Outcomes Evaluation Toolkit

September 15, 2021

— A Beacom Research Fellows Resource | Augustana Research Institute

Gedion Alemayehu
Grace Bucklin
Charlotte Berg
Annie Olson

Beacom Research Fellows

Augustana Research Institute
Augustana University
Sioux Falls, South Dakota

This report was made possible by a gift from Miles and Lisa Beacom to support the Beacom Research Fellows Program at Augustana University. The Beacom Research Fellows Program partners Augustana students with community organizations to complete collaborative research projects on behalf of the organization. Fellows make rigorous research, data management, analysis, and reporting more broadly available to organizations in the Sioux Falls area that are working to improve the quality of life.

Table of Contents

Introduction and Overview

This report begins with a brief guide to the principles of outcomes evaluation, covering the various purposes that evaluation may serve and the special significance of outcomes evaluation for monitoring program performance. It then walks through the process of identifying outcomes that are aligned with a program's goals and objectives, defining measurements or indicators to record and track those outcomes, and the steps involved in designing and implementing ongoing outcomes monitoring.

The following section of this report contains a set of tools and templates that mission-driven organizations may find useful in designing and implementing outcomes monitoring. The tools in this section include a logic model template, a checklist to guide organizations through operationalizing outcomes, a checklist to assess the quality of evaluation questionnaires, and a simple template that can be used to track key outcomes across multiple programs. Combined with the first section (the brief guide to outcomes evaluation), these tools constitute an Outcomes Evaluation Toolkit.

Section 1: A Brief Guide to Outcomes Evaluation

There are several reasons why nonprofit organizations should perform program evaluation regularly and in a timely manner. Some of the most important reasons for evaluation include the following:

- Program evaluation and outcome measurements can help an organization accomplish its mission more effectively and improve its success over time.
- The needs of the community change over time. Program evaluation helps target the most important community needs.
- Funders and regulators often require evidence-based data; program evaluations will help to measure outcomes and tell an appealing story demonstrating empirical evidence of the extent of improvement accomplished.
- Program evaluations give staff important feedback. Program leaders can keep a record of accomplishing what they have set out to achieve and get satisfaction for the work they do.
- Results from program evaluations might lead to new ways of doing work under each program and the overall organization, spurring innovation and improvement.

Given the range of purposes that evaluation can serve, it is no surprise that there are different types of evaluations intended to serve these different purposes. At the most basic level, evaluations can be categorized as formative, summative, or outcomes evaluations.

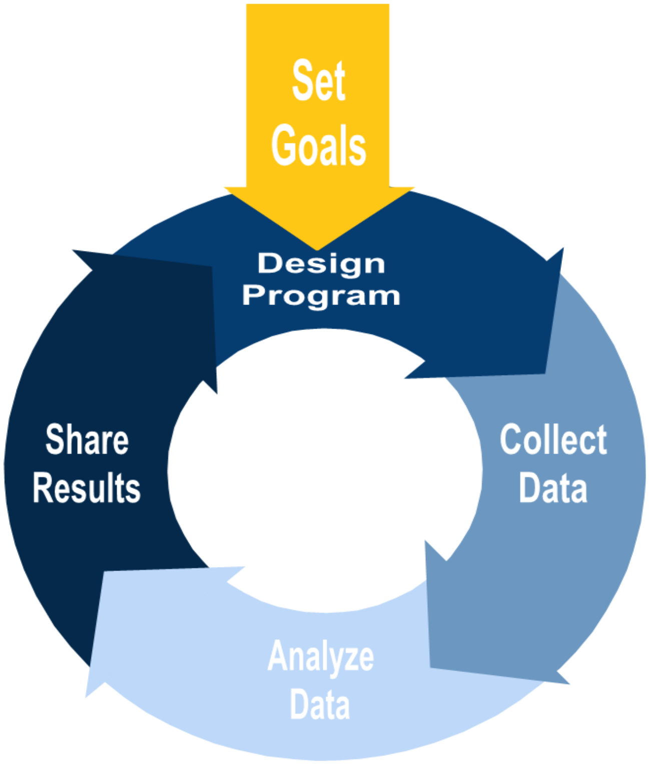
Formative evaluations are intended to help shape the development and implementation of programs. This type of evaluation can help management and staff understand how to improve a program. Often, the audience for formative evaluation is internal, and the goal is to provide information that will help form or reform a program. Typically, formative evaluations emphasize producing immediately actionable recommendations, and even preliminary results may be shared frequently in order to begin to implement change as early as possible. Within this vein, needs assessment, feasibility studies, implementation assessments, process evaluations, and ongoing program monitoring of inputs and outputs may be useful for continuously adjusting and improving programs.

Summative evaluations are meant to yield a definitive statement of findings, often a judgment about whether or not a program is effective. Typically, summative evaluation is undertaken once a program is fully formed. Unlike formative evaluation, this type of evaluation does not look at how the program is carried out, but rather, whether it has its intended effect. Often, the audience for summative evaluation includes both internal and external stakeholders, such as funders or regulators. In most cases, summative evaluations lead to a final report of findings, and methods and results may be held to a higher standard of scrutiny than with formative evaluation. In this vein, a formal impact evaluation may try to rigorously disentangle the causal effects a program has on participants, after accounting for all of the other influences (outside of the program) that might also be affecting changes in participants' lives.

Outcomes evaluations fall somewhere between formative and summative evaluation. Outcomes evaluation runs a spectrum from continuous program monitoring for improvement (which takes more of a formative approach, like a frequently updated performance dashboard), to more summative evaluation (a summative grant report, annual report to the community, or published research study).

The Evaluation Cycle

Effective evaluation is a cycle. To make the most of evaluation, organizations must close the loop and complete the cycle, ensuring that the results of evaluation are used to improve programs. Evaluation results that gather dust on a shelf are not worth the effort of collecting in the first place. Evaluations should be carefully designed to ensure that the results are trusted, reported clearly, and actionable. The diagram below illustrates the evaluation cycle.



To begin, program leaders should agree on the overarching mission and the specific program goals and objectives. As they design a program, they should also explicitly outline a program theory that explains how the services provided by the program will help achieve the desired goals. From there, leaders should select indicators to measure progress toward goals, then collect and analyze data. Once data has been collected, it should be shared with relevant stakeholders so that it can be used to improve the program and update the underlying program theory as needed. Then, the cycle begins with another round of data collection to inform the next round of program improvements.

