AMERICAN
EVALUATION
ASSOCIATION
www.eval.org

Research, Technology & Development
Topical Interest Group

Evaluating Outcomes of Publicly-Funded Research, Technology and Development Programs: Recommendations for Improving Current Practice

Version 1.0

Prepared by the Research, Technology and Development Evaluation Topical Interest Group of the American Evaluation Association (AEA)

February 2015

For information or comments, contact Gretchen Jordan [Gretchen.Jordan@comcast.net]

Research, Technology & Development Evaluation Topical Interest Group, American Evaluation Association, 2015

ACKNOWLEDGEMENTS

We wish to thank the volunteers from the Research, Technology & Development Evaluation Topical Interest Group of the American Evaluation Association who contributed to this paper. Specifically, we gratefully acknowledge the team leaders, Gretchen Jordan (360 Innovation LLC) and Dale Pahl (U.S. Environmental Protection Agency). Appreciation is also extended to the team members: Liza Chan (Alberta Innovates – Health Solutions); Kathryn Graham (Alberta Innovates – Health Solutions); Deanne Langlois-Klassen (Alberta Innovates – Health Solutions); Liudmila Mikhailova (CRDF Global); Juan Rogers (Georgia Tech); Rosalie Ruegg (TIA Consulting, Inc.); Josh Schnell (Thomson Reuters); Robin Wagner (Centers for Disease Control and Prevention); Madeleine Wallace (Windrose Vision LLC), and Brian Zuckerman (Science and Technology Policy Institute).

We also wish to thank the numerous reviewers who graciously provided invaluable comments on earlier drafts that served to strengthen this paper and to the RTD evaluation practitioners who shared their thoughts with us at workshops. Providing written comments were: Erik Arnold, Technopolis Group; Frederic Bertrand, Independent Consultant; Mark Boroush, U.S. National Science Foundation; Irwin Feller, Pennsylvania State University; Laura Hillier, Canada Foundation for Innovation; Jean King, University of Minnesota; Jordi Molas-Gallart, Spanish Council for Scientific Research; Lee Kruszewski, Alberta Research and Innovation Authority; Al Link, University of North Carolina at Greensboro; Steve Montague, Performance Management Network; Cheryl Oros, Independent Consultant and Liaison, Evaluation Policy Task Force, American Evaluation Association; Richard Riopelle, Ontario Neurotrauma; Anita Schill, National Institute for Occupational Safety and Health; and Bill Valdez, U.S. Department of Energy.

As reflected by the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, we encourage readers of this paper to expand on this work by adding to the materials in the Appendix and sharing their comments and suggestions with the Research, Technology & Development Evaluation Topical Interest Group of the American Evaluation Association. For information or comments, contact Gretchen Jordan (Gretchen.Jordan@comcast.net).

Contents

1. Introduction

Effective guidance and tools for program evaluation have been long sought by the legislative and executive branches of the United States federal government and other governments to inform evaluation policies and practices. In the United States, the resulting federal directives that began in 1993 create the context and requirements for a prominent role for evaluation of the performance and results of federal programs. Frequently, these evaluations include the objectives of measuring the return on public investments, demonstrating accountability, and increasing the effective use of taxpayers' money. The prominent role for evaluation and these common evaluation objectives are especially relevant to federal research, technology and development (RTD) programs because scientific leadership and innovation are seen as keys to solving many pressing problems and improving national competitiveness (America COMPETES Act, 2007).

The evaluation of federal RTD programs is not without significant challenges and this became evident when federal agencies and their oversight organizations began responding to federal evaluation policies. One challenge, as compared to other evaluation domains, relates to the nature and timing of RTD progress as it is usually unpredictable and the translation of research into societal outcomes occurs through complex processes that involve many actors downstream of the RTD program. Other challenges, which are not unique to RTD programs, include access to data; data and analysis quality; and the synthesis of data to inform decisions and policies.

There are now many years of experience about the evaluation of RTD programs in the United States and around the world, most of that gained in the past twenty years due to increased requirements. This experience has led to a growing body of individual studies and guidance documents. Given these relatively recent advances, an overview that summarizes the current status of RTD program evaluation policy and practice would be of benefit to RTD program managers and practitioners who evaluate federal RTD programs.

Research, technology and development programs are complex and diverse. Management, and therefore evaluation, of RTD programs in the United States is also decentralized. Consequently, available documentation about evaluation practices within RTD mostly addresses individual aspects of evaluation for a specific type of program or outcome without providing larger context and guidance on when evaluation designs or methods can or cannot be appropriately applied. Given the diversity of what is being evaluated in what context, the findings of individual studies can seldom be synthesized or aggregated to look at questions across programs, such as what works better and why. The ability to synthesize findings is also limited without the use of more common language about the bigger picture of RTD and its possible outcomes even when appropriate evaluation design and methods are applied.

This paper provides a summary of current policy and practice in the evaluation of publicly-funded RTD programs with a focus on programs in the United States. Additionally, it provides recommendations for further improvements in current practices. The content for this paper was derived from the body of existing literature and the authors' collective experience as RTD evaluation practitioners. It benefited from two rounds of written expert peer review as well as feedback received from evaluators and RTD program managers at multiple workshops.

1.1 Purpose and Scope

The purpose of this paper is to engage a broad audience – including managers of RTD programs, RTD program design and evaluation professionals, and government decision and policy makers – in a dialogue about current RTD evaluation practice (including performance measurement) and how it might be further improved. Improvement will come in part from the establishment of a consensus on a common evaluation language and practice that is broadly implemented within and across publicly-funded RTD programs. The larger objective of this paper is to improve RTD evaluation so as to be better able to inform RTD program improvements and, in turn, contribute to improved program outcomes.

