Chapter 1: Key Concepts and Issues in Program Evaluation and Performance Measurement

Preview Questions

1. What is the difference between a policy and a program?
2. What are the key concepts for program evaluation?
3. What are key questions to ask when evaluating programs?
4. What is the difference between formative and summative program evaluations?
5. What is the difference between *ex ante* and *ex post* program evaluations?
6. How do you analyze cause and effect when evaluating programs?
7. What is the process for conducting a program evaluation?

Reading Summary

Introduction

* Evaluation is a structured process for judging a policy or program.
* There are many methods for conducting evaluations.
* Challenges in conducting evaluations:
  + Often random experiments are not feasible.
  + The program may already be implemented.
  + There may be no baseline data for comparison.
  + Mechanisms for collecting data on outcomes may not be in place.
* There is some debate as to whether or not performance measurement is really a part of the evaluation field.
  + Some consider it just a part of organizational management.
  + The principles, practices, and challenges of performance measurement are similar to program evaluation.
* Program evaluation and performance measurement inform performance management.
* Performance management is also referred to as results-based management.
* Results-based management part of the new public management (NPM) movement.
* NPM is a public sector reform approach that emphasizes business-like practices.
* Performance management cycle is the framework for organizing evaluation activities.
  + State objectives
  + Develop strategies (i.e., design program)
  + Align management systems (i.e., implement program)
  + Measure and report performance
  + Make decisions regarding the program (e.g., accountability phase)
* Policies versus programs
  + Policies are statements of intended outcomes or objectives (ends) and the methods (means) by which governments or their agents will achieve them.
    - New policies are often modifications of existing policies.
  + Programs are sets of related activities intended to achieve one or more related outcomes or objectives.
* Conceptualizing program evaluation
  + Fitting round pegs into square holes.
  + The round pegs may go into the square holes but there will be gaps.
  + Evaluators must adapt the tools to each unique situation.
  + Conclusions and recommendations produced by an evaluation are influenced by what the evaluator brings to the table (i.e., tacit knowledge).
* There is a gray area between pure program evaluation that has a minimum level of methodological sophistication and program review.
* There is no one dominant view of what evaluation methods are correct.
* Important points about program evaluation
  + Often relies on triangulating evidence from different points of view.
  + Typically uses data collection and analysis methods that are well-understood.
  + Recommendations often intended to improve a program.
  + Evaluator options are often constrained.
  + The methodology chosen must be defensible.
  + The conclusions and recommendations must be credible and useful.
  + The evaluator must use his or her own professional judgement throughout the process.

Definition of a program

* A group of related activities intended to produce at least one specific outcome.
* An open systems approach is used to describe and model programs.

Program evaluation key concepts

* Program evaluations enable us to:
  + Determine to what extent, if any, a program produced the observed outcomes.
  + Determine to what extent, if at all, are the observed outcomes consistent with the intended outcomes.
* Program evaluation compares intended outcomes with observed outcomes.
* Effectiveness is an indication of whether or not a program produced its intended outcomes.

Key evaluation questions

* What is the need for the program?
* What is the rationale for the program?
  + Is the program relevant?
  + Does the program fit with priorities and policies?
* Is the program appropriate?
  + i.e., Does the program structure make sense?
* Is the program being implemented as intended?
* Is the program efficient?
  + Technical efficiency (i.e., is the ratio of program outputs to program inputs acceptable?)
* Is the program responsible for the observed outcomes (Type 1 effectiveness)?
* Is the program achieving its intended objectives (Type 2 effectiveness)?
* Is the program cost-effective?
  + Are program outcomes greater than program costs, usually expressed in dollars.
  + ***What’s the difference between technical efficiency and cost-effectiveness?***
  + ***What is the difference between an output and an outcome?***
* Is the program cost beneficial?
  + Economic efficiency (i.e., is the ratio of program outcomes (benefits) to economic costs acceptable?)
* Is the program adequate?
  + Visibly affecting the overall issue being addressed.