When designing an evaluation or choosing what to measure, program leaders should think ahead to how the proposed data will inform or influence action. Rethink any data elements that are not actionable or do not help leaders derive insights about how a program is working. The remainder of this section offers specific recommendations for defining program goals, operationalizing outcomes, collecting data, and reporting results with an eye toward program improvement.

Process, Outcomes, and Impacts

Evaluations can be categorized based on how they will be used--that is, formative outcomes for changing and improving a program, and summative outcomes for rigorously measuring program results. Evaluations can also be categorized based on the aspect of a program they evaluate. In particular, evaluations can be categorized as process, outcomes, or impact evaluations.

Process evaluations, sometimes known as implementation evaluations, are usually formative. Fundamentally, process evaluations review how well a program has been implemented and can help an organization improve how a program operates. They focus mainly on the operation of the program, addressing the steps taken as inputs, activities, and outputs. Process evaluations tend to be more qualitative in nature, often asking open-ended questions about how a program is working or why an organization does things one way rather than another.

Process evaluations are often done early during program implementation, but they should be continued as part of continuous quality improvement. Indeed, process evaluations can help throughout any stage of program development: When a program is first implemented, they can be used to ensure that the program follows statutory requirements, professional design, and customer expectations. In later stages of program development, process evaluations can help to determine if the implementation of the programs meets customer expectations. A vital piece of the overall evaluation endeavor, process evaluations are important and can improve the quality of services, the efficiency with which they are offered, staff satisfaction, client satisfaction, and sustainability.

Outcomes evaluations are also one of the most common types of program evaluation. As their name indicates, they assess the outcome of a certain program. Unlike process evaluations, outcomes evaluations do not explain *why* a program works, but they are useful in determining *whether* it works (to achieve certain goals) or *what* works (by comparing programs or changes over time). It is also useful in identifying whether there are certain groups for whom a program works or does not work well (i.e., when, for whom it works).

In defining outcomes, it is important to understand the difference between "outputs," "outcomes," and "impacts." Project results can be classified into three categories:

1. Outputs
2. Outcomes
3. Impacts

Both outcomes and outputs entail the result obtained at the end of the program. However, outputs are an aspect of program delivery, whereas outcomes are observed among participants or beneficiaries of a program. For example, the number of coats distributed is an output; the reduced incidence of frostbite and hypothermia among recipients is an outcome.

Additionally, outcomes may range from proximal outcomes that are closely connected to a program, to more remote outcomes that are farther away from a program's immediate effects in terms of location, time, or scale. That is, outcomes are not always seen instantly at the end of the program; they can be intermediate to long-term effects. On the other hand, outputs are the results that happen immediately as a function of the program.

Impacts may refer to the broadest, most remote or long-term outcomes (e.g., community-wide effects), or to the net effects of a program after alternative causes have been accounted for. Impacts are more difficult to measure and evaluate than either outputs or outcomes; impact evaluation is briefly discussed below but is beyond the scope of this document.

Organizations tend to measure outputs instead of outcomes because they are usually easier to measure. Many types of outputs are measured as part of doing business (e.g., the number of people registered for lessons or the number of counseling appointments booked). Outcomes, on the other hand, may take extra effort to measure because they might not occur at the time or place where a program delivers services. Also, organizations may feel they have less control over outcomes and be resistant to measuring them out of concern they cannot truly drive change.

Despite these difficulties, outcomes are important to measure because they show the effect and importance of outputs—not only what an organization does or produces, but why it matters (and why potential donors should support it).

In addition to making sure outcomes and outputs are separated, it is also important to account for the difference between intended outcomes and unintended outcomes. Consideration of unintended outcomes in the evaluation process can help better improve and understand the programs' success as well as customer satisfaction.

Beyond Outcomes: Impact Evaluation and Cost-Benefit Analysis

Outcomes evaluation is often used for continuous quality improvement—it is more about ongoing monitoring of metrics that are practical and informative for program management rather than a deep, methodologically rigorous study of a program's causal effects. Impact evaluation digs deeper, disentangling program effects from alternative causes and rigorously measuring the impact of a program on participants or the broader community. An extension of impact evaluation, cost-benefit analysis measures the return on investment of a program, comparing the cost of running a program to the value it creates.

In essence, both impact evaluation and cost-benefit analysis extend outcomes evaluation to ask: (a) what are the larger impacts of the program's outcomes (i.e., long-term or community-wide, beyond immediate outcomes for participants); (b) to what degree can those impacts be directly attributed to the program rather than alternative causes; and (c) how does the benefit created by the program compare to the cost of running the program?

Impact evaluation and cost-benefit analysis both require more intensive data collection and typically use more complex methods to isolate a program's effects from the multiple causes that might influence an observed outcome. Both are beyond the scope of this document and are only briefly described here.

Impact evaluation is different from both outcome and output evaluation mostly due to its long-term nature. The impact might not even be achieved in the lifecycle of the program. These types of evaluations are designed to assess the net effect of the program, after accounting for outside forces that might also affect the impact of interest. For example, a job training program that tries to increase employment among participants would want to distinguish the effects of job training from community-wide changes that also affect participants' employment, like a major new employer coming to town or an economic recession. Since there could be multiple external factors related to a certain impact, it is hard to pinpoint the program's impact on its participants, and doing so can be costly; hence, impact evaluations are not very common. Well executed impact evaluation is the gold standard for establishing the existence and size of a program's effect on participants, net of alternative causes. Because impact evaluations require careful planning in advance, they ought to be done only after a program has matured.

Cost-benefit analysis is sometimes considered a sub-part of outcome evaluations because the outputs of the program are an integral part of performing these evaluations. Cost-benefit evaluations attempt to measure and compare the outputs of a program to the cost of the program. However, benefits are extremely hard to quantify and thus cost-benefit evaluation is not one of the most common program evaluation types nonprofit organizations utilize. Cost-benefit analyses may attempt to monetize the outcomes of a program in order to put them in a common currency to compare with costs (e.g., the monetary value to society of increasing the high school graduation rate from 70% to 80% could be measured based on the increased earning potential of graduates and reduced risk of incarceration). This type of evaluation is sometimes referred to as a social return on investment calculation.