This paper covers evaluation of all aspects of research, technology, development and diffusion/deployment type efforts.

This paper has a relatively broad scope but is limited to publicly-funded RTD given the resource constraints of the authors. Publicly-funded RTD, by its nature, plays a different role in the innovation ecosystem than industrial RTD efforts. In particular, publicly-funded RTD serves multiple stakeholders, is generally characterized by large spillover effects that benefit both producers and consumers, and often broadly enhances the nation's capacity for further innovation (Martin & Tang, 2007; Hall et al., 2009).

Another focus for the scope of this paper is the program level of analysis[1], with particular attention paid to the monitoring and evaluation of: program outputs; progress towards achieving near, mid and longer outcomes/impacts; and a program's contribution to outcomes/impacts. This paper uses the terms 'outcomes' and 'impacts' interchangeably or together (outcomes/impact) while recognizing the importance of the entire sequence of outcomes, from early progress to ultimate outcomes or goals (see Glossary).

This paper's scope also takes into consideration evaluations over the life cycle of the program, whether it was prior to program commencement, during program implementation, or after the

[1] Within this paper, a program implies an entity with a stated budget and objectives that is comprised of multiple projects and their associated activities.

conclusion of the program. This is done with the intention to cover, and indeed to link, all aspects of research, technology, development and diffusion/deployment type efforts. As a result, the evaluation of "innovation" programs is included, with innovation defined as a new product, process, or organizational practice that is entering the "market." An example of innovations associated with organizational practice could be represented by a new way of delivering health care or by a different way of organizing research such as through strategic clinical networks.

2. Policy Implications and Relationship to the AEA Evaluation Roadmap

This paper, through its focus on evaluation practices and policies that are specific to publicly-funded RTD programs, supplements previous recommendations from the American Evaluation Association (AEA) as described in its *Evaluation Roadmap for a More Effective Government* ("AEA Roadmap") (AEA, 2013).

The AEA Roadmap emphasizes that "…there is a strong case to be made for a commitment to evaluation as an integral feature of good government, whether the goal is better performance, stronger oversight and accountability, or more data-informed and innovative decision making." To guide the development and implementation of evaluation programs in federal agencies, the Roadmap provides 17 recommendations in the areas of scope and coverage; management; quality and independence; and transparency. While the authors of the current paper take all of these recommendations as being fundamental to RTD evaluation practice, this paper expands on two of these recommendations for publicly-funded federal RTD programs:

- Build into each new program and major policy initiative an appropriate evaluation framework to guide the program or initiative throughout its life; and
- Promote the use and further development of appropriate methods for designing programs and policies, monitoring program performance, improving program, operations, and assessing program effectiveness and cost.

A third area of emphasis was added as the paper evolved, namely the desirability of more use of common frameworks appropriate for RTD:

- The RTD community should move toward the utilization of agreed upon evaluation frameworks tailored to the RTD program type and context in order to learn from synthesis of findings across evaluations.

Discussion within this paper about current practices in publicly-funded federal RTD programs in the areas of evaluation frameworks and the use and development of methods provides arguments for how improvements can be made. The paper lays the groundwork for moving toward a common evaluation framework, indicators and design. The related conclusions and specific recommendations are summarized at the end of this paper (see *Conclusions*).

3. Evaluation Context

The context of the evaluation is a key consideration that determines which evaluation approach and methods to use. It is therefore important for context to be included as an explicit component within this review of common practices in the evaluation of publicly-funded RTD programs.

3.1 Requirements

For federal programs in the United States, the Government Performance and Results Act (GPRA) Modernization Act of 2010 (GPRAMA, 2010) and Office of Management and Budget (OMB) Circular A-11 (OMB, 2013) contribute significantly to the evaluation context. The GPRAMA, as indicated by both OMB and the Government Accountability Office (GAO, 2013), modernizes the federal government's performance management framework by retaining and amplifying several aspects of the GPRA of 1993 (GPRA, 1993) while also addressing some of its limitations. In particular, while GPRA 1993 established strategic planning, performance planning, and performance reporting as a framework to guide federal agencies in reporting publicly on their progress to achieving their missions, GPRAMA places heightened emphasis on the requirements of government-wide and agency priority-setting and cross-organizational collaboration to achieve shared goals.

Part 6 of OMB Circular A-11 (OMB, 2014) describes the federal performance framework, strategic and annual plans, the performance management cycle, and the role of program evaluation. It describes program evaluation as "individual, systematic studies to assess how well a program is working to achieve intended results or outcomes." It also highlights that "evaluations can help policymakers and agency managers strengthen the design and operation of programs and can help determine how best to spend taxpayer dollars effectively and efficiently."

Specific guidance on evaluation is also included in the annual budget priority memo sent jointly by OMB and the Office of Science and Technology Policy (OSTP) (2010). For example, for fiscal year 2012, it was indicated that "agencies should develop outcome-oriented goals for their science, technology and innovation activities, establish timelines for evaluating the performance of these activities, and target investments toward high-performing programs in their budget submissions."

In summary, congressional and executive requirements for federal leaders view program evaluation as an important tool – to be used in conjunction with goals, measurement, analysis, and data-driven reviews – to improve results of programs and the effectiveness and efficiency of agency operations. These requirements also (i) describe in general terms a role for program evaluation and (ii) indicate that, in addition to the program's near-term performance goals and indicator, the program's long-term goals and objectives should be taken into consideration when selecting an evaluation approach for federal programs. Based on this, longer term goals, objectives, performance goals, and indicators have been included in the evaluation framework and logic model for RTD programs that are described in Section 6 of this paper. Additional information about federal requirements for evaluation is provided in Appendix E.

