. The reasons for doing it this way is because no conclusions should be reached before all of the results have been digested -- to do so would mean each outcome was evaluated in isolation of every other outcome. Your chance of understanding what all of the research can show would be minimized. Also, if there are inconsistencies among the results, you would be put in an untenable position to make a conclusion only to contradict it in a subsequent section.
Taking a step beyond -- forming conclusions

When all the results are in, now is the time to make sense of it all. Each analysis should be related back to a corresponding research question and hypothesis. If your results supports your assumptions and answers your questions without any reservations, you should reaffirm that all went as planned. If the outcome was not as expected, then you'll need to explain any mitigating circumstances or give logical reasons why.

After relating your findings back to your original questions and research problem, you should then tie it into the theoretical framework presented in the related literature section. How does your research fit into what is already known? How does your findings expand upon or go beyond what is known? What will be the impact of your study?

Within the conclusions section (preferably at the end) would be a subsection focusing on the limitations of your research findings. Here is where you would discuss any deviation from the original proposed project and any unexpected outcomes. You also would remind the reader of the limitations mentioned earlier in the study.

Making recommendations

After you've presented the results and conclusions, the logical question that would be asked by the reader (or the researcher) is, "What now?" In the recommendation section, you would discuss how to overcome the limitations of your research -- how you (or others) would do it differently next time around. You would also discuss what sort of follow-up research, change in services or change in policy would be recommended in light of the results of your study.

The report format

In general, the final report should be written in the same style and format as was the proposal. Expanding your original format, you should include additional subheadings to represent the results, conclusions and recommendations section. A general outline of a final research report might look like the following:

- Cover sheet
- Title page
- Table of contents
- List of Tables
- List of Figures
- Abstract
- Introduction
  - Background
  - Purpose of the study
  - Research approach
- Statement of the problem
  - Nature of the problem
  - Significance of the problem
  - Need for the study
Theoretical framework

Methodology
Research design
Target population
Sampling procedure
Discussion of variables
Description of data sources
Data collection methods
    Instruments
    Procedures
Data analysis techniques

Results
Conclusions
Recommendations
Summary
Appendices
    Sample instruments
    Tables

Attention to detail

Think of the appearance of your final report as a statement about you and your organization. Attractive, well-organized, good visuals -- these are the qualities that your final report should have. The following are some guidelines for improving the appearance of the final report:

- Always use standard, 8 1/2 x 11 inch white paper
- All pages, except the first page, should be numbered
- All tables, charts and graphs should be consecutively numbered
- Left margin should be at least 1 1/4 inches, top, bottom and right margins should be at least 1 inch
- Use standard typestyles (fonts)
- Right margin should be ragged, not justified, if typestyle is not a proportionally spaced font
- Single-space the abstract (unless directed otherwise)
- Do not use abbreviations. Place acronyms in parentheses following their full description
- Tables should be interspersed with text only when they are essential to the understanding of the text
Writing an Evaluation Proposal I

OBJECTIVES

- Given an abstract, create an appropriate title
- Identify the requirements of the cover, cover page, transmittal letter and vitae.
- Given a general program problem, create a formal "research problem."
- Given a sample evaluation proposal, create a table of contents, list of illustrations, and list of appendices.
- Given a sample evaluation project, identify its assumptions, limitations and delimitations (scope).
- Define and contrast "evaluation questions" and "hypotheses."

Keywords

transmittal letter, abstract, problem statement,. assumptions, limitations, delimitations, evaluation questions, hypotheses

What is an evaluation project?

A evaluation project is a set of self-contained., planned operations for doing evaluation. Self-contained means that the project is complete, stands on its own and can be done independently of any other activities. A project has a defined set of goals that must be accomplished within a specific time frame. In the analogy of making a covered dish, your cooking project included all of the activities directed towards producing your recipe that met all of the criteria within the time frame allotted.

Sometimes people may use the terms project and program interchangeably. While there are obvious similarities between the two, projects are generally considered to be of limited duration while programs may be ongoing or have the potential to be.

What is an evaluation proposal?