Formative and Summative Program Evaluations

* Formative evaluations focus on how a program can be improved where the existence of the program is not in question.
* Summative evaluations focus on whether a program produced the intended outcomes.
* Evaluations are not nearly so binary in actual practice.
* Third type focuses on acquiring a better understanding about the factors that underlie public problems and the fit of programs designed to address them.

*Ex Ante* and *Ex Post* Evaluations

* *Ex ante* evaluations are done when programs have been operating for some time.
* *Ex post* evaluations are done before programs are implemented.

Analyzing Cause and Effect

* Three conditions of causality:
  + Program occurred before the observed outcomes.
  + The outcomes tended to occur when the program occurred.
  + There are no other plausible explanations for the observed outcomes.
* Not always possible to rigorously rule out other possible explanations.
* The relationship between a program and observed outcomes is probabilistic.
* All program evaluations require a substantial number of judgement calls by the evaluator.

Program Evaluation Process (see Table 1.1 on p. 26)

* Evaluation assessment questions
  1. Who are the clients for the evaluation?
  2. What is driving the evaluation?
  3. What resources are available to do the evaluation?
  4. What has been done previously (i.e., what do we already know)?
  5. What are the characteristics of the program (i.e., the logic of the program)?
  6. How will environmental factors influence the evaluation?
  7. What are the appropriate research design alternatives?
  8. What information sources are available?
  9. Which evaluation alternative is likely to be least problematic?
  10. Should the evaluation proceed?
      + ***What are the reasons that would cause an evaluation not to proceed?***
* Evaluation study steps
  1. Develop the measures
  2. Collect the data
  3. Analyze the data
  4. Write the report
  5. Disseminate the report
  6. Implement changes to the program as appropriate based on the evaluation
     + Improvements
     + Scale (increase or decrease)
     + Scope
     + Eliminate and replace
     + Eliminate without replacement

Preview Questions

1. What is a logic model?
2. How are logic models constructed and used to evaluation policies and programs?
3. What are the different types of logic models?
4. What are program logics?
5. How do you construct program logics?
6. What are program technologies?
7. How do you determine program objectives, program environments, and organizational objectives?
8. How do you use logic models for performance measurement?
9. What are the strengths and limitations of program logic models?

Reading Summary

Introduction

* Program logic models are visual representations of programs.
  + Show how program is supposed to work in theory.
  + Resources🡪Activities🡪Outputs🡪Outcomes
  + Communicate key parts of program and their intended relationships
* Means-end relationships are causal relationships among multiple factors
  + One factor causes the other
* Programs are means-end relationships.
* The open-systems metaphor is a way of conceptualizing and describing programs as similar to biological or engineering causal relationships that are bounded and both affect and are affected by the environment in which they operate (i.e., open systems).
* Implications of open-systems metaphor for describing programs
  + Programs have a conceptual boundary that cannot be directly observed, which affects how we perceive and model the programs.
  + The purpose of programs is to accomplish objectives that society values.
  + There is a causal relationship between program structures, activities, outputs, and outcomes.
  + Programs operate in environments, which both constraint and create opportunities.

Basic Logic Modeling Approach

* Goal is to describe programs in a way that helps develop measures of program activities, outputs, and outcomes.
* Program outputs are the immediate results of program activities.
* Program outcomes are the changes in the environment relevant to the issue the program is intended to address.
* Time-related sequence is implied in logic models.
* Not all effects of a program are observable immediately following the completion of the program or program activities.
* Sample program logic model shown in Table 2.1 on p. 46.
* Program constructs are words or phrases that describe the key features of a program.
* Limitations
  + Does NOT specify linkages among specific activities, outputs, and outcomes.
  + It only categorizes and describes activities, outputs, and outcomes but does not specify causal relationships.