Congressional and executive requirements for federal leaders view program evaluation as an important tool.

3.2 Data and Other Evaluation Challenges

The characteristics of RTD programs create challenges for evaluation. As noted in studies by the National Academy of Sciences and others, the typical setting and measuring of specific, quantitative and timed performance targets may not be appropriate for RTD activities and programs because of the unpredictable nature and timing of research progress. Additionally, there is generally an extended period of time between completion of the research activities and the subsequent emergence of the intended social or economic outcomes. It is frequently stated that it takes an average of 17 years for research evidence to reach clinical practice (Morris, Wooding, & Grant, 2011). RTD activities also typically involve multiple actors who build on each other's work, with the translation of research into societal outcomes often occurring through complex processes that involve many actors who have different roles in the innovation ecosystem and contribute in varying degrees. This complexity makes determination of attribution to economic and social benefits (what would have happened in the absence of the intervention) a difficult undertaking (see Section 5.2).

The histories of many federal RTD programs predate the requirements for formal evaluation. As a result, some programs have been challenged by the need to retrofit a performance management framework to existing operations and to develop plans that systematically guide all aspects of evaluation. Evaluation planning and the systematic evaluation of a program are important as they can assist in preventing piecemeal responses to external demands for information about progress or results. However, this requires overlapping efforts in real time to plan, monitor, and implement data collection and evaluations studies and build supporting databases that store data over time.

Data collected routinely by a single program or even a single federal agency will seldom be sufficient to capture program outcomes because researchers receive funds from multiple organizations and over time may change names or employment. In part, these challenges may be eased by recent initiatives focused on open unique identifiers for scientific and other academic authors. For example, ORCID[2] is a non-profit community-based effort that provides a registry of persistent unique identifiers for researchers as well as methods for linking the researchers to digital research objects. Funders also have the opportunity to integrate ORCID identifiers into their research workflows, such as grant application processes and grant progress reporting protocols. Combined with efforts by research organizations and publishers, systematic and consistent embedding unique identifiers into critical funding workflows creates the possibility of linking a researcher's contributions across his or her career. In Europe, CERIF is gaining support as an open European standard for the exchange of information about research; this may assist in the sharing of data between separate systems[3].

Permission to access data can be difficult. Concerns about burdening researchers' time with administrative, non-research tasks can lead research managers to deny access. Similarly, program managers may be hesitant to burden partners with data collection and data collection instruments used by U.S. federal agencies must be approved by the OMB under the Paperwork Reduction Act in order to protect citizens from unnecessary requests. In the private sector, data access challenges often relate to the protection of proprietary data. The response rate for surveys is another data challenge, with reductions in response rates further adding to difficulties of data collection.

Data quality also depends on the context in which it is applied, also referred to as the "fitness for user needs", because information is context dependent.

As reflected by the expression "garbage in, garbage out", data quality is essential and often requires considerable effort. Data cleaning can account for 50% to 80% of a data analysts time (OECD, 2015) and is affected by the structure of the data. Data that has been structured data and includes appropriate meta data typically requires less cleaning as exemplified by the Web of Science and Scopus publication databases that carefully differentiate author names and institutions. Data quality also depends on the context in which it is applied, also referred to as the "fitness for user needs", because information is context dependent. Factors affecting data quality include timeliness, relevance, coherence, interpretability, accuracy, credibility, and accessibility (OECD, 2015).

[2] See http://orcid.org/organizations/funders
[3] See http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1

Another data challenge relates to the capacity to analyze data as this capacity influences the information that can be extracted or interpreted from the data. While this capacity is partially determined by available data structures (e.g., meta data, links between data sets, etc.) and technologies, it is also affected by the pre-existing knowledge and skills of the analysts. The recent proliferation of “big data” also has important implications for statistical agencies (OECD, 2015). Specifically, in addition to potential errors caused by poor data quality or inappropriate use of data, errors can result from unexpected changes in the environment in which the data is collected. This is particularly true when the data analytics are automated (OECD, 2015). It also happens when the data is collected, structured and analyzed without enough information (or program theory) on how the data (such as inputs and outcomes) are linked.

This paper deals with technical issues only, and both the questions evaluators are asked to study and the interpretations and uses of findings concerning program effectiveness and/or efficiency are political/policy matters. As stated by Carol Weiss in her paper *Where Politics and Evaluation Research Meet*, "…the policies and programs with which valuation deals are the creatures of political decisions. They were proposed, defined, debated, enacted, and funded through political processes, and in implementation they remain subject to pressures – both supportive and hostile – that arise out of the play of politics" (Weiss, 1973).

There is also the challenge of looking across evaluation studies to draw broad conclusions on what worked under what circumstances and why. A review by RAND Europe examined 40 years of studies of how scientific research drives innovation and social-economic benefits that often result from that (Marjanovic, Hanney, & Wooding, 2009). They concluded that there are seven persistent challenges to carrying out research evaluation that can provide a robust evidence base that can inform policy:

- Apparent contradictions between the conclusions of various studies due to differences in study design such as types of innovations studied and timeframes considered;
- Biases in the selection of cases to examine in research;
- A lack of clarity and unity in the definitions of explored concepts (across studies), such as discovery, invention and innovation;
- Unclear descriptions of study methodology and techniques for data collection and analysis with associated difficulties in the ability to repeat them;
- The challenge of setting boundaries in research for data collection and analysis, including defining the starting and finishing lines;
- Challenges in impact attribution; and
- Issues of sector idiosyncrasies with respect to innovation processes.