Measuring Outcomes for Program Improvement

The rest of this document focuses on outcomes evaluation, beginning with how to identify outcomes and define them in order to make them measurable. As introduced above, outcomes evaluation measures whether a program achieves its intended effect. To identify which outcomes to evaluate, an organization first needs to reach agreement on the intended effect of a program. Why does the organization run this program? How does this program relate to the organization's overarching mission? What are the goals and objectives of the program and, if they're achieved, how would that help the organization as a whole achieve its mission?

Identifying Outcomes: Establishing a Program Theory to Align Program Activities and Goals

Goals and objectives are the vehicles that carry the mission of the organization. Before jumping to creating a survey or other measure of outcomes, organizations should take the time to outline each program's goals and objectives, plus how they relate to the organization's mission. This process is part of outlining a program theory.

All programs are built on a program theory, or theory of change: assumptions about how the program will produce the desired effects on participants or the community. Sometimes, that program theory is explicit: it may be outlined in an administrative handbook, a strategic plan, or other program materials. Often, however, program theories are implicit: they are not written down or recorded anywhere, and although the program might be built on reasonable assumptions about how it will achieve its goals, program staff may have different interpretations of how and why the program works.

Before embarking on outcomes evaluation, program staff should explicitly outline a theory of change for the program to be evaluated. This explicit theory will describe the "program as intended," or the idealized version of the program—how the program is expected to function, processes it is expected to follow, and outcomes those processes are expected to produce. Operationally, it is important for organizations to take the time to get this theory of change right, because if the logic underlying the program is not sound, the program is unlikely to be effective regardless of how well it is run. In terms of evaluation, an explicit theory of change helps identify the outcomes most closely tied to program operations, which are the best candidates for evaluation and monitoring.

Logic Model: A Visual Depiction of Program Theory

Logic models are a useful framework for formalizing a program's theory of change. Logic models visually depict the logic connecting program activities to program goals, or outcomes. Program goals should be aligned with the assessed needs of the community that relate to the organization's mission. The process of developing a logic model can help guide organizations toward selecting the most appropriate outcomes for evaluation. By visually depicting the relationships between activities/outputs and outcomes, a logic model can encapsulate the underlying theory of change. To fully tie them together, a logic model can be coupled with a statement of the underlying theory of change, a hypothesis, or causal explanation of how the work a program does results in the goals it is intended to achieve (e.g., "By providing mentorship and financial assistance to women re-entering the workforce, program XYZ aims to improve self-efficacy, employment status, and earnings for participants, which will in turn increase financial independence for participants and their families.").

Working on an explicit theory of a change and a logic model offers organizations an opportunity for staff to come together and ensure everyone is on the same page about program goals, activities, and the outcomes that matter most. Organizations may find it fruitful to use an interactive process: Draft a theory of change and a logic model, bring it to the group, discuss, update, and repeat until everyone agrees.

Begin to draft a logic model by outlining each program's goals, or intended outcomes. Keep in mind that outcomes are program effects, things that can be observed among the target population of a program or in the social conditions a program is meant to change. Outcomes are not characteristics of the program itself. Measures of program services are better characterized as activities or outputs, not outcomes. For example, receiving a box of food from a food giveaway is an output of the *program*, not an observable condition of the *beneficiary*. In this case, the outcome would be the benefit to the recipient of the food box—e.g., meeting caloric or nutritional needs.

The next step in building a logic model is to work backwards from the desired outcomes for a program to the activities or outputs the program will use to achieve those outcomes. This step offers an opportunity for program staff to revisit a program's core activities or outputs and ask whether they are logically related to the stated goals. If not, the organization may need to fundamentally rethink what the program is trying to achieve (goals) or how it is trying to achieve those goals (activities/outputs). In the example logic model below, a food pantry that provides nonperishable goods may successfully meet its goal of reducing food insecurity but is unlikely to increase consumption of perishable items such as fresh fruits and vegetables; this organization should revisit its goals (e.g., revise them to increase consumption of any fruits or vegetables including canned or dried) or revisit its activities/outputs (e.g., find a way to include fresh fruits and vegetables in food boxes).

The example logic model below is intentionally simplified. Some organizations may find it useful to outline the inputs they provide for a program (e.g., the resources, time, staff, training, etc. that they commit to the program, i.e., their investment). It is also possible to separate activities and outputs (activities are things the program does, whereas outputs are the result of the activity). When focusing on outcomes, however, it may be adequate to combine them.

Example logic model

Program	Activity / Outputs	Intended Outcome (Goal or Objective)	Outcome Measures
Food pantry	Activity: Provide households with a box of nonperishable foods up to once per week Output: Number of households served / boxes distributed per week	Reduce food insecurity	Percentage of households with “low” or “very low” food security
		Increase consumption of fresh fruits and vegetables	Percentage of people consuming 5+ servings of fresh fruits and vegetables daily

When determining which outcomes will be the focus of evaluation and program monitoring, ask whether the candidate outcome is realistic (as a result of program actions). In other words, is there a reasonable connection between program activities and the intended outcome? That connection could be based on published research, anecdotal or personal experience, or logic, but program leaders should agree that it is reasonable.

In general, long-term impacts do not make good outcomes for program monitoring because they cannot be traced back to the program directly. The farther away a goal is from program activities, the more likely that other factors will influence it, making it difficult to disentangle those various forces. Because of this, some organizations may find it useful to separate initial (or proximal) outcomes from intermediate or longer-term outcomes. For example, an initial outcome of a nutrition program may be that participants are knowledgeable of nutrition guidelines, whereas longer-term outcomes are that participants actually change their diets and improve nutrition-related health markers. In general, longer-term outcomes are more difficult to measure and more difficult to attribute directly to a program’s activities rather than to intervening factors; more proximal outcomes may be easier to measure and can act as indicators that a program is performing as expected by the theory of change and on track to achieve longer-term outcomes.