Categorizing and Specifying Intended Causal Linkages

* Program inputs are the resources required to operate a program.
* Program components are related groups of activities in the program.
* Implementation objectives indicate what needs to occur to cause the program to begin producing outputs.
  + Not the same as program objectives.
  + Do NOT mix outcome-focused language with implementation objectives.
* Sample program logic model with implementation objectives and linking constructs shown in Table 2.2 on p. 48.
* Successful program implementation does not automatically lead to achievement of program outcomes.
* No need to evaluate program efficiency or effectiveness if implementation objectives are not achieved.
* Two kinds of negative evaluation results
  + Program failure is when program implementation is faulty or inadequate.
  + Theory failure is when the program outputs do not produce the intended outcomes.
* Linking constructs are factors that connect program outputs to program outcomes.
  + Non-program activities that must occur after a program output is produced to result in the intended program outcomes.
  + Linking constructs are bridging variables.
  + Not all program logic models have linking constructs.
* Environmental factors can influence the likelihood of success for a program.
* Logic models generally have the following features:
  + One or more components
  + At least one implementation objective for each component
  + At least one output for each component
  + Outcomes may be categorized as short-term, medium-term, and long-term
  + Linking constructs depending on how the program operates
  + Each linking construct is connected to one or more outputs
  + Each output is connected to one or more short-term outcomes
  + One-way causal arrows to avoid overcomplicating the model
* Logic models are templates which may not fit all situations well.
* Open-systems logic models work best when:
  + Objectives are clearly stated
  + Program has been implemented
  + Program has a track record
  + Administrative responsibility for the program is within the organization in which the program is operated
* Developing logic models is an iterative process

Flow Charts

* Flow charts can be used to represent program activities
* Flow charts depict how a program is intended to operate
* Generally used as a complement to program logic models

Constructing Program Logics

* Iterative, qualitative process
  1. Review documentation
  2. Interview program managers
  3. Interview stakeholders
  4. Prepare a draft logic model
  5. Obtain feedback on the draft logic model from program managers and stakeholders
  6. Revise the draft logic model as appropriate based on the feedback
  7. Repeat steps 5-6 as necessary
* Characteristics of program objectives
  + Specify the target population and domain in which outcomes should occur
  + Specify the direction of intended change (i.e., increase or decrease)
  + Specify the magnitude of the expected change
  + Specify the time-frame over which the change will occur
  + Specify measureable outcomes
* Politics is the allocation of values within an organization
* Politics is intrinsic to the environments in which programs operate
* Objectives are political statements by nature.
* It’s important to secure agreement on what a program is actually intended to accomplish before the evaluation begins.

Program Technologies

* Core technologies are causal relationships that define the competencies of an organization that are useful in accomplishing program and organizational objectives.
* Core technologies are combinations of knowledge, technique, and experience
* Instrumentally perfect technologies work every time
* Social programs often really on imperfect core technologies
* Program evaluators and managers should consider who should be held accountable if the program does not achieve its objectives.
  + ***Is this a productive approach?***
  + ***Does it negatively impact innovation and risk taking in program design?***
* Being held accountable for outcomes is problematic because causality is often unclear and the potential effects of environment.

Program Theories and Program Logics

* Theory-driven evaluations use logic models to unpack program structures so that one can test linkages in the model.
  + Understand how key constructs in the logic model (i.e., working theory) are linked to each other and whether observed empirical patterns correspond to the expected linkages.
* Meta-analysis is the synthesis of the results from multiple evaluations for a given type of program.
* Meta-evaluation is the evaluation of multiple completed evaluation projects.
* Ad hoc systematic review synthesize evaluations in a particular field to describe underlying patterns, trends, and key findings.
* Realistic evaluations seek to develop program knowledge based on CMOs associated with program success and failure.
  + Understand why programs work within specific contexts.
* Context-mechanism-outcomes (CMOs)
  + Cause and effect in programs are mediated by the contexts in which programs are implemented.
  + Contextual factors
    - Program-related conditions
    - Organizational conditions
* Program logic models are specific to context.
* Theory-driven evaluations are not yet common practice.

Program Objectives, Program Environments, and Organizational Objectives

* See Table 2.4 on p. 74.
* Programs are open-systems embedded with organizations which are themselves open-systems.
* Behavioral goals are goals that reflect the self-interest of the individuals and the organization.
* Program objectives can conflict with behavioral goals (e.g., organizational objectives).