3.3 Progress Has Been Made in RTD Evaluation

Multiple formal and informal efforts have been made since 1993 to enhance the understanding of outcome evaluations for research and to improve evaluation practices. During the past decade for example, federal agencies and independent expert committees convened by the National Academy of Sciences' National Research Council applied the concepts of results-based management to describe how federal RTD programs are designed to function, the outcomes to which the RTD programs contribute, and the approaches to evaluate them (National Academies of Sciences, 2000, 2007, 2009; COSEPUP, 1999, 2008; National Science and Technology Council, 1996, National Research Council, 2007, 2008). Highlighted within the recommendations from the Committee on Science, Engineering and Public Policy in its publication titled *Evaluating Federal Research Programs: Research and the Government Performance and Results Act* (COSEPUP, 1999) was that:

- Research programs should be described in strategic and performance plans and evaluated in performance reports;
- The use of measurements needs to recognize what can and cannot be measured and that the misuse of measurement can lead to strongly negative results;
- Federal agencies should use expert review to assess the quality of the research they support, the relevance of that research to their mission and the leadership of the research and that each agency should develop clear guidance on expert review processes;
- Both research and mission agencies should describe their goal for developing and maintaining adequate human resources and that human resources should become a part of the evaluation of a research program;
- A formal process should be established to identify and coordinate areas of research that are supported by multiple agencies; and
- The science and engineering community can and should play an important role in GPRA implementation.

The use of measurements needs to recognize what can and cannot be measured and that the misuse of measurement can lead to strongly negative results.

Progress in RTD evaluation is also reflected by the RTD community's increased focus on the assessment of social and economic outcomes and impacts in spite of the challenges that it presents. As emphasized by GAO in 2012, it must be appreciated that the driving force behind the timing and design of evaluations for RTD outcomes are the characteristics of the research program itself (GAO, 2012). Whereas an evaluator might readily measure the effectiveness of an applied research program by whether it met its goal to improve the quality, precision, or efficiency of technologies or processes, such immediate and concrete goals are usually not associated with basic research programs. Rather, evaluation of the effectiveness of basic research programs would measure less concrete outcomes such as advancing knowledge in a field and building capacity for future advances through investments in training students or the development of useful tools or supports for the scientific community. Multiyear investments in basic research might also be expected to eventually influence innovations in technology that then yield social or financial value, such as energy savings or security.

At a macro level of analysis, a number of seminal economic studies have demonstrated that RTD contributes to economic growth (Solow, 1957; Abramowitz, 1956; Baumol, 1986; Romer, 1990). While macroeconomic studies provide invaluable context for public policy, many related questions of interest to decision-makers regarding specific outcomes and impacts from RTD program and portfolio investments are best addressed through evaluations aimed at the program (micro) level of analysis.

A review of various performance measurement and evaluation guides and evaluation studies provide a few major observations about additional progress that has been made in the understanding and practice of RTD evaluation (Ruegg & Feller, 2003; Rogers, Youtie, & Kay L, 2012[4]). One such observation is that expert judgment remains a primary method for assessing the quality and significance of the scientific and technical outcomes of basic and applied research. However, expert judgment has often progressed to take into consideration, or be complemented by, additional data and analysis about outputs and outcomes. Sources for this additional data may include, but is not limited to, publications, patents, and network analysis. Another observation is that RTD programs typically use a variety of mixed methods (i.e., case studies, social network analysis, statistical and econometric techniques, cost-benefit analysis, etc.) to evaluate the economic and social welfare consequences of program outcomes. For example, RTD program evaluations often couple expert judgment with defined protocols such as Technology Readiness Levels (Mankins, 1995) or Stage Gate (Cooper et al., 2002) to determine technical progress, current status of a technology, or current contextual conditions. Another major observation relates to the selection of the method(s) used for the evaluation. Specifically, the methods for evaluating the outcomes of publicly-funded RTD programs depend on the level of analysis and the questions to be addressed.

Feasibility and appropriateness of the study design are the focus of another major observation about progress that has been made in the understanding and practice of RTD evaluation. In particular, a current concern is that the application of experimental design to outcome

[4] Additional information and examples of technology and development program evaluation methods are available in a U.S. Department of Energy overview of evaluation methods and in a recent evaluation handbook (Ruegg & Jordan, 2007; Link & Vonortas, 2013).

evaluations of RTD programs is favored by OMB (OMB 2013) and others even though this “gold standard” for determination of attribution of outcomes is typically infeasible for the study of RTD programs (GAO 2012). Even complex demonstrations of attribution of outcomes such as comparison pre- and post-action or against a control group are not considered "best suited for" research programs by the GAO as per Table 5.1 in its 2012 report. On the other hand, quasi-experimental and non-experimental designs using a counterfactual approach and mixed methods have been found to be feasible approaches for programs in the development or diffusion parts of RTD. To summarize, experimental study design is a powerful approach when it is feasible and fits the evaluation objectives, but not being able to implement this approach should not discourage RTD programs from conducting outcome evaluation.

4. Recommendations for Planning and Implementing Evaluation in RTD Programs

4.1 Recognize Evaluation as a Management Tool To Be Used Across the Program Life Cycle

The ***first area of recommendations*** in this paper focuses on building into each new program and major policy initiative an appropriate evaluation framework to guide the program or initiative throughout its life.

Evaluation is a valuable management tool that can be used to inform decision making at every point in the life cycle of a program. Specifically, evaluation can inform answers to questions about program planning, implementation, progress, outcomes/impacts, and learning and redesign (see Table 1).