One tool to use when defining program goals for outcomes evaluation is SMART. The SMART criteria for goals are as follows:

- **Specific**: Intended outcomes should be defined in specific terms. Avoid vague goals like “make life better for people in our community”—they do not effectively communicate the value of a program and what specifically it does (as compared to any other effort that might “make life better”). Plus, vague terms and vague measures are too easy to fudge; they do not truly hold a program accountable.
- **Measurable** (or observable): Measurable outcomes can be observed using information that is cost-effective and practical to collect. That is, measurability is not just about whether a goal can be quantified or asked about in a survey, but whether an organization has access to the right people at the right time to measure the outcome. For example, a preschool program could intend to increase high school graduation rates, but not have the capacity to track participants for more than a decade in order to learn whether or not they graduated. Consider in advance how, when, and from whom data will need to be collected, and whether that data is likely to be accessible. At the very least, measurable outcomes should be able to answer questions like how much, how many, or when observers will know a goal has been met.
- **Achievable**: When setting a target for an outcome or program goal, consider whether there is a good match between activities or outputs and the outcome. Is the outcome achievable given the scale of the program? For example, a program that serves a dozen students per year is unlikely to achieve significant change in the unemployment rate among young adults across a state. Take into consideration the program’s current resources and constraints (funding, staffing, available volunteers, program length, etc.) when deciding whether an outcome is achievable.
- **Relevant**: Organizations should focus their efforts on measuring outcomes that are relevant and important. Ask whether there is a good match between the goal or outcome and the organization’s mission. Is this outcome worth the time it takes to measure, track, and report? Is it important to the community? To funders? To participants? Is it altogether consistent with and aligned with other program goals, or will pursuing this goal result in the organization working against itself?
- **Time-bound**: Goals should be mapped to the time in which they are intended to be achieved. For outcomes measurement, it defines the period of time in which you will observe (or fail to observe) a successful outcome. For example, by the end of the program, participants will be able to do X or pass Y assessment.

Once a theory of change and logic model have been created, the next step is to identify outcome measurements--specific, measurable indicators used to monitor program outcomes.

Outcome Measures: Operationalizing Goals

To evaluate outcomes, an organization must be able to measure those outcomes. That is, outcomes evaluation requires that program goals be transformed into measurable indicators, or operationalized. For every outcome of interest, organizations should define an outcome measure, or indicator, that will be used to collect and track data. The process of matching outcomes and measures is called operationalization.

Operationalized goals are those that have been defined in such a way that they can be measured, reported, and tracked over time. To operationalize goals, consider both the definition of the goal or outcome and the data available to measure it. For ongoing outcome monitoring, organizations should strive to pick outcome measures that are practical to collect routinely, are informative, and can drive action for program management.

In choosing which outcome to operationalize and how, consider a measurement's ability to distinguish a program's effect on an outcome net of alternative causes, the proximity of the measure to program activities, multiple dimensions of outcomes, and available data. Each of these elements is discussed in more detail below.

Net Effects

It can be difficult to sift through the many factors that influence an outcome. Here are some basic tips for designing an evaluation that is better able to measure a program's effects net of alternative causes (adapted from Rossi, Lipsey, and Freeman 2004):

- Measure change, not levels. Measuring the level of an outcome at a program's end is less informative than measuring change in that outcome. Change in an outcome that occurs over the course of participation speaks more directly to the effects of a program's activities. For program monitoring, organizations should strive to measure change in an outcome by measuring outcome levels at the end of the program and comparing them to some indicator of outcome levels at the beginning of the program (e.g., a pre-test). This makes it possible not only to report the outcome level at program completion but also to calculate the change in outcome level, showing the amount of benefit that might have been achieved by the program.
- Use a comparison group. Measure change for participants as compared to change among similar non-participants during the same time period. It is possible that participants could change outcome levels during a program due to factors outside

the program's control, such as aging's effect on cognitive development, a global recession, or a tech boom's effects on employment rates. The question is whether program participants saw change above and beyond what would have occurred if they had not participated in the program. Knowing that requires counterfactual inference to estimate what the outcome change might have been for participants if they had not been part of a program; it is impossible to know for sure since they were in fact part of the program. Instead, try to triangulate with available data and knowledge of local conditions, or work with an expert evaluator to design methods that can help isolate a program's effect by comparison to a control group.

Proximity to Activities

In choosing outcome measures, organizations should focus on proximal outcomes. An organization might think about outcomes on a continuum, from those most immediately affected by the program to long-term but somewhat nebulous impacts that are important but cannot easily be traced to a single program's effects. It is often easier and more relevant to measure those outcomes most immediate, or proximal, to a program's outputs even though they may seem more limited. If a program's theory of change is correct, then successfully improving proximal outcomes should cascade down the continuum of outcomes to affect more distant ones. In other words, proximal outcomes may not be the most important in terms of broad social impact, but they are most directly affected by the program. Evaluating these outcomes can be more informative than making assumptions about how to isolate program effects on longer-term, more distant outcomes. If a program fails to improve even the most proximal outcomes, it is unlikely the program is having much effect on more distant outcomes.

In general, it is also more practical and reliable to measure outcomes that are closer to program activities, rather than long-term outcomes that must be measured in the distant future, on a larger geographic scale, or among a broader population. For example, a 10-week mentorship program aims to increase middle-school-aged girls' feelings of self-efficacy. That outcome is proximal to program activities and can be measured with pre- and post-tests. Down the continuum, improved self-efficacy could lead to girls' increased participation and engagement in school, improved academic achievement, and increased employment and earnings—those more distant outcomes and impacts could be measured in the future, but would be harder to trace back to the program's effects given the many forces that could intervene in participants' lives over the ensuing years.

Multiple Dimensions

Many outcomes are multidimensional and could be measured in a variety of ways. For example, consider "increased employment" as an outcome. It could be measured as:

- Level of employment (unemployed, temporary, part-time, full-time)
- Number of hours worked during a given period
- Type of employment (industry or occupation)
- Stability of employment (time at current job)
- Earnings during a given period

Consider whether it makes sense to evaluate multiple measures of an outcome or to focus on some that are most important. Regardless, it can be useful to flesh out the full range of possible measures of an outcome in order to better understand the possible ways in which the outcome could be operationalized.

Available Data

Basic sources of outcomes data include observations, administrative records, questionnaires or interviews, standardized tests or assessments, and physical measurements. While many organizations develop their own tools for collecting outcomes measurements, others adopt existing tools or measures. In either case, operationalizing an outcome requires applying standard criteria or processes to data: Often, those criteria refer to conventional or standard measures—for example, definitions of employment used by federal agencies or diagnostic criteria used in medicine. In other cases, criteria may be established by an organization to suit their needs and the constraints of available data.

Existing Measures

In picking a way to operationalize an outcome of interest, consider whether there is already a standard definition or tool for measuring that outcome. Review prior research on similar programs and look for the outcomes they measured and any validated tools or common measurements they used. Conducting a literature review like this is a way to find relevant outcomes that might otherwise be overlooked and to see how outcomes have been measured—including, sometimes, finding available, validated assessments or evaluation tools that can be adopted.

