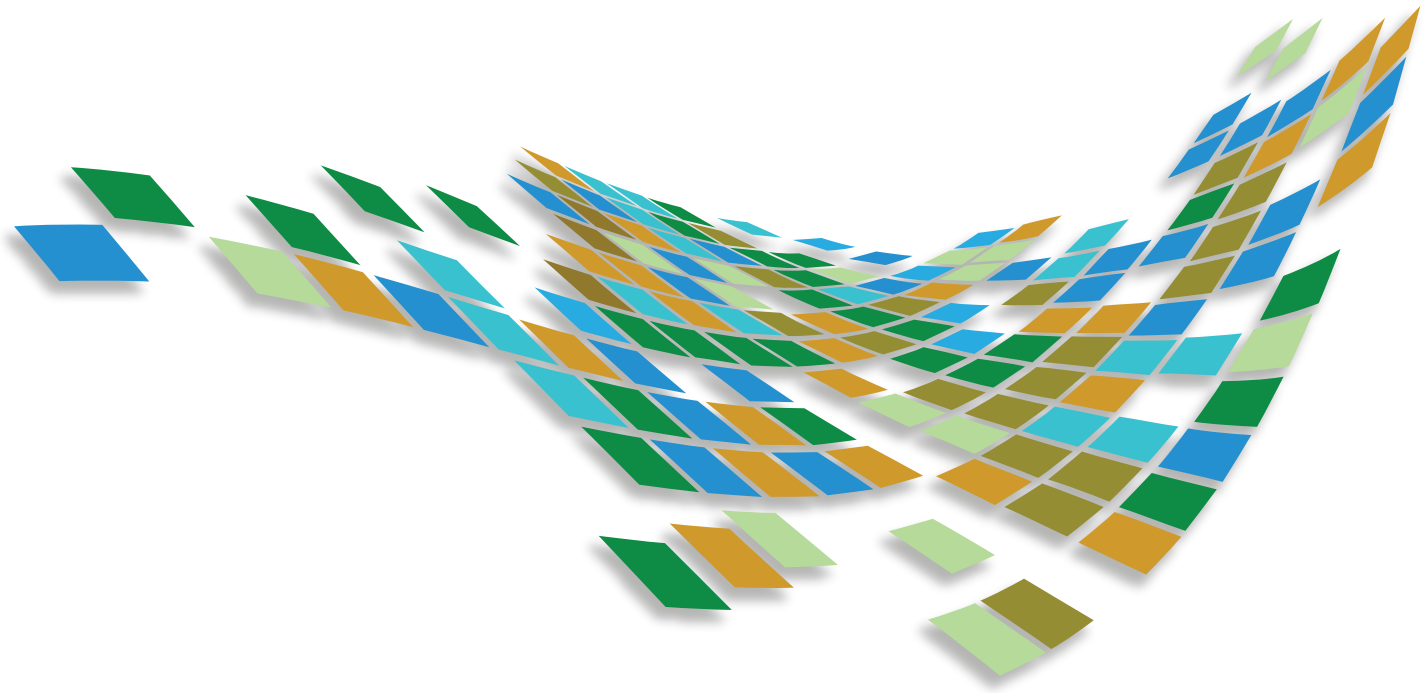
Building Smarter Data for Evaluating Business Assistance Programs

A Guide for Practitioners



U.S. Small Business Administration

May 2017

ABOUT THIS GUIDE

The U.S. Small Business Administration (SBA) is publishing this guide on behalf of the Evaluating Business Technical Assistance Program (EBTAP) working group. The Council of Economic Advisers and Office of Management and Budget (OMB) formed the working group in 2013 to explore opportunities and overcome challenges in using administrative data to evaluate the impact of business assistance programs. The Council of Economic Advisers and OMB chair the working group. Members include representatives of federal programs within the U.S. Department of Agriculture (USDA), U.S. Small Business Administration (SBA), U.S. Department of Labor (DOL), U.S. Department of Commerce (DOC), and statistical agencies including the U.S. Census Bureau, DOC and the Economic Research Service, USDA. This guide was prepared by a subcommittee of the EBTAP working group, chaired by **Giuseppe Gramigna**, SBA chief economist, and **Susan Helper**, Economics and Statistics Administration, chief economist through July 2015.

Collaborators on this guide included program practitioners, economists, statisticians, and performance and policy staff from several federal agencies. The federal agencies and members of the EBTAP subcommittee that authored this guide and provided technical content are listed in alphabetical order.

U.S. Department of Agriculture	**Cynthia Nickerson**, Economic Research Service (Chair of the EBTAP working group and senior economist on the Council of Economic Advisers, July 2013–May 2015)
	Timothy Park, Economic Research Service (Chair of the EBTAP working group and senior economist on the Council of Economic Advisers, June 2015–June 2016)
	John Pender, Economic Research Service
	Tim Wojan, Economic Research Service
U.S. Department of Commerce	**J. David Brown**, Center for Economic Studies, Census Bureau
	Christine Heflin, Office of Performance, Evaluation, and Risk Management
	Susan Helper, Office of the Chief Economist, Economics and Statistics Administration (Co-chair of the EBTAP subcommittee through July 2015)
	Cassandra Ingram, Office of the Chief Economist, Economics and Statistics Administration
	C.J. Krizan, Center for Economic Studies, Census Bureau
	Paul Marck, Quality Program Staff, Census Bureau
	Samantha Schasberger, Economic Development Administration (through August 2015)
	Kenneth P. Voytek, Manufacturing Extension Partnership, National Institute of Standards and Technology
U.S. Department of Labor	**Jonathan Simonetta** (through July 2016)
U.S. Small Business Administration	**Giuseppe Gramigna** (Co-chair of the EBTAP subcommittee)

ACKNOWLEDGMENTS

The EBTAP subcommittee would like to thank the following people who provided reviews, comments, and other valuable contributions to this guide. Any errors in the guide are solely the responsibility of the authors.

Katharine Abraham, Director, Maryland Center for Economics and Policy, University of Maryland

Sarah Ali, formerly Rural Development, U.S. Department of Agriculture

Jon Baron, Laura and John Arnold Foundation

Melissa Creech, Census Bureau, U.S. Department of Commerce

Byron Crenshaw, Census Bureau, U.S. Department of Commerce

Debra Delay, International Trade Administration, U.S. Department of Commerce

Catherine Dibble, Aiki Labs

John Eltinge, Bureau of Labor Statistics, U.S. Department of Labor

Harris Eppsteiner, Council of Economic Advisers

Lucia Foster, Census Bureau, U.S. Department of Commerce

Christine Geter, Census Bureau, U.S. Department of Commerce

Scott Gibbons, Employment and Training Administration, U.S. Department of Labor

Elzie Golden, Census Bureau, U.S. Department of Commerce

Harry Hatry, Urban Institute

Ron Jarmin, Census Bureau, U.S. Department of Commerce

Christa Jones, Census Bureau, U.S. Department of Commerce

Shawn Klimek, Census Bureau, U.S. Department of Commerce

Harry Knight, Office of Performance, Evaluation, and Risk Management, U.S. Department of Commerce

Sam Le, Office of General Counsel, U.S. Small Business Administration

Heather Madray, Census Bureau, U.S. Department of Commerce

Rochelle Martinez, Statistical and Science Policy Branch, Office of Management and Budget

Demetra Nightingale, Chief Evaluation Office, U.S. Department of Labor

Bruce Purdy, Office of Women's Business Ownership, U.S. Small Business Administration

James Rivera, Office of Disaster Assistance, U.S. Small Business Administration

Barry Robinson, Office of the General Counsel, Economic Affairs, U.S. Department of Commerce

Dan Rosenbaum, formerly Economic Policy Division, Office of Management and Budget

Alexis Solano, Rural Development, U.S. Department of Agriculture

Janet Sweeney, Census Bureau, U.S. Department of Commerce

Andrea Taverna, Council of Economic Advisers

Hampton Wilson, Census Bureau, U.S. Department of Commerce

CONTENTS

EXECUTIVE SUMMARY

This guide is designed for policy makers and program managers who make critical decisions concerning business assistance program design, delivery, and information collection. The guide identifies critical data and best practices that support the use and improvement of administrative data and other existing data sources for rigorous impact evaluations.[1]

Evidence building is increasingly important to federal programs, including government business assistance programs that are the focus of this guide. Agencies that provide business assistance compete for scarce resources. Managers need findings from well-designed evaluations to support assertions about program impact. Further, they need to demonstrate that a program provides a better return-on-investment than alternative approaches. As decision makers look for new and better ways to measure progress and program impact, reliable evidence will be needed. Evaluators may be able to develop this evidence at lower cost using data from existing records.

Surveys are often the default approach for collecting data for evaluations. However, they are expensive and a burden to the business community. Further, the approach often does not collect all the data necessary to measure whether the impact could have occurred without assistance.[2] Alternatively, program data combined with other sources of government or private data can provide the necessary information on both assisted businesses and similar businesses that did not receive assistance. The combined data may permit the evaluator to expand the scope of the analysis to include more years and consider additional research questions. In essence, this approach can reduce evaluation costs and increase quality, and does not add to survey burden on businesses and taxpayers.[3]

When programs are launched, the information and protocols that facilitate linking program data on assisted businesses to other data sources (e.g., census, survey, and administrative data from other programs) should be anticipated. Lack of this advanced planning has been one of the greatest impediments to using existing datasets effectively for impact evaluation. This guide addresses this recurring problem by providing 18 data collection and design practices needed to take advantage of external data sources. In summary, the guide recommends that program managers:

- Identify administrative data needed for both program service delivery and eventual impact evaluation at the beginning of a program or pilot. This prevents the need for expensive, after-the-fact additional data collection. Similarly, for existing programs, early assessments of the quality and availability of administrative data and actions needed to remedy data deficiencies can increase the value of administrative data for eventual evaluation.
- Solicit the input of evaluation experts early in the process of developing data plans. These experts can help identify the best methodology for measuring program impact and the most cost-effective ways to assemble the data necessary to support high-quality evaluations.
- Explore whether linking program administrative data on assisted businesses to secondary data of government agencies or commercial sources is a viable option. This linkage, which requires sufficient unique applicant/participant-level identifying information in all datasets, can increase evaluation quality and reduce the need to conduct post-service surveys. Ensure that sufficient security procedures are in place to protect the data and their confidentiality.
- Engage departmental attorneys and policy officers, including privacy, confidentiality, and security officers, early in the process of developing data plans. This can avoid problems and delays that arise when data collection and sharing for evaluation purposes are treated as separate or after-the-fact considerations.

[1] Administrative data are "Information [records] kept by business establishments, institutions, and governments primarily for their own purposes in running their business or program." "Questions and Answers When Designing Surveys for Information Collections," Office of Management and Budget, January 20, 2006.

[2] When only businesses that received assistance are surveyed, analysts often cannot conclusively determine whether the assistance had an impact relative to businesses that did not receive assistance (i.e., a control group). Many types of outcome data—on both businesses receiving assistance and businesses that do not—may be available in secondary sources.

[3] See Chapter 7 of the *2014 Economic Report of the President*, Washington, D.C., U.S. Government Printing Office, 2014.

SUMMARY OF BEST PRACTICES

BEST PRACTICE 1:
Have One Plan for All Data Needs

Design one system to collect the necessary data for program administration, impact evaluation, and other evidence-building strategies. Identify and implement relevant data security and privacy and confidentiality requirements (see Best Practices 12 and 18). This can save time and reduce overall costs.

BEST PRACTICE 2:
Develop a Program Theory and Logic Model

At program inception, or at a minimum in advance of major data collections, generate a program theory (i.e., statement of what actions will cause what impact) to identify critical data needs for performance measurement and impact evaluation and to assess the feasibility of a useful evaluation.

BEST PRACTICE 3:
Check for Statistical Power

When making decisions about developing data for an impact evaluation, consider whether the expected magnitude and variability of the impact, and the sample size available for analysis (e.g., the number of those assisted that also have control group counterparts) permit the use of statistics to measure if the program had the intended effect (i.e., whether there is sufficient statistical power). Evaluation experts can help assess the statistical power of different evaluation approaches and their data needs, including whether combining program administrative data with secondary data (e.g., from federal statistical agencies) can improve the reliability of an evaluation.

BEST PRACTICE 4:
Determine Data Segments

While adequate data could be developed to make statistical determinations about all the businesses assisted, the data may not be effective in the analysis of important subgroups (e.g., age, size, and industry). With an evaluation expert, assess if the data for these subgroups are large enough to produce reliable estimates about the impacts of interest (e.g., jobs, revenue, and exports) for each of these subgroups. This analysis may help fine tune programs.

BEST PRACTICE 5:
Assess Alternative Impact Evaluation Methodologies

In deciding on evaluation methods (Randomized Control Trial or Quasi-Experimental Design), consider their different data needs against currently available data or data that could be obtained going forward. It is essential to engage evaluation experts upfront and explicitly consider the feasibility, strengths, and weaknesses of each evaluation method given available data, including information about applicants that did not receive services. Consult with evaluation experts, attorneys, and policy officers about the potential uses and permissibility of retaining data on applicants, including those that did not ultimately receive services. Rejected applicants may serve as a high-quality control group; the success of firms assisted by the program is compared with those that did not receive the help. See Best Practice 11 about informing both program applicants and participants about the use of their data for statistical research and evaluation purposes.

BEST PRACTICE 6:
Collect the Indispensable Data

Administrative data needs for evaluation may include Unique and Supplemental Identifiers for applicants and participants; participant-level data on the nature, intensity, and timing of program services (i.e., the treatments); and participant and applicant characteristics (e.g., size and age of firm). Investigate the possibility of accessing secondary data, like those housed at statistical agencies, for both participants and control groups. These secondary sources can provide a broader range of data on firm characteristics, as well as a broad range of data on outcomes for both groups.

BEST PRACTICE 7:
Collect Pre- and Post-Assistance Data on Impact

Access or collect pre- and post-treatment outcome data for the firms that received services and for control groups. This is central to identifying changes in outcomes (e.g., jobs, revenue, or exports) potentially attributable to the program. The data can come from a secondary data source or, if necessary, from a post-service survey of both the treated and the control groups. If surveys are the only means to collect outcome information, specify in service award conditions that post-service survey participation is a requirement for receiving assistance, and that the survey data will be used/shared to evaluate the program. In addition, having pre- and post-treatment observations at different periods for both groups is necessary to estimate both short- and long-run effects of the program (i.e., outcomes).

BEST PRACTICE 8:
Identify Other Assistance Provided

Evaluations can distinguish a program's impact from the impact of other programs when they include data on related services provided by other entities to the treated and control groups. This information, if possible to collect, helps ensure that changes in outcomes are not attributed only to a single program, if services from multiple programs contributed to the change.

BEST PRACTICE 9:
Create a Data Dictionary

Establish and maintain a data dictionary documenting data item definitions and changes, how data are collected (e.g., retain example forms and instructions), and relationships between key data items. Describe each data item and note the valid values/time periods. Describe any new records or revisions to existing records, including when the changes were made.

BEST PRACTICE 10:
Keep Records on Program Changes

Maintain historical records detailing the program at inception and over time. This documentation may include information on the original program, as well as changes to program design, eligibility criteria, legislation, service area, factors affecting program participation, available services, services provided, and delivery methods. Ensure these records also include information on criteria used to determine when/if an applicant/participant received services. This is particularly important for business assistance programs that provide advisory services where an evaluator will want to know if assistance was quick advice or intensive consulting. Records of decision rules used to select participants for program services, such as threshold levels or scores on various criteria, are also helpful.

BEST PRACTICE 11:
Check if Data Can Be Shared

Consult with general counsel and use guidance from OMB's M-14-06 to determine whether applicant/participant data can be shared with researchers for statistical analysis, including impact evaluation. If the data on past services cannot be shared, work with general counsel to evaluate whether language relevant for data sharing can be changed to permit sharing data on assistance provided in the future. In rulemaking language and other settings, ensuring the protection of applicants'/participants' privacy and confidentiality is important for sustaining the ability to share and use the data for statistical analysis in a secure environment.

BEST PRACTICE 12:
Involve Attorneys and Policy Officers, Including Privacy and Confidentiality Officers, Sooner Not Later

Hold early discussions with agency attorneys and policy officers to familiarize them with the data needs for evaluation, which could include Unique Identifiers such as social security numbers (SSNs) and employer identification numbers (EINs). Make sure the use of these data is in line with relevant legal authorities and that the requirements for data security, privacy, and confidentiality are documented. Discuss sharing and linking data for statistical analysis, including impact evaluation. Ensure Privacy Notices, System of Record Notices, Privacy Impact Assessments, and other documentation adequately address data access and sharing activities. Early discussions can avoid later problems and delays that arise when data collection and sharing are treated as separate or after-the-fact considerations.

BEST PRACTICE 13:
Put Data Requirements in Contracts and Grants

Ensure that grants and contracts with service providers include data collection, retention, record keeping, and data sharing provisions, as well as data security, privacy, and confidentiality protections. In the agreements, identify important data to collect and allow the data to be shared and used for statistical analysis and impact evaluations while employing robust privacy and confidentiality safeguards.

BEST PRACTICE 14:
Create a Way to Link Data

Unique and Supplemental Identifiers for applicants/participants are critical to linking program and secondary data. Preferably, these identifiers can be collected, retained, and shared for statistical analysis and impact evaluation purposes, but this requires both legal and technical considerations. Review forms used to initiate assistance (intake) and service agreements to ensure adequate informed consent to share and link data. Ensure these uses are properly documented in the System of Records Notice. Additionally, review these forms to ensure sufficient Unique and Supplemental Identifiers for applicants/participants to make data linkage feasible.

BEST PRACTICE 15:
If Necessary, Generate a Unique Identifier

Where confidentiality agreements preclude the collection, retention, or sharing of applicant/participant employer identification numbers (EINs) and social security numbers (SSNs), at a minimum, generate and retain a unique identification number created by the agency. This facilitates intra-agency data linking, making it possible to create a complete record of all the assistance the firm received from the agency.

BEST PRACTICE 16:
Reduce the Risk of Data Errors

Design and implement data collection systems, such as a client relationship management (CRM) system, with autofill or drop-down lists to minimize input errors. Provide clear guidance to database managers about the importance of entering data consistently. Implement quality control procedures to detect data entry and coding errors and inconsistencies. Include flags to identify edits and updates. Identify potential data and analysis problems with staff that create the data. Implement adequate quality control procedures early in the data collection process to avoid having to apply costly remedies to mitigate data quality problems. Demonstrate to data entry staff how the data are being used.

BEST PRACTICE 17:
Consider Cost/Benefit of Data Retention

Since historical data are often critical in impact evaluations, it is important that database retention policies consider the value of retaining historical records for evaluations with confidentiality protection.

BEST PRACTICE 18:
Protect the Data

Consult with lawyers and the information technology team to identify data security needs and policies. Create systems and practices that meet the needs, comply with policies, and preclude unauthorized access to data or disclosure of person- or business-specific data.

ORGANIZATION OF THIS GUIDE

The body of this guide provides high-level insights to help program staff develop their data strategy. A series of concise supplemental sections, referred to throughout the paper as supplementals, contain critical technical information and definitions that program decision makers may wish to review and share with other stakeholders (e.g., evaluators, legal experts, etc.) to further inform data collection strategies.

Chapter I explains how impact evaluations are complementary to other evidence-building activities, such as ongoing performance measurement, and demonstrates the value of designing data collections that fulfill the needs to track program performance and enable impact evaluations. Two decision trees are included to assist readers as they investigate the usefulness of their existing program administrative data to conduct impact evaluations. These decision trees also direct readers to relevant chapters and supplementals that will help them develop their data collection strategies. The first decision tree can be used to analyze the usefulness of existing administrative data for retrospective impact evaluations (i.e., when services have already been delivered). The second decision tree is useful for identifying data needs for evaluating future services.

The first step in both decision processes is developing a program theory, which outlines the causal linkages from program concept, resources, processes and services to outcomes. Chapter II provides an overview of program theory and an example of a logic model—a graphical representation of a program theory—for a business technical assistance program. Chapter III, supported with more explanation in Supplemental II, identifies several key considerations when determining whether a program is suitable for an impact evaluation using statistical methods. Suitability is based on factors such as program size and expected impact, and is best assessed before initiating an evaluation. The benefit of investing in administrative data improvements is smaller if evaluation findings will not provide reliable evidence of program impacts.[4]

Chapter IV and Supplementals III, IV, and V summarize key impact evaluation concepts and the two main approaches, Randomized Control Trials and Quasi-Experimental Designs, for conducting high-quality evaluations. Chapter V, along with Supplementals I, IV, and V, identifies critical types of data that enable administrative data systems to be used in high-quality evaluations. Chapter VI briefly describes several other methods for building evidence, explaining that they have different strengths and can be complementary to one another. Some of these methods have data needs that overlap considerably with those of impact evaluations and can thus benefit from administrative data system improvements suggested in this guide. Chapter VII identifies many of the data challenges faced by program managers who inherit a program and ways of dealing with the challenges. The concluding chapter, Chapter VIII, explains how making initial investments to improve administrative data to support evaluation can save evidence-building time and costs in the long run. Two additional supplementals describe key issues that arise when considering an impact evaluation: legal issues related to access and sharing of data with evaluators (Supplemental VI) and considerations when choosing an evaluator (Supplemental VII).

[4] Note that administrative data improvements can often benefit other evidence-building methods, as discussed in Chapter VI.

CHAPTER I. INTRODUCTION

Many federal agencies are exploring better ways to design administrative data collection to support program impact evaluations that can be used to improve programs.[1, 2] This guide primarily aids those interested in using administrative and other types of collected data to meet these evidence needs. The guide does not cover other important issues relevant for evaluations, such as determining evaluation priorities, addressing ethical concerns, and ensuring transparency and independence.[3]

Data needed for impact evaluations are not always collected or retained in the normal course of program operations. When these data are absent from program databases, administrative data are less useful for these evaluations. In addition, their absence can lead to less reliable and costlier evaluations when agencies have to conduct follow-up surveys or collect additional data needed to measure program impacts.[4] Nonetheless, with some advance planning these data weaknesses can often be remedied.

Impact evaluations use statistical methods to estimate program impacts. Impact refers to the difference between what actually occurred and what would have occurred in the absence of program services. The purpose of this report is to help agencies identify the critical data that, if collected, enable rigorous impact evaluations of business assistance programs. By following the best practices provided, agencies may not be limited to using follow-up surveys to develop data for impact evaluations, but can also use administrative and other secondary data that are already being collected.

Typically, some of the same data needed for other ways of building evidence are also required for impact evaluations (see Chapter VI). This may be particularly true for performance measurement. However, the purposes of performance measurement and impact evaluations differ—with the former focused on metrics that inform day-to-day program management (Figure 1). Impact evaluation provides more definitive answers about whether a program achieves the outcomes decision makers intended when they initiated the program. The intended outcomes of business assistance can include adding or retaining jobs, increasing the short- or long-term viability of a business, and/or continuing or increasing a business' contribution to the economy through exports. Impact evaluation uses statistical research methods to determine if these impacts can be attributed to the program. The most cost-effective data collection and administrative data systems consider data needed for performance measurement and impact evaluation, as well as the data security and privacy and confidentiality requirements, at the outset (see Supplementals I and VI).

BEST PRACTICE 1: Have One Plan for All Data Needs

Design one system to collect the necessary data for program administration, impact evaluation, and other evidence-building strategies. Identify and implement relevant data security and privacy and confidentiality requirements (see Best Practices 12 and 18). This can save time and reduce overall costs.

[1] An *impact evaluation* is defined by the Millennium Challenge Corporation and the U.S. Agency for International Development as, "An independent study that measures the changes in income and/or other program objectives that are attributable to a defined intervention. Impact evaluations require a credible and rigorously defined counterfactual that estimates what would have happened to the beneficiaries absent the project." Katherine Farley, Sarah Lucas, Jack Molyneaux and Kristin Penn, "Principles Into Practice: Impact Evaluations of Agriculture Projects," Millennium Challenge Corporation, October 2012, <https://assets.mcc.gov/reports/paper-2012001116901-principles-impact-evaluations.pdf>.

[2] OMB has recently directed agencies to manage data so it can better support statistical uses, including evidence building and evaluations. "Guidance for Providing and Using Administrative Data for Statistical Purposes," OMB's M-14-06, 2014.

[3] See U.S. Department of Labor Evaluation Policy, 2013, <www.dol.gov/asp/evaluation/EvaluationPolicy.htm>.

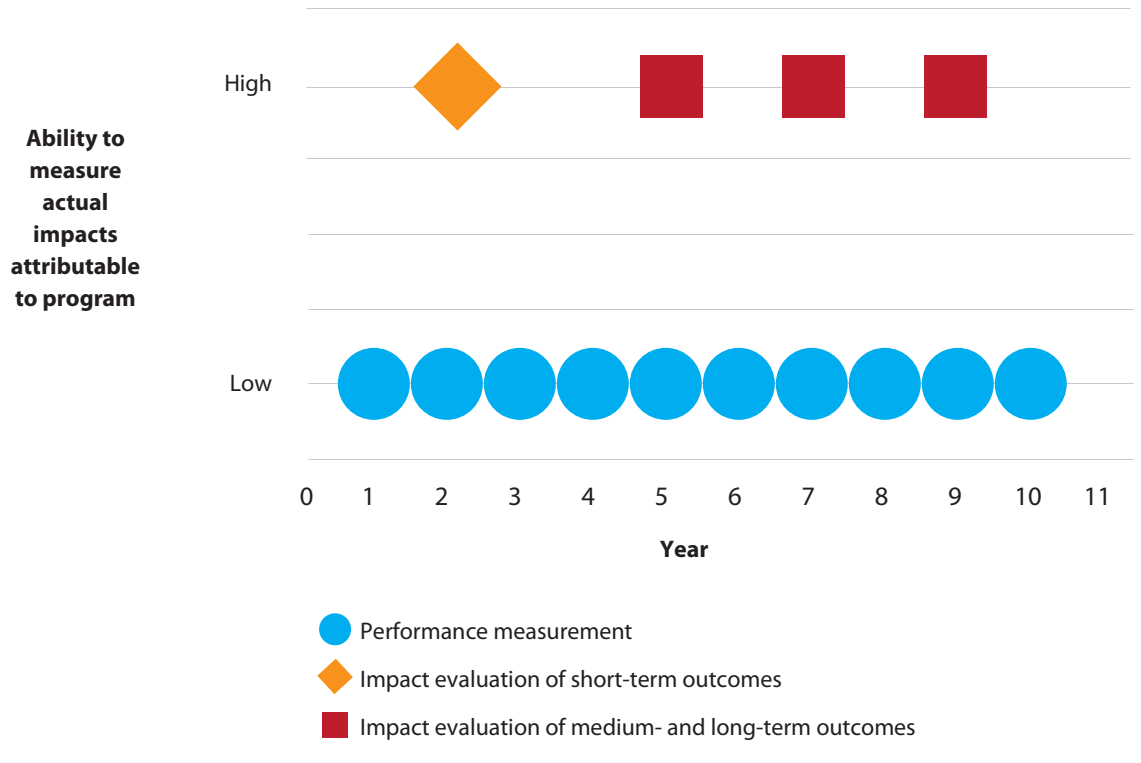
[4] See, for example, Bruce D. Meyer, Wallace K. C. Mok, and James X. Sullivan, "Household Surveys in Crisis," *Journal of Economic Perspectives*, Vol. 29, No. 4, pp. 199–226, 2015, who find that survey respondents systematically underreport the amount of public assistance they have received.

Since it is very difficult to collect accurate data on services delivered in the past, the longer agencies wait to assess the usefulness of their data for conducting an impact evaluation and take the steps needed to perform an evaluation, the costlier the evaluation, which can limit an agency's options.

The circles in Figure 1 show that performance measurement can start almost immediately, because it measures numbers of clients served and other "outputs." It can also measure short-run outcomes that occur soon after service delivery. In contrast, impact evaluations are more episodic and are used to assess the short-run (diamond) and long-run (squares) outcomes. Impact evaluations may not be feasible in the earliest years of a program because it may take more time for data on the impact to become available. Further, programs with small numbers of participants in any one year may require multiple years of service delivery to achieve an adequate sample size (see Chapter III and Supplemental II) that would be needed for a reliable impact evaluation.

Figure 1.

Stylized Depiction of Timing of Performance Measurement and Impact Evaluation in an Existing Program



A. Background: Impact Evaluation, Its Components, and Importance

Program managers often face questions about the impact of their program, where impact refers to the difference between what actually occurred as a result of program services and what would have occurred in their absence. At issue is whether their program caused the relevant intended outcome. Establishing that a program caused an impact (i.e., a change in outcome) is not straightforward. However, periodic research, following sound methodology, builds a body of reliable evidence for decision-making.

Consider the following simple example of how chemists might evaluate the impact of a new chemical substance. The purported impact of the chemical substance is that it makes water fizz with bubbles. A researcher designing an experiment to test the impact of the chemical substance might consider two identical beakers of water. The chemist would then drop the chemical substance in the beaker on the right and leave untouched the beaker on the left. The chemist can then compare the outcomes across the two beakers to understand the impact of the new chemical substance. If, for example, the beaker with the chemical substance bubbles, while the untouched beaker on the left does not, the chemist may conclude that the chemical substance has an impact on the water: bubbles. By contrast, if the beaker on the left also generates bubbly water, the chemist cannot persuasively conclude that it was the new chemical substance that led to the bubbles. Something else caused the impact.

What is the most important point in this simple example? In order to evaluate the effect of the new chemical substance, the chemist needed a control—the second, untouched beaker of water—to understand what happens in the absence of the new chemical substance (i.e., the counterfactual). The impact of a program is the difference between what outcomes are achieved by a program (i.e., the introduction of the chemical substance into the beaker), and what outcomes would have occurred in the absence of the program (i.e., the untouched beaker). Unfortunately, unlike some other scientists, social scientists evaluating actual programs do not have a controlled lab environment to observe the counterfactual, so researchers estimate these likely outcomes using a variety of methods.[5]

Critical Concept

Impact is the difference between what actually occurred and what would have occurred without program services. It is not a simple before-and-after measure of change.

Impact evaluations are powerful tools that help agencies understand whether their programs are having the intended effects. When done rigorously, they provide evidence that apparent program outcomes, such as additional jobs or business revenue, can be attributed to the program being evaluated. When compared to other types of evidence such as performance metrics, impact evaluations are considered the strongest type of evidence about program impacts. See Chapter VI and the Evidence Continuum developed by the Corporation for National and Community Service for further information.

[5] If well-executed, this beaker study is applicable to the particular setting where the test occurred, but it may or may not apply more generally to other settings, such as when using a different water source. See Chapter V and Supplementals IV and V for discussion of data needs for separate impact analysis under different settings. For further discussion on the importance of transparency and opportunities to replicate findings, see Klaus F. Zimmermann, "Evidence-Based Scientific Policy Advice," IZA Policy Paper No. 90, 2014, <http://ftp.iza.org/pp90.pdf>.

Impact evaluations are most helpful when performed throughout the lifecycle of a program, regularly and consistently, rather than sporadically or randomly.[6] For example, a unique set of circumstances at one point in time may lead a program to have an economic impact that is not typical. If the impact is demonstrated to be in the same direction in different periods, the evidence on whether the program works is much stronger. Moreover, bodies or portfolios of evidence from multiple studies, often employing different methods, generally provide stronger evidence with conclusions that can be generalized across a program when it operates in different places and under different conditions.

Other types of evidence can complement impact evaluations and help explain if and/or how programs achieve desired outcomes. For example, focus groups and other qualitative methods can help clarify key constructs in the program theory such as the hypothesized cause and effect relationship between program activities and outcomes (see Chapters II and VI). Focus group conversations and case studies can also clarify the key features of the program that are critical to achieving results and can be used to refine measures of success. In short, impact evaluations are part of a broader evidence-building tool set but are particularly critical to understanding program impacts.

[6] See, for example American Evaluation Association, "An Evaluation Roadmap for a More Effective Government," 2013, <www.eval.org/p/cm/ld/fid=52>.

B. Navigating the Guide Using the Decision Trees

This guide is structured to help program managers and evaluators better understand whether their administrative data collections, or plans for such collections in new programs, will be useful in an impact evaluation, alone or in conjunction with external, secondary data. The following two decision trees (Figures 2 and 3) include questions that can arise in developing or improving a data collection plan and direct readers to answers in the relevant chapter of this guide. Figure 2 can be used to analyze the usefulness of already collected administrative data in existing (ongoing) programs. Figure 3 is useful for analyzing data needs for future data collections for ongoing programs, new programs, and pilots that will deliver services in the future.

Figure 2.

Decision Tree—Existing (Ongoing) Programs: Are a program's administrative data fit for use in an impact evaluation?

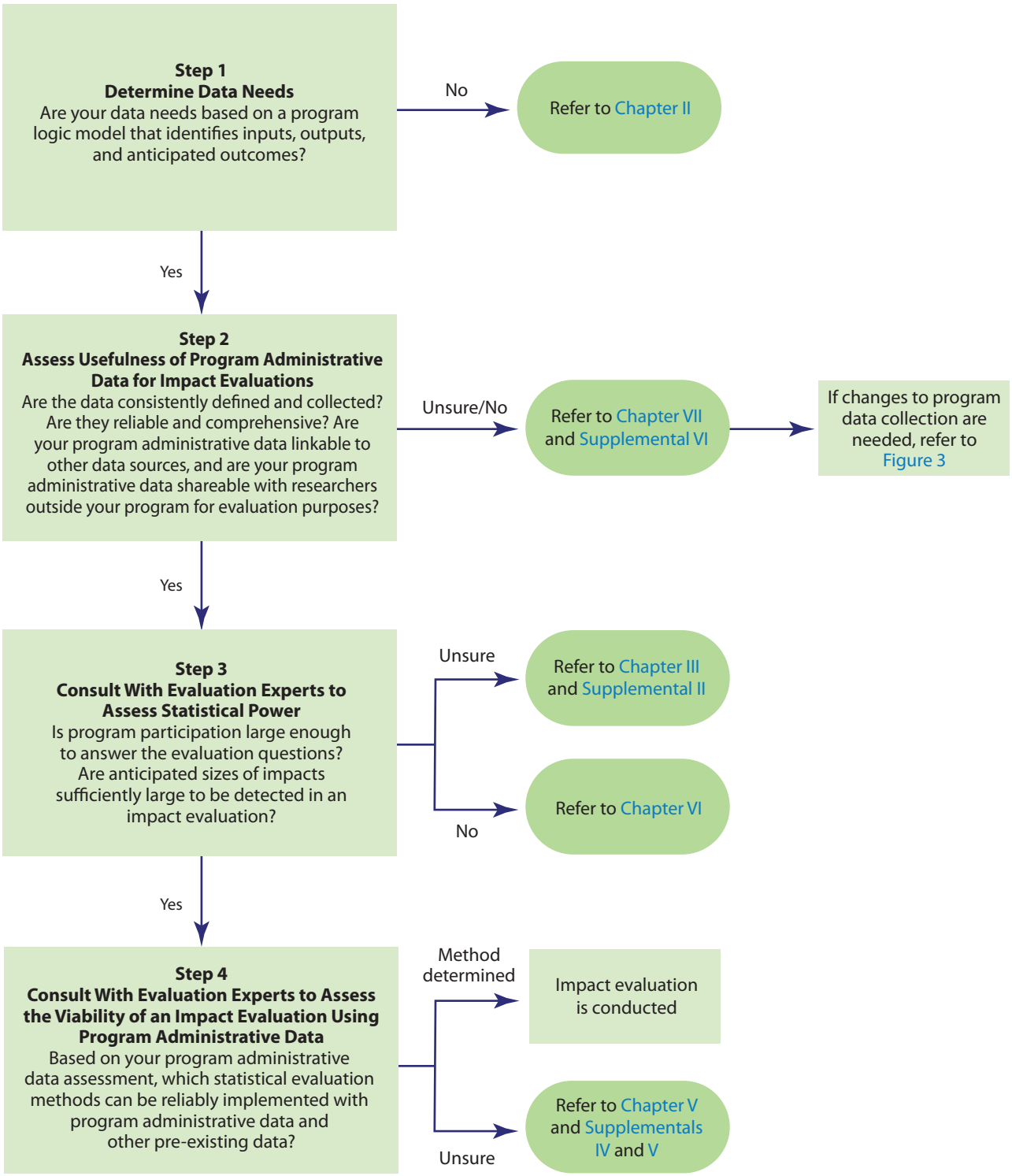
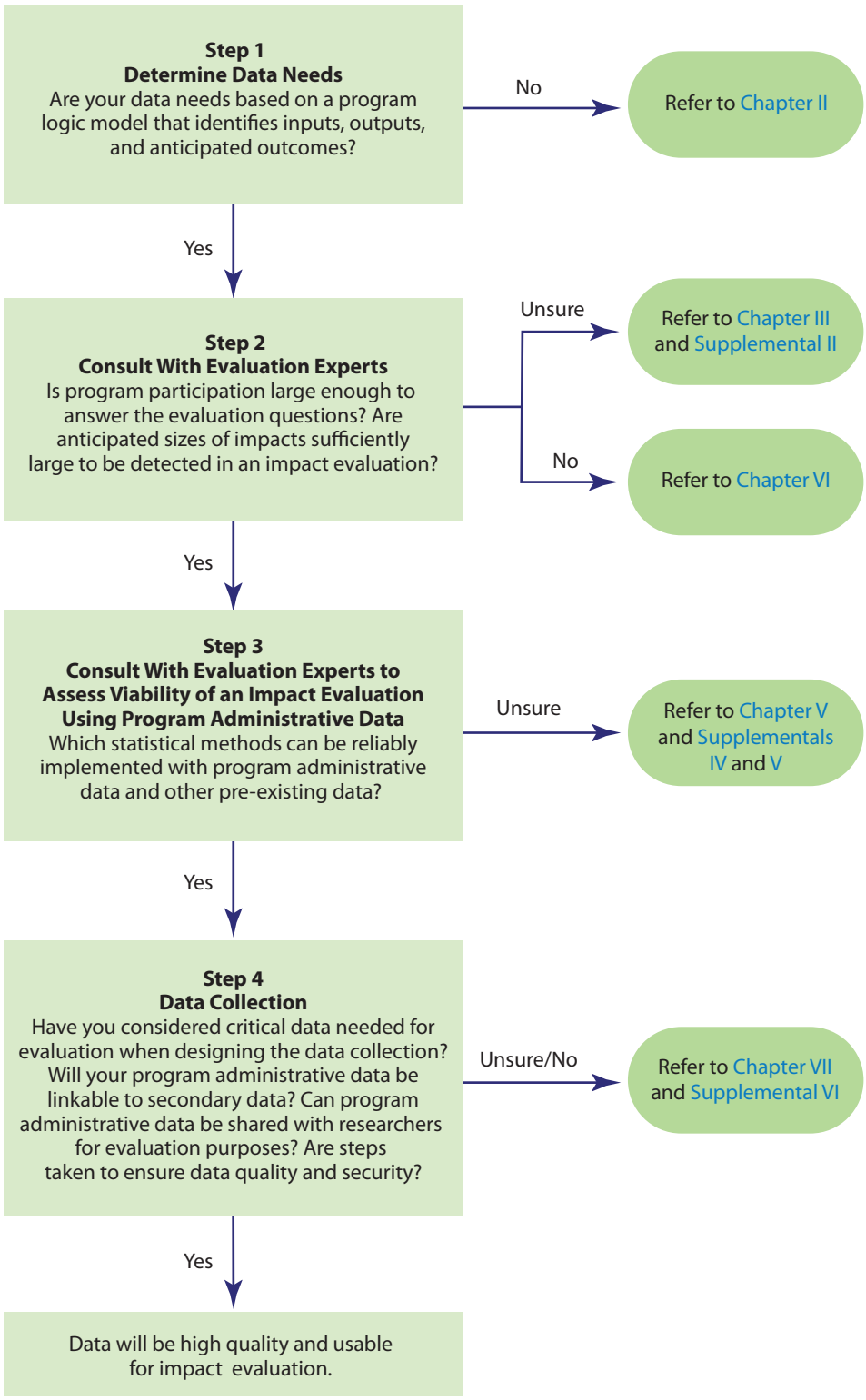


Figure 3.

Decision Tree—New Data Collections: How can new administrative data collection be designed for use in an impact evaluation?



CHAPTER II.
PROGRAM THEORY AND LOGIC MODELS

Government-based business assistance programs are designed to help alleviate business challenges related to either start-up or continuing operations. To understand what data are needed to measure program impact, it is important to identify how the program expects to achieve outcomes. Program managers and evaluators have used three intrinsically related tools to help with this task: a program theory, a logic model table or matrix, and a logic model graphic.

A. Program Theory Is a Set of Hypotheses About How a Program Affects Outcomes

Experience reveals differences in stakeholder opinions about how program activities achieve outcomes. They even can disagree on the outcomes the program should achieve. Program mission statements may not be specific about the intended outcome of a program. To design a program evaluation, it is essential to be clear on the cause, effect, and outcome assumptions of a program.

A first step in achieving clarity is to define a set of hypotheses about how program activities lead to outcomes (i.e. the program theory). These hypotheses are often posed as **if-then** statements. The following is an example of an **if-then** statement for a business assistance program: **If** a potential entrepreneur is provided a specific type of entrepreneurial training, **then** the entrepreneur will start their own business.

The **if** and **then** assertions should be supported by solid evidence, accepted theories, or at a minimum well-established programmatic experience.[7] For instance, providing financial training to individuals who want to start a business might be supported by a body of research in peer-reviewed journals that shows potential entrepreneurs are deterred from starting a business if they do not know how to set up and maintain their accounts. Alternatively, small business counselors might report through a survey that lack of financial acumen is the concern most frequently mentioned by their clients.

A well-developed program theory may reflect that change occurs in stages and over time. The program may provide training, assist the would-be entrepreneur in gaining experience with the skills, and then provide an opportunity for the neophyte to work with an established business to experience how the skills contribute to success. After the three stages, the assisted party is ready to start their own business.

The following **if-then** example statements provide a sample set of hypotheses (i.e., program theory) that trace the linkages (i.e., causal mechanisms) for a typical business assistance program from inputs to long-term impacts:

The following **if-then** example statements provide a sample set of hypotheses (i.e., program theory) that trace the linkages (i.e., causal mechanisms) for a typical business assistance program from inputs to long-term impacts:

- **If** a set of inputs (e.g., trained staff, meeting space, computer equipment, curriculum materials) are available from the XYZ Business Creation program, **then** it can provide a set of activities or services to individuals wishing to start their own business. This program **capacity hypothesis** details what the program can provide and to whom.
- **If** potential entrepreneurs receive services from the XYZ program, **then** they would increase the knowledge, skills and abilities needed to create and maintain a business. This program **causality hypothesis** suggests a mechanism that links program services to the recipient's behavioral changes.
- **If** individuals increase their business creation knowledge, skills, and abilities, **then** they have a greater chance of starting their own business. This **impact hypothesis** links the recipient's behavioral changes to the program's short-term impacts.
- **If** these individuals have a greater chance of starting their own business, **then** they have a greater chance of hiring workers, thus having a broader impact on the local business community. This **impact hypothesis** links the program's short-term impacts to its long-term impacts.

[7] Chapter VI discusses evidence gathering from surveys and focus groups that can help inform the **if** and **then** assertions.

A good program theory includes sufficient relevant details on providers, activities, services, and recipients to answer the following questions:

1. If the service is provided, then what should be the results achieved by those receiving the service?
2. Why should the activity lead to the postulated results?
3. What are the underlying assumptions about how these changes occur?
4. What is the body of existing research and data supporting the assertion that the activity or service will lead to these results?

If a strong evidence base on this program does not exist, then evidence from similar programs or published research will suffice. Thus, a good program theory provides the logical framework and describes the causal mechanisms of how a program is expected to achieve impacts.

B. A Logic Model is a Mapping of the Linkages Between a Program and Outcomes/Impacts

Armed with the program theory developed above, the second step in this process is to develop a logic model, a graphical representation of the hypothesized causal linkages between program activities and expected outcomes. A logic model is a visual tool that depicts why a program is expected to work and for whom. It also depicts how a program leads to the expected outcomes. Using a logic model is critical to increase the likelihood that a program operates effectively. The logic model also serves as the foundational step for data development and impact evaluation. Logic model uses include:

- Providing a basis for an evaluation design.
- Specifying data items (e.g., independent, dependent, and contextual variables) needed for evaluation.
- Facilitating performance management by guiding development of useful performance measures (see Chapter VI).
- Illustrating important features to stakeholders.
- Managing programs.

The following components are generally used in developing a logic model:

- **Inputs**: Resources or materials used by the program to provide its services.
- **Activities**: Services provided by the program.
- **Outputs**: Quantifiable amount of service provided (e.g., classes attended, people served, number of hours of services received, financing, etc.).
- **Outcomes/impacts**: Any behavioral, economic, or other change occurring as a result of receiving these services.
- **External/exogenous factors**: Factors beyond the control of the program manager such as the local economy or labor market conditions.

Table 1 provides an example of a generic logic model table. Table 2 provides examples of specific activities, outputs, and outcomes for the hypothetical XYZ Business Creation Program, including measures that could be collected to both track program performance and enable an impact evaluation.

Table 1.

Generic Logic Model

Inputs	Activities	Outputs	Immediate outcomes	Intermediate outcomes	Long-term outcomes
In order for the program to operate, the following things are needed and provided.	In order to address the stated problem, the following activities are needed and provided.	The intervention or program will create the following immediate service delivery outputs.	The following immediate changes in behavior will occur as a result of program participation/ completion.	After a certain period, the following changes will occur as a result of program participation/ completion (6 months to 2 years).	After a longer period, the following changes will occur over a specified period of time as a result of program participation/ completion (2+ years).
External or exogenous factors					
These are factors that are beyond the control of the program manager, such as the local economy or labor market conditions. These can have a significant effect on outcomes of the program.					

Table 2.

Example Logic Model and Performance Measures for the XYZ Business Creation Program

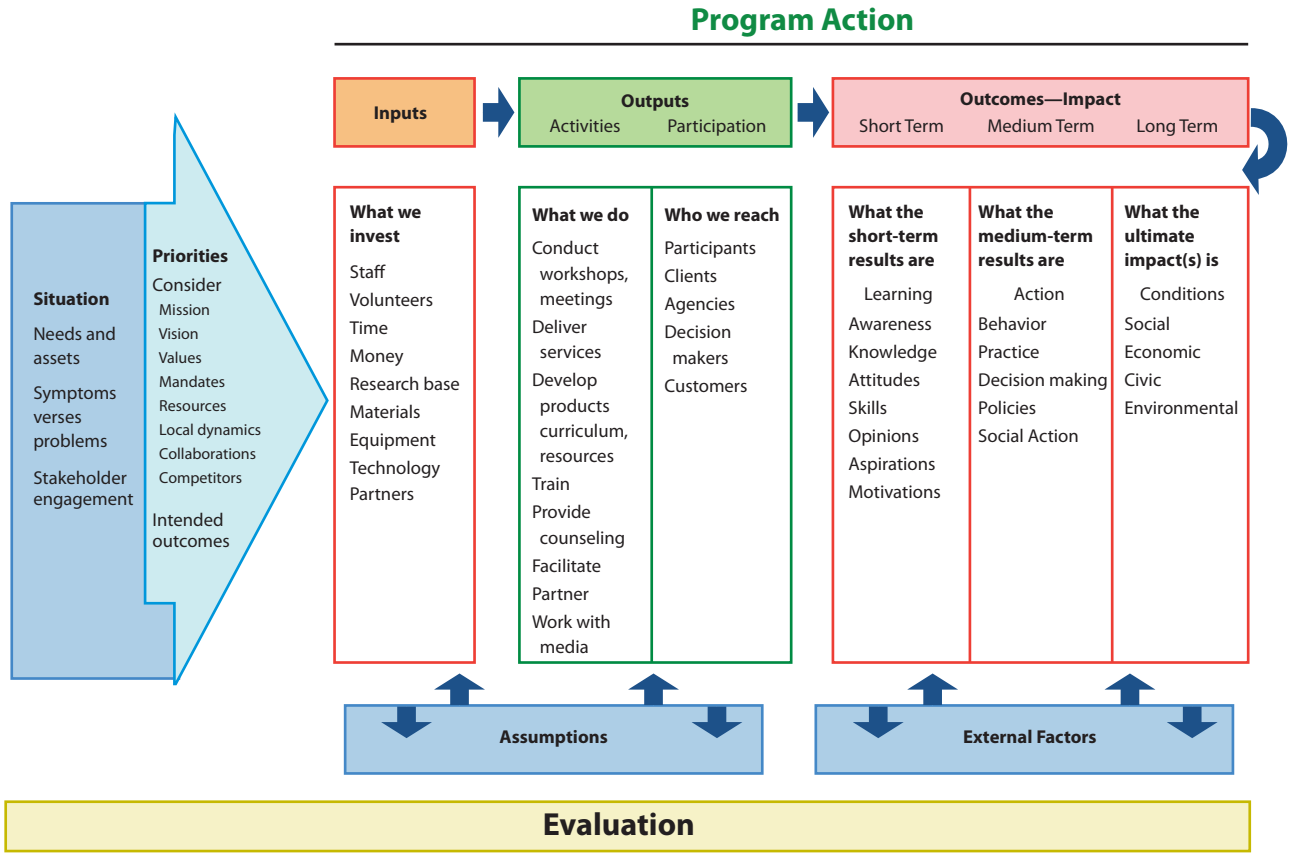
Inputs	Activities	Outputs	Immediate outcomes	Intermediate outcomes	Long-term outcomes
The XYZ federal program provides: Physical space. Pipeline of potential entrepreneurs. A ready curriculum. Trained curriculum facilitators. Monetary resources of $250,000 for training services to be provided by ABC training services.	ABC training services offers a 1-week "boot camp" to 2,000 potential entrepreneurs on how to start a small business, including the development of a bankable business plan.	ABC training services enrolls 300 participants. Of these 300 participants: 250 complete the training within 1 year.	3 months after completing the training: 50 individuals take no further action. 200 individuals develop a bankable business plan. Of these 200 individuals: 150 start a business.	1 year after completing the training: 100 firms are still in business. Of these 100 firms: 90 recorded revenue. 75 hired an additional employee. 10 were inactive, but still filed corporate taxes.	3 years after completing the training: 50 firms are still in business. Of these 50 firms: 45 have employees. 25 recorded a profit within the 3 years. 5 were inactive, but still filed corporate taxes.
External or exogenous factors					
The above results are influenced by several factors. Among these are a strong local labor market, strong local economic activity, ample available capital and credit, and additional small business support resources from local entities.					

A good practice is to supplement the logic model table with a graphical logic model detailing how the various parts of the model fit together and are related to one another. The University of Wisconsin-Extension has developed an excellent example of a graphical logic model (Figure 4).

When building logic models, evaluators and program analysts often use the following strategies to facilitate a clearer understanding of a program theory.

- **Use clear and plain language**. Breaking down into simple steps how and why a program works will make it easier to decide which program components to measure and what data to collect.
- **Start by identifying outcomes and then work backwards**. One effective way to build logic models is to start at the expected outcomes and work backwards to fill in the other columns. Many program experts can clearly articulate the program's expected immediate and long-term outcomes, which can make the link backwards to inputs and activities more intuitive than proceeding from activities to outcomes.
- **Clearly link program activities and outcomes**. Make sure that a program's services and their intended

Figure 4.
Example Graphical Logic Model



Note: For more information, see <www.uwex.edu/ces/pdande/evaluation/evallogicmodel.html> and <https://fyi.uwex.edu/programdevelopment/files/2016/03/lmguidecomplete.pdf>.
Source: UW-Extension Program Development and Evaluation Logic Model, <http://www.uwex.edu/ces/lmcourse/>, (c) 2002 Board of Regents of the University of Wisconsin System.

outcomes are clearly linked, sequenced, and logically connected. Most importantly, verify that these linkages reflect reality. An inaccurate depiction of how a program is intended to work can lead to an impact evaluation of little value or an ineffective intervention to improve a program.

- **Use existing resources**. Federal agencies, colleges, universities, and private entities use logic models and make their literature and tools publicly available for beginners and experts.
- **Revise and revise again**. Are the program's most important services included? Are the program's intended outcomes significant to stakeholders (e.g., target community, broader population, and government entities)? Are the outcomes clear and realistic? Are the connections in the logic model feasible and appropriate? What is the evidence base that corroborates these connections?
- **Evidence evolves the model**. As cause and effect relationships are tested with data and evaluations, use study results to refine the program theory and the logic model. Doing so should lead to improved program performance.

Developing a program theory and utilizing logic models (tables and graphical logic models) at program inception (and throughout program implementation) identifies the information required for program management, implementation, and evaluation. These tools can minimize the need for expensive additional data collection after the fact. Early implementation of these cornerstones provides a solid foundation for improving the results of assistance programs and creates the capacity to learn what works and what does not as conditions change.

BEST PRACTICE 2: Develop a Program Theory and Logic Model

At program inception, or at a minimum in advance of major data collections, generate a program theory (i.e., statement of what actions will cause what impact) to identify critical data needs for performance measurement and impact evaluation and to assess the feasibility of a useful evaluation.

Suggested Additional Readings

- Many DOL agencies use logic models to illustrate program operations, which help inform their agency's operating plans. See an excellent discussion of this by Dave Frederickson, "At the Department of Labor We're Using our Strategic Plan to Manage. No, Really," 2012, <www.dol.gov/_sec/stratplan/20120928-DOLPublicManagerArticle.pdf>.
- The Centers for Disease Control and Prevention provides a variety of evaluation resources that include a fair number of logic models. See <www.cdc.gov/EVAL/resources/index.htm>.
- The Department of Justice provides excellent resources on logic model use and development. See <www.ojjdp.gov/grantees/pm/logic_models.html>.

CHAPTER III.
UNDERSTANDING A PROGRAM'S SUITABILITY FOR IMPACT EVALUATION

A. Assessing Statistical Power

Impact evaluations are most useful when researchers are confident that substantive changes in outcomes can be detected if they exist. An impact evaluation with a high likelihood of detecting a substantive change has adequate statistical power. Statistical power should be assessed prior to deciding to proceed with the evaluation. The results of evaluations conducted with low statistical power may be useless. Worse, they may be misinterpreted. Decision makers could modify or terminate a program that is having a positive effect that simply has not been detected.[8] Case studies, for example, may be preferable to an underpowered study that appears to have rigor but does not in fact have dependable findings. This important topic is explained further in Supplemental II.

The three fundamental determinants of statistical power are:

- **Mean effect size**: How big is the expected impact, on average, that you are seeking to measure?
- **Variance in effect size**: How much heterogeneity in impacts is expected across program participants?
- **Sample size**: How many participants will be studied?

[8] Statistical power refers to the probability of correctly rejecting the null hypothesis (e.g., the hypothesis that a program had no impact) when this hypothesis is false (i.e., the program did have an impact). If an evaluation has low statistical power, it is not possible to distinguish between cases where the program truly has no impact, versus a case where the program does have an impact but the evaluation is not powerful enough to detect it.

When a program's effect is likely to be small, detecting the effect may require a very large sample size. If a program is a small pilot program or has limited funding, it may not be reaching a large enough number of participants to permit an appropriately powered evaluation. Even if the program reaches a large population of businesses, achieving the necessary statistical power may prove expensive if the impact is small and a very large sample size is required. A larger sample will also be required if the impacts are highly variable across participants. For example, programs serving businesses in both rapidly growing and declining sectors might expect vastly different impacts in different industries.

In the case of programs with few participants (such as a program that began delivering services very recently), programs with very modest interventions or impacts, or programs with highly variable impacts, they may not be ripe for a reliable impact evaluation.

B. Statistical Power for Subgroups

Often evaluators want to measure the impact on subpopulations, e.g., whether technical assistance provided to midsized firms has more impact on employment than assistance provided to large firms. When this is the case, the statistical power of the evaluation for each subpopulation should be considered. The overall sample size needed for estimates pertaining to subpopulations is larger than the sample required for measuring impact for the population as a whole.

C. Increasing Statistical Power

An evaluation approach that links and combines program administrative data and/or data housed at federal statistical agencies may increase the sample size, thereby increasing statistical power and improving the reliability of an evaluation. For example, an evaluation of the impact of a "small business incubator" in one location may not have adequate statistical power. However, using business data housed at the Census Bureau that include many incubators across the country may provide sufficient statistical power for more reliable evidence.

BEST PRACTICE 3: Check for Statistical Power

When making decisions about developing data for an impact evaluation, consider whether the expected magnitude and variability of the impact, and the sample size available for analysis (e.g., the number of those assisted that also have control group counterparts) permit the use of statistics to measure if the program had the intended effect (i.e., whether there is sufficient statistical power). Evaluation experts can help assess the statistical power of different evaluation approaches and their data needs, including whether combining program administrative data with secondary data (e.g., from federal statistical agencies) can improve the reliability of an evaluation.

CHAPTER IV.
IMPACT EVALUATIONS: AN OVERVIEW OF DESIGNS AND REQUIREMENTS

A. Two Impact Evaluation Approaches: RCT and QED

To generate credible conclusions, impact evaluations require data from both treatment and control groups. The treated group includes the businesses that received program assistance. The control group includes businesses that did not receive any assistance, but are otherwise similar to the treated businesses.[9] Programs regularly collect data from participants and generate their own program administrative data. However, most do not collect data on nonparticipants or retain sufficient information about rejected applicants to create a control group.[10] Thus, secondary data (i.e., data collected by federal agencies or entities other than the program) are often necessary to create a control group for assessing impacts. The datasets produced or housed by federal statistical agencies such as the Census Bureau and the Bureau of Labor Statistics (BLS) are some of the most useful for this purpose.

[9] Data on program participants alone can be sufficient to evaluate impacts of alternative levels of service or implementation approaches if the alternative levels of service or implementation approaches are randomly assigned.

[10] Rejected applicants can sometimes, but not always, serve as a good control group. They work well, for example, when the rejection is due to a randomization device, as in an RCT design, or when it is based on an observable application score or an eligibility requirement that is continuous in nature (e.g., owner age), in which case a Regression Discontinuity Design (RDD) can be employed (see Chapter V for an example). If, however, the rejection is based on noncontinuous eligibility requirements that may be correlated with the outcome of interest, such as the owner being a member of a minority group, use of such a control group could introduce confounding factors that would make it difficult to draw meaningful conclusions about the impact of the program.

Evaluators have a number of research design options. Their best choice depends in part on program characteristics and data availability. Although there are many different options, this report focuses on two basic evaluation designs that can produce credible estimates of program impacts:

- **Randomized Control Trial (RCT)**: Program participants (the treated group) and nonparticipants (the control group) are randomly selected (see text box "Examples of RCTs and QEDs" and Supplemental IV).
- **Quasi-Experimental Design (QED)**: The control group is deliberately selected to have characteristics as similar as possible to the treated group; except that only the treated group has received program services (see text box "Examples of RCTs and QEDs" and Supplemental V).

RCTs and QEDs differ in many ways (see Supplementals IV and V), but two differences particularly germane to this guide are when the approaches can be used and their specific data requirements. While RCTs are widely viewed as the best means of estimating impacts attributable to a program, they cannot be used to evaluate past services that were delivered via nonrandom selection (e.g., using application scores). In such situations, QEDs are generally the best option. Some data items are desirable for an RCT, but essential for a QED.

Examples of RCTs and QEDs

Example of an evaluation using a Randomized Control Trial (RCT) design:

- Jacob Benus, Theodore Shen, Sisi Zhang, Marc Chan, and Benjamin Hansen, "Growing America Through Entrepreneurship: Final Evaluation of Project GATE," Final Report, IMPAQ International, December 2009. The DOL and the Small Business Administration created a demonstration pilot called Project GATE, designed to help people create, sustain, or expand their own business. Applicants were randomly assigned to either the treatment group that received program services, or a control group that did not receive services. The data used in the evaluation were a combination of administrative data and survey data. The study can be found at <http://wdr.doleta.gov/research/FullText_Documents/Growing%20America%20Through%20Entrepreneurship%20-%20Final%20Evaluation%20of%20Project%20GATE.pdf>.

Examples of impact evaluations using Quasi-Experimental Designs (QEDs):

- C.J. Krizan, "Statistics on the International Trade Administration's Global Markets Program," U.S. Census Bureau, CES 15-17, September 2015. To conduct their first program evaluation study, ITA partnered with internal researchers at CES who were already highly familiar with the international trade data housed at the Census Bureau. The researchers used name, address, and website information to link the Global Markets program treatment data to the administrative data on all exporting businesses. This allowed them to construct control variables (groups). The study found evidence that receiving counseling was positively correlated with increased exports and, to some extent, employment growth. The published results of the study also contain a "Lessons Learned" section that discusses examples of matching, identifiers, and other data issues encountered during the evaluation. The results can be found at <https://www2.census.gov/ces/wp/2015/CES-WP-15-17.pdf>.

- Clifford A. Lipscomb, Jan Youtie, Sanjay Arora, Andy Krause and Philip Shapira, "Evaluating the Long-Term Effect of NIST MEP Services on Establishment Performance," U.S. Census Bureau, CES 15-09, March 2015. The National Institute of Standards and Technology's Manufacturing Extension Partnership (MEP), an ongoing program, has been evaluated several times using QEDs. MEP has hired contractors, who had access to MEP program data for research purposes and formed control groups by linking the MEP data to Census Bureau business data for the analysis. They estimated difference-in-differences regressions comparing assisted establishments' change in productivity from before to after assistance receipt to nonrecipient control establishments' productivity change over the same time period. See the latest evaluation results at <http://www2.census.gov/ces/wp/2015/CES-WP-15-09.pdf>.

CHAPTER V.
DATA NEEDS FOR IMPACT EVALUATIONS

A. When to Assess Data Needs

An assessment of the viability of RCT and QED methodologies and data needs can be done after a program is implemented. However, it is often more expensive to collect needed data after initiation of service delivery and even more so after completion. For example, collecting data from previous participants is subject to Paperwork Reduction Act reviews that may increase administrative effort and time for the evaluation. Obtaining the participants' informed consent, often required to use participant data for evaluation, can take several months or may not be possible as projects age. Past participants have little incentive to respond, if they no longer participate in the program. Some are hard to reach or have gone out of business. If failed businesses are not in the sample, a significant bias is introduced in the study. Decision makers may abandon an evaluation when increased costs, delays, and bias are considered.[11]

Early evaluation assessment/design should include clear definition of the impacts to be evaluated, and the data needed to estimate those specific impacts. Decision makers may also need to delineate how impact data will be segmented. For instance, policy makers may want to know whether a program as a whole, or only some program components, is having the intended effects on outcomes. Policy makers may need data on business age, size, or industry, etc., to learn if impacts are different for these subgroups. In general, the greater the number of impacts and population segments introduced into the evaluation, the more data will be required.

[11] Early consideration of important types of data to collect, and maintaining the quality of the data, are consistent with recent OMB guidance, which directs program agencies to manage "high value" datasets so they are accessible and of sufficient quality. The guidance notes, "[Program agencies] can do this most efficiently and effectively by integrating these considerations into data collection and management for programmatic purposes, rather than treating them as separate or after-the-fact considerations. This includes, for example, collecting and retaining data items that would facilitate evaluation and analysis of the data" ("Guidance for Providing and Using Administrative Data for Statistical Purposes," OMB's M-14-06, February 14, 2014).

BEST PRACTICE 4: Determine Data Segments

While adequate data could be developed to make statistical determinations about all the businesses assisted, the data may not be effective in the analysis of important subgroups (e.g., age, size, and industry). With an evaluation expert, assess if the data for these subgroups are large enough to produce reliable estimates about the impacts of interest (e.g., jobs, revenue, and exports) for each of these subgroups. This analysis may help fine tune programs.

B. Data Requirements for Different Methodologies

The data required to conduct RCTs are also required for QEDs. QEDs have additional data requirements that are not absolutely necessary for RCTs, but the additional data (e.g., pretreatment characteristics) can strengthen confidence in RCT impact estimates. QEDs require more information to build confidence that the control group is otherwise similar to the treated group. The control group in a QED is designed to be comparable to the treated group in key characteristics; the assumptions on what attributes are key may be incorrect or incomplete. There may be unobserved traits that systematically differ across the treated and control groups that affect impact. To minimize this problem, data on more characteristics and a larger sample will be needed (see Supplementals IV and V).[12]

[12] For example, a larger sample permits selection of controls that on average are more similar to the treated firms. Having a large sample does not completely eliminate concerns about impact estimate bias—it can minimize systematic differences in observable characteristics across treated and control firms, but not unobservable ones. See Supplemental V for a hypothetical example of how unobservable and systematic differences between treated and control groups can confound impact estimates.

Whether RCT or QED is used when designing the data collection, consider:

- **The necessity of data on the identities of the applicants/participants, and the nature, intensity, location, and timing of the services provided (treatment)**. Information on the type and intensity of services will permit researchers to test and learn about the relative performance of different types and levels of services. Information on timing is needed to know whether an outcome is pre- or post-treatment, and whether the impact is a short- or longer-run effect. Note that identifiable data are often subject to privacy and confidentiality statutes.
- **Requirements to have data on nontreated, but eligible entities for creating control groups**. Typically, such information comes from secondary datasets. The more comprehensively a business dataset covers a relevant population, the better. A dataset that underrepresents or excludes a category of entrepreneurs/firms can bias the evaluation findings.

  There is one situation in which data on nontreated entities is not needed—when an RCT is assessing impacts of different levels of program service or alternative implementation approaches. In this case, a planned variation experiment can be done where program participants are randomly assigned to a control or treatment group. The control group participants receive either a baseline level of service or implementation approach, and the treatment group participants receive a different level or approach. Here, the data for the evaluation can be collected solely from program participants. This avoids what can be a challenging and costly problem of obtaining data from entities that did not participate in the program.

  To create a control group, QEDs require data on key factors affecting selection into a program and potentially the level of outcomes and impacts. For example, firm age, size, geographic location, economic sector, and owner characteristics are needed for both treated and control groups in QEDs. They are also desirable for RCTs. Often qualified, but denied, applicants can be used to build high-quality control groups (see text box "The Value of Retaining Information on All Program Applicants Using Ranking or Scoring to Select Participants").

The Value of Retaining Information on All Program Applicants Using Ranking or Scoring to Select Participants

The Department of Energy (DOE) selects firms to receive grants from its Small Business Innovation Research (SBIR) program by ranking the applications according to a numerical scoring system. Because DOE retains information on all firms that apply (in line with its authorities and under careful data security and confidentiality safeguards), as well as information on firms receiving grants, a researcher was able to use an evaluation design that compares firms immediately around the award cutoff point. The design is called regression discontinuity design. Using these data, the researcher found strong evidence that a DOE-SBIR Phase I award of $150,000 approximately doubles a firm's chance of subsequently receiving venture capital investment, and that these DOE grants do not crowd out private capital. This research provides robust evidence about what is working well in DOE's SBIR program. Another part of this study identified the need for further research to better understand why a different part of DOE's SBIR program is not having its intended effects.

Sabrina Howell, "Financing Constraints as Barriers to Innovation: Evidence from R&D Grants to Energy Startups," 2015, <http://scholar.harvard.edu/files/showell/files/howell_innovation_finance_jmp_jan17.pdf>.

More information on the SBIR program administered by the SBA in collaboration with the 11 largest federal agencies is available at <www.sbir.gov>.

- **Baseline information on outcomes prior to providing program services**. Some evaluation approaches can use this information to compare changes in outcomes before and after treatment for both treated and control groups. See Supplemental V for an example of a before-and-after comparison.
- **Requirements for at least one period of post-service data on outcomes (e.g., jobs added, revenues increased, and new businesses established)**. Conduct an assessment of the accessibility, cost, and fitness for use of different sources of outcome data. Consider survey vs. secondary data from alternative sources, such as data held by statistical agencies. Multiple post-treatment observations are needed to estimate both short- and longer-run effects.
- **Data bias**. For example, self-reported outcome data collected directly from surveys of program participants may be subject to biases (see footnote 4). This is especially true if respondents or service providers have incentives, but no penalties, to overstate or understate their information in order to increase expected future benefits from the program. Using secondary data sources (such as those housed by statistical agencies) linked to the program's administrative data can reduce this bias. If a survey is used, a comparison with the secondary data can validate the accuracy of the survey responses.
- **Including data on services provided by other federal agencies (or other entities) to program participants and control group members**. Otherwise, changes in outcomes could be incorrectly attributed only to the program of interest, potentially biasing the estimates of that program's impacts. Identifying all assistance provided is difficult because there is no centralized data source or universal ability to link datasets (see the "Challenges in Using Secondary Data" section below).

BEST PRACTICE 5: Assess Alternative Impact Evaluation Methodologies

In deciding on evaluation methods (Randomized Control Trial or Quasi-Experimental Design), consider their different data needs against currently available data or data that could be obtained going forward. It is essential to engage evaluation experts upfront and explicitly consider the feasibility, strengths, and weaknesses of each evaluation method given available data, including information about applicants that did not receive services. Consult with evaluation experts, attorneys, and policy officers about the potential uses and permissibility of retaining data on applicants, including those that did not ultimately receive services. Rejected applicants may serve as a high-quality control group; the success of firms assisted by the program is compared with those that did not receive the help. See Best Practice 11 about informing both program applicants and participants about the use of their data for statistical research and evaluation purposes.

BEST PRACTICE 6: Collect the Indispensable Data

Administrative data needs for evaluation may include Unique and Supplemental Identifiers for applicants and participants; participant-level data on the nature, intensity, and timing of program services (i.e., the treatments); and participant and applicant characteristics (e.g., size and age of firm). Investigate the possibility of accessing secondary data, like those housed at statistical agencies, for both participants and control groups. These secondary sources can provide a broader range of data on firm characteristics, as well as a broad range of data on outcomes for both groups.

BEST PRACTICE 7: Collect Pre- and Post-Assistance Data on Impact

Access or collect pre- and post-treatment outcome data for the firms that received services and for control groups. This is central to identifying changes in outcomes (e.g., jobs, revenue, or exports) potentially attributable to the program. The data can come from a secondary data source or, if necessary, from a post-service survey of both the treated and the control groups. If surveys are the only means to collect outcome information, specify in service award conditions that post-service survey participation is a requirement for receiving assistance, and that the survey data will be used/shared to evaluate the program. In addition, having pre- and post-treatment observations at different periods for both groups is necessary to estimate both short- and long-run effects of the program (i.e., outcomes).

BEST PRACTICE 8: Identify Other Assistance Provided

Evaluations can distinguish a program's impact from the impact of other programs when they include data on related services provided by other entities to the treated and control groups. This information, if possible to collect, helps ensure that changes in outcomes are not attributed only to a single program, if services from multiple programs contributed to the change.

C. Linking to Secondary (i.e., External) Data Sources

Linking and combining program administrative data with secondary data can increase the feasibility and validity of evaluations and may reduce their implementation cost and time.[13] Secondary data are perhaps the best and least expensive tools for obtaining reliable, comprehensive data on applicant/participant characteristics and short- and long-term outcomes for both treatment and control groups. Federal agencies produce numerous secondary data sources as part of their ongoing operations. These may be their own program administrative data (i.e., collected primarily for the purpose of administering a program, usually during delivery of a service, or are required for reporting taxes); survey data (i.e., a large survey conducted by a statistical agency to produce statistics); or a combination of both (e.g., the Census Bureau's Business Register, made up of Internal Revenue Service (IRS), Social Security Administration, and BLS administrative data, supplemented by existing survey data from Census Bureau censuses and surveys). Nonfederal entities such as state and local governments may also provide very useful secondary data. Commercial vendors, such as credit bureaus, can provide transactional data. For more information about the Census Bureau's Business Register and commercial data from Dun & Bradstreet, see the text box "Examples of Secondary Data Sources With Firm-Level Information: Some Benefits and Limitations." Building a program administrative dataset that includes Unique Identifiers and learning about relevant secondary data sources can set the stage for an effective and cost-efficient impact evaluation.

[13] Linking to external data may present legal questions and considerations around participant privacy, which should be carefully considered and coordinated with general counsel and technical staff (see Chapter VII and Supplemental VI).

Examples of Secondary Data Sources With Firm-Level Information: Some Benefits and Limitations

Program agency administrative data generally only include information on program participants, or at most, applicants not served (e.g., those rejected). Data with comprehensive coverage of the relevant population can be used to identify high-quality control group samples for QED studies.

- Census Bureau data may be available for the entire life of the firm, allowing evaluators to assess both short- and long-run effects. The Census Bureau's Business Register covers close to the entire economy, including all nonfarm businesses filing taxes. The Census Bureau's Business Register data come mainly from IRS tax filings and are supplemented with data from Census Bureau censuses and surveys and from other federal agencies.
- The Census Bureau's Business Register can be used to select control groups for QED studies with linked program administrative data and Census Bureau data. The Census Bureau's Business Register can also be used to measure nonresponse bias in surveys collected for evaluation purposes (see Supplemental VII). For more information on the Census Bureau's Business Register and its use as a bridge to other relevant business data, see Bethany DeSalvo, Frank F. Limehouse, and Shawn D. Klimek, "Documenting the Business Register and Related Economic Business Data," CES Working Paper No. 16-17, 2016, <http://www2.census.gov/ces/wp/2016/CES-WP-16-17.pdf>.
- Dun & Bradstreet (D&B) is an example of a credit bureau with voluntary business enrollment. Typically, businesses enroll with D&B when they would like to attract external financing, so it is not a comprehensive list of all businesses. Therefore, it can lead to biased coverage and other measurement problems. While this dataset covers most of the U.S. economy, studies have shown that the D&B data have less complete coverage than Census Bureau data, especially among young and small businesses, and D&B does not accurately identify entry and exit. See Steven J. Davis, John C. Haltiwanger, and Scott Schuh, *Job Creation and Destruction*, Cambridge, The MIT Press, pp. 70–72, 1996, for a summary of the literature on D&B measurement issues.

D. Challenges in Using Secondary Data

Using secondary data is likely to be less expensive compared to conducting a new survey. Also, in a recent memorandum, OMB noted that linking administrative data to other data sources is consistent with the Paperwork Reduction Act, because linking minimizes the burden associated with federal government information collections.[14] However, there may be circumstances when those data cannot fully meet a program's evaluation needs (e.g., when data on an outcome of interest is unavailable). A significant challenge in using secondary data is the necessity to link the secondary data with the dataset on treated participants (and also on eligible, but not treated, applicants in an RCT). Program administrative data that do not contain sufficient identifying information to allow for timely, high-quality linkage may reduce, or preclude, the use of outside secondary data sources. In such cases, evaluation costs could be significantly greater.

[14] "Guidance for Providing and Using Administrative Data for Statistical Purposes," OMB's M-14-06, February 14, 2014.

The most efficient and reliable method for this linking process is to use Unique Identifiers, which consist of business identifying information (BII), such as employer identification numbers (EINs) or numeric identifiers from the Data Universal Numbering System (DUNS), and personally identifiable information (PII), such as social security numbers (SSNs), combined with Supplemental Identifiers, such as applicant/participant name, address, and if available, website or e-mail address (see text box "Applicant/Participant-Level Identifiers"). If Unique Identifiers are not available, the ability to link datasets is severely curtailed. Using only Supplemental Identifiers often results in a significantly lower match rate between data sources (see text box "Case in Action: Linking Data—SBA 7(a) and 504 Loan Programs"). However, Supplemental Identifiers are useful for linking even if Unique Identifiers are present. Unique Identifiers are sometimes entered in error—some businesses use multiple EINs, and data contain some missing values for the Unique Identifiers.

Applicant/Participant-Level Identifiers

i. Unique Identifiers = Unique Business Identifiers and Unique Personal Identifiers

   a. Unique Business Identifiers = EINs or DUNS.

   b. Unique Personal Identifier = SSNs.

ii. Supplemental Identifiers = Applicant/participant name, address, telephone number, and website.

   a. Business Supplemental Identifiers = Business name, address, e-mail address, telephone number, website, etc.

   b. Personal Supplemental Identifiers = Person name, address, e-mail address, telephone number, website, etc.

Note:
Business Identifying Information = Unique Business Identifiers and Business Supplemental Identifiers
Personal Identifying Information = Unique Personal Identifiers and Personal Supplemental Identifiers

Case Study: Linking Data—SBA 7(a) and 504 Loan Programs

As part of a research project with benefits to the Census Bureau, Census Bureau researchers have linked SBA 7(a) and 504 loan program recipient data to the Census Bureau's Business Register using an EIN, business name, and business address. When linking only by business name and address, 44 percent of the loans matched to the employer business register, while 79 percent matched when also including an EIN (another 7 percent matched to the nonemployer business register). Once the link to the business register was established, the data were merged with the Longitudinal Business Database (LBD), containing establishment-level data on all nonfarm businesses with payroll employment from 1976–2012. The LBD provides a comprehensive set of potential control firms, variables that can explain selection into treatment, and a long-time series before and after treatment for both treated and potential control firms. Besides recipient identifying information, the loan program data contain the program type, loan disbursement date, loan amount, and other firm, owner, loan, and lender characteristics—data which are valuable in increasing the quality of the analysis. See J. David Brown and John S. Earle, "Finance and Growth at the Firm Level: Evidence from SBA Loans," *Journal of Finance*, forthcoming, <http://ftp.iza.org/dp9267.pdf>; and J. David Brown, John S. Earle, and Yana Morgulis, "Job Creation, Small vs. Large vs. Young, and the SBA," in *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, John Haltiwanger, Erik Hurst, Javier Miranda, and Antoinette Schoar (eds.), Chicago and London: University of Chicago Press, forthcoming, <www.nber.org/papers/w21733>.

The time and costs required for accessing secondary data should be weighed against other options before embarking on an evaluation design that uses these data. For example, Census Bureau business micro datasets contain data protected by both Title 13 and Title 26 of the U.S. Code. Projects that use these datasets must be approved by both the Census Bureau and the IRS. In addition, in keeping with statutory confidentiality protections, program and statistical agencies may place restrictions on what kinds of aggregate outcome statistics can be disclosed at the end of the study. Early consultation with evaluation experts and relevant statistical agency staff can ensure that evaluations are completed in a timely manner and in accordance with all relevant statutes and disclosure avoidance protocols.

CHAPTER VI.
OTHER METHODS FOR BUILDING EVIDENCE

While the focus of this guide is on the types of administrative data needed to support impact evaluations, impact evaluation methods are not the only approach that can build evidence about a program. This chapter summarizes several other ways programs build evidence. Some of the methods discussed use the same types of data as impact evaluations, reinforcing the value of advance planning for all kinds of evidence building. The other methods summarized here have different strengths and address somewhat different questions than impact evaluations, complementing rather than substituting for impact evaluations. Using these other evidence methods can strengthen a manager's understanding of the connections between program implementation, outputs, and outcomes specified in the program theory, and they can help inform impact evaluation design. Indeed, the strongest evidence generally comes from a portfolio of high-quality studies using a variety of methods.[15] Note, however, that these other methods are not substitutes for impact evaluations in determining whether measured changes in outcomes can be attributed to the program.

A. Performance Measurement

Performance measurement and impact evaluation play a symbiotic role. While much of the same data are needed for both, they have different objectives. As defined by the Government Accountability Office, performance measurement provides "ongoing monitoring and reporting of program accomplishments, particularly progress toward pre-established goals."[16] Performance measurement occurs during and shortly after program implementation and helps managers with day-to-day decision making. Impact evaluation results provide more definitive answers to complex causal questions that are harder to answer, e.g., whether the impact would have occurred without the program.

Data collected for performance measurement enable program decision makers to understand the costs (inputs), and the activities of a program. They also can measure the direct products and services delivered by a program (outputs). Sometimes, performance data also includes the results of those products and services—but only for program participants (intermediate outcomes). With these data, performance measurement helps paint an early picture of implementation, tracking progress on steps identified in the logic model. For example, a business assistance program that provides training in writing business plans may measure the number of staff deployed to deliver services to businesses (inputs and activity), the number of businesses receiving services (outputs), and the number of businesses completing business plans after receiving training (intermediate outcomes). Because performance measurement does not track whether nonparticipants write business plans at the same rate as participants, it provides, at best, suggestive evidence on whether the program inputs and activities contribute to changes in the number of businesses writing business plans on average (intermediate outcomes).

B. Participant Surveys and Focus Groups

Agencies often survey program participants to gather information about how a program is working. Participant-only surveys can be very useful for helping program managers understand how, rather than whether, a program contributes to changes in participants' knowledge or performance. This feedback can help program managers learn which types of assistance activities are likely to be most helpful and how service delivery can be improved. Focus groups, case studies, and post-service surveys can clarify participant perceptions of key elements of the program, how success can be defined and measured from participants' perspectives, and which outcomes are most relevant for each particular service. Post-treatment surveys of just participants are also useful to gauge the degree to which participants think the service is responsible for different intermediate and long-term outcomes. This type of information can be especially useful to determine what outcomes to focus on for a subsequent impact evaluation with secondary data.

[15] See Chapter 7 of the *President's Budget for Fiscal Year 2017*, Washington, D.C., Office of Management and Budget, 2016, accessed on April 28, 2016 at <www.whitehouse.gov/sites/default/files/omb/budget/fy2017/assets/ap_7_evidence.pdf>.

[16] Performance Measurement and Evaluation: Definitions and Relationships," GAO-11-646SP, <www.gao.gov/products/GAO-11-646SP>.

Surveys and focus groups also may be a useful follow-up to an impact evaluation. For example, if an impact evaluation shows minimal or no impact, or no impact on a specific category of assisted business (e.g., service-sector businesses), this may be due to a flaw in the program theory (e.g., training about new business practices may affect productivity, but not employment) or service delivery problems. Surveys and focus groups can be used to explore whether the delivery process or program theory is flawed.

However, surveys have significant limitations. For example, a survey that excludes nontreated entities cannot provide conclusive evidence of program impact. In addition, surveys often rely on respondent perceptions, which may be subject to recall and other biases. For a discussion on the use of survey data in impact evaluations, see Supplemental I.

C. Input-Output Models

Input-output (IO) models are often used in policy development to simulate the potential total employment or income impact of a program. They also may be one of the only tools available to get an estimate of a program's potential impact when the program is new and/or in the absence of sufficient statistical power to conduct impact evaluations. It is important to realize that IO models are not intended to estimate the same effects as impact evaluation models (i.e., IO models estimate total effects that include direct, indirect, and induced effects, whereas impact evaluations usually focus on only the direct effects). Nonetheless, program managers and evaluators may use IOs in conjunction with impact evaluations.

IO models start with estimates of the direct effects of the program, and then add downstream effects that derive from the initial change. For example, the **Total Employment Impact** is comprised of the following three parts:

- **Direct Employment Impact**: The change in employment at the firm that received government services (e.g. a restaurant).
- **Indirect Employment Impact**: The employment generated from the economic activity associated with direct employment (e.g., the increased employment at firms supplying the restaurant).
- **Induced Employment Impact**: Employment generated from spending resulting from the increase in direct and indirect employment (e.g., the increased employment at local theaters to accommodate additional spending by persons newly hired by the restaurant and its suppliers).

It would be a mistake to compare estimates of program impacts obtained from an impact evaluation and an IO model if the impact evaluation only measures direct employment impacts. Nevertheless, because the quality of estimates of total employment impact depends in part on the quality of estimates of direct employment impacts, evaluations of direct employment impacts can serve an important role in verifying the reasonableness of estimates of direct employment impacts used in IO models. Furthermore, an appropriately-designed impact evaluation can also be used to measure indirect and induced employment or income impacts.

The ratio of the total employment impact to the direct employment impact is commonly referred to as the employment multiplier. Since multipliers estimated by input-output models are always greater than one and may reasonably be as large as four for some industries, the employment impact of a program will usually be underestimated if only direct employment is considered.[17] However, because IO models assume that prices are fixed and that labor supply is unlimited, they tend to overstate the total employment effect. This is a reason to use appropriately-designed impact evaluation methods to measure multiplier effects.

[17] Multipliers of less than one are possible if substitution effects or factor market constraints are taken into account, as in a computable general equilibrium model.

D. Models of Expected Impacts

Other types of economic models may be used to predict the range of expected impacts of some outcomes. These results can be helpful in estimating whether an impact evaluation is likely to have sufficient statistical power (see Chapter III and Supplemental II). Two of the most common are computable general equilibrium (CGE) models and agent-based models. In the case where a statistically powerful test is not feasible—perhaps because the critical impacts take a long time to show up in the data and the size of the effect varies substantially, or where no control group is possible, such as when evaluating a policy that is applied universally—these computational methods may provide useful forecasts of program impacts.

E. Limitations for Measuring Impact

The methods outlined in this chapter can provide some evidence of program impact, but they have real limitations. For example, it is difficult to determine if the models of expected impacts make realistic predictions. Findings based on participant surveys may be subject to reporting or perceptual biases and lack high-quality control groups, limiting confidence in whether the reported outcomes can be attributed to the program. And, the population in case studies is often too small and may not be representative of the target population, making it difficult to generalize findings to the larger population.

Suggested Additional Readings

- J. Doyne Farmer and Duncan Foley, "The Economy Needs Agent-Based Modelling," *Nature*, Vol. 460, No. 6, pp. 685–686, 2009.
- David Mulkey and Alan W. Hodges, "Using Implan to Assess Local Economic Impacts," University of Florida, IFAS Extension, 2000.
- Fedrico Pablo-Marti, et al., "MOSIPS Agent-Based Model for Predicting and Simulating the Impact of Public Policies on SMEs," in T. Gilber, M. Kirkilionis, and G. Nicolis, eds., *Proceedings of the European Conference on Complex Systems 2012*, 2013.
- Eliecer E. Vargas, Dean F. Schreiner, Gelson Tembo, and David Marcouiller, "Computable General Equilibrium Modeling for Regional Analysis," *The Web Book of Regional Science*, Regional Research Institute, West Virginia University, 1999.
- Robert K. Yin, "The Case Study Crisis: Some Answers," *Administrative Science Quarterly*, Vol. 26, No. 1, pp. 58–65, 1981.

CHAPTER VII.
OVERCOMING DATA CHALLENGES

In many cases, managers inherit a program and face certain common challenges conducting an impact evaluation, particularly if they plan to incorporate secondary data. This chapter identifies many of these challenges and offers suggestions for addressing them.

A. Create and Retain Sufficient Program Data Documentation

A good understanding of the program and how it changed over time is critical to the evaluation process. The viability of impact evaluations is limited if programs do not have sufficient documentation. Best practices for program documentation include: (1) developing and retaining sufficient documentation on the retained data, (2) building and retaining sufficient institutional knowledge about the program and its evaluation options, and (3) consistently collecting data on important program features and elements.

BEST PRACTICE 9: Create a Data Dictionary

Establish and maintain a data dictionary documenting data item definitions and changes, how data are collected (e.g., retain example forms and instructions), and relationships between key data items. Describe each data item and note the valid values/time periods. Describe any new records or revisions to existing records, including when the changes were made.

BEST PRACTICE 10: Keep Records on Program Changes

Maintain historical records detailing the program at inception and over time. This documentation may include information on the original program, as well as changes to program design, eligibility criteria, legislation, service area, factors affecting program participation, available services, services provided, and delivery methods. Ensure these records also include information on criteria used to determine when/if an applicant/participant received services. This is particularly important for business assistance programs that provide advisory services where an evaluator will want to know if assistance was quick advice or intensive consulting. Records of decision rules used to select participants for program services, such as threshold levels or scores on various criteria, are also helpful.

B. Collect and Retain Sufficient Applicant/ Participant-Specific Data

An ongoing program may not have collected the types of data that are essential or highly useful for analyzing program impacts—or if they were collected, the information may not have been electronically recorded. The importance of these key data for building rigorous evidence about what works cannot be overstated: without these items, the usefulness of program administrative data for high quality analysis is much more limited.

These problems are particularly common when services are delivered by third-party service providers, such as Manufacturing Extension Partnership grantees, Minority Business Development Centers, and Small Business Development Centers. Grantee agreements best support using administrative data for impact evaluations when they require that third-party providers collect and retain applicant/participant-level data and make them available to the grantor agency for statistical research purposes. Lack of such data collected at the applicant/participant level will make identifying results or outcomes challenging or may even render it impossible. Best Practice 5 addresses this challenge.

In some cases, federal or agency-specific laws restrict or prevent the sharing of applicant/participant-specific data with outside entities. For more information, see Supplemental VI. OMB recently directed agencies to manage data so they can better support statistical uses, which include evaluation. OMB also directed agencies to "find solutions that allow data sharing to move forward in a manner that complies with applicable privacy laws, regulations, and policies" ("Guidance for Providing and Using Administrative Data for Statistical Purposes," OMB's M-14-06, 2014).

As previously noted, using program administrative data in impact evaluations frequently requires linking program data with secondary data held by other federal agencies, and possibly data from the private sector. This process requires that program administrative data include identifiable information on program applicants and participants, as well as obtaining applicants' and participants' informed consent.[18, 19] Also, the same identifiers need to be present in both the program data and the secondary data to which they are to be linked. If no PII or BII (e.g., SSN, EIN, or DUNS number) is included in the program administrative data, it is much more difficult to link the program data to databases outside the program agency, which constrains opportunities to develop high-quality control groups. Name and address alone are often insufficient to achieve high data-linkage rates; Unique Identifiers are extremely useful and often result in higher (and highly accurate) data-match rates and lower data-linking time and costs.

Where confidentiality precludes including PII or BII in program databases, program agencies can develop alternative ways to accurately link participant information—e.g., a unique identification number created by the agency. This would facilitate intra-agency data linkage, making it possible to consolidate multiple assistance activities by the agency for the same firm.

BEST PRACTICE 11: Check if Data Can Be Shared

Consult with general counsel and use guidance from OMB's M-14-06 to determine whether applicant/participant data can be shared with researchers for statistical analysis, including impact evaluation. If the data on past services cannot be shared, work with general counsel to evaluate whether language relevant for data sharing can be changed to permit sharing data on assistance provided in the future. In rulemaking language and other settings, ensuring the protection of applicants'/participants' privacy and confidentiality is important for sustaining the ability to share and use the data for statistical analysis in a secure environment.

[18] "Informed consent refers to a person's agreement to allow data to be provided for research and statistical purposes. Agreement is based on full exposure of the facts the person needs to make the decision intelligently, including any risks involved . . . Informed consent describes a condition appropriate only when data providers have a clear choice. They must not be, or perceive themselves to be subject to penalties for failure to provide the data . . . " (G.T. Duncan, T.B. Jabine, and V.A. de Wolf, editors, *Private Lives and Public Policies*, National Research Council, 1993, available at <www.nap.edu/catalog/2122/private-lives-and-public-policies-confidentiality-and-accessibility-of-government>.

[19] Any program activity or evaluation using identifiable information should have robust confidentiality and security protocols, developed in consultation with technical or information technology staff.

BEST PRACTICE 12: Involve Attorneys and Policy Officers, Including Privacy and Confidentiality Officers, Sooner Not Later

Hold early discussions with agency attorneys and policy officers to familiarize them with the data needs for evaluation, which could include Unique Identifiers such as social security numbers (SSNs) and employer identification numbers (EINs). Make sure the use of these data is in line with relevant legal authorities and that the requirements for data security, privacy, and confidentiality are documented. Discuss sharing and linking data for statistical analysis, including impact evaluation. Ensure Privacy Notices, System of Record Notices, Privacy Impact Assessments, and other documentation adequately address data access and sharing activities. Early discussions can avoid later problems and delays that arise when data collection and sharing are treated as separate or after-the-fact considerations.

BEST PRACTICE 13: Put Data Requirements in Contracts and Grants

Ensure that grants and contracts with service providers include data collection, retention, record keeping, and data sharing provisions, as well as data security, privacy, and confidentiality protections. In the agreements, identify important data to collect and allow the data to be shared and used for statistical analysis and impact evaluations while employing robust privacy and confidentiality safeguards.

BEST PRACTICE 14: Create a Way to Link Data

Unique and Supplemental Identifiers for applicants/participants are critical to linking program and secondary data. Preferably, these identifiers can be collected, retained, and shared for statistical analysis and impact evaluation purposes, but this requires both legal and technical considerations. Review forms used to initiate assistance (intake) and service agreements to ensure adequate informed consent to share and link data. Ensure these uses are properly documented in the System of Records Notice. Additionally, review these forms to ensure sufficient Unique and Supplemental Identifiers for applicants/participants to make data linkage feasible.

BEST PRACTICE 15: If Necessary, Generate a Unique Identifier

Where confidentiality agreements preclude the collection, retention, or sharing of applicant/participant employer identification numbers (EINs) and social security numbers (SSNs), at a minimum, generate and retain a unique identification number created by the agency. This facilitates intra-agency data linking, making it possible to create a complete record of all the assistance the firm received from the agency.

C. Ensure Sufficient Data Quality

Data quality problems arise for a number of reasons, including a lack of consistent or clear data definitions and even simple typos. Data entry into agency databases is often done by multiple staff in multiple program offices, which may result in errors and inconsistencies. Quality problems also arise when records are revised without an indicator reflecting that a change has been made. Inconsistent data hamper the proper identification of who received services, the number and types of service, and the timing of services rendered. Data quality problems are one of the most common impediments to program performance measurement and impact evaluation. To improve data entry quality, data entry staff should be informed about the value and regular use of the data.

BEST PRACTICE 16: Reduce the Risk of Data Errors

Design and implement data collection systems, such as a client relationship management (CRM) system, with auto-fill or drop-down lists to minimize input errors. Provide clear guidance to database managers about the importance of entering data consistently. Implement quality control procedures to detect data entry and coding errors and inconsistencies. Include flags to identify edits and updates. Identify potential data and analysis problems with staff that create the data. Implement adequate quality control procedures early in the data collection process to avoid having to apply costly remedies to mitigate data quality problems. Demonstrate to data entry staff how the data are being used.

D. Establish Data Retention, Revision, and Security Policies

Data retention and record keeping policies can place limitations on impact evaluations. Sometimes data retention policies call for the destruction of historical data after a short period of time. Also, many agencies overwrite datasets to minimize data storage costs, not realizing the value of keeping older data. Sometimes, individual data records may be overwritten to maintain current records, but this process may destroy critical program data. For example, in instances where a service provider merges with another and undergoes a name change, some agencies replace the older provider name with the new name. In these instances, it becomes impossible to tell which provider actually rendered the assistance.

Finally, and significantly, a program agency collecting confidential data, such as applicant/participant-level data, must have appropriate data storage security in place. Specialized data security training may also be required. Appropriate data protection must be in place for any evaluation, especially those involving personally identifiable information.

How long should data be retained?

It is important to retain program data for sufficiently long periods to enable researchers to detect, with sufficient confidence, whether important program goals are being achieved. However, in accordance with laws such as the Privacy Act of 1974, data owners may have data retention policies that require the destruction of data when they are no longer necessary for their original purposes, as a way to protect client confidentiality. If program evaluations are not among the routine uses, data retention periods may be short. Short data retention periods may prevent the calculation of long-run statistics on program impact, as well as replication studies taking advantage of new methodologies and supplementary data sources developed after the completion of the initial study. Agencies should consult with evaluation experts to help identify which historical data to retain and then consult with attorneys and policy officers to determine if policy revisions and updates to the System of Records Notices can be justified. One way to address this issue for new data collections is to make sure program evaluation is listed as one of the data's program uses, among the original purposes for which the data are collected.

BEST PRACTICE 17: Consider Cost/Benefit of Data Retention

Since historical data are often critical in impact evaluations, it is important that database retention policies consider the value of retaining historical records for evaluations with confidentiality protection.

BEST PRACTICE 18: Protect the Data

Consult with lawyers and the information technology team to identify data security needs and policies. Create systems and practices that meet the needs, comply with policies, and preclude unauthorized access to data or disclosure of person- or business-specific data.

CHAPTER VIII. CONCLUSION

Increasingly, government agencies are called upon to use rigorous impact evaluations to promote learning about what works in government programs and use the evidence to continually improve programs to achieve better outcomes. And, they are asked to do so at least cost and burden to taxpayers. Agencies are responding by looking for new ways to utilize program administrative data and secondary data sources for impact evaluations, thereby reducing reliance on surveys when possible. This guide can serve as a practical tool to help agencies identify important data-related practices and the critical data that need to be collected and retained. This will allow agencies to effectively use their administrative data for rigorous impact evaluations. While focused primarily on the data needs for evaluating business technical assistance programs, the vast majority of the recommended data practices will be useful in building other types of evidence.

In some cases, modifying program administrative data collections to be more useful for impact evaluations may entail substantial effort. Nevertheless, the redesigned program administrative data can substantially lower the cost of future evaluations. In each subsequent year, evaluators can add an additional year of data and update evaluation findings at relatively little cost. This is far less expensive than conducting a new post-service survey every time an impact evaluation is needed. Program agencies are best able to capitalize on these savings when they ensure impact evaluations are sufficiently documented and when input is obtained from key stakeholders about how best to improve future evaluations.

SUPPLEMENTAL I. EXAMPLE DATA LISTS

This supplemental provides examples of data items that are often collected by program agencies in the course of program administration, data items that can be found in secondary data collected by other agencies or commercial data vendors, and types of data items that may only be available via post-service surveys (of both participants and nonparticipants) that can be helpful in an impact evaluation of business assistance programs. Below are three tables of data items, separated by source: data items found in administrative data, secondary data, and from surveys. The lists are not meant to be exhaustive, and some data items are not applicable to all programs. The lists expand upon data concepts highlighted in Chapter II and Supplementals IV and V. In addition to the microlevel data listed here, it is also important to collect information on program objectives, how the program is administered, and program costs.

Applicant and participant identifying information is listed in all three tables. As noted in Best Practice 14, unique identifying information is critical for linking program administrative data records to secondary data (statistical agency, commercial vendor, and other sources) and survey data. In addition, applicant and participant characteristics such as industry codes can facilitate quality linkage when multiple businesses have the same name or address. The higher the quality and quantity of identifying information and characteristics, the greater the share of records that can be linked with confidence. The absence of this information limits or possibly closes off use of secondary sources for evaluation. **Before designing any program data collection (or making changes to existing data collections), it is advisable for program agencies to work with evaluation experts, legal experts, and statistical agencies or commercial data vendors to ascertain what secondary data could be used for evaluation (see p. 28).**

Examples of Impact Evaluation-Relevant Data Found in Program Administrative Data

The program administrative data listed in Supplemental I: Table 1 include information on both applicants for program assistance and the clients actually receiving services (participants). In addition to information on program participants, it can often be helpful to retain as much information as possible, within all relevant legal authorities and while implementing robust privacy safeguards, on applicants that do not receive services. They may be good candidates for control groups that are central to a rigorous impact evaluation.

Examples of Impact Evaluation-Relevant Data Found in Secondary Data Sources

Secondary data are useful for selecting "control" firms in an impact evaluation and obtaining a variety of pre- and post-treatment outcomes on both participant and control firms (Supplemental I: Table 2.). Obtaining information on outcomes both pre- and post-treatment is particularly useful for benchmarking purposes.

Supplemental I: Table 1.

Program Administrative Data

Applicant/Participant Identifying Information
Applicant/participant Unique Identifiers (e.g., SSN, EIN, or DUNS number, in separate fields if more than one)
Applicant/participant name, street number, street name, city, state code, zip code, telephone number, e-mail address, and website address (in separate fields)
Service Provider Identifying Information
Service provider Unique Identifier (e.g., SSN, EIN, or DUNS number)[1]
Service provider name, street number, street name, city, state code, zip code, and website (in separate fields)
Service provider—service rendered location—Unique Identifiers (e.g., SSN, EIN, or DUNS number)
Service provider—service rendered location—name, street number, street name, city, state code, zip code, and website (in separate fields)
Other Applicant/Participant Information
Applicant/participant (individual): credit scores, individual identifiers (SSN), gender, race, ethnicity, disability status, military/veteran status, age, years of education, and years of employment
Applicant/participant (firm): firm date of incorporation, employment levels (hours worked or numbers), annual receipts, annual profits, net worth, sources of credit received independently of business assistance programs, primary industry (NAICS code), legal form of organization, credit score, application score, number of owners, and owner Unique Identifiers (SSN)
Applicant's forecast about number of jobs created and number of jobs retained as a result of the assistance*
Nature, Intensity, and Timing of Treatment (Logic Model "Activities")
Type of service requested (e.g., a specific loan or grant, or entrepreneurial training in business plans, accounting, marketing, legal issues, logistics, partnerships, supply chain systems, or exporting)
Dollar value of requested service
Referral type for requested service
Duration of requested service
Date service request made
Type of service delivered (see above list for type of service requested)
Dollar value of delivered service
Duration of service delivered
Date of service delivery
Outputs Data (Logic Model "Outputs")
Training program: number of persons trained
Loan or grant program: number of loans (grants) disbursed and total dollar amount of loans (grants) disbursed

*Jobs created or retained information is obtained at the time of application for assistance. As discussed below, this may suffer from reporting bias, and forecasts are less reliable than realized outcomes. It would be preferable to use actual pre- and post-treatment employment numbers from secondary data to measure employment change.

[1] Service provider information is only necessary for programs that fund third-party providers that render services to the agency's clients.

Supplemental I: Table 2.

Secondary Data

Applicant/Participant Identifying Information	
Applicant/participant Unique Identifiers (e.g., SSN, EIN, or DUNS number, in separate fields if more than one)	
Applicant/participant name, street number, street name, city, state code, zip code, telephone number, e-mail address, and website address (in separate fields)	
Pre- and Post-Treatment Outcome Data (Logic Model "Intermediate and Long-Term Outcomes")	
Outcome being evaluated	**Example data items**
Changes in workers	Number of payroll employees, number of nonpayroll employees, payroll, compensation of nonpayroll employees, and demographics and earnings of each employee
Changes in business size and scope	Annual sales, number of establishments, and industry codes of the establishments[1]
Change in productivity	Value of annual output (preferable) or sales, annual hours worked, value of capital inputs, and value of material inputs
Change in financial performance	Profits, net worth, short-term debt, long-term debt, loan defaults, and credit rating
Business formation	Year of business entry
Business survival	Year of merger or acquisition, year of bankruptcy, and year of business exit
Changes in exports	Total value of exports and value of exports by destination country
Innovation	Number of patents, number of trademarks, number of trade names, R&D spending, and number of science and technology employees
Information About Similar Services Provided by Other Business Assistance Entities to the Applicant or Participant	
Name of entity providing service	
Type of service delivered (see above list for type of service requested)	
Dollar value of delivered service	
Duration of service delivered	
Date of service delivered	
Data on Factors Influencing Eligibility/Selection Into Treatment and Post-Treatment Outcomes	
Primary industry (NAICS code); year of entry; credit rating; number of employees; annual sales; net worth; geography; and owner employment history, race, ethnicity, and citizenship	

[1] The industry codes of the establishments can capture changes in the scope of the business.

Examples of Impact Evaluation-Relevant Data Collected Via Post-Treatment Surveys

Surveys are particularly useful for collecting outcome data not typically available in secondary sources, such as immediate outcomes of a business technical assistance program like completing a business plan. Surveys are used even when the information needed is available from secondary sources, because surveys may provide more immediate feedback (e.g., participants can report changes in numbers of employees after receiving services, while post-treatment employment data from secondary sources may become available only after several months or perhaps years). To be useful in evaluating the impact of program services (versus no service), a survey would need to obtain information from both participants and a nonparticipant control group. This requirement distinguishes these surveys from other types of surveys of program participants that agencies commonly conduct—such as post-treatment surveys of participants only.[1] Since participant-only surveys by definition exclude control entities, such surveys are generally not suited for measuring program impacts.

However, research has shown that self-reported survey data may be less accurate than administrative data.[2] It can also be costly to collect survey data on a large enough sample to achieve satisfactory statistical power (see Supplemental II), and identifying good control group businesses to survey is challenging. Program agencies may not have access to an up-to-date and accurate business list that has sufficient detail and coverage to support the selection of control businesses with characteristics similar to those of the program participants. While surveys are an important tool in the evidence-building toolbox for other reasons (see Chapter VI), the weaknesses in the quality of survey data are one of the main reasons agencies are increasingly assessing the merits of using secondary data sources for impact evaluations. However, comparisons of outcomes reported in survey data to secondary data sources have identified cases where reporting biases are minimal. Expert survey design can increase the accuracy of self-reported data.

Supplemental 1: Table 3 lists basic identifying information that is included in a survey. It also provides examples of the types of outcomes that are difficult or impossible to analyze using administrative data in combination with secondary data—conducting a survey may be the only feasible data collection method for testing these types of outcomes.

[1] Nonetheless, such participant-only surveys can be useful for learning about other aspects of program operation, such as learning which activities participants think contribute the most to different intermediate and long-term outcomes; see Chapter VI.

[2] See, for example, Bruce D. Meyer, Wallace K. C. Mok, and James X. Sullivan, "Household Surveys in Crisis," *Journal of Economic Perspectives*, Vol. 29, No. 4, pp. 199–226, 2015, who find that survey respondents systematically underreport the amount of public assistance they have received.

Supplemental I: Table 3.

Survey Data

Applicant/Participant Identifying Information
Applicant/participant Unique Identifiers (e.g., SSN, EIN, or DUNS number, in separate fields if more than one)
Applicant/participant name, street number, street name, city, state code, zip code, telephone number, e-mail address, and website address (in separate fields)
Nature, Intensity, and Timing of Treatment
How and when service was delivered, intensity of engagement with the program, and type of service
Applicant/Participant Characteristics
Applicant/participant(s') age, race, ethnicity, gender, disability status, military/veteran status, and years/type of education
"Control" Entity Identifying Information
Control entity Unique Identifiers (e.g., SSN, EIN, or DUNS number, in separate fields if more than one)
Control entity name, street number, street name, city, state code, zip code, telephone number, e-mail address, and website address (in separate fields)
"Control" Entity Characteristics
Control entity age (if an individual) or date of incorporation, race (owner's race if a business), ethnicity, gender, disability status, military/veteran status, and years/type of education

Post-Treatment Outcome Data (Logic Model "Immediate, Intermediate, and Long-Term Outcomes")

Outcome being evaluated	Example data items
Business plans	Number of completed business plans in a specific time frame (months, quarters, etc.)
Loan applications	Number of loan applications submitted in a specific time frame (months, quarters, etc.)
New suppliers	Number of contacts made with potential new suppliers in a specific time frame (months, quarters, etc.)
New customers	Number of contacts made with potential new customers in a specific time frame (months, quarters, etc.)
Supply network	Number of new suppliers used in a specific time frame (months, quarters, etc.)
Customer network	Number of new customers in a specific time frame (months, quarters, etc.)
New markets	Amount of sales in new markets in a specific time frame (months, quarters, etc.)
New products	Amount of sales of new products in a specific time frame (months, quarters, etc.)
Capital access	Amount of financing received in a specific time frame (months, quarters, etc.)
Technical networks	Number of new persons in technical network in a specific time frame (months, quarters, etc.)
Business networks	Number of new persons or firms in business network in a specific time frame (months, quarters, etc.)
Environmental efficiency	Percentage change in NOx emissions in a specific time frame (months, quarters, etc.) relative to a previous period of the same timeframe
Energy efficiency	Percentage change in electricity usage in a specific time frame (months, quarters, etc.) relative to a previous period of the same timeframe
Change in financial access	Source of loan, loan amount, loan term, loan monthly payment, loan origination date*
Change in credit history	Current loan status, loan default amount, loan default date, loan paid in full date*
Change in contracting performance	Number and dollar value of contracts submitted (awarded)*

*Some programs collect these data on participants as part of their program administration. In such cases, the survey's role would be to collect this information on control groups, as well as participant loans and contracts not part of the regular program administrative data collection.

SUPPLEMENTAL II. QUESTIONS TO DISCUSS WITH EVALUATION EXPERTS ABOUT A PROGRAM'S SUITABILITY FOR IMPACT EVALUATION

Agencies interested in using administrative data to analyze the impact of programs may determine that additional information will need to be collected. Before agencies embark on revising administrative data collections to gather information needed for an impact study, it is helpful to assess whether the resulting data (existing administrative data plus new data collected) will be sufficient to support high-quality impact evaluations.[3] An important part of this assessment is determining whether—even with the best data at hand—program impacts are likely to be detected, if they exist, and that program impacts won't be detected if they don't exist. This involves ensuring that an evaluation would have high statistical power. Statistical power is the probability that an evaluation will detect an effect when there is an effect to be detected. Weak statistical power limits the informative value of the findings. Evaluation experts can be very helpful in making this determination.

To understand the importance of assessing the likely statistical power of a prospective evaluation, it is helpful to consider that every impact evaluation has four possible outcomes:

Outcome 1: The program produces the desired impact, and the evaluation detects it.

Outcome 2: The program has little impact, and the evaluation correctly fails to detect a substantial impact.

Outcome 3: The program produces the desired impact, but the evaluation fails to detect it.

Outcome 4: The program has little impact, but the evaluation erroneously detects a substantial impact.

Outcomes 1 and 2 provide reliable evidence of program impacts and can be used to inform agency decisions. Outcomes 3 and 4 provide erroneous evidence of program impacts that, if used, may lead to poor decisions.

[3] The focus here is on impact evaluation. This discussion, however, should not be seen as dismissing other reasons to collect program administrative data, such as to support performance measurement.

Careful design of an evaluation can minimize the likelihood of outcomes 3 and 4, so it is important that design issues are fully resolved before compiling or collecting data. Factors associated with outcomes 3 and 4 are discussed in turn below.

Outcome 3 can be avoided by ensuring that the evaluation has high statistical power.

The three fundamental determinants of statistical power are:

1. **Mean effect size**: How big is the expected impact, on average, that you are seeking to measure? If the average effect size of a program is thought to be small, then detecting an effect will require a larger sample size than a program with a substantial anticipated effect.
2. **Variance in effect size**: How much heterogeneity is expected in impacts across program participants? If the effect of the program across clients is expected to be highly variable, then this would also require a larger sample size relative to a program where effects are thought to be more uniform.
3. **Sample size**: How many participants will be studied? Establishing the maximum possible sample size requires an assessment of the data likely to be available for an evaluation—the number of evaluation participants with complete data and which have controls with complete data. If effects are to be estimated for subsets of the program (e.g., separately by service type, by service provider, or by different client types), estimate the sample size for each of these subsets. Note that linking and combining data from programs with secondary data at federal statistical agencies may increase the sample size, thereby increasing statistical power and improving the reliability of an evaluation.

Past evaluations of the program, evaluations of other similar programs, case studies of this program, or other client feedback can be used to obtain estimates of the mean and variance of the effect size.

Outcome 4 can be avoided by accepting only a stringent (high) significance level to minimize the chance that a significant effect is due to sampling error. There may also be more complex considerations such as reverse causality that may contribute to false positives. Evaluation experts can help assess these more complex considerations.

Is it possible to obtain statistically valid results from a pilot or new program with a relatively small number of clients?

A pilot program with a relatively small number of clients may not provide a large enough sample size to support a statistically powerful test. In these circumstances, the evaluation may have a low probability of detecting a program impact even if it exists. However, statistically powerful tests with small pilot programs are possible if the average anticipated effect of the program is large and if the effects are thought to be relatively uniform. For example, a pilot program to assess whether school performance is improved by supplying corrective lenses to school children with poor eyesight may support a statistically powerful test with a relatively small number of study participants. In contrast, a pilot program directed to increasing exports by providing technical assistance may be difficult to evaluate with a small sample given the vicissitudes of international business.

Suggested Additional Readings

- Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NJ, Lawrence Erlbaum Associates, Inc., 1977.
- Timothy R. Wojan, Jason P. Brown, and Dayton M. Lambert, "What to Do About the 'Cult of Statistical Significance'? A Renewable Fuel Application Using the Neyman-Pearson Protocol," *Applied Economic Perspectives and Policy*, Vol. 36, No. 4, pp. 674–695, 2014.

SUPPLEMENTAL III. IMPACT EVALUATION—KEY CONCEPTS

An impact evaluation uses statistical methods and data to estimate the impacts of a program or implementation approach, including estimation of the degree of confidence that can be assigned to the estimated impacts. *Impacts* of a program refers to the differences in the distribution of outcomes (usually the focus is only on differences in the mean outcomes) between what actually occurred and what would have occurred in the absence of the program (when evaluating an entire program). Alternatively, the distribution of outcomes may be compared between a baseline implementation approach and a new implementation approach. Since what would have happened in the absence of the program or under an alternative implementation approach (the *counterfactual*) is not observable, some method is necessary for estimating it. Impact evaluation uses statistical methods to estimate the counterfactual situation.

Types of Validity in an Impact Evaluation

The need to compare observed outcomes to an unobserved counterfactual is one of the major challenges to the *validity* of impact evaluations. *Validity* refers to the extent to which a statement is logically consistent and empirically supported. The validity of an impact evaluation refers to the extent to which its conclusions are logically consistent and supported by the data. A *high quality* impact evaluation is one that overcomes threats to the validity of the conclusions to the maximum extent possible.[4]

Two broad types of validity are discussed in the impact evaluation literature—internal validity and external validity. *Internal validity* refers to the approximate validity with which we can infer that a causal relationship exists or doesn't exist between two variables (Campbell and Stanley, 1963).[5] We want to know whether implementation of a program (P) causes changes in some outcome measure (Y); for example, whether a business assistance program causes firms to be more successful. *External validity* refers to understanding how representative the impacts found are likely to be for other settings, populations, and time periods (Campbell and Stanley, 1963). For example, we would like to know whether the relationship between P and Y found in one impact study is likely to be true for other ways of measuring Y, or for other populations or time periods.

A third type of validity to consider is *construct validity*, which refers to how clearly defined the underlying construct is, and how well the measures used in a study represent the underlying construct. In order to assess how the program P affects the set of outcomes Y, both P and Y have to be defined and operational measures for these constructs developed. For a business assistance program, what exactly is *the program*? Depending on how *the program* is defined, it may have broader or more targeted impacts, and the potential for conducting an impact evaluation may be greatly affected. For example, a narrow program construct would be a program that strictly provides businesses with marketing research, with an intended outcome of sales in new markets. A broader construct would be providing various types of technical assistance to increase the volume of business and employment levels. The broader construct is more difficult to define and measure.

Some outcome indicators may be highly correlated, suggesting that they reflect a common underlying construct (e.g., plant expansion and increased employment), while others may be less correlated and represent different constructs (e.g., stock price increase and employment levels). A clear understanding of the underlying constructs (cause and effect relationships) that the measures represent is necessary to be able to generalize knowledge from the evaluation.

[4] There can never be 100 percent certainty that the conclusions of an impact evaluation are true. The best that can be hoped for is that the conclusions are unbiased (the estimated impact equals the actual impact in expectation) and that the degree of uncertainty about the estimated impacts can be kept "small" relative to the size of impact that one seeks to detect.

[5] Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, Chicago, IL: Rand-McNally, 1963.

SUPPLEMENTAL IV. RANDOMIZED CONTROL TRIALS AND THEIR DATA NEEDS AT A GLANCE

A *randomized experiment* (also known as a randomized control trial) is an experiment in which the treated/assisted group and the nontreated group (the control group) are selected by some random process.

A *Randomized Control Trial (RCT)* is often considered the method that most closely approximates the counterfactual (the outcome a treated firm would have achieved without treatment). RCTs often have the strongest design and provide the most definitive evidence of impact. However, they are not feasible in some situations, such as retrospective analysis of impacts in ongoing programs. There is thus no single "gold standard" impact evaluation method that is appropriate for all contexts. The most useful and valid evaluations will use the tools best suited to the situation and often include combinations of different methods.

Leveraging Secondary Data in RCTs

- Leveraging secondary data for an RCT reduces the need for new data collection efforts, a major contributor to the required time and expense of evaluations.
- Example: The Department of Labor and the Small Business Administration evaluated whether a demonstration pilot called Project GATE helped improve participants' well-being and resulted in business creation by using an RCT that leveraged secondary data.
  - Applicants were randomly assigned to either the treatment group that received program services or a control group that did not receive services.
  - Secondary data, such as quarterly wage records and unemployment insurance, were used in the analysis.
  - To read the full report on this evaluation, see Jacob Benus, Theodore Shen, Sisi Zhang, Marc Chan, and Benjamin Hansen, "Growing America Through Entrepreneurship: Final Evaluation of Project GATE," 2009, available at <http://wdr.doleta.gov/research/FullText_Documents/Growing%20America%20Through%20Entrepreneurship%20-%20Final%20Evaluation%20of%20Project%20GATE.pdf>.

Benefits of RCT Approaches

Random selection ensures no systematic differences between treatment and control groups in factors that may affect outcomes (e.g., entrepreneurial ability and ambition). In other words, randomly selecting who participates in the program avoids the introduction of biases. If program participants are selected based on their past business success, it is not clear if that selection factor or the program assistance caused a positive outcome. Random selection reduces the chances that factors other than the program caused the observed impacts. RCTs are widely viewed as the best means of addressing differences between the characteristics of program participants and nonparticipants that may affect outcomes.

Limitations and Potential Issues

A common objection to using an RCT approach is a concern with denying assistance to those in the control group. However, nearly all programs face budget constraints and can only enroll a certain number of businesses each year. In these cases, determining which businesses are eligible and then randomly selecting from that set of eligible businesses is a defensible selection method. No business gets special priority provided eligibility requirements are met.

RCTs are widely viewed as the best means for assuring "internal validity," (or whether a causal conclusion can be justified), because they generally address *selection bias* (differences in the characteristics of program participants and nonparticipants that affect outcomes) more completely than other methods. However, RCTs can result in other biases, such as changing the nature of the pool of firms that would have received the treatment, thus changing the impacts of the program ("randomization bias"); causing members of the control group to seek substitutes for the program ("substitution bias"); spillovers of the program effects onto nonparticipants; and nonresponse and attrition. Other limitations may include RCT feasibility, cost, and ability to answer some important questions. Note that these limitations also can be issues in other evaluation methods. Supplemental IV: Table 1 summarizes the data requirements for RCTs.

Supplemental IV: Table 1.

Data Requirements for Randomized Control Trials

Types of data needs[1]	Status
Nature, intensity and timing of treatment	Required
Information about similar services provided by other agencies[2]	Required
Participant identifying information[3]	Required
Participant characteristics	Desirable
Post-treatment outcome data	Required
Pretreatment outcome data	Desirable
Multiple pre- and post-treatment outcome observations	Desirable
Frame for choosing controls	Required
Data on factors influencing both selection into treatment and post-treatment outcomes	Desirable
Data on factors influencing selection into treatment, but no direct effect on post-treatment outcomes	Unnecessary
Data on variables determining eligibility for treatment	Unnecessary

[1] See Supplemental I for examples of the specific data items.

[2] If it is likely that a program's participants received other services (even if receipt of such services cannot be confirmed), it would be advisable to indicate in the evaluation report that other programs may have contributed to the estimated impact.

[3] Client-level identifying information (e.g., name, address, EIN/DUNS) is only required if the evaluation plan will include linking program data with other program or secondary data sources.

Suggested Additional Readings

- Vetan Kapoor, Michael Taylor, and Ana Boltik, "How Low-Cost, Lightweight RCTs Can Improve Program Evaluation," 2014, available at <http://www2.deloitte.com/content/dam/Deloitte/us/Documents/public-sector/us-fed-4-lowcostlightw-rcts-final-12122014.pdf>.
- Corporation for National & Community Service, "Evaluation Plan Guidance: A Step-by-Step Guide to Designing a Rigorous Evaluation," 2013, available at <www.nationalservice.gov/documents/social-innovation-fund/2014/social-innovation-fund-evaluation-plan-guidance>.
- Coalition for Evidence-Based Policy, "Checklist for Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence," 2010, available at <http://coalition4evidence.org/wp-content/uploads/2010/02/Checklist-For-Reviewing-a-RCT-Jan10.pdf>.

SUPPLEMENTAL V. QUASI-EXPERIMENTAL EVALUATION DESIGNS AND THEIR DATA NEEDS AT A GLANCE

A *Quasi-Experimental Design (QED)* is an alternative evaluation design in which the treated (assisted) group and the control group are as similar as possible on salient factors, but with the critical exception that the treated group is provided with some form of assistance or intervention. QED methods are particularly useful for retrospective analysis of programs that are already underway, when it is not possible to randomly assign individuals or businesses to treatment or control groups.

Despite the lack of randomization, a carefully designed QED may still permit strong conclusions regarding impact. Researchers using a QED approach use different methods to address the risk that the treatment and control groups vary in ways other than whether program services were received or not. Different types of quasi-experimental methods can be combined, which may result in an even more rigorous evaluation approach.

QEDs still require data to understand the group that is assisted and the control group; the nature, intensity, and timing of treatment; the program outputs; and pre- and post-treatment outcomes relevant to the program's goals. More information on different types of QEDs follows along with a table that summarizes data needs for the different approaches.

Contrasting Randomized Control Trials (RCTs) and QEDs

The fundamental difference between these two methods lies in how treatment and control groups are formed. In an RCT, selection into treatment versus the nontreated control group is determined solely by random choosing from among applicants. Thus, selection into treatment and control groups is fully *exogenous* to entities' choices to apply to the program and to nonrandom selection for treatment among the applicants. There are not likely to be differences in the characteristics of program participants and nonparticipants that will affect outcomes (i.e., "selection bias")—the only systematic difference between the groups is that the treated group received services (treatment).[6]

QED methods are generally used in situations where the treated group is already formed. The treatment group in a QED study is not formed by a random process, but is based on the decision to apply and applicant qualifications.[7] The control group is chosen for comparability with the treated group on key characteristics. Thus, selection into treatment and control groups is *endogenous* to the program.

This difference matters, because simply selecting a control group in a QED study from nonparticipating entities may involve "selection bias" (entities either applied and were rejected or did not choose to apply): self-selection implies that participants (the treated group) may differ from nonparticipants (the control group) in several ways, not just whether treatment has been received or not. Thus, QEDs must carefully form control groups in ways that avoid selection bias.

- **Hypothetical example**: A business technical assistance program aims to provide entrepreneurial services to minority business owners to help expand their supply networks. However, those choosing to participate in the program (the treated group) have fewer business connections than those who choose not to participate. If business connections are unobservable to the researcher, but they influence the outcome being assessed (e.g., the amount by which their supply networks expanded), an evaluation that did not consider this difference between program participants and nonparticipants is likely to result in a biased estimate of the impact of the program, and the program may appear less effective than it actually is.

[6] This assumes that those randomly selected for treatment and control groups decide to participate in the program and the study. If some do not participate, or they stop participating after some time, there can be nonresponse bias or attrition bias. These are forms of selection bias resulting from nonrandom decisions to not participate or to exit the study, despite an initial random assignment. Addressing these issues requires similar data to those collected by QEDs to address the selection bias issue.

[7] Bias from the decision to apply can be removed by using rejected applicants as the control group, though differences in applicant qualifications will remain.

QED Methods Typically Used in Impact Evaluations

Regression Discontinuity Designs

If a program uses a scoring or ranking system to determine eligibility for a program, then a regression discontinuity design can be considered as a possible evaluation protocol. This approach uses the scoring or ranking system to set up viable treatment and control groups immediately above (the treated group) and below (the control group) the eligibility threshold. By focusing the analysis on those observations on either side of the threshold (i.e., subsets that are likely to be most similar, except that those above the threshold received the treatment), it may be possible to estimate a local average treatment effect. The identifying assumption is that the small differences in scores or ranks between the two groups do not affect the outcomes substantially other than via treatment receipt, so treatment can be thought of as being randomly assigned across the entities in the two groups.

- **Hypothetical example**: A study estimates the impact of a loan on entrepreneurs' performance, where the loan is awarded only to entrepreneurs with credit scores of 700 or above. Since individuals scoring 700 or above would be expected to perform better on average than those scoring below if no loans were given to anyone or if loans were given to all, a simple comparison of all loan recipients versus all nonrecipients would provide little information on the effect of the loan.

  Instead, the comparison is made among those who score very near 700. The treated group might be those scoring 700–720, and the control group has scores between 679–699. One might expect little difference in outcomes across these two groups if no loans were given to either group or if loans were given to all. Comparing outcomes across these two groups would thus provide information on whether the loan is associated with improved entrepreneurial performance for recipients who have credit scores near 700.

  Benefits: This method can be straightforward to implement, and it is valid if the reasons for establishing the specific threshold are compelling.

  Limitations: The method assumes that the ranking is applied consistently and captures the essential features associated with outcomes. If there are too few clients on either side of the eligibility threshold, this evaluation design would lack statistical power and may not yield conclusive results of impact. Also, the estimated impacts are not necessarily applicable to clients with scores further from the threshold, e.g., 600 or 800.

  Data needs: This method requires data about the application process and on all applicants (treated and nontreated). It is important to consider the need for data on applicants, particularly when developing the Federal Funding Opportunity (FFO) notice or launching a program. Applicants will need to be notified in the FFO that the information on their application will be used for program evaluation.

Difference-in-Differences Estimation

Difference-in-differences estimation essentially requires comparison of four outcome data points. Outcome data are collected for both program participants and a control group before and after program implementation, when it is believed impacts from the program have been realized. The effect of the treatment is the observed change in the treatment group minus the change in the control group (the "normal" difference).

- **Real-world example**: Holzer et al. (1993),[8] estimate difference-in-differences regressions to assess whether the receipt of employee job training grants affects output quality, sales, employment, and wage levels in manufacturing firms in Michigan. They compare grant recipient firms' changes in quality, sales, employment, and wage levels after versus before assistance receipt to nonrecipient applicant firms' changes over the same time period. The authors estimate that output quality

[8] Harry J. Holzer, Richard N. Block, Marcus Cheatham, and Jack H. Knott, "Are Training Subsidies for Firms Effective? The Michigan Experience," *Industrial and Labor Relations Review*, Vol. 46, No. 4, pp. 625–636, 1993.

and employment increase after grant receipt, but they do not find statistically significant sales or wage effects.

Benefits: The method is intuitive and straightforward to implement. If outcome data can be constructed from large, secondary administrative datasets, the evaluation may be very low cost and empirically robust.

Limitations: Data collection may be required prior to program implementation if secondary data do not capture the outcome variables of interest. Whether findings are compelling depends on the reasonableness of the assumption that the treated and control groups were following similar paths prior to the intervention (called the "parallel trends" assumption).

Data needs: Outcome data measured prior and after treatment for both treated and control firms are necessary for this method.

Matching Estimators

The matching estimator method uses various business characteristics (e.g., pretreatment outcomes, firm size, owners' demographic status, and firm geography) to predict the probability of an outcome occurring. Participants are matched to nonparticipants with similar predicted treatment probabilities. This method can be used to address potential selection bias. Such matching estimators are referred to as propensity score estimators.

Exact matching is sometimes done using characteristics thought to be particularly influential to treatment propensity and subsequent outcomes, such as firm age, industry, or firm geography. Matching within some range on pretreatment outcomes (e.g., employment, sales, or productivity) can also be used. The logic is simple: if the correlation in pretreatment outcomes of matched pairs is close, then any observed post-treatment difference between treated and control pairs can be attributed to the treatment. Combining matching with difference-in-differences estimation can be a particularly effective strategy.

- **Real-world example**: Regional development programs, such as the Appalachian Regional Commission, the Delta Regional Authority, and the Tennessee Valley Authority are examples where selection into the program is based on geography.[9] These regional authorities were established to address the poor economic performance of disadvantaged areas. Thus, simply comparing their performance to other regions would likely bias results.

  To overcome potential selection bias, in a study on the impact of the Delta Regional Authority (DRA), Pender and Reeder (2011) used a quasi-experimental matching approach to select control counties. This approach entailed matching individual counties within the regional authority to similar counties elsewhere in the same region as well as in the Southeast. The criteria used to assess good matches were the similarity in trajectory of critical economic indicators prior to the start of the program (parallel trends). Counties performing similarly with respect to employment growth, earnings growth, poverty rates, and other indicators provided plausible "control counties" for their regional authority counterparts.

  Limitations: This method requires that observable characteristics do a good job of explaining program participation and that potential controls that closely resemble those participating in the program are available. If one were evaluating a program providing loans to low-income, blind entrepreneurs in rural communities, it might be difficult to identify a large enough group of similar (low-income, rural, blind entrepreneurs) who did not participate in the program to serve as a control group. In addition, this approach assumes that unobservable differences between participants and nonparticipants do not affect outcomes.

Supplemental V: Table 1 summarizes the data requirements for QEDs.

---

[9] For studies on the these programs see Andrew Isserman and Terance Rephann, "The Economic Effects of the Appalachian Regional Commission: An Empirical Assessment of 26 Years of Regional Development Planning," *Journal of the American Planning Association*, Vol. 61, No. 3, pp. 345–364, 1995; John Pender and Richard Reeder, *Impacts of Regional Approaches to Rural Development: Initial Evidence on the Delta Regional Authority*, USDA-Economic Research Service ERR-119, p. 73, 2011, available at <www.ers.usda.gov/webdocs/publications/err119/7407_err119.pdf>; and David Freshwater, Timothy R. Wojan, Dayan Hu, and Stephan Goetz, "Testing for the Effects of Federal Economic Development Agencies," TVA Rural Studies Working Paper 97-02, University of Kentucky, 1997, accessed April 27, 2016, at <www.uky.edu/Ag/AgriculturalEconomics/pubs/tvaFreshwater97-02.pdf>.

Supplemental V: Table 1.

Data Requirements for Common Quasi-Experimental Design Methods

Types of Data Needs[1]	Regression Discontinuity Designs	Difference-in-Differences	Matching Estimators
Nature, intensity, and timing of treatment	Required	Required	Required
Information about similar services provided by other agencies[2]	Required	Required	Required
Participant identifying information[3]	Required	Required	Required
Participant characteristics	Desirable	Desirable	Desirable
Post-treatment outcome data	Required	Required	Required
Pretreatment outcome data	Desirable	Required	Required
Multiple pre- and post-treatment outcome observations	Desirable	Desirable	Desirable
Frame for choosing controls	Required	Required	Required
Data on factors influencing both selection into treatment and post-treatment outcomes	Unnecessary	Desirable	Required
Data on variables determining eligibility for treatment	Required	Unnecessary	Unnecessary

[1] See Supplemental I for examples of the specific data items.

[2] If it is likely that a program's participants received other services (even if receipt of such services cannot be confirmed), it would be advisable for the evaluation report to indicate that other programs may have contributed to the estimated impact.

[3] Participant identifying information (e.g., name, address, EINs/DUNS) is only required if the evaluation plan will include linking with other program or secondary data sources.

Suggested Additional Readings

- Coalition for Evidence-Based Policy, "Which Comparison-Group ("Quasi-Experimental") Study Designs Are Most Likely to Produce Valid Estimates of a Program's Impact?: A Brief Overview and Sample Review Form," 2014, available at <http://coalition4evidence.org/wp-content/uploads/Validity-of-comparison-group-designs-updated-Feb-2012.pdf>.
- Christopher Ordowich, David Cheney, Jan Youtie, Andrea Fernandez-Ribas, and Philip Shapira, "Evaluating the Impact of MEP Services on Establishment Performance: A Preliminary Empirical Investigation," CES Working Paper 12-15, 2012, available at <http://www2.census.gov/ces/wp/2012/CES-WP-12-15.pdf>.
- Philipp Brandt and Josh Whitford, "Fixing Network Failures? The Contested Case of the American Manufacturing Extension Partnership," *Socio-Economic Review*, forthcoming.
- Kenneth P. Voytek, Karen L. Lellock, and Mark A. Schmit, "Developing performance metrics for science and technology programs: The case of the manufacturing extension partnership program," *Economic Development Quarterly*, Vol. 18, No. 2, pp. 174–185, 2014.
- Jan Youtie, "An Evaluation of the MEP: A Cross Study Analysis," in Philip P. Shapira and Charles W. Wessner, eds., *21st Century Manufacturing: The Role of the Manufacturing Extension Partnership Program*, Washington, D.C., National Academies Press, pp. 390–427, 2013.

SUPPLEMENTAL VI.
LEGAL AND POLICY CONSIDERATIONS

Agencies face several legal considerations when designing programs and data collections, particularly regarding obtaining and sharing client-level administrative data for research and evaluation purposes. For example:

- Laws may constrain sharing or be unclear about whether applicant or participant data collected by local service providers (grantees/cooperators) can be shared with the program agency and evaluation experts.
- Agency staff members are sometimes unsure if program data, especially Unique Identifiers and other sensitive data that are critical for linking to other datasets, can be shared with statistical agencies or contractors for statistical research and evaluation purposes.
- Data sharing arrangements often are documented in a Memorandum of Understanding (MOU), which may require legal assistance to develop.

Guidance from OMB—in particular M-14-06—and a department's general counsel office can help program managers navigate these legal considerations.[10] Obtaining advice from attorneys is critical, because a number of statutes and regulations are potentially relevant to data sharing, including the Privacy Act, the Trade Secrets Act, and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[11, 12, 13] There may also be agency- or case-specific statutes to consider.

There are many examples of successful data sharing arrangements with statistical agencies and outside contractors. For example:

- The Department of Commerce's International Trade Administration Global Markets Program approached the Census Bureau's Center for Economic Studies (CES) to help it obtain statistics on the performance of its client firms versus the general business population and understand the impact of its technical assistance activities. See page 68 for details of the arrangement. The results of the study can be found in C.J. Krizan, "Statistics on the International Trade Administration's Global Markets Program," CES Working Paper No. 15-17, 2015, available at <http://www2.census.gov/ces/wp/2015/CES-WP-15-17.pdf>.
- Data on the Small Business Administration's 7(a) and 504 loan programs were shared with the Census Bureau's CES for statistical research purposes with benefits to the Census Bureau, enabling a study on the impact of the loan program on business employment, survival, and productivity. Results can be found in J. David Brown and John S. Earle, "Finance and Growth at the Firm Level: Evidence from SBA Loans," *Journal of Finance*, forthcoming, <http://ftp.iza.org/dp9267.pdf> and J. David Brown, John S. Earle, and Yana Morgulis, Forthcoming, "Job Creation, Small vs. Large vs. Young, and the SBA," in *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, John Haltiwanger, Erik Hurst, Javier Miranda, and Antoinette Schoar (eds.), Chicago and London: University of Chicago Press, forthcoming, <www.nber.org/papers/w21733>.
- The National Institute of Standards and Technology (NIST) employed outside contractors to evaluate the Manufacturing Extension Partnership (MEP) program. NIST shared program administrative data on MEP recipients with the Census Bureau, and the outside contractors used the MEP data linked with Census Bureau data at a secure federal research data center facility, measuring the effects of MEP assistance on establishment productivity, employment, sales, and firm survival. Results of the study can be found in Clifford A. Lipscomb, Jan Youtie, Sanjay Arora, Andy Krause, and Philip Shapira, "Evaluating the Long-Term Effect of NIST MEP Services on Establishment Performance," CES Working Paper No. 15-09, 2015, available at <https://ideas.repec.org/p/cen/wpaper/15-09.html>.

[10] OMB's M-14-06 "encourages federal departments and agencies to promote the use of administrative data for statistical purposes and provides guidance in addressing legal and policy requirements for such uses, including the need to continue to fully protect the privacy and confidentiality afforded to the individuals, businesses, and institutions providing the data" ("Guidance for Providing and Using Administrative Data for Statistical Purposes," OMB's M-14-06, 2014).

[11] 5 U.S.C. §552a.

[12] 18 U.S.C. §1905.

[13] See Title V of the E-Government Act of 2002, Public Law 107–347.

OMB's M-14-06 and the general counsel office can assist with determining the conditions under which data may be shared, including what new notice (if any) about the data sharing needs to be communicated to program participants.

Best Practices for Improving Access to Data

Check if Data Can Be Shared (see Best Practice 11**)**: Consult with general counsel and use guidance from OMB's M-14-06 to determine whether and how applicant /participant-specific data can be shared with researchers for statistical analysis, including impact evaluation. If the data cannot be shared, look for opportunities to address this barrier. Authorizing and rulemaking language may provide the ability to share and use the data for statistical analysis in a secure environment, ensuring applicants'/ participants' privacy and confidentiality.

Involve Attorneys Sooner Not Later (see Best Practice 12**)**: Hold early discussions with agency attorneys, as well as privacy and confidentiality officers, to familiarize them with the data needs for evaluation, including Unique Identifiers such as social security numbers (SSNs) and employer identification numbers (EINs). Discuss sharing and linking data for statistical analysis, including impact evaluation. Early discussions can avoid problems and needless delays that arise when data collection and sharing are treated as separate or after-the-fact considerations.

When to Talk to General Counsel About Sharing and Linking Data for Evaluation Purposes

Working with general counsel offices early in evaluation planning stages can help familiarize them early on to the project scope, which could help avoid delays later in the process. For example, lack of familiarity with the intended data uses could result in general counsel citing certain statutes and rules as reasons that data sharing agreements cannot be approved, when on closer inspection those rules may not be applicable to the particular context. For example, the Computer Matching and Privacy Protection Act (CMPPA) may at first glance seem relevant when reviewing the legality of requests to link program and statistical agency data for statistical impact studies. According to the Office of Management and Budget's final guidance interpreting this law, the CMPPA restricts certain automated matching using databases containing federal personnel records or matching to make decisions about the rights, benefits, or privileges of specific individuals under federal benefits programs. Automated record linkages that produce anonymized, aggregated data products or support statistical research are not barred by the CMPPA (see 54 FR 25818). These statistical linkages may be used to inform decisions about the general implementation of federal programs that may impact federal beneficiaries as a class or group.

SUPPLEMENTAL VII: WORKING WITH OUTSIDE RESEARCHERS TO CONDUCT EVALUATIONS USING LINKED PROGRAM AND SECONDARY DATA

Many different approaches can be employed to study program impacts, and different researchers bring their own unique strengths to the program research process. Historically, agencies have hired contractors and conducted surveys of program participants (the "treated" group) and businesses that did not receive program services (the "control" group) to build evidence and evaluate program impacts. Increasingly, agencies are realizing that more credible, and potentially cost-effective, impact evaluations can be done when client-level data held by program agencies are linked to other already existing datasets across the federal government, and possibly across the private sector, that contain critical additional information on clients receiving services, individuals or businesses that could be used in control groups, and outcome data for both groups.

Also, program agencies are exploring opportunities to work directly with researchers in statistical agencies to statistically analyze program impacts, rather than hiring outside (nonfederal) contractors. Statistical agencies have access to, and familiarity with, data that can be of high value in building evidence and some have capacity to conduct studies of program impacts. Statistical agencies have a long history of conducting high-quality data linkages and have significant expertise conducting statistical analyses using secondary data for research purposes (see text boxes: "Using Census Bureau Microlevel Data" and "Example: Working With the Census Bureau to Build Evidence of Program Impacts Using Administrative Data"). Another major benefit of partnering with statistical agency researchers is that after the initial investment in data sharing, data linkages, and statistical impact analyses, there is greater opportunity for replication and follow up studies to be done more efficiently, and at lower cost.[14]

Regardless of their affiliation, it is important that the staff conducting the studies have strong training and experience in the central technical areas, including statistical methods, information technology for data capture and management, and substantive features of the business assistance program. When contractors are used for impact evaluations, the federal employee overseeing the work should be qualified to assess the proposed methodology and recommend improvements.

For program agencies considering an evaluation to build evidence of program impacts, the main steps involved in the evaluation process are as follows:

1. **Consult evaluation experts** as early as possible (ideally early in the program's life, and well in advance of an evaluation) to plan data collection decisions. Among other things, evaluation experts can help identify what impacts can be efficiently evaluated using administrative and other secondary data, and which require the use of a survey.
2. **Identify and come to agreement with a research team** (with statistical agency researchers, outside contractors, or a hybrid). If an outside contractor will be used for data linking and impact evaluation, the program agency can expect to negotiate the terms of the contract, including cost. Similarly, if statistical agency staff will do some or all of the work, then the program agency may need to negotiate an agreement with the statistical agency, including costs in most cases (one with costs requires a reimbursable agreement).
3. **Work with counsel and agency management to develop an *Interagency or Other Special Agreement (IOSA)* that governs the terms of data sharing**. The IOSA documents terms and conditions governing data access and use when program agencies provide data that are not publicly available to statistical agencies and/or outside contractors for statistical research and evaluation purposes. If statistical agency staff will be involved in the project implementation, then the data sharing IOSA and reimbursable agreement could be the same agreement.

[14] Giving other statistical agencies and external researchers the opportunity to conduct replication studies is a valuable research quality control mechanism (Klaus F. Zimmermann, "Evidence-Based Scientific Policy Advice," IZA Policy Paper No. 90, 2014, available at <http://ftp.iza.org/pp90.pdf>). To make it easier to interpret studies on the program and to facilitate replication studies, it is important that there be transparency in program implementation, data collection and editing, and how the research studies are conducted.

4. **Conduct the study and disclose the aggregated statistical results.** While statistical agency staff can provide aggregate data and analysis of programs, to maintain objectivity, statistical agency researchers cannot translate the implications of the results into specific policy recommendations. For example, statistical agency researchers can state that the program had an impact of a particular magnitude and statistical significance, given a set of identifying assumptions, but they are not permitted to give advice on what program changes should be made as a result. Program agencies could seek the assistance of their departmental chief evaluation office or chief economist office to help translate the findings into actionable advice. The ability of statistical agencies to greatly improve the rigor of an evaluation approach and the quality of data linked to a specific program suggests that working with these agencies is a powerful step that can advance evaluation efforts.

Using Census Bureau Microlevel Data

The Census Bureau houses a wealth of business data relevant for program impact evaluation, ranging from primarily administrative data (from IRS, SSA, BLS, and state labor departments) to censuses and surveys it conducts on a regular basis. A number of requirements must be met to access Census Bureau data:

- Evaluations involving linking program data to Census Bureau business data require approval by the Census Bureau and IRS.
- Approval requires the following conditions to be met:
  - Demonstrates scientific merit.
  - Demonstrates a need for microlevel data.
  - Feasible.
  - Poses no risk of disclosure of confidential information.
  - Provides an approved benefit to the Census Bureau.
  - Presents no policy concerns, no conflict of interest, and no financial gain.
  - No regulatory, administrative, or enforcement purpose affecting any individual may result from access to and linkage of any individual records.
- All output is reviewed to ensure no disclosure.
- There are two paths to using Census Bureau microlevel business data.
  - **Reimbursable project**: In this case, researchers in the Census Bureau's Center for Economic Studies (CES) undertake the empirical work. If CES researchers will be conducting the data linking and analysis, they can assist with drafting the project proposal. Results are published as a CES Working Paper.
  - **Federal statistical research data center project**: In this case, non-CES researchers (other federal government researchers, academic researchers or third-party contractors hired by the program agency) conduct the empirical work, which includes linking the datasets and performing the impact analysis. The non-CES researchers develop the project proposal following the guidance available on the CES website.
- If Census Bureau microlevel business data were used in the analysis, a post-project report for IRS needs to be written (including a summary of findings, and documenting how the research benefits the Census Bureau).

Keys to Successfully Working With Statistical Agencies or Outside Contractors

- **Establish a realistic timeline**: Statistical agencies maintain a wealth of supplementary data on clients as well as firms that could serve as control groups. However, each project must undergo a review before being approved. Often program agencies (or research partners) are surprised by the length of time the approval process can take. In particular, approval time may be longer for projects involving comingled Census Bureau and IRS data—regardless of whether Census Bureau staff will conduct the statistical analysis or whether an outside contractor proposes to link program data to Census Bureau-held data through a federal research data center. The research implementation can also take longer than anticipated, e.g., due to unforeseen data challenges or feedback that suggests going in a different direction. It is advisable to set a project end date that allows for these possibilities.[15]
- **Identify potential delays early on**: Well in advance of when an evaluation is needed, the program agency, statistical agency, and partners/outside contractors should work together to identify potential delays that may occur at the time the evaluation will be implemented. In addition to program data sharing considerations, this may include developing ways the project can benefit statistical agency programs. Parties to data sharing agreements should strive to keep each other informed about any issues that could delay the project approval process. They should work to standardize and streamline the approval process to facilitate future efforts.
  - Specifying in award conditions that outside researchers may be part of the federal evaluation program team can prevent unanticipated problems providing access to data for an impact evaluation.
  - If including outside researchers in an evaluation team, begin the process of ensuring that they have access to necessary data early.
- **Help researchers understand the data**: If program agency staff are unable to directly engage in program research studies, it is especially important that the program agency provide data and program documentation to the researchers doing the studies and make program agency staff available to answer questions. This also applies to statistical agencies when their data are used by external researchers for such studies.
- **Allow sufficient time to develop the Interagency or Other Special Agreement (IOSA)**: For program agencies with limited or no experience sharing data with other agencies or contractors, the process of developing an IOSA for data sharing can seem daunting the first time an agency goes through it. Although the first data sharing IOSA can take many months, identifying language to which all parties agree is a critical investment that pays off with repeated projects. Reviewing the IOSA template with each subsequent use will be important, however, to assure the language reflects current data format requirements and other practices. Other tips for data sharing agreements include:
  - Adopt the framework for a model interagency agreement that is provided in OMB guidance M-14-06. Templates can also be used to speed up the process of developing agreements with nonfederal entities.
  - Ensure IOSA contains all required elements, including the parties; legal and programmatic authority; duration or period of agreement; purpose of the activity, including goals and anticipated benefits; use and/or limitations of data use; expected data elements and quality; roles and responsibilities for data protection, including data security and privacy and confidentiality; means of data transfer; data retention and record keeping; penalties for disclosure and protocols for a data breach; relevant disclaimers; project reporting requirements; administrative contacts; funding and cost information if relevant; resolution of conflicts; process for modifying, reviewing, or canceling the agreement; and agency signatories.
  - Keep IOSA language current. Consult researchers who will perform the work before signing the agreement.
  - Establish clearly the boundaries of what the outside researchers can do with agency data, such as how they need to protect data.
  - For projects requiring future data, include provisions in the IOSA for data updates.

[15] The Census Bureau's Global Markets program impact study discussed in the box below encountered both of these types of delays, resulting in the need to apply to the IRS for a project extension.

- Consider Freedom of Information Act (FOIA) implications of data sharing. It may be appropriate to include language in the IOSA about how the agencies will address FOIA requests involving these data.
- Include IT security staff in the IOSA drafting process for sections on IT security language, electronic data transfer, and National Institute of Standards and Technology (NIST) language.
- Track and monitor data sharing agreements, including their duration, legal requirements, and terms and conditions.

Example: Working With the Census Bureau to Build Evidence of Program Impacts Using Administrative Data

The International Trade Administration's Global Markets (GM) program has taken several steps over the past couple of years to collect more rigorous evidence on its overall impact and performance. After commissioning outside experts to help it develop both a logic model and evaluation methods to be piloted, GM approached the Census Bureau's Center for Economic Studies (CES) to help it obtain statistics on the performance of its client firms versus the general business population and understand the impact of its technical assistance activities.

After the Department of Commerce Office of the General Counsel (OGC) confirmed that the GM data could be shared with CES, CES provided GM with a cost estimate for the project, which was to be conducted in two phases. In the first phase, CES would attempt to match the GM data to Census data. If the first phase resulted in a sufficiently high match rate, the data would be analyzed and firm outcome statistics would be calculated comparing GM client performance to the general business population. With the cost estimate and project outline determined, CES developed a project proposal that described the work in detail, including how it would be conducted, how long it would take, and how it would benefit the Census Bureau. Once approval to proceed with the project was obtained both internally from the Census Bureau and externally from IRS, an Interagency Agreement/Memorandum of Understanding (MOU) between the two agencies was drafted. (A template for MOUs is at <www.census.gov/about/business-opportunities/resources/iosa.html>.) When the MOU was given final approval by OGC and the heads of each agency, the data were transferred and the project began. See the results of the statistical impact study at <https://ideas.repec.org/p/cen/wpaper/15-17.html>.

CES researchers have also worked with other agencies such as the Small Business Administration and NIST to provide results on the outcomes of their client firms.