Agreeing upon and clearly stating the questions an evaluation is to answer is an important early step in commissioning or planning any evaluation. While unique evaluation questions may be developed for an individual program based on its specific context, objectives and requirements, there are also several questions that are commonly asked by program managers. Several of these common questions have been captured as evaluation “criteria” in the OMB document *Research and Development Investment Criteria of Relevance, Quality and Performance* (OMB 2003).

Evaluation can inform answers to questions about program planning, implementation, progress, outcomes/impacts, and learning and redesign.

The ability to answer a broad range of questions throughout the program life cycle enables evaluation to serve multiple purposes. The primary purposes discussed in the United States are accountability and program improvement. A recent report by RAND Europe adds two other general purposes and defines these four purposes (Guthrie et al., 2013):

- Accountability: to show that money and other resources have been used efficiently and effectively, and to hold researchers to account;
- Advocacy: to demonstrate the benefits of supporting research, enhance understanding of research and its processes among policymakers and the public, and make the case for policy and practice change;
- Allocation: to determine where best to allocate funds in the future, making the best use possible of a limited pool of funding; and
- Analysis (program improvement and learning): to understand how and why research is effective and how it can be better supported (or allocated), feeding into research strategy and decision making by providing a stronger evidence base.

Table 1. Examples of RTD Program Evaluation Questions Posed by Government Leaders

Stage in the Program Life Cycle	General Questions	Evaluation “Criteria”	More Detailed Questions
Planning	What will the program do, when and why?	• Program implementation design • Evaluation plan exists	• What are the planned end outcomes and strategies for achieving them? • How can we measure progress, success? • What part of this can be achieved within the timeframe of the performance assessment?
	Are we doing the right thing?	• Relevance	• How are the planned outcomes of the program aligned with the organization’s larger strategic goal(s)? • What is the program’s “critical link” with these outcomes: how will specific contributions from the RTD program be transferred and used to achieve the intended outcome(s)?
Early/Mid Implementation	Are we doing it the right way?	• Economy • Efficiency • Quality • Performance (early)	• What is the program’s progress toward the “critical link” with outcomes? • Are the “right” (varyingly defined) investigators applying and receiving awards? • Are RTD activities and partnerships proceeding as expected? • What are the performance measures (indicators) that demonstrate this progress?
Mid/End of Implementation	What has been the outcome/impact?	• Effectiveness • Performance • Value For money	• When transfer and use occur, what are the intermediate outcomes that must be accomplished (beyond the scope of the program) to achieve the broad societal goal(s) that the agency aims to accomplish? • How will the program demonstrate the program’s achievement of end outcomes, and contribution to these end outcomes (impact) that are the agency’s broad societal goal(s)?
Learning/ Redesign	What do we do next?	• Use of evaluation findings	• Can this program be replicated in other situations and if so, which ones? • How have the evaluation findings been used to improve, expand, redirect or discontinue activities and the likely results of that change?

4.2 Use Different Types of Evaluations to Answer Different Questions

There are a few broad types of evaluations that are distinguished by when they occur, the purpose of the evaluation, and the kinds of questions asked. Current observations on these different types of evaluation for RTD programs are provided below.

Prospective Outcome/Impact Evaluation
Prospective outcome/impact evaluations are exercises conducted prior to program completion and, in some cases, prior to the program start, to project the future outcome of a program based on several assumptions and forecasts. As a result of these projections, prospective outcome/impact evaluations necessarily entail greater uncertainty than retrospective outcome evaluations do. The trade-off for this uncertainty is that prospective evaluations can be used in a formative sense to test potential outcomes/impacts based on alternative program designs, implementation approaches, and other factors affecting outcomes/impacts.

Monitoring Outputs (also referred to as Tracking or Performance Measurement)
Monitoring outputs of ongoing RTD programs can provide early progress information that is useful for assessing the extent to which program targets are being met and whether the program is 'on track' for achieving the intended results. To do so, monitoring involves the systematic and routine collection, analysis, and feedback of data about the program or other entity of interest (e.g., monitoring may identify the achievement of key technical goals). In the longer term, monitoring is expected to provide important data for periodic in-depth evaluation. Consequently, monitoring has a complementary and facilitative role in RTD program evaluation.

Process Evaluation with Short Term Outcomes (also referred to as Formative Analysis)
Process evaluation with short term outcomes is typically done while a program is under way but after the completion of one program iteration (cycle). This enables determination about whether program processes are working as intended and if they are having the desired effect. Formative in nature, process evaluation therefore provides the opportunity to identify the need for mid-term corrections in program processes. For example, a grant program may investigate responses to its call for proposals, allowed time for proposal submission, selection criteria, proposal review and feedback, funding practices, and other program processes to assess perceived fairness, transparency, participation rates, and other factors that affect outcomes. The program may then use the information acquired through the process evaluation to inform changes to program processes prior to the next iteration of the program.

Traditionally, formative analysis has been an integral feature of process evaluation and less so in outcome/impact evaluation. However, as noted earlier, there appears to be a growing demand on the part of managers of publicly-funded RTD programs for the inclusion of formative analysis within impact evaluations. One possible explanation for this demand is that program managers and other stakeholders increasingly expect all evaluations to inform decisions and policies. With formative analysis in both process and outcome/impact evaluation, the main distinction is that formative process evaluation informs on-going program investment decisions and policies whereas formative impact evaluation informs future program decisions and policies. A discussion of how formative analysis may be incorporated into impact evaluation is provided in a recently prepared U.S. Department of Energy guide on retrospective impact evaluation (Ruegg et al., 2013).