There are advantages and disadvantages to adopting existing measures for outcomes. On the one hand, using existing, widely used measures makes it possible to compare program outcomes to other programs, and to communicate to funders or policymakers who may be familiar with standard measures. On the other hand, existing measures may be a poor fit with a program's goals or impractical to administer. Organizations should avoid adopting an existing measure just because it exists and seems similar to an outcome of interest. Be sure measures faithfully represent the outcomes important to the organization.

In some cases, multiple existing measures may fit the intended program goals. In that case, take into account how easy it will be to administer an assessment or collect data (e.g., how long it takes, whether participants can self-administer, or whether it requires program staff to assess them individually).

Another potential source of existing measures is administrative data. These data, collected in the course of doing business, can sometimes perform as outcome measures too. Many organizations have administrative data collected for another purpose that would also be relevant for measuring an outcome. For example, an organization might collect information in case notes and just need to formally record it in a database or spreadsheet so it can be more readily measured and summarized (e.g., employment outcomes for participants in a mentoring program).

Creating New Measurements

Sometimes organizations may find it necessary to design a new measurement. Ideally, this should be done in consultation with experts, but that is not always possible. Frameworks exist to guide organizations in developing their own outcome measurements. For example, sample indicators (generic but ready to be adapted to specific programs) are available from the Urban Institute's taxonomy of outcomes and common outcome framework.[1]

When creating new outcome measurements, consider a measurement's reliability, validity, and sensitivity:

Reliability: A reliable measurement produces the same results again and again whereas an unreliable measurement is prone to measurement error. For example, a bathroom scale is reliable if it reports the same weight every time you step on it (assuming your true weight remains constant). Reliability means that from person to person, season to season, you can be sure this indicator is measuring the same thing in the same way.

More subjective survey questions like "Did this program make your life better?" may have lower reliability (they could depend on someone's mood, for example). In general, recall-based measures that ask people to remember something from the past are less reliable than measures that do not depend on memory. Reliability can also be affected by differences in the way an evaluation is administered (who does an interview, where participants fill out a questionnaire), so evaluations should be administered the same way every time. Various other factors, which cannot be controlled by program staff, can also affect reliability (e.g., participants' interest in the survey). Consider possible sources of bias (e.g., whether participants will feel comfortable being critical of a program if the survey is administered by a coach or mentor). To avoid such problems, strive to identify measurements that are less subject to such influence.

[1] Available online at https://www.urban.org/sites/default/files/2015/04/10/taxonomy_of_outcomes.pdf and https://www.urban.org/research/publication/building-common-outcome-framework-measure-nonprofit-performance/view/full_report

A measurement’s reliability can be measured. One way to check the reliability of a measurement is to administer it two or more times during a period and under conditions where the underlying outcome should not change, then compare the results to see whether the measurement also did not change. This type of check is tough with surveys because people remember their prior response, and it influences their response on subsequent surveys. Alternatively, reliability can be checked by using multiple measurements of the same outcome (e.g., ask the question multiple times in different ways) to look for internal consistency. Organizations that are designing their own outcome measurements can work formally with a statistician to test reliability; at the very least, organizations should think critically about possible factors that could influence the reliability of a proposed outcome measurement.

Validity: Valid measurements truly measure what they are meant to measure. Validity can be harder to achieve than one might expect. For example, it is notoriously difficult to develop a valid measurement of criminal activity: although arrest records and incarceration seem like good candidates, neither is an entirely valid measure of criminal activity. Arrest depends on police observing criminal activity and choosing to respond, and incarceration depends on decisions made by a judge.

In surveys, respondents’ diverse interpretations of vague or unclear questions can jeopardize validity. For example, if a survey asked participants, “Did this program make you feel better?” some might interpret that to mean physically better, while others interpret it to mean emotionally better. Likewise, survey questions may lose validity if respondents do not have the knowledge they need to answer. For example, if a survey asked parents whether children had increased self esteem because of a program, could parents provide a valid assessment?

To overcome concerns about validity, an organization can use multiple measures of an outcome of interest. Consider different dimensions or different data sources to triangulate the outcome of interest; when all move in the same direction, that is evidence that the underlying outcome is changing. Organizations can also address the validity of measures by conferring with key stakeholders about whether they believe a measure is valid and finding one that funders, regulators, and other stakeholders agree on.

Of course, validity might also be a concern when adopting existing measures. Organizations should be careful to choose existing measures that are valid for the outcome they wish to measure. For example, if a program aims to increase academic achievement, an IQ test, which is intended as an ability test rather than an achievement test, might not be a valid measure.

Sensitivity: Sensitive measures are able to detect changes in the outcome at a scale that is relevant to likely program effects. Using a very broad assessment to measure change because of a program with very targeted interventions may not be sensitive enough.

Adopting measures that are not sensitive enough is like trying to measure the width of a microbe with a ruler.

When considering whether a measure is sensitive enough, organizations should consider whether the indicator is likely to respond directly to possible program effects. Avoid using existing data that is too broad to be useful. For example, there is readily available data about unemployment rates for metropolitan areas, but a program that serves only a small fraction of people in that geographic area is unlikely to affect those rates.

Program completion rates can also affect sensitivity. Including outcome measures for non-completers (i.e., people who did not truly participate in the program) may not yield a good measure of the program's effect on those who *did* complete it. Therefore, organizations may choose to exclude people who drop out of a program from outcome measures. Instead, they may measure dropout rates as a separate indicator of performance measure. By excluding people who do not complete the program from outcomes measures, organizations can separate measures of service utilization rates from outcome measures, then address them both as needed.

Cautions: Perverse Incentives and Corrupted Indicators

Outcome measures are meant to influence action, but organizations should exercise caution in choosing outcome measures that motivate productive action. Any outcome used for accountability, program monitoring, and performance review is likely to receive attention—not only from managers, funders, and policymakers, but also from program staff. If staff know an outcome is being given attention and used to drive decisions about a program, they will rationally try to maximize that outcome. It is important to choose outcomes and indicators with care so that maximizing those outcomes truly does improve program performance, not distort it. Two potential pitfalls of outcome measures deserve special mention: perverse incentives and corrupted indicators.