- In general, a proposal is a written offer or request put forward by one person, company, or agency to do work for another in order to solve a problem in a particular way using a specific plan. A proposal is also used to request compensation for the work to be done.
- Proposals are usually prepared for a very limited audience. The purpose of a proposal is to convince the source of the compensation that the person or company has the capability and expertise to carry out the specific tasks identified in the proposal.
Proposals describe the specific tasks to be done, how the tasks will be managed, when it will occur, what will be produced and how much it will cost.

An evaluation proposal is a type of research proposal whose intent is to receive funding to complete an evaluation study. All research proposals have common characteristics:

- They define a problem or need.
- They define a set of goals or outcomes.
- They request financial support or compensation.
- They form the basis of an agreement or contract.
- They define an interrelated set of tasks and activities.
- They are based on or relate to previous research activities.
- They specify information to be collected and analyzed.
- They specify how information will be collected and analyzed.
- They require the expenditure of human and material resources to accomplish their goals.
- The results of the research will be of benefit to others.
- There is a time frame within which the work must be done.
- They supply material that supports the importance of the topic chosen and the appropriateness of the methods used.
- They have a formal written structure but not necessarily a universally applicable or specific format.
- They have a protocol that must be followed.

Even when a proposal may not be required -- for example, when doing an in-house evaluation project for one's job, it is a good idea to generate one anyway and to receive approval for the proposed work before it is actually done. Remember, a proposal is an agreement between you and the recipient of the evaluation -- it takes the mystery and misery out of doing unapproved work.

Parts of the proposal

The following is a brief synopsis of each major part of the proposal.

Cover

Believe it or not, the inclusion of a cover is not universally accepted. Some authors say that all formal proposals have covers; others say that the proposal should be clipped together to facilitate copying. I would think a cover that would allow the removal of pages would be acceptable.

The labeling should be plain, conservative and attractive. Include on the cover the title, name of the agency or individual submitting the proposal and the date.
Title page

A separate title page or cover greatly enhances the appearance of a proposal. The title page should include the title of the proposal, the name and address of the organization or individuals submitting the proposal, the date, and the name of the funding agency to whom the proposal is being submitted. Title pages are not necessary for proposals sent to government funding agencies that provide forms as part of their application material.

Abstract

Abstracts are concise summaries of a evaluation project. In a minimum number of clearly written words, they describe, in plain language, the program being evaluated, the problem or need addressed by the program, the purpose and methodology of the proposed evaluation and, for completed evaluations, the results, conclusions and recommendations.

Table of contents

A table of contents is included if the proposal contains five or more sections and if it runs 10 or more pages.

List of illustrations

When the proposal contains several tables and figures (charts, graphs), you probably should have a separate heading for them. If your proposal only has tables, then this would become a "List of Tables."

Letter of transmittal

A letter of transmittal (1) carries the name, address, and phone number of the organization transmitting the proposal; (2) indicates why the proposal is being sent to that particular funding agency; (3) lists a brief statement about the organization's interest in the project and its capability and experience and (4) notes the name of the person to contact for further information.

Introduction

The introduction presents a succinct line of inquiry, identifies the main evaluation constructs and indicates how the proposed evaluation model fits into an existing body of theory. The introduction describes the setting of the program and the evaluator's understanding of the problem.
Background of the problem

In this section, the evaluator would discuss what is already known about the problem and how this knowledge gave rise to the development of the program to be evaluated.

Stating the problem

The purpose of the problem statement is to explain the nature and importance of an identified need. The statement of the problem translates an observed need into measurable terms and makes it amenable to evaluation.

Rationale for the evaluation

The rationale or need for the evaluation is the justification for using evaluation results to continue a program unchanged, to modify it or to eliminate it entirely.

Purpose of the evaluation

Here, the evaluator describes what the project will accomplish and how the evaluator will attempt to determine program effectiveness. This section should include specific evaluation goals in the form of measurable objectives.

Evaluation questions

Evaluation questions are what the evaluator attempts to answer by doing the evaluation. The evaluation questions are derived from the purpose and delineate exactly what the evaluator will determine.

Delimitations and limitations

Since no single evaluation has the resources to do everything and the ability to control for all internal and external factors, there will be limits to what it can accomplish. The limitations section lists all of the constraints that impinge upon the evaluator and the operation of the project: financial, procedural, political, statistical, and so forth. Delimitations refer to the boundaries within which the project will operate or investigate the program. The boundaries may be a certain group of people, a certain geographic locale, a certain age group. The boundaries delimit the range of generalization that can be made from the evaluation.
Assumptions

An assumption is like a belief -- it is acceptance of something being true without having direct proof. For example, one assumption that the evaluator might make is that the program managers will cooperate with the evaluator. Another assumption might be that the data to be collected is a normally distributed variable.

Providing definitions

The evaluator should define all technical terms and any terms with special meaning. It's important that the evaluator be very precise in the use of language -- avoiding the use of words that have more than one meaning.

Related evaluations

Since other evaluators may have published their work in regards to this problem or to similar programs, the evaluator should always include a thorough review of the relevant literature.

Theoretical framework

A discussion of the theoretical framework describes all that is known about a problem and indicates how the evaluator will use that knowledge to design the evaluation and to support the approach taken.

Methodology

This is the procedure section wherein the evaluator would describe the (a) program and the target population served; (b) sampling method; (c) instruments; (d) evaluation design; (e) procedures for collecting, recording, and coding data and (f) methods for analyzing the data.

Time and work schedule

A time and work schedule highlights when and how the major tasks of a project will be done. All activities must be completed within a set time frame -- creating a time and work schedule maximizes the use of one's time, serves as a planning tool for managing the project and demonstrates to the funding agency when certain products and accomplishments will be expected.
Facilities and equipment

The evaluator should indicate the facilities and equipment required to conduct the evaluation and how the evaluator will acquire them. This section could be used to describe the capabilities of the resources available to the evaluator for completing the project. Funding agencies would rather have expensive equipment and facilities donated as in-kind items than have to pay for them. Funding agencies also will look more favorably on a project that makes use of specialized facilities or facilities that have a positive reputation in the area of evaluation.

Personnel

Just as proposal writers should feature the special capabilities of their facilities or physical resources, they should also spotlight the background and experiences of the personnel who will be working on the evaluation. Qualifications of staff would go here in narrative form (with complete resumes and vitae in the appendix) with special emphasis placed on achievements on prior evaluation activities (track record) and any special recognition awarded to individuals.

Letters of reference or support

One way of documenting the need for the evaluation is to engender support for the study. Including letters of support for the project and letters of reference for the investigators from community leaders or influential individuals will greatly enhance the credibility of the project and the project's leadership.

Budget

The budget section lists in detail all of the direct and indirect expenses of doing the project along with the justifications for the amounts requested. Items listed in a typical budget include salaries, equipment, communication, supplies, travel and per diem, consultant fees and indirect costs such as utilities, work space, custodial services, maintenance, and fringe benefits.

Do the abstract first (and last)

Abstracts are concise summaries of a evaluation project. The reason why the abstract is written in plain, nontechnical language is because the abstract is meant for the general public. Abstracts presented with published evaluation reports are written in past tense. In contrast, abstracts written for an evaluation proposal will be written in future tense. Abstracts should be prepared before the first chapter of the proposal is completed.
Doing an abstract first will help focus the evaluator on the specific objectives of the evaluation and the means for accomplishing those objectives. In addition, the abstract serves as a basis for obtaining prior approval for doing the actual evaluation. Also, many agencies require a "letter of intent" which is, in effect, a formal abstract that describes the proposed project.

**Introduction and background information**

In some types of proposals, the abstract replaces the introduction section. Generally, though, it's a good idea to include this section because it serves as a prelude to the problem statement by establishing the context of the problem. This section acquaints the reader with the events or information that lead up to the identification of the problem. In a well-written introductory section, the problem statement often becomes self-evident or predictable.

One way to understand this process is to think of two stories you've read about or seen on television. In the first story, the ending is a total surprise. In the second story, the ending is just as expected given the flow of the story. In the second story, the ending makes a statement about what the story was about. So, too, should the background information to an evaluation project lead to a predictable problem statement which, in turn, confirms what the background information has said.

**The statement of the problem**

As was mentioned earlier, a problem is a discrepancy between what is known and not known, between actual and ideal situations, or between potentially related bodies of knowledge (theory, sets of observations). The statement of the problem translates this gap into measurable terms; i.e., to make it amenable to evaluation.

Often, what stakeholders will identify as the problem is actually the symptom of a discrepancy and not the actual discrepancy. In stating the problem, the evaluator should elaborate on the how the problem has come to be by providing both quantitative and qualitative information on its scope and extent. Providing supportive data helps to highlight the importance of the problem and the importance of the proposed evaluation.

**The purpose of the program**

Once you've clearly defined the problem, then you will need to specify how the program has addressed the problem. Basically, the purpose provides answers to the implicit questions raised by the statement of the problem. The purpose defines why the program was started. In this section, you should also list the goals and objectives of the program; i.e., what was it intended to accomplish? What will be the outcomes and the benefits?
The objectives of a program may be drawn from a more general set of goals provided by the funding agency. When funding agencies provide a set of objectives, the writer is expected to include in the proposal a well-defined set of activities and methods showing exactly how each of the objectives will serve as the basis for evaluating the effectiveness of the program.

**The need for the evaluation**

The need for the evaluation project should be evident from the statement of the problem and the purpose of the program. Another synonym for this section would be the rationale for the evaluation. An evaluation rationale is a collection of statements documenting how an evaluation will contribute to understanding both the problem and the program. Imagine a jigsaw puzzle: supplying a missing piece of a puzzle not only helps to explain the meaning of the pieces that interconnect with it -- putting the piece in place helps to clarify the meaning of the entire picture. The piece "fits" the puzzle at two levels: the local level (the pieces it interconnects) and the global level (all of the pieces and the picture it represents).

In evaluation, the puzzle is known as the "theoretical framework." Except that now we need to change analogies from a flat object (puzzle) to a multidimensional object (a house). To build a house, one would start with a firm foundation to support the weight of the entire house. Then, a frame is built upon which will go the various components. Each part of the frame is designed to support a particular part of the house: a window, a door, a wall, a ceiling. In evaluation, the theoretical counterpart to the house foundation is a foundation of knowledge (an accumulation of facts and observations). Upon this foundation rests the various theories of how the "house" (or, in this case, human behavior) functions as a whole. The theories in turn support observations which we recognize as examples of the theory in action.
LESSON THIRTEEN

Writing the Evaluation Proposal II

OBJECTIVES

- List the contents of a Procedures or Methods section.
- Distinguish between various types of sampling giving examples of when they would be applicable.
- Determine sample size given desired precision.
- Given a description of an evaluation project, list the procedures in order of occurrence.
- Develop time and work schedule for completing procedures.
- Define reliability and validity.
- Given evaluation problem, select an appropriate design.

Keywords

sample, random selection, random assignment, stratified sample, cluster sample, systematic sample, reliability, validity, treatment, control, exploratory, descriptive, field studies.

Choose your method

The design of the evaluation, the type of data collected, the method for collecting data, and the means for analyzing the data represent the most difficult part of the proposal writing. The primary reason is because proposal writers have had the least amount of training and direct experience applying that training to real-world problems.

For the reason mentioned above, many writers approach this section with fear and loathing. Partly because it is a difficult task to do and partly because it may involve the use of statistics and mathematics -- two subjects not high on anyone's best-loved subjects list. What often happens in this section is that the writer will present in barest detail what methods will be used. When comes the moment to conduct and analyze the study, the analysis may not be doable.

This is the one section of the proposal that detail and precision cannot be ignored. The main reason why attention to precision and detail is critical is because the project must be replicable -- it must be possible for others to repeat it. Otherwise, people would have to accept on faith alone that your project accomplished what you said it accomplished.

In the methods section, you will describe the population you are collecting data from, the definitions of concepts that you will use, the data collection techniques and instruments, the analytical instruments, any statistical tests, description of variables and data categories, the control of variables, the plan for analyzing the data and presenting the results.