Retrospective Outcome/Impact Evaluation

Retrospective outcome/impact evaluations are generally conducted after completion of the program or, at the very least, after a sufficient period of time has passed to have enabled the associated outcomes and ultimate impacts to occur. Scientific and technical outcomes can be assessed every few years but socio-economic outcomes seldom occur that quickly.

Retrospective impact evaluation measures what the program accomplished against a baseline, which is usually done by comparing the program's outcomes/impacts against its goals. Nonetheless, it is possible to simply observe what has unfolded over time by comparing to a more general baseline such as the "state of the art" at the beginning of the measurement period. This is done in some evaluations that follow up with program participants after five years. Retrospective outcome/impact evaluation also seeks to determine and document evidence about what part of the observed outcomes/impacts resulted from the actions of the program being evaluated as opposed to rival explanations. The challenges this presents for RTD programs is discussed throughout this paper.

A change is occurring in retrospective evaluation which traditionally has been summative in nature. That is, retrospective outcome/impact evaluation has recently begun to incorporate formative analysis features (Jordan et al., 2014).

4.3 Plan Evaluations Around a Logical Framework

Evaluation planning, which is one part of the performance management process, is needed to accomplish the systematic evaluation of a program. In the absence of planning, evaluation activities inevitably amount to piecemeal responses to external demands for information about progress or results. Evaluation planning is therefore intended to organize the evaluation activities according to a logical framework that describes the program logic or theory of change, which are the intended activities and strategies for achieving the goals. Building the evaluation plan around the logical framework assists in ensuring that the indicators used to answer the evaluation questions link to the intended resources, activities, strategies and goals of the program. In addition to taking the program's goals and resources into account, evaluation planning requires that responsibilities, approaches, metrics/indicators, data requirements, data collection and analysis methods, and reporting mechanisms be identified.

Developing a logic model can help design or redesign a program.

The elements of a performance measurement and evaluation framework are illustrated in Figure 1. At the top is a program logic model that describes the inputs, activities, and outputs produced with partners for customers, who then take those outputs and apply them to produce a chain of outcomes. As depicted in the middle of the figure, all of the logic model elements are done in a context of driving and restraining influences. The bottom of the figure highlights the need to identify indicators for each of the logic model elements as well as the questions associated with why, how, and in what context the observed results occurred or are expected to occur.

In addition to effectively capturing the key indicators to investigate when measuring performance, the process of developing a logic model, particularly one that captures the program theory, can help design or redesign a program through the consideration of current circumstances (Funnell & Rogers, 2011). Logic models also assist in building a common understanding of expected program performance among staff responsible for program delivery. Additionally, simplified versions of the logic model can be used to describe the program to external stakeholders. Several resources for developing logic models and defining performance indicators are provided in Appendix C.

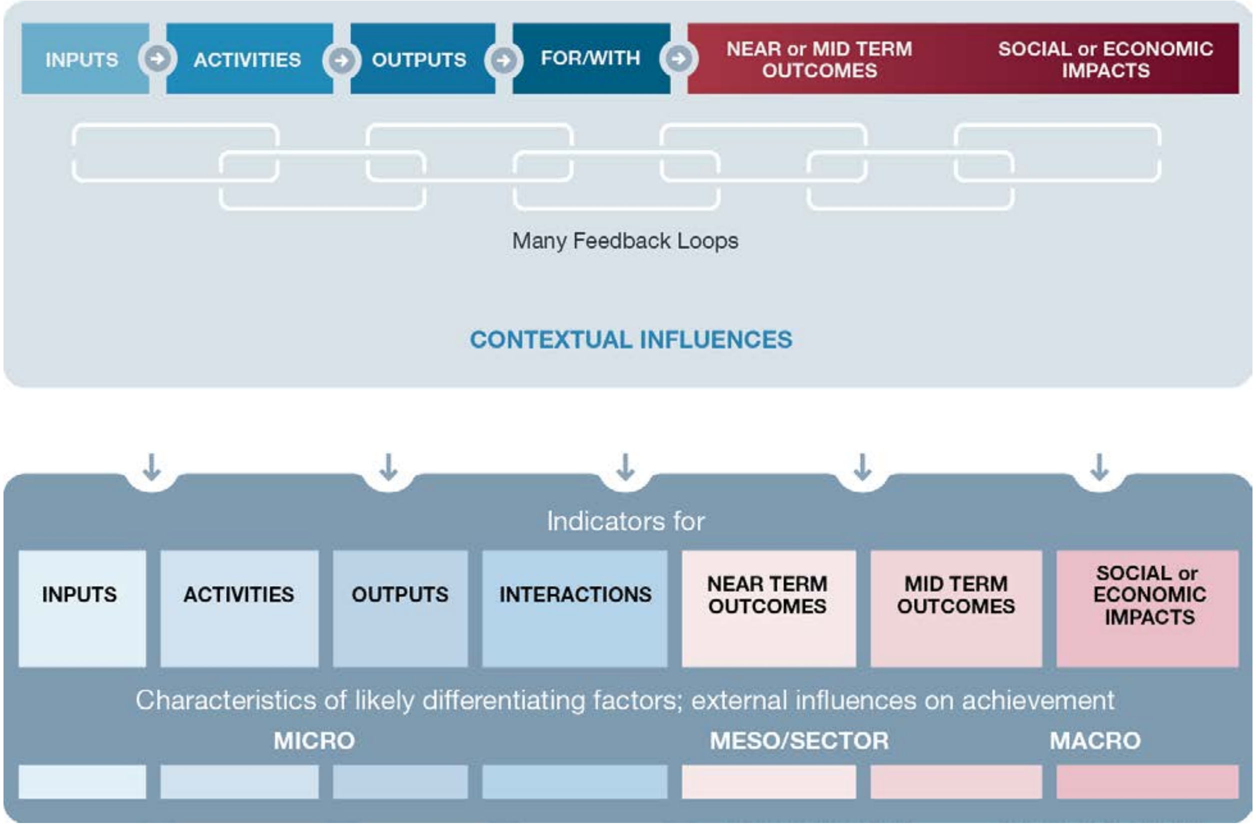


Figure 1. A Framework Where Indicators Flow From a Logic Model and Accompanying Context