Perverse incentives arise when outcome measures are misaligned with program goals. Inappropriate or incomplete outcome measures can create misaligned incentives for staff. As staff try to maximize the outcome measure, they engage in actions that either do not affect program goals or may even work against them.

In one often cited example, a state's goal was to increase placements for foster children. To measure progress toward this goal, administrators measured the number of new foster homes licensed. In response, program staff quickly licensed new homes, but without attention to the quality of those homes (i.e., foster parents' skills and abilities). As a result, the number of foster homes increased (and the outcome measure suggested program success), but quality placements for children did not (Affholter 1994).

Corrupted indicators occur when it is possible to move the needle on an outcome measure without actually taking programmatic action (e.g., providing more or better services). Program staff naturally want to make their programs look good, and when outcome measures are not tightly tied to program performance, they may find ways to inflate outcome *measures* without actually improving program outcomes. Ambiguity in the definitions of outcomes and indicators make it more likely they will be corrupted. Therefore, it is important to choose outcomes and measures that are clearly defined. Working with an external evaluator can help ensure that data is collected consistently and faithfully.

A Note on Customer Satisfaction

On face, customer satisfaction is a promising choice as an outcome measure: It is closely tied to program services and measured only among people who participate in the program. Logically, customer satisfaction is related to whether participants feel like they benefited from a program, and satisfied customers could be considered a positive outcome in itself.

However, there are some things to consider before adopting customer satisfaction as the sole or primary outcome measure for a program:

1. Satisfaction measures should be specific. Common measures of customer satisfaction, like the net promoter score,[2] are too general to tie to specific program benefits. Instead, ask participants for feedback about specific benefits they may have experienced as a result of a program. For example, ask, “Has the food you received from the food pantry been helpful to you in maintaining your health and nutrition?”
2. Program participants might not be equipped to recognize, assess, acknowledge, or even be aware of program benefits. For example, participants may not have the information or knowledge they need to accurately answer a question such as, “Has the food you received from the food pantry helped you maintain adequate levels of iron and vitamin D?”
3. Questions about satisfaction are prone to bias, especially acquiescence and social desirability bias. Participants might over-report benefits for a variety of reasons, including documented forms of bias common in survey response. Acquiescence or agreement bias refers to the tendency of respondents to agree with statements or answer questions in the affirmative. Social desirability bias is the tendency of respondents to over-report what they believe surveyors want to hear, or what they believe would be considered desirable behaviors or viewpoints. Careful phrasing of questions can help reduce bias, but bias may always be some concern in customer satisfaction survey evaluations.

[2] Net promoter score is frequently used in market research and customer satisfaction surveys. The score is usually based on responses to a single survey question, which asks customers how likely they would be to recommend a program to a friend.

4. When possible, measure an outcome directly. If customer satisfaction is the intended outcome of a program, then a measure of customer satisfaction is a direct measure. But more often, outcomes are tied to mission or program goals such as reducing food insecurity, increasing financial self-sufficiency, building self-efficacy, or the like. In that case, customer satisfaction may be an indirect measure of outcomes. Better, though, to measure the outcome directly if possible.

Questions about customer (or client or volunteer) satisfaction may be useful for formative evaluation. They provide immediate feedback about how things are going and can be used to course correct as needed. Focusing on customer satisfaction can also help improve processes to maximize participant retention and program completion.

Conducting Outcomes Evaluation

Once goals and outcome measures have been defined, organizations must embark on the process of evaluating or monitoring those outcomes. Outcomes evaluation proceeds in three steps: collecting data, examining the results, and reporting the results. As described above, in ongoing outcomes monitoring for program improvement, these steps are cyclical. Once results are reported, they should drive action (e.g., program modifications), then a new cycle of outcome measures should be collected, examined, and reported. This section discusses each step in turn.

Collect Data

Data collection is one of the most vital stages of outcome measurement. With the right evaluating questions, data can be exceptionally powerful in not only measuring the outcome of the program, but also telling an appealing story about the most important parts of the program. In collecting data, however, it is necessary to think carefully about how different data collection strategies might affect data quality as well as the staff effort required to collect and manage the data. A few tips are offered here. For additional guidance in collecting data, consult the resources listed at the end of this report.

- Consider when, how, and by whom data should be collected—and how different data collection strategies might affect results or response rates. This can vary depending on the population an organization (or program) serves. Often, there are tradeoffs in the ease with which data can be collected and the comprehensiveness of that data. Online surveys are easy to administer and collect. But do all participants use email? Do they have internet access? Is technology accessible for them? If not (or if some do not), avoid online surveys. Do most participants read English easily, or do written evaluation tools need to be printed in additional languages or read out loud?

- Get feedback on any newly created outcome measures. Consult with stakeholders on whether they consider the measure valid and useful. When creating a new survey questionnaire, conduct a pilot test with participants. Do they understand the questions? Are they interpreting questions the same way program staff do? Do they find the questionnaire cumbersome or time consuming?
- Communicate to participants the purpose of the evaluation. Make it clear that the evaluation is meant to evaluate the program and the organization, not the individual. Some participants may be more willing to respond when they understand that the information will be used to improve the program, not to judge them.
- Avoid collecting data that will not be reported or used. This “nice to have” data collection puts an undue burden on both participants and staff. Instead, focus time and effort on collecting high quality, complete data that will be useful for measuring the outcomes of interest and driving decision-making.
- Collect relevant data about participant demographics or important characteristics. This type of information is not a direct measure of outcomes, but it can be used alongside outcome measures to analyze whether there are certain groups for whom the program works or does not work.
- Transform data as needed to make it more useful. For example, data can be both qualitative and quantitative. However, using quantitative measures such as “percent of” can help represent a qualitative outcome in a more meaningful and measurable manner.

Examine Results

Once data has been collected, it is time to do an analysis. The data analysis process consists of several steps, such as cleaning and visualizing, but it is important to always keep in mind the goals and objectives of the program in each step of the analysis process.

Data analysis should begin with a simple summary of the data collected, then compare outcomes across groups, and finally consider the context around the outcomes data.

- Summarize data with simple statistics such as mean (average) or mode (most common value). It might also be useful to calculate the percentage of values above or below a significant threshold. Graphical summaries, or data visualizations, can help reveal patterns.
- Set a benchmark or threshold for success and compare results to that threshold. At the participant level, a program might count the number of participants who started below the threshold of success and surpassed it after participating in a program. For example, a test preparation program could set a score that constitutes success, then measure the proportion of participants who scored below that level on a pretest and above that level on a post-test. To benchmark, set success thresholds with reference to similar programs and the outcomes expected based on those

programs' successes or to past program performance. For more guidance on benchmarking, see Keehley et al. (1996).

- Compare within groups to measure change. As described above, outcomes monitoring should not only look at outcome levels after participants complete a program. Ideally, those levels should be compared to participants' outcome levels from before they started the program (pre-test/post-test style), or to outcomes for a control group of people who are similar in important ways to participants but did not take part in the program. Either strategy helps indicate change in the outcome.
- Compare outcomes across groups, over time as a program changes, across sites if applicable, or along other dimensions in order to identify whether the program is performing similarly for all groups and under various circumstances.
- Consider the context around outcomes data. Outcomes are sensitive to factors outside of program performance, and these factors need to be taken into account and/or reported along with outcomes to aid in interpretation. For example, report participants' exposure to other influences, changes in regional conditions, or other relevant factors that might be affecting participant outcomes. Even if done informally, a narrative description can help contextualize results for interpretation. Additional factors to consider as context include the following:
    - Changes in client mix
    - Demographic or economic trends
    - Change in community context (e.g., available referral services)
    - Relevant program process or utilization trends (e.g., dropout rate)

Report Results

Finally, report results internally and/or externally to stakeholders. Consider who needs to see the results and how the results will be used and integrated into program management. It is not enough to conduct regular surveys or track outcomes if that information is not used to monitor performance and improve program quality. Organizations can find ways to take time to discuss outcomes reports and reviews in monthly staff meetings, for example, or dedicate a part of the annual report to program outcomes like they would financial reports.

Section 2: Tools and Templates

A. Logic Model Template

Logic models can be as simple or complex as organizational needs demand. The template on the following page is an intentionally simplified version of a logic model. It can be used to begin conversations within an organization about program outcomes for evaluation. The elements of the model are briefly described below; for additional detail, refer to the "Brief Guide to Outcomes Evaluation" section on "Identifying Outcomes."

Although inputs logically come first, when filling out a logic model, it can be more effective to begin from intended outcomes (i.e., the goals of a program). Then, work backwards to fill in the outputs necessary to achieve those outcomes, the activities necessary to produce those outputs, and the inputs necessary to enable those activities. Finally, return to the intended outcomes and begin to develop outcome measures, or indicators, to evaluate progress toward achieving them.

Inputs are the resources an organization invests into a program, such as staffing, office space, and materials. Inputs enable activities.

Activities are the service that programs (and their staff) do, such as making home visits, counseling participants, tutoring students, or packing food boxes. Activities produce outputs.

Outputs are the product a program produces. Often, outputs are countable. For example, outputs might include number of home visits conducted, number of counseling sessions held, number of students tutored, or pounds of food delivered. Together, activities and outputs are intended to achieve program goals, or outcomes.

Intended Outcomes are the goals of a program. When stating intended outcomes, apply the SMART criteria: goals should be specific, measurable, achievable, relevant, and time-bound.

Outcome Measures are the indicators used to measure outcomes, or operationalized goals. Consult the "Brief Guide to Outcomes Evaluation" and the Operationalizing Outcomes Checklist for further guidance on selecting outcome measures.

Inputs	Activities	Outputs	Intended Outcomes	Outcome Measures

B. Operationalizing Outcomes Checklist

Effective outcome measures are aligned with program goals. Both the goals and the measures should be realistic. Outcome measures should be useful and informative for guiding program improvement, and they should also be practical to collect. Apply the checklist below to ensure outcomes and outcome measures meet these criteria.

□ **Program goals and objectives are clear and well defined.**

Avoid vague statements. Be clear, specific, and concrete. Programs will be held accountable to these goals, so they must be framed in a way that makes it clear whether or not they have been achieved.

□ **Program goals and objectives are feasible.**

Outcomes should be realistic given the size and scope of the program. Is there reason to believe that the specific program activities and outputs will directly lead to attaining the intended outcomes? Avoid unrealistically lofty goals beyond a program's influence. Consider community context and what is achievable given local constraints.

□ **The program's theory of change is plausible.**

Check the assumptions behind the entire program theory, testing each causal link from program inputs through intended outcomes. What is the evidence base for believing this program will work? What need is the program intended to address, and is there evidence that can demonstrate the program's activities and outputs intervene at the correct point to meet that need and produce the intended outcomes?

□ **The program's intended beneficiaries are clearly defined.**

Specify who the program is meant to affect. Who are the participants? Outline any eligibility criteria (e.g., age, income, or geographic restrictions). Do intended beneficiaries include groups beyond participants? If so, define those groups.

□ **Outcome measures align with program goals, and they are measured among intended beneficiaries.**

Outcomes should be observable among the intended beneficiaries of a program. Ensure that they are well aligned with program goals to avoid creating perverse incentives or corruptible indicators. For each outcome and outcome measure, ask whether it is possible to improve that measure without making genuine progress on the underlying goal.

□ **Outcome measures are practical to collect.**

Consider how much time and effort will be required to collect outcome measures. Who will be responsible for collecting these measures? How frequently will measures be collected and updated?

□ **Outcome measures are valid and reliable.**

Outcome measures should consistently measure things in the same way (reliability) and truly measure the outcome they are meant to measure (validity). Have outcome measures been formally tested for validity and reliability? If not, are they nevertheless accepted by stakeholders as valid and reliable?

□ **Outcome measures separate the program's net effects from alternative causes.**

As much as possible, outcome measures should try to isolate the net effects of a program from alternative causes of change. Focus on measures of more proximal outcomes and measures of change as opposed to static snapshots of outcome levels. Contextualize outcomes as needed to aid interpretation.

C. Questionnaire Design Checklist

If existing data and evaluation tools are not a good fit for a program's intended outcomes, an organization may need to design new data collection tools. Typically, those take the form of interview questions or survey questionnaires. This checklist offers tips for designing effective survey questionnaires. Additional guidance can be found in the resources listed at the end of this report.

□ **The questionnaire begins with an introduction that sets the context for the evaluation.**

Use the introduction to motivate respondents to complete the questionnaire. Explain how the information they provide will be used. Emphasize that the questionnaire is meant to evaluate the program, not the individuals taking the questionnaire.

□ **The first question is interesting, related to the program, and easy to answer.**

Avoid asking dull fact-based or sensitive questions early in a questionnaire. Instead, begin with questions that are engaging and easy to answer. This strategy helps build rapport with respondents and encourages them to complete the survey.

□ **The questionnaire includes questions about program outcomes.**

If the questionnaire is intended to measure outcomes, it must include questions that directly relate to intended outcomes. Questions should be reviewed to ensure they are reliable and valid measures of intended outcomes.

□ **The questionnaire includes questions about program processes or implementation.**

If the questionnaire is intended to evaluate a program's processes or implementation, it must include questions that directly relate to respondents' perceptions of process and implementation.

□ **The questionnaire includes questions about participant demographics or other relevant characteristics.**

In addition to collecting information about program performance and outcomes, be sure to include questions that make it possible to compare outcomes across relevant groups. These questions can also be used to compare the mix of participants in one period to another or to compare to the local population. Information should be kept confidential and reported in an aggregated manner.

□ **The questionnaire contains relatively few open-ended questions, and most are located toward the end of the survey.**

Open-ended questions typically take longer for respondents to answer, and they also require more effort for staff to analyze. Locating them toward the end of the survey can help ensure respondents answer other questions first. Open-ended questions can be useful for collecting rich, qualitative feedback, but they should be used sparingly and with attention to the time it will take to process on the back-end. Open-ended questions are also good for questions about amounts: it is better to ask respondents to enter a number than to fit their response into a series of ranges.

□ **Closed-ended questions have answer choices that are mutually exclusive and exhaustive.**

Mutually exclusive answer choices do not overlap; a respondent is able to clearly distinguish between different options. Exhaustive answer choices cover the entire range of possible responses to a question. Pay special attention to scales, which need to cover the entire range of possible sentiment: typically, scales should use five or seven points.

□ **Question wording is clear and simple.**

In general, questions should use simple, familiar words and avoid technical jargon. Choose specific, concrete wording over words with ambiguous meanings. Avoid single or double negations.

□ **Questions have been reviewed to reduce the risk of bias (e.g., from leading questions, double-barreled questions, and questions prone to acquiescence bias or social desirability bias).**

Avoid leading questions, which suggest certain answers over others. Avoid double-barreled questions, which ask about more than one thing at once. Avoid question phrasing that asks respondents to agree/disagree or respond yes/no (acquiescence bias makes respondents more likely to agree or respond yes). Carefully phrase questions about sensitive topics where respondents might be subject to social desirability bias (i.e., over-reporting what they believe is a desirable behavior or attitude).

□ **The questionnaire has been pilot tested to ensure it is easy to administer, understood as intended by respondents, and takes an acceptable amount of time to complete.**

Before adopting a new questionnaire as a program evaluation tool, pilot test it. Ideally, the pilot test should be with program participants (i.e., people who are similar to those who will respond to the questionnaire for evaluation purposes). Ask for feedback about the questionnaire's clarity. Measure the amount of time it takes to complete, generally aiming for no more than five to ten minutes. Consider any difficulties that arise in administering the survey.

D. Outcomes Tracking Example

The outcomes tracking example below offers a simple template for a dashboard to monitor outcomes across multiple programs. Organizations could maintain a similar dashboard in a shared spreadsheet that program staff can update each period and managers can review regularly.

This example tracks outcomes in absolute numbers and percentages. Comparing both numbers and percentages can help contextualize large changes in percentages that can occur with very small numbers of participants, and absolute numbers can also put into perspective the number of individuals affected by a program.

The value in the change column compares the previous period value to the current period value. Change values can be color-coded to quickly flag values that are increasing or decreasing. Likewise, the previous period values and current values can be color-coded to highlight those that are above or below a benchmark value.

Program	Intended Outcome	Outcome Measure	Benchmark	Previous Period Value	Current Value	Change
Food Pantry	Decrease food insecurity	Number and percent of participants who increased food security to "high or marginal" on the USDA 6-point scale	120	125	129	+4
			67%	70%	72%	+2%
	Reduce incidence of iron deficiency anemia in children	Number and percentage of children with iron deficiency anemia who increase iron to normal levels	17	16	18	+2
			85%	80%	90%	+10%
Youth Mentoring	Re-engage disconnected youth in school or work	Number and percent of participants (ages 16 to 24) who gained or maintained employment or school enrollment during the previous period	600	480	536	+56
			75%	60%	67%	+7%
	Improve social relationships	Number and percent of youth who indicate improved relationships compared to before program start	720	704	672	-32
			90%	88%	84%	-4%

References and Additional Resources

Evaluation (General)

Affholter, D.P. 1994. “Outcome Monitoring.” In J.S. Wholey, H.P. Hatry, and K.E. Newcomer (eds.), *Handbook for Practical Program Evaluation* (pp. 96-118). San Francisco: Jossey-Bass.

Fine, Allison H., Colette E. Thayer, and Anne Coghlan. 2000. “Program Evaluation Practice in the Nonprofit Sector.” No*nprofit Management and Leadership*, 10: 331-339

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach*. Thousand Oaks, CA: Sage Publications.

Thayer, Colette E., and Allison H. Fine. 2001. “Evaluation and outcome measurement in the non-profit sector: stakeholder participation.” *Evaluation and Program Planning*, 24(1): 103-108.

Survey and Questionnaire Design

Babbie, Earl R. 1990. *Survey Research Methods* (Second Edition). Cengage Learning.

Dillman, Don A., Jolene D. Smyth, Leah Melani Christian. 2014. “The Fundamentals of Writing Questions” (Chapter 4) and “How to Write Open- and Closed-Ended Questions” (Chapter 5) in *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Wiley.

Fowler, Floyd J. Jr. 1995. *Improving Survey Questions: Design and Evaluation*. SAGE Publications, Applied Social Research Methods Series Volume 38.

Iarossi, Giuseppe. 2006. *The Power of Survey Design: A User’s Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*. World Bank Publications.

Krosnick, Jon A. and Stanley Presser. 2009. “Question and Questionnaire Design.” *Handbook of Survey Research* (Second Edition) James D. Wright and Peter V. Marsden, Eds. San Diego, CA: Elsevier.

Benchmarking

Keehley, P., S. Medlin, S. MacBride, and L. Longmire. 1996. *Benchmarking for Best Practices in the Public Sector: Achieving Performance Breakthroughs in Federal, State, and Local Agencies.* San Francisco: Jossey-Bass.