Logic Models for Performance Measurement

* Program evaluation and performance measurement are complementary.
* Logic models are important tools for developing performance measures.
  + Frame discussions about what to measure.
* Key performance measures are also called key performance indicators (KPIs).

Strengths and Limitations of Program Logics

* Program logics do the following:
  + Categorize work
  + Depict causal relationships
  + Distinguish what is in the program from what is in the environment
* Organizational logic models are logic models for whole organizations that link organizational goals to objectives to strategies to performance measures.
  + Useful when trying to understand authority and responsibility
* Logic models can become an impediment to change if we don’t remember that they are imperfect representations of reality.
* General limitations of logic models
  + Some programs don’t lend themselves to logic models
    - e.g., lots of change in the environment
  + They are time-limited
* Developmental evaluations are alternatives to formative and summative evaluations and are used in situations where programs and environments are in flux.

Preview Questions

1. What is research design?
2. Why is experimental design important?
3. How do you use experimental designs to evaluate programs?
4. What are the four kinds of validity in research designs for evaluations?
5. What are quasi-experimental designs?
6. How do you test causal linkages in program logic models?
7. What is the relationship between research designs and performance measurement?

Reading Summary

Introduction

* Understanding the logic of research designs helps us develop defensible judgements about the extent to which programs are responsible for actual outcomes.
* All experimental research designs are about external comparisons that have a control group that does not receive the treatment.
* Implicit research designs focus on comparisons internal to the treatment (e.g., dosage).
  + Also called case studies.
* Four types of validities
  + Statistical conclusions validity
  + Internal validity
  + Construct validity
  + External validity
* Internal validity is important for assessing causal linkages.
* Performance monitoring uses research designs to make comparisons over time.
* Performance results can help with conducting program evaluations.
* Research designs are only one type or component of evaluation design.
* There is debate about whether experimental designs should be regarded as the gold standard for evaluation.
* Experimental and quasi-experimental evaluation approaches may miss something fundamental in human interactions.
  + Some believe the assumption about being able to objectively measure human attributes is flawed.

Defining what is research design

* Research design is fundamentally about examining the linkage between the program and observed outcomes.
* To test causal linkages in logic models one must be able to isolate each linkage in turn and hold all other linkages constant.
* Research design tends to focus on the main linkage due to resource constraints.
* Main linkage is between the program as a whole and the observed outcomes.
* Performance monitoring is concerned with the link between the program as a whole and intended outcomes.
* Patched-up research designs consist of several research designs that are combined in an effort to counterbalance the weakness of any individual research design.
* Randomization controls factors that interfere with determining causal linkages.
* Experimental research designs work best where:
  + *Ex ante* evaluation designs
  + Sufficient resources for control groups
  + Feasible to do random assignment to treatment and control groups
  + Can sustain group assignments long enough
* Experimental design types
  + Before-after designs
  + After-only designs
* For n<30, pretesting can establish that the two groups are really equivalent.
  + Pretesting can produce adverse effects in the evaluation
  + Solomon Four-Group Design mitigates problems caused by pretesting (see Table 3.1 on p. 99)
    - Two groups that are pretested
    - Two groups that are not pretested
* Non-experimental designs DO NOT have explicit comparisons built into them.

Why Experimental Design is Important

* Issues with experimental designs
  + Deception of human subjects may be necessary
  + Control group does not receive benefit of treatment
  + Opportunities to use experimental and quasi-experimental are limited and emphasis on them sets an unrealistic expectation.

The Four Kinds of Validity in Research Designs for Evaluations

* Establishing causal relationships
  + Cause variable precedes the effect variable (temporal asymmetry)
  + Effect variable varies with the cause variable (covariation)
  + No other plausible explanation of covariation (no rival hypotheses)
* Statistical conclusions validity 🡪 internal validity 🡪 construct validity 🡪 external validity
* Establishing cause and effect
  + Statistical conclusions validity 🡪 correctly using statistical tests of significance
  + Internal validity 🡪 ruling out rival hypotheses
* Generalizing cause and effect of variables to constructs
  + Construct validity
* Generalizing from program to other settings
  + External validity
* Possible threats to internal validity
  + External factors coincide with the implementation of a policy or program (history)
  + Observations for the unit of analysis change with the age of the unit of analysis (maturation)
  + Testing familiarity (testing)
  + The way in which variables are measured changes (instrumentation)
  + Extreme pre- test scores tend to regress toward the mean in post-test (statistical regression)
  + Differences in the treatment and control group in characteristics other than the ones of interest (selection)
  + Units of analysis drop out over the course of the evaluation (attrition)
  + Correlation does not mean causation (ambiguous temporal sequence)
  + Additive effects of multiple threats to internal validity (selection-based interactions)
* There can be internal validity problems with the implementation of experiments.
* Construct validity
  + A construct is an idea or concept.
    - e.g., user satisfaction, perceived safety
  + Measurement requires translating constructs into observable variables that can be measured.
  + Construct validity consists of working backwards from measurements of observable variables to constructs.
    - How valid are the measures of the constructs?
    - Is the cause variable what we think it is?
    - Have we correctly defined the cause variable?
    - e.g., classroom and home visits vs. classroom, home visits, and sibling reinforcement
    - e.g., server training vs. server training and bar manager influence
  + Arise from the way a policy or program is implemented.
  + Construct validity threats
    - Treatment group communicating with control group (treatment diffusion)
    - Control group receives elements of the treatment to appease them (compensatory equalization)
    - Control group changes behavior to compete with treatment group (compensatory rivalry)
    - Control group changes behavior in a negative manner because of perceive unfairness (resentful demoralization)
    - Behavior of unit of analysis changes because of knowledge about being part of the evaluation (e.g., Hawthorne effect)

External validity

* Factors that limit generalization o fa policy or program.
  + Units of analysis have unique features
  + Treatments have unique features
  + Patterns of outcomes (e.g., outcome definitions) are unique to environment
  + Setting has unique features
  + Pre-existing features in the environment influence program implementation and therefore program outcomes (context-dependent mediation)

Quasi-Experimental Designs

* Research designs that DO NOT include random assignment to treatment and control groups.
  + Threats that would have been mitigated by random assignment may be present.
* Focus on eliminating the most likely (i.e., most plausible) threats to validity.
* See Table 3.4 on p. 120.
* Before-after comparisons and comparative time-series can be coupled with statistical analyses, which compensate for selection threats to internal validity.

Important Concepts

* Triangulation is reducing the uncertainty of a proposition by generating confirmatory evidence from two or more independent evaluation processes.
* Counterfactual is what would have happened if a treatment had not been implemented.

Testing the Causal Linkages in Program Logic Models

* Evaluations that isolate causal linkages in program logic models are expensive.
* Each measure selected for a construct implies a research design.
* We can measure constructs but usually we cannot test the connections between constructs.
  + Must measure both ends of a linkage.
  + The cause and effect variables must be measured with the same unit of analysis.

Research Designs and Performance Measurement

* Using administrative data
  + Minimizes time
  + Minimizes cost
  + Raises questions about data validity and reliability
* Output data often available internally to organization
* Outcome data often requires additional resources
* Performance measurement systems
  + Describe observed outcomes
  + Indicate whether observed outcomes are consistent with intended outcomes
  + Usually not enough resources to establish that observed outcomes were the result of program activities
* Program managers have more control over outputs than outcomes.
  + Typically program managers are less willing to own the observed outcomes.
* Using performance measures to determine outcomes presumes that causality between outputs and outcomes is established.

Preview Questions

1. What is the difference between reliability of measures and validity of measures?
2. What are units of analysis?
3. What are levels of measurement?
4. What are typical sources of data for evaluations?
5. How do you use surveys when conducting evaluations?

Reading Summary

Introduction

* Must decide which constructs will be measured and how constructs will be measured.
  + Some constructs more important than others
  + Constructs convey the meaning assigned to elements of a logic model
  + Constructs may sound different but mean the same thing
* Two aspects of measurement
  + Collecting relevant data
  + Translating constructs into observables to produce valid and reliable data
* Focus on measuring outputs and outcomes
* Rarely find an evaluation study based on only one method of data collection
* Environmental factors can affect both program and outcomes
* Must measure environmental factors that could be plausible rival hypotheses or mediating factors
* Outputs more controllable than outcomes for program managers
* Proxy measures are assessable measures that substitute for others measure of interest that are not assessable.

Reliability and Validity of Measures

* Measurement procedures are used to translate constructs into variables.
* Measurement procedures are called measurement instruments.
* Measurement reliability is whether a measurement result is repeatable for a given situation (i.e., precise).
  + Inter-coder reliability is the extent to which different coders will assign the same value to the same qualitative data.
  + Split-half reliability is the extent to which to sets of Likert statements intended to measure the same construct are consistent.
  + Cronbach’s alpha is used to determine the extent to which a group of survey questions are a reliable measure of a construct; it takes on values between 0 (no reliability) and 1 (perfect reliability).
* Few measurement instruments are completely reliable.
* Statistical analysis assumes that variables are measured reliably (i.e., without error).

Understanding Measurement Validity

* Measurement validity is whether we are measuring what we intend to measure (i.e., accurate).
* Must have validity to have reliability; but can have reliability without validity.
* Must use judgement to assess measurement validity.

Types of Measurement Validity

* Measurement validity is not the same as research design validity.
* Measurement validity is a component of construct validity.
* Construct validity
  + Measurement validity
    - Relationship between single measure (variable) to single construct
      * Face validity
      * Content validity
      * Response process validity
    - Relationship between multiple measures to single construct
      * Internal structure validity
    - Relationship between a measure-construct pair and other pairs
      * Concurrent validity
      * Predictive validity
      * Convergent validity
      * Discriminant validity
* Face validity is whether a measure (variable) appears to be measuring the construct it is intended to measure on the face of it.
* Content validity is whether a measure of a construct is consistent with the full theoretically relevant range of construct content.
* Response process validity is whether respondents to a measurement instrument engage sincerely.
* Internal structure validity is whether a group of items that are collectively meant to be a measure of the same construct are related.
* Concurrent validity is whether a new measure of a construct correlates to an existing validated measure of the same construct.
* Convergent validity is whether measures of two or more theoretically related constructs correlate with one another.
* Discriminant validity is whether measures of two or more constructs that are theoretically unrelated do not correlate with one another.

Units of Analysis

* The cases or unit of focus for an evaluation.
* They have various characteristics or attributes that are measured (observed).

Levels of Measurement

* Nominal simply groups units of analysis into categories.
* Ordinal groups units of analysis into categories and the categories have a sequence to them.
* Interval groups units of analysis into categories, the categories have a sequence to them, and the interval of the category sequence conveys meaning.
* Ratio groups units of analysis into categories, the categories have a sequence to them, the interval of the category sequence conveys meaning, and the interval scale has a true zero indicating the absence of the characteristic being measured and no measurement below zero can exist.
* Parametric statistics are statistical methods used for interval and ratio measures.

Sources of Data

* Existing organization records
* Observational data collected by program evaluators
* Surveys administered by program evaluators and managers
  + Fact-based warm-up questions
  + Fact-based program-related questions
  + Opinion-based program-related questions
  + Overall rating of the program
    - Last to minimize bias ratings of individual elements and help the respondent fully recall the program experience.
  + Demographic information
    - Some may view as too personal

Using Surveys to Estimate the Incremental Effects of Programs

* Surveys are measurement instruments.
* Pilot survey questions to evaluate the appropriateness of question wording, layout, etc.
* Event calendars (tables with time across top horizontal and events along the vertical) can help respondents accurately recall events of interest.
* Remind respondents that accuracy is important.
* Train those that administer surveys about the meanings of questions in the survey instrument.
* Retrospective pretests establish baselines retrospectively.
  + Addressed response-shift bias of conventional before-after testing.
  + May be susceptible to change response bias (i.e., expectation of program effect).
  + Tend to be more valid is administered separately from other post-tests.
* Survey designs are not research designs.