A few broadly applicable evaluation frameworks have been developed for use in RTD program outcome evaluations. Two examples of evaluation frameworks are the Canadian Academy of Health Sciences (CAHS, 2009) Impact Framework and the Framework for the Review of Research Programs of the National Institute for Occupational Safety and Health (NIOSH) (National Academies, 2007). There are also some schemes for evaluating research or research organizations that may be called frameworks but are not frameworks in the sense that the term is used this paper. These include: the Excellence in Research for Australia (ERA) framework; the Research Excellence Framework (REF) used in the United Kingdom for assessing and comparing university research departments; and STAR METRICS[5] in the United States.

The histories of many federal RTD programs predate the requirements for formal evaluation. For some federal RTD agencies, this has necessitated overlapping efforts in real time to plan, monitor, and implement evaluations studies as well as build supporting databases. These efforts were necessary because existing databases and tracking mechanisms were often unsuitable for real-time data compilation in support of evaluation.

A notable exception to the above efforts that required retrofitting a performance management framework to existing programs is provided by the former Advanced Technology Program (ATP) that was operated by the National Institute of Standards and Technology (NIST) (Ruegg & Feller, 2003). The ATP provides a good example of what can be done to build a robust monitoring and evaluation system. It is also an example of how even well-planned and implemented RTD evaluation is only one source of information for policy decisions.

The ATP, which was established by the Omnibus Trade and Competitiveness Act of 1988 and was several years prior to GPRA requirements for evaluation, developed a comprehensive monitoring and evaluation plan. This plan was enabled by a director that was supportive of evaluation; a budget that permitted resources to be allocated for evaluation purposes; a specific Congressional mandate that required the reporting of program outcomes by a stated date; advice from leading theoreticians and practitioners in evaluation; and an internal staff charged with making it happen.

The ATP's initial evaluation planning centered on establishing what to measure; designing databases to capture program activities, participants, outputs, and outcomes as they (unpredictably) unfolded; and establishing monitoring activities. Leading economists and evaluators from other disciplines were also invited to participate in the evaluation planning activities. Additionally, on-line data collection instruments and other supporting mechanisms were developed and implemented. At the outset of the evaluation planning process, each

[5] An initiative led by the National Institutes of Health (NIH), the National Science Foundation (NSF), and the White House Office of Science and Technology Policy (OSTP) focused on developing inter-agency capability to monitor the effects of federal RTD investments on employment, knowledge generation, health and other outcomes.

funded project was analyzed to identify its key technical goals and metrics as well as technical and other dimensions of progress that could be monitored throughout the funding period and 5-years after completion. A star-rating system based on program goals was used to characterize intermediate outcomes for each completed project and for the overall portfolio of projects. In the longer term, ATP pioneered benefit-cost evaluation of technology portfolios, compiled extensive databases to serve its evaluation needs, and conducted studies not only of impact evaluation but also those aimed at improving an understanding of program dynamics, such as analyzing the determinants of successful collaboration.

4.4 Use of a Common Framework is Desirable

Several challenges in trying to use RTD evaluations to study broad issues of science, technology and innovation policy were highlighted in Section 3.2. Similar challenges are seen when trying to look across similar programs to compare or aggregate program results and investigate what worked better in what circumstances. A methodology for looking across evaluation studies is called synthesis evaluation (GAO, 1992) and it includes techniques that address the challenges found in a previous study (Marjanovic, Hanney & Wooding, 2009).

A major element of evaluation design to enable synthesis across studies is the use of standard logical frameworks and design approaches as it creates necessary cohesion between studies. Another important aspect is to strategically plan studies using standard frameworks and design approaches to purposefully investigate gaps in knowledge. Synthesis, aggregation, and the subsequent answering of new broad questions of interest to program managers and policy makers would be possible if RTD evaluators conducted studies with the foreknowledge that the studies would be synthesized or aggregated and therefore used a standardized framework and initiated studies to fill knowledge gaps.

5. Recommendations: Use of Appropriate Methods

The ***second area of recommendations*** in this paper is the use and development of appropriate methods for designing programs and policies, monitoring program performance, improving program operations, and assessing program effectiveness and cost.

Development of appropriate methods aligns with requirements at an agency level in GPRMA, OMB and OSTP guidance (GPRMA, 2010; OMB, 2013; OMB & OSTP, 2013). This section is intended to highlight key points being made about methodologies within current discussions among RTD evaluators. As such, it is not intended to repeat or summarize the large body of literature written on methodology. More detailed information on methodologies can be found in Ruegg and Feller (2003), Ruegg and Jordan (2007), and Link and Vonortas (2013).

5.1 Clarify Purpose and Questions before Deciding on a Method

Recommendations highlight the use and development of appropriate methods for designing programs and policies, monitoring program performance, improving program operations, and assessing program effectiveness and cost.

The combination of the evaluation purpose and the program theory for the program being evaluated determine the questions to be asked and answered in the evaluation (see Figure 2). As previously highlighted (see Section 4.1), the four main evaluation purposes are accountability, advocacy, allocation, and analysis (program improvement and learning). For example, the evaluation purpose may be to demonstrate impacts that can be attributed to the program activities (accountability) or it may be to improve the targeting of program resources in order to increase impact (program improvement and learning). The second element, program theory, provides critical information that needs to be taken into consideration when formulating the evaluation questions through its description of the intended impacts, the strategies for achieving the impacts, and other likely influencing factors. Once developed, the evaluation questions inform selection of the required method(s) and design.

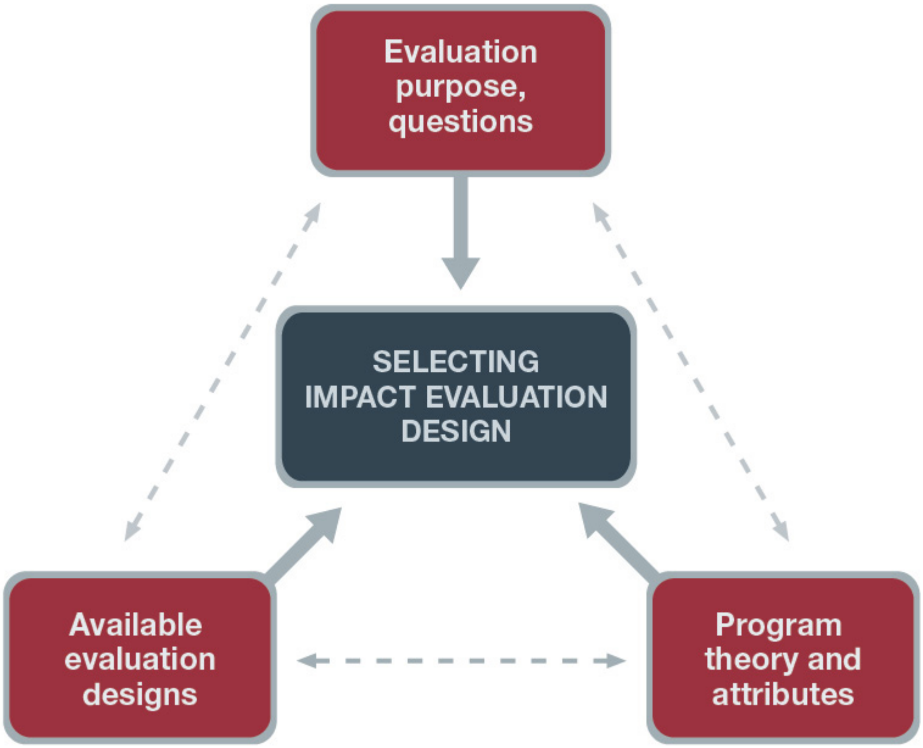

Evaluation purpose, questions
SELECTING IMPACT EVALUATION DESIGN
Available evaluation designs
Program theory and attributes

Figure 2. Factors to Consider When Selecting an Outcome/Impact Evaluation Design

Adapted with permission from ACIAR (Figure 6 in Mayne & Stern, 2013)

To illustrate how questions can precede methods, suppose that the pressing question is whether a program's economic benefits have exceeded its costs. In this case, a benefit-cost methodology would be appropriate. Another example is a question that asks whether a program has stimulated collaborative activity, where RTD evaluators are beginning to use social network analysis. For this question, the network analysis should be applied as a pre-post design in order to compare changes in the network; once just before or close to the beginning of the program intervention and again after the passage of time. Naturally, evaluations may entail many other questions and methods and any given evaluation may entail multiple questions and multiple methods.

The choice of design and method for demonstrating program outcomes depends on the questions asked and the context of the program being evaluated.

The choice of design and method for demonstrating program outcomes depends on the questions asked and the context of the program being evaluated. Although this is recognized as the current best practice in evaluation, it is not uncommon to see requests for a particular method such as bibliometrics or randomized controlled trials to be used without consideration of context or the question to be answered. Once the evaluation questions are defined, additional factors may also influence the choice of research design and method and how these are applied. Notably, budget and timeframe constraints frequently affect what is done and the level of effort because some methods may be too costly or require a time period that exceeds what is available. Further, as discussed in the next section, conditions may also limit the type of study design that is feasible and this may subsequently affect the choice of evaluation methods or approaches used.

5.2 Choosing a Design for Outcome Evaluation, Attribution

A research design is the logical approach to inferring answers to the evaluation questions from the data collected. Since the questions asked and evaluation circumstances differ, this could be as simple as reporting the findings of a carefully constructed peer review or as complex as inferring outcomes attributed to an intervention from the results of an experimental or quasi-experimental design. Methods such as tabulating descriptive measurements and finding statistical significance of a relationship between variables are usually not thought of as research design, but they are since they are a logical approach to inferring answers to questions. Similarly, comparison of a program against a standard or against expectations is one of the most common research design approaches for assessing outputs and outcomes of basic and applied research programs despite its inability to demonstrate formally what outputs and outcomes would have happened without the program (GAO, 2012). These commonly used approaches to inferring answers to descriptive evaluation questions and to comparing results against targets are not discussed further in this paper because they are relatively straightforward and well documented elsewhere.

Research design is more demanding when the objective is to determine if a program has caused part or all of an observed outcome/impact. Determining if and how much of an observed outcome/impact was caused by the program is commonly referred to as estimating program attribution or assessing "additionality." For example, an evaluation may estimate that a stated percentage of an observed impact was caused by a given program intervention if the intervention had accelