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.

**Chapter 5: Applying Qualitative Evaluation Methods**

Preview Questions

1. What are the different approaches to qualitative evaluation?
2. What are the basic concepts of qualitative evaluation design?
3. How do you design and conduct qualitative program evaluations?
4. How do you assess the credibility and generalizability of qualitative findings?
5. What is the connection between qualitative evaluation methods and performance measurement?

Reading Summary

Introduction

* The wide variety of approaches for qualitative evaluations is what makes using them challenging.
* The criteria for assessing qualitative evaluations are questioned by some qualitative evaluators.
* Qualitative evaluations emphasize textual sources consisting of words.
  + Narratives
  + Organization textual material
  + Non-numerical sources
* Qualitative approaches are less structured than quantitative methods.
  + Useful for data that doesn’t easily reduce to numbers.
  + Useful for exploratory work and participatory (i.e., empowerment) evaluations.

Comparing and Contrasting Different Approaches to Qualitative Evaluation

* Qualitative approaches emerged in the 1970s.
* Qualitative evaluation methods is a subset of qualitative research methods.
* Paradigms are belief systems that shape
  + what we consider to be important when conducting research
  + how we perceive the phenomena we’re researching
  + whether we can even perceive particular phenomena
* Paradigms are incommensurable (i.e., perception in one paradigm not translatable to others)
* The difference between quantitative and qualitative evaluation is at the level of methods, not epistemology (i.e., beliefs about how knowledge is obtained).
* Examples of epistemology
  + Objectivism assumes that objects have meaning independent from human perceptions (i.e., intrinsic meaning).
  + Constructionism believes that objects are not meaningful independent from human perceptions (i.e., they do not have intrinsic meaning).
  + Social constructionism believes that social contexts are critical for understanding how meanings and realities are generated.
  + Antirealist ontology believes that reality consists only of ideas in the human mind.
  + Constructivism focuses on the way individuals interpret the world.
  + Positivism believes that perceptions are factual.
  + Postpositivism believes that the world can’t be viewed independently of the observers’ standpoint.
  + Interpretivism believes that the description of an object is always shaped by the culture of the person describing it.

Pragmatism as a Response to the Philosophical Divisions among Evaluators

* Evaluators should focus on what works situationally.
  + Methods are independent of any presumed underlying philosophical assumptions.
* Methods are techniques and procedures used to gather and analyze data.
  + Sampling
  + Direct observation
  + Interviews
  + Focus groups
  + Statistical analysis
  + Content analysis
* Methodologies are the strategies for achieving a research objective which drive our choice of methods.
  + Experimental research
  + Survey research
  + Ethnography
  + Action research

Alternative Criteria for Assessing Qualitative Research and Evaluations

* Being clear about theoretical assumptions helps clarify the criteria readers should use in assessing our research findings.
  + People find assessing qualitative research difficult because it’s not clear which criteria should be used.
  + No single accepted way of assessing qualitative research.
* There is broad acceptance that qualitative and quantitative methodologies and methods are complementary.

Basics of Qualitative Evaluation Designs

* Appropriate applications for qualitative evaluation approaches
  + Understanding program processes and operation
  + Comparing programs
  + Studying program changes over time
  + Studying how a program was implemented
  + Developing logic models for a program
  + Assessing whether a program should be evaluated.
  + Studying program outcomes
  + Studying individualized outcomes
  + Studying a program’s effectiveness in preventing an undesired phenomenon

Comparing and Contrasting Qualitative and Quantitative Evaluation Approaches

* Evaluation is about finding a fit between the characteristics of a particular situation and the methodologies and methods used to study the situation.
* Interpretive qualitative evaluations emphasize unique human experiences
  + Don’t try to impose categories or structures on human experiences
  + Builds upward to find patterns
  + Evaluators are the principal measuring instrument
* Inductive qualitative evaluations start with data and then construct understandings of the programs.
* Holistic approach to program evaluation
  + Discovering themes in the data
  + Weighing the themes
  + Verifying themes with stakeholders
  + Documenting the findings and conclusions
* Naturalistic designs
  + Work with the program as it is
  + Work with stakeholders as they naturally interact with program
  + Use the same words used by program stakeholders
* Most qualitative approaches avoid linear cause and effect logic.
* Outcome mapping documents the performance of a program over time and estimates whether the program contributed to the outcomes observed.

Designing and Conducting Qualitative Program Evaluations

* Clarify the evaluation purpose and questions
  + Begin qualitative data collection without a fixed agenda (i.e., unstructured approach) to ascertain issues, concerns, and problems is appropriate but is generally costs more.
* Identify research designs and appropriate comparisons
  + Implicit designs are often used.
  + Implicit designs are evaluations that occur after a program has been implemented and don’t have non-program comparison groups.
* Two broad comparison types
  + Case-by-case analysis where each case is a unit of analysis
    - A case may include more than one individual
  + Across cases analysis

Mixed-Methods Evaluation Designs

* Mixed-methods evaluation designs combine qualitative and quantitative sources of data.
* Assumes that designs that combine methods is more credible.
* Four factor framework (see Table 5.6 on p. 206)
  + Timing 🡪 will qualitative and quantitative data be collected concurrently or sequentially
  + Weighting 🡪 what priority will be given to each type of data
  + Mixing 🡪 when and how the evaluator brings together the qualitative and quantitative data
  + Theorizing 🡪 ways that theories frame the project
* Six primary mixed-methods strategies
  + Quantitatively led sequential explanatory design
    - Collect and analyze quantitative data before qualitative data
    - Greater weight given to quantitative data.
    - Used to explore or explain puzzling findings from the quantitative analysis
  + Qualitatively led sequential explanatory design
    - Collect and analyze qualitative data before quantitative data.
    - Identify issues and questions that drive quantitative efforts.
  + Concurrent triangulation approach
    - Collect and analyze qualitative and quantitative data concurrently.
    - Determine degree of convergence.
    - Weaknesses in one method offset by strengths of another method.
    - Findings that are consistent across multiple sources are considered more reliable.

Identifying Appropriate Samples in Qualitative Evaluation

* Qualitative sampling strategies deliberately select cases.
  + Purposeful sampling (also called theoretical sampling)
* Total number of cases is limited.
* Purposeful sampling methods
  + Snowball sampling (also called chain sampling) 🡪 ask for referrals
    - Sample is large enough once you reach saturation where themes and issues begin to repeat themselves across individuals interviewed.
  + Opportunistic sampling 🡪 participants recruited to explore issues that come up in earlier interviews.
  + Stratified purposeful
  + Maximum variation

Collecting and Coding Qualitative Data

* Qualitative data collection instruments are structured to some extent.
  + Even informal conversational interviews typically use a topic guide.
* Standardized open-ended questions that are pre-planned as always asked in the same order.
  + Can also contain closed-ended questions.
  + Placing closed-ended questions at the beginning may establish a short answer response pattern.
* Limitations of data collection instruments
  + Structured
    - Missed opportunities
    - Imply a particular point of view
  + Unstructured
    - Costs
* Interviewing is a principal means of collecting qualitative data.
  + See Table 5.8 on p. 211.
* Qualitative findings must be clear, credible, and relevant
* Coding qualitative data
  + Become familiar with the data
  + Identify emerging issues
  + Summarize and organize the data using a thematic coding chart (see Table 5.10 on p. 215)
  + Classify and interpret the data
    - Identify patterns, explanations, and hypotheses

Assessing the Credibility and Generalizability of Qualitative Findings

* Fewer cases but higher quality and completeness of information
* Increasing robustness of qualitative data and findings
  + Check for representativeness
  + Check for researcher effects
  + **Triangulate data sources**
  + Determine whether some data sources are more credible than others
  + Check for sample bias
  + Use extreme cases to calibrate findings
  + Seek explanations for anomalous findings
  + Look for negative evidence that do NOT support your conclusions
  + Check for internal consistency of findings using if-then statements
  + Check for alternative explanations of findings
  + Replicate findings in a comparable setting
  + Check for rival explanations using the data
  + **Confirm summarization with participants.**

Connecting Qualitative Evaluation Methods to Performance Measurement

* Forcing the use of performance measurements that rely on numbers may not capture key outcomes
* Most Significant Change (MSC) approach
  + Alternative to quantitative logic models
  + Ask participants to identify positive and negative changes over a fixed time period.
    - Indicate which change is the most significant and why.
    - Partially participatory
  + Useful in situations where data collection capacity and resources are limited.

The Power of Case Studies

* Often easier to convey key findings using case examples.
* Order of increasing intuitive meaning and relatability: Data 🡪 Tables 🡪 Graphs 🡪 Narratives
* Single cases often treated as though they contain more evidence than just one case.
  + Must be methodologically defensible.

**Chapter 6: Needs Assessments for Program Development and Adjustment**

Preview Questions

1. What should be considered when conducting needs assessments?
2. How do you conduct a needs assessment?

Reading Summary

Introduction

* Needs assessments examine potential unmet needs of a population in a given environment.
* Needs assessments are becoming more common as a part of evaluation-related activities.
* Usually performed to inform program development and planning or to fine-tune ongoing programs.
* Can be resource intensive.

General Considerations Regarding Needs Assessments

* In terms of evaluation, needs are gaps between what is and what is desired.
* Needs assessments are about
  + Defining the gap
  + Assessing the extent of the gap
  + Prioritizing the needs to address
  + Determining an plan of action to address the needs
* Group-Level Focus for Needs Assessments
  + Needs assessment can focus on various levels
    - Individual
    - Group (i.e., population)
    - Service provider
  + System
  + Methodologies used usually are NOT sensitive enough for individual-level decisions.
  + Sometimes tools meant for individuals can be useful for group level needs assessments.
* Needs assessment fits into the strategic planning phased of the performance management cycle.
* Needs assessments are often conducted as programs are being developed or modified.
* Periodic assessment of the need for a program has become part of core evaluation practices.
* Types of needs
  + Normative needs are defined by professionals or experts according to some acceptable standard (similar to prescribed needs).
  + Felt needs are determined by having individuals or groups confirm the absence or presence of a state (sometimes referred to as perceived needs)
  + Expressed needs are extrapolated from the actual update of a service, service wait list, or requests from individuals.
  + Comparative needs are determined by comparing units of analysis in comparable situations.
  + Maintenance needs come to light only after a program is withdrawn.
* Translating needs into policies and programs involve choices based on values.
* Needs assessments are inherently political.

Steps in Conducting Needs Assessments

* Can be led internally or externally
* Keep needs assessment projects simple
  + Basic enough to incorporate into ongoing organizational monitoring as part of the strategic planning process.
* Three phases of needs assessment
  + Pre-assessment
  + Assessment
  + Post-assessment
* Needs assessments can be iterative.
* Needs Assessment Committee (NAC)
  + Helps achieve organizational commitment to the needs assessment findings
  + Encourages the implementation of the recommended changes.
* Pre-Assessment Phase
  + Relies on existing information and informal interviews.
  + Determines whether to proceed with a more formal needs assessment.
  + Can be facilitated by an external advisor.
  + Focuses the needs assessment
    - Determine questions to identify the nature of the need.
    - Understand the target population
      * Sociodemographic characteristics
      * Geographic scope
    - Understand the strategic context of the program
      * Helps sets the parameters for the scope of the problem.
    - Inventory existing programs that address the need.
    - Identify duplication of services and co-services
    - Determine the resources available for the needs assessment
      * Assessment approach
      * Data collection and analysis
      * Facilitation of collaboration, planning, and priority setting
      * Report writing and dissemination
      * Operational expenses
    - Summarize the pre-assessment findings
      * 5-10 page summary
        + Scope and size
        + Subdividing geographies or topics
        + Usefulness of collection more information
        + Need for collecting more data
        + Need for further preliminary data collection
        + Formation and make-up of a needs assessment committee (NAC)
        + Resources the organization can commit to the needs assessment
        + Time requirements
        + Importance to the organization
    - Using available sources to learn about preliminary as-is and should-be conditions.
      * Literature reviews
      * Organizational data (e.g., waitlists, referrals)
      * Government reports
    - Make Go/No-Go decision
* Assessment Phase
  + Can be time consuming and costly
  + Recommended agenda for coordination meeting
    - Was enough learned from the pre-assessment that make new data unnecessary?
    - Have we developed a good sense of the need?
    - Is there agreement as to priorities?
    - Should the focus shift to root causes?
    - What criteria should be used for developing or selecting solutions?
    - Do we know enough to skip a full assessment and begin post-assessment implementation?
  + Must have the resources and expertise to identify and analyze needs.
  + Combination of methods is often used.
  + Surveys are a principal means of obtaining new data.
    - Pretest before full implementation.
    - Make necessary arrangements to protect respondents’ privacy.
    - Most respondents will NOT rank more than six choices.
  + Must have a defensible sampling procedure.
    - Random is ideal.
    - Other types of sampling procedures
      * Systematic sampling using a skip factor.
      * Stratified sampling.
      * Proportionate stratified sampling.
      * Disproportionate stratified sampling.
    - Cost is a function of the sample size.
      * Construct sample size based on estimated proportion having the perceived need and the amount of error that is acceptable.
  + Measurement validity is about overcoming bias.
    - Using surveys is not straightforward; must adjust for bias.
    - Needs assessments are subject to many of the threats to measurement validity.
    - Warning respondents that survey contains fictitious and real options can discourage response bias.
  + The risk associated with NOT addressing a need should be considered when setting priorities.
  + Examples of criteria for setting priorities
    - Magnitude of problem
    - Cost-effectiveness of solving the problem
    - Trend direction of the problem
    - Magnitude of difference compared to similar situations
    - Available funds.
    - External directives
    - Severity of consequences for not addressing the need
    - Acceptability to community
    - Potential for prevention
    - Political pressure
  + Needs assessment report
    - Executive summary
    - Introduction
    - Methods and lines of evidence
    - Findings
    - Discussion (i.e., interpretation of findings)
    - Conclusion and recommendations
      * Adjustments to existing programs are easier to implement than new programs or greatly expanding existing programs.
    - Appendixes

Post-Assessment Phase: Communicating and Implementing a Needs Assessment Action

* Recommendation is to reconfigure the NAC to include additional managerial representation.
* Build a support coalition and engage stakeholders throughout the process so there are few surprises.
* Communication content and style depends on the audience.
* Releasing needs assessment results with recommendations in sufficient time to be included in budget development is ideal.
  + Work backwards from the time of year for budgeting when planning needs assessments

**Chapter 8: Performance Measurement as an Approach to Evaluation**

Preview Questions

1. Why is performance measurement being emphasized more?
2. What is the purpose of performance measurement?
3. What metaphors are used to explain and understand performance measurement?
4. What is the difference between performance measurement and program evaluation?

Reading Summary

Introduction

* Performance measurement complements program evaluation in assessing the effectiveness of policies and programs.
* The deficits, debts, and taxes of the 1970s and 1980s gave rise to results-focused performance measurement as a principal means for demonstrating accountability to the public.
* There are three basic metaphors for explaining and understanding performance measurement.
* Performance measurement describes program results.
  + Measure to what extent intended results have been achieved.
  + Compare planned outcomes to target outcomes in an effort to describe accomplishments.
* Program evaluation seeks to explain why observed results of a program occur.
* Performance measurement is nearly a universal expectation in public and nonprofit sectors.
* Drivers of performance measurement:
  + Globalization
  + Public debt
  + Citizen dissatisfaction with public services
  + Public service inefficiencies
  + Advances in information technology
* Emphasis of process compliance being supplemented or supplanted by emphasis on achieving objectives.
* Public sector evaluative criteria are intended to link resources to outcomes in a manner that is analogous to private sector measures of success.

Growth of Performance Measurement

* Performance measurement historically connected with financial accountability.
  + Accounting provided the framework for measuring inputs and determining efficiency and productivity.
* Performance measurement began in the United States in the 1960s.
  + Planning, programming, and budgeting systems (PPBS)
  + Bureau of Municipal Research created by New York City, New York in 1907
    - Instituted the annual “Budget Exhibit”
* Comparable bureaus of municipal research in 16 northeastern U.S. cities by 1916
* Roughly 89 local governments in the U.S. were issuing performance reports by 1941.
* Conceptualized clusters of administrative activities did not need to coincide with organizational units.
* Programs are being conceptualized as open systems that convert resources into results.
* Goal/objective-driven vision for governments led to performance management-related views of government functioning.
* Successors to PPBS:
  + Zero-based budgeting
  + Management by objectives (MBO)
* Public accountability for results was a key feature of government restructuring of the 1970s and 1980s.
* New Zealand considered exemplar of public sector reforms aimed at reducing the scale and scope of government.
* Osborne & Gaebler (1992) principles of sustainable government:
  + Steer rather than row
  + Citizens should participate in ownership and control of public services
  + Competition among public sector service deliverers is beneficial
  + Governments should be mission-driven, not rules-driven
  + Funding should be tied to outcomes, not inputs
  + Performance information should be used to improve results
  + Governments should focus on citizen needs, not interest groups or bureaucracy needs
  + Fund generation should also be a focus in the public sector, not just spending
  + Governments should focus on anticipating and preventing problems rather than fixing problems after the fact
  + Government should use participatory and decentralized management approaches
  + Government should use market mechanisms to achieve public purposes
* Public choice theorists argued that private sector microeconomic models could be applied to the public sector.
* New Public Management (NPM) framework emerged in the 1990s.
  + Accounts for much of the current widespread emphasis on performance measurement in the public sector.
* Government Accounting and Standards Board (GASB) was established in 1984 to establish standards of financial accounting and reporting for local and state governments.
  + Championed user-friendly non-financial performance reporting as a component of overall reporting
  + Service Efforts and Accomplishments (SEA)
* Performance measurement and reporting has been legislated by most states.
* Congress passed the Government Performance and Results Act (GPRA) in 1993 to mandate results-based management in U.S. federal departments and agencies.
  + Replaced by the Government Performance and Results Act Modernization Act in 2010.
* No comparable legislation to GPRA in Canada but similar expectations.

Performance Measurement Purposes

* General two purposes for performance measurement systems:
  + Accountability
  + Performance improvement
* NPM champions two themes that seem contradictory at times:
  + Government managers must be accountable for results
  + Government managers must have freedom to manage
* Approach to driving public sector performance:
  + Mandate focus on intended outcomes
  + Measure performance towards achieving intended outcomes
  + Report publically on whether intended outcomes were achieved
* Performance measurement is primarily summative
  + Usually top-down initiatives

Metaphors that Support and Sustain Performance Measurement

* Organizations and machines
  + Organizations are designed to produce tangible outputs and results
  + Rooted in the scientific management movement of the early 1900s
  + Emphasis on use of time and motion studies to break down processes
  + Emphasis on quantitative measurement
  + Performance measurement systems rely on visual heuristics to provide managers with an at-a-glance way of assessing organization performance, much like the instruments on the dashboard of a machine
* Government as a business
  + Guides efforts to infuse business practices into government
  + Emphasizes clearly stated objectives, planned and managed programs, efficiency, and attention to bottom-line results
  + NPM champions these same principles
* Organization as open systems
  + Biological metaphor
  + Goal of homeostasis, which is maintaining a steady-state in relation to fluctuations in the environment.
  + Biological organisms operate within certain parameters
  + Indicating the health of the government analogous to indicating the health of a person
  + Suggests that we have a good understanding of cause-and-effect linkages in programs
  + Use of logic models is indicative of the adoption of the opens systems metaphor

Comparing Program Evaluation and Performance Measurement

* Program evaluation knowledge and skills can be adapted to performance measurement systems.
* Performance measurement is an approach to evaluation.
* Program evaluations are projects with a beginning and an end, similar to a research project.
* Program evaluation vs. performance measurement
  + Episodic vs. ongoing
  + Issue specific vs. general focus
  + Customized measures and lines of evidence vs. routinized measures and data
  + Attribution needs to be proven vs. attribution being assumed
    - Counterfactual condition is what would have happened without a program
    - Performance measurement focuses on construct validity
    - Core technology of a program affects the likelihood that observed outcomes can be tracked back to the program.
    - Performance measurement can be used more confidently to measure outcomes of programs with high-probability program technologies.
  + Targeted resources vs. resources embedded as part of organizational infrastructure
  + *Ex ante* negotiated purposes vs. *ex post* evolving purpose

**Chapter 9: Design and implementation of performance Measurement Systems**

Preview Questions

1. What are the key steps in designing and implementing a performance measurement system?
2. How is performance measurement used for public accountability?

Reading Summary

Introduction

* Perspectives on public sector organizations:
  + Rational/technical view emphasizes systems and structures
  + Political/cultural view emphasizes dynamics of people interacting to accomplish tasks
* There are 12 steps to designing and implementing performance measurement systems.
  + Some steps align with the rational/technical view while others align with the political/cultural view
* Another metaphor is organizations as political systems.
* Politics is about the formal and informal processes used to allocate scarce resources among competing values.
* Some environments are more conducive to sustaining performance measurement systems than others.
* There must be a demand for performance information to sustain a performance measurement system (i.e., supply of performance information).

Key Steps in Designing and Implementing a Performance Measurement System

* Steps are guidelines
* Steps:
  + Identify the organizational champion of the change
    - Provide continuing support for the process
    - Multiple purposes
      * Managerial decision making
      * Organizational alignment
      * Transparency and accountability
    - Leadership commitment affects performance information use
    - Process of building performance measurement can increase uncertainty for managers
  + Understand what performance measurement systems can and cannot do
    - Performance measurement only describes what is happening; they do not explain why it is happening.
    - Cannot conclude the observed outcomes are due to a program unless we have additional information that excludes other factors in the environment.
    - Six strategies to reduce uncertainty about cause-and-effect linkages
      * Develop intended-results chain
      * Assess existing research that supports the results chain
      * Assess alternative explanations for observed outcomes
      * Prepare the performance story
      * Seek additional evidence
      * Revise the performance story, if necessary
      * Conduct program evaluation if performance story is insufficient
  + Establish multichannel communication pathways to facilitate information sharing, problem identification, and problem solving
    - Include intended users
    - Top-down communication serves to clarify direction, provide a framework and timelines for processes, specify resource availability, and affirm importance
    - Bottom-up communication questions or seeks clarification of definitions, timelines, resources, and direction.
    - Horizontal communication provides examples, shares solutions, and offers informal support.
  + Clarify the expectations for the intended uses of the performance information that is created
    - Use performance measurement for internal performance improvement to obtain buy-in
    - Public reporting should not be the primary reason for developing a performance measurement system
    - The potential to use performance information to criticize elected official and bureaucrats creates an incentive to limit reporting of information that may reflect negatively on the government
    - Andrew Ladley’s “Iron Rule of the Political Contest”:
      * The opposition is intent on replacing the government
      * The government is intent on remaining in power
      * Elected officials want to get re-elected
      * Party leadership is dependent on retaining the confidence of colleagues
      * Retaining the confidence of colleagues is dependent on the first three principles
    - Using a system designed for formative purposes for summative purposes changes the incentives for those involved.
      * Increases the likelihood of gaming the system
      * Programs that begin with formative intentions often evolve to summative purposes over time.
  + Identify the resources available
    - One time infusions of resources are NOT sufficient to sustain the system.
    - Training of staff is important.
    - Ensuring that the system is sustainable requires managing with performance data.
  + Understand the organization history around similar initiatives
    - The previous experience of managers will affect their willingness to support current efforts.
    - Understand why previous effort were or were not successful.
    - Losing employees involved in previous efforts is a liability if those efforts were successful but can be an asset if those previous efforts were unsuccessful.
  + Develop logic models for the programs and identify key constructs to be measured
    - General program objectives that were created to mollify competing expectations complicates performance measurement.
    - Well-formed objectives:
      * State an expected change
      * State a magnitude of expected change
      * State a target population for the expected change
      * State a time-frame for achieving the expected change
    - Logic models useful for identifying constructs that are candidates for performance measurement.
    - Three major foci for performance measures:
      * Efficiency
      * Quality
      * Effectiveness
    - It’s worthwhile to include constructs for environmental factors in performance measurement systems to help with addressing the attribution questions.
    - Likelihood of successfully implementing performance measures in various environments:
      * Coping organization
        + Tasks change frequently
        + Results not visible
        + Low likelihood
      * Production organizations
        + Simple, repetitive tasks
        + Results are visible and countable
        + High likelihood
      * Craft organizations
        + Mix of professional knowledge and skill unique to tasks
        + Visible outcomes
      * Procedural organizations
        + Rely on processes to produce outputs
        + Outputs are visible and countable
        + Outcomes are less visible
  + Identify any constructs that apply beyond the single program
    - Organizational logic models are logic models for the entire organization that link business units to organizational goals.
    - It’s possible to have programs that are not meeting their objectives while overall the organization is meeting its objectives.
    - Individual objectives 🡪 group objectives 🡪 program objectives 🡪 organizational objectives
  + Involve prospective users in reviewing logic models and constructs
    - Participation in validating constructs increases likelihood that the performance measurement results will be useful.
  + Measure the constructs that have been identified
    - Translating constructs into observables involves measurement.
    - Issues when using secondary data sources:
      * Whether existing data be adapted to fit the constructs.
      * Whether existing data sources sufficiently cover the constructs that need to be measured (i.e., content validity).
      * Whether existing data sources allow triangulation of measures of key constructs.
      * Can existing data sources be manipulated by stakeholders (i.e., response process validity).
    - Issues when using primary data sources:
      * Whether resources will be available for data collection on an ongoing basis.
      * Whether the methodology ensures data credibility.
      * Determining who will collect and report data to ensure its credibility.
      * Having a rationale for each measure that is included or excluded.
    - Performance measurement is about finding measures that plausibly connect constructs with data.
    - Internal structure validity, concurrent validity, predictive validity, convergent validity, and discriminant validity are generally beyond the scope of performance measurement situations.
    - Typically 10 to 20 performance measures in a report.
    - Desirable to have both quantitative and qualitative measures in performance measurement systems.
      * Single cases can only illustrate.
  + Record, analyze, interpret, and report the performance data
    - Prospective users of performance information should be offer scenarios of various interpretations of the information.
    - Ambiguous measures tend to weaken the credibility of the performance measurement system.
    - It’s important to determine the kinds of comparisons that the performance information will be used to make.
    - The ratchet effect is the tendency for performance targets to be lowered over time if agencies fail to meet them.
    - Buy-in of the performance measurement system by managers is an incremental process.
  + Regularly review user feedback and make changes as appropriate
    - Develop channels for user feedback.
    - Create an expert review panel of person who do NOT have a stake in the system.
    - External auditing helps ensure the long-term credibility of the system.

Performance Measurement for Public Accountability

* General purposes of performance results:
  + Public accountability
  + Improve performance
* Elected officials and the pubic are the primary recipients of performance reports.
* The assumption is that public accountability incentivizes organizations to become more efficient and effective.
* The political culture can influence the quality and use of performance information.

Summary

* Managerial involvement is important for the success of performance measurement systems.
* Six of the 12 criteria are more important:
  + Sustained leadership
  + Effective communication
  + Clear expectations of the system
  + Sufficient resources
  + Identifying key program and organizational constructs
  + Valid measures in which stakeholders have confidence

**Chapter 10: Using Performance Measurement for Accountability and Performance Improvement**

Preview Questions

1. What are the different ways performance results can be used?
2. What needs to be considered when implementing and sustaining performance measurement systems?

Reading Summary

Introduction

* Intended uses versus actual uses of performance information.
* Incentives and organizational political factors are important to understanding how performance information is actually used.
* Public performance reporting is often risky.
  + Political cultures are risk-averse.
  + Political culture is the set of values, attitudes, beliefs, and behavioral propensities that characterize the relationship among individuals, groups, and institutions in a society.
* Performance measurement does NOT replace program evaluation.
  + Performance measurement focuses and informs program evaluation.
* Accountability historically focused on the process by which organizations made decisions.
* Public sector motivations are central to understanding how to design well-performing organizations and processes.
  + Historical assumption was that public servants were motivated to serve the public interest.
  + Public choices theorist and New Public Management (NPM) proponents believe that public servants are motivated by incentives, rewards, and sanctions.
* Using performance reporting to induce performance improvement assumes that performance results have real consequences for the organization.

Using Performance Results

* There is little research on whether and to what extent elected officials make use of performance information.
* Legislator Expected Versus Actual Uses of Performance Reports in British Columbia, Canada
  + Performance information was either under-utilized or not used at all.
  + Performance information was less useful to opposition members as critics of government policies and programs.
  + Suggest that an important link in the performance management cycle is weak.
* High-Stakes Uses of Performance Measures
  + “Targets and terror” approach to performance management used public rating and ranking systems.
    - Challenged the reputation of each organization.
  + Four criteria for effective ranking systems
    - Ranking system must be established for the organizations in a sector.
    - Ranking results need to be published and disseminated widely (i.e., “named and shamed”).
    - Ranking results need to be easily understood by the public and stakeholders.
    - Use cyclical rankings to determine whether performance has improved.
  + Can have unintended adverse effects.
* Assessing the “Naming and Shaming” Approach to Performance Management in Britain
  + What gets measured matters
    - What is not or cannot be measured doesn’t matter and simply gets neglected
    - “Hitting the target and missing the point”
  + Picking key performance measures often misrepresents the complexity of the work.
  + Gaming the performance measure can be a real problem.
    - Leads to lack of credibility.
  + Categories of gaming behavior
    - Ratchet effects are where organizations try to negotiate performance targets that are easy to attain.
    - Threshold effects are when performance targets produce organizational behaviors that distort the range of work activities.
    - Output distortions are where performance results are adjusted to align with expectations.
* Reassessing the Performance Management Cycle: Incentives and Organizational Politics
  + Performance management cycle is a normative model
    - It display s intended relationships
  + Must police other behaviors through audits to ensure credibility of the performance measurement system.
    - Audits are costly and can affect the sustainability of the system.
  + Unintended adverse effects are more likely when:
    - Stakes are high
    - Political culture is adversarial
    - Other interest groups are openly critical
    - Performance results are highly visible
  + Pattern for public performance measurement system development
    - Indicators become more numerous and comprehensive.
    - Formative approaches become more summative.
    - Summative approaches become linked to incentives and sanctions, which increase the pressure for gaming.
    - Indicators become more complex and more difficult for non-experts to understand.
    - Ownership of the performance system becomes more diffused.
    - Trust in the performance data and interpretations decline.
  + It’s possible to use a low-stakes approach to performance measurement.
* Use of Performance Measures in a Non-Adversarial Political Environment
  + Measures developed by managers for their own use.
  + Performance data used to improve public accountability AND performance.
  + Performance measurement systems generally not useful or sustainable if managers are not involved in developing them.
* Using Performance Information from Management: Encouraging Internal Uses
  + Decouple externally reported performance results from internally used performance information.
* Increasing uses of Performance Information by Elected Officials: Supply and Demand
  + Little evidence that performance information affects appropriations.
  + Audit strategy
    - It assumes that increasing credibility of information will increase its use
  + Legislation and training
    - The production and review of performance information has been legislated in many cases.
    - It assumes that increasing the availability of performance information will increase its use.
    - Possibilities for training political decision makers are likely limited.
  + Drive to measure and report performance results is entrenched and unlikely to change even it use of such information is limited.

Three Additional Considerations in Implementing and Sustaining Performance Measurement Systems

* The Central Influence of Performance Measurement in Public Organizations
  + Results-based management requirements were layered on top of existing process-focused requirements.
  + Effect has been to centralized organizational decision making.
  + Targets don’t result in more managerial latitude; they impose more control on managers.
* Attributing Outcomes to Programs
  + Attribution becomes more of a problem the further down the logic chain from program outputs.
  + Low-probability technologies are much more vulnerable to external influences.
  + Holding managers accountable for outcomes may result in gaming behaviors.
* The Levels of Analysis Problem
  + Conflating organizational, program, and individual performance is an issue.
  + One level of analysis is not necessarily equivalent to the other.
  + To fully measure performance you must measure it at each level in the organization.

**Chapter 11: Program Evaluation and Program Management**

Preview Questions

1. What are the issues with joining evaluation and management?
2. Can evaluators and managers partner in evaluation?
3. What are the limits and opportunities associated with manager involvement in evaluation?
4. How do you achieve objectivity in program evaluation?
5. What are the commonly accepted criteria for high-quality evaluations?

Reading Summary

Introduction

* The relationship between program managers and evaluators is influenced by evaluation purposes and organizational contexts.
* Many evaluators are part of the organizations in which they do evaluations.
* Formative and summative evaluations each have their own incentives that can bias manager involvement in evaluations of their own programs.
* Program evaluation is situation-specific.
* Managers are more likely to view formative evaluations as friendly rather than summative evaluations.
  + Managers more willing to cooperate with formative evaluations.

Issues Associated with Joining Management and Evaluation

* Evaluation and organization are somewhat contradictory.
* Managers need to balance stability of their programs with evidence-based evaluation.
* Idealized evaluator characteristics
  + Questions assumptions
  + Skeptical
  + Detached
  + Views organizations and programs as means to ends; not ends in themselves
  + Relies on evidence
  + Focused on the social needs that the program serves rather than organizational needs
* Idealized manager characteristics
  + Committed to their programs
  + Advocates for what their programs do
  + Do not want their commitments curtailed
  + Do not want their resources reduced
* Challenges associated with testing programs experimentally before launching them when evaluation is an internal function.
  + Different stakeholders may have different values and expectations.
  + May upset prospective clients that want the program but can get it because they are part of the control group.
* One solution is to make program evaluation an external function.
  + Evaluators part of agency not under administrative control of the organization’s managers.
    - Likely to face considerable objection from managers.

Evaluators and Managers as Partners in Evaluation

* Utilization-focused evaluation seeks to produce evaluations that managers and other stakeholders will actually use, which means developing a working relationship between evaluators and managers.
  + Answers the question of whose values will frame the evaluation.
* Participatory evaluation
* Six stages of developing internal evaluation capacity
  + Ad hoc evaluations focused on single programs
  + Regular evaluations that describe program processes and results
  + Program goal setting, measurement of program outcomes, program monitoring, adjustment
  + Evaluations of program effectiveness, improving organizational performance
  + Evaluations of technical efficiency and cost-effectiveness
  + Strategic evaluations including cost-benefit analyses
* The six stages assume that it is possible to transform an organizational culture to embrace evaluation.

Building an Evaluative Culture in Organizations

* Key features of an evaluative culture
  + Engages in self-reflection and self-examination
  + Engages in evidence-based learning
  + Encourages experimentation and change
* Role of evaluation encompasses knowledge management for the organization
* Evaluative cultures are learning organizations
  + Scan and anticipate change
  + Develop an ability to question, challenge, and change operating norms and assumptions
  + Allow an appropriate strategic direction to emerge
* Double-loop learning is learning that uses evidence to critically assesses existing organizational goals and priorities and provides options for adopting new goals and objectives.
* Five building blocks for creating learning organizations
  + Systemic problem solving using evidence
  + Experimentation and evaluation of outcomes before broader implementation
  + Learning from past performance
  + Learning from others
  + Treating knowledge as a resource to be widely communicated
* Developmental evaluation
  + Alternative to formative and summative evaluations
  + Views organizations as co-evolving in complex environments
  + Organizational and program objectives may be in flux
  + Evaluators take on the role of organizational development specialists
* To build and sustain an evaluative culture
  + Managers need adequate autonomy to manage for results
  + Managers need to be able to adjust operations as they learn what works and what doesn’t work.
  + Consistent with the New Public Management (NPM) movement.
  + Requires performance expectations to replace existing process controls rather than simply being layered on top of them.
  + Impacts willingness to take risks.
    - Acknowledge below-par performance in order to learn and improve.

Limitations and Opportunities to Manager Involvement in Evaluations

* Program managers are increasingly expected to participate in evaluating their own programs.
* Involving managers in summative evaluations is different than involving them in formative evaluations.
* Standards for evaluation quality from the Joint Committee on Standards for Educational Evaluation
  + Utility
  + Feasibility
  + Propriety
  + Accuracy
  + Accountability
* With external evaluations, organizational staff don’t have the knowledge or expertise to continue for themselves.
* Evaluating for accountability is typically summative.
* Organizational managers are likely to be skeptical of evaluations whose results will become public and may affect their programs even if the stated intention of the evaluation is to improve the program (i.e., formative).
* Over-reliance on external evaluations could invite corruption at worst or incompetent evaluations at best.
  + Administrators that find themselves in political conflicts over their program may be willing to pay for friendly evaluations.
  + This would discredit the evaluation field.

Striving for Objectivity in Program Evaluations

* Technical credibility
  + Technically competent, defensible, and transparent
  + Objective
  + Must also be perceived as technically competent and objective
* Evaluation is the linking of theory and practice
* Being an evaluator is different from being an evaluation consultant
  + Evaluator relies on validity and objectivity
  + Evaluation consultant is undoubtedly influenced by interactions with clients and stakeholders and their biases
* Evaluator as “expert witness” called to “testify” about a program.
  + Expert witnesses can be purchased.
* Schriven (1997) definition of objectivity is:
  + With basis
  + Without bias
* Objectivity can mean that the evidence and conclusions can be verified and confirmed by people other than the original authors and the conclusions must follow from the evidence.
  + **This sounds more like reproducible**.
* Objectivity is about the ability to scrutinize and replicate methods and repeat findings.
* Objectivity does not confer truth.
  + Truth is agreement among those who ascribe to a common paradigm.
* Repeatability is rarely applicable to program evaluation.
  + Instead there is an audit trail for evidence and findings.
* In practice objectivity is conferred when results can be corroborated or disconfirmed.
  + Objectivity is predicated on process.

Criteria for High-Quality Evaluations

* Several professional associations with stakes in program evaluation have developed codes and ethics or best practices.
  + Little specific focus on objectivity as a criterion.
* Scientific realism 🡪 emphasize objectivity and independence
* Social constructionism 🡪 emphasize balancing different realities
* American Educational Research Association (AERA) specifically includes objectivity as a criterion.
* AERA criteria
  + Logical, evidence-based chain of reasoning
  + Methods appropriate to the question being investigated
  + Observational and experimental designs with instruments that provide reliable and generalizable findings
  + Data and analysis adequate to support findings
  + Clear, detailed explanations of procedures and results
  + Adherence to professional norms of peer review
  + Dissemination of findings
  + Access to data for re-analysis and replication
* Other quality-related criteria are more relevant for many evaluation associations
  + Accuracy and credibility of evaluation information
  + Honesty and integrity of evaluators and the evaluation process
  + Validity and reliability of evaluation information
  + Declaring and avoiding conflicts of interest
  + Impartiality in reporting findings and conclusions
  + Evaluator independence and competence
  + Transparency
* Asserting objectivity in evaluation is not feasible because the evaluation field has many different epistemological and methodological stances.
  + Accounting and management consulting have more unified views of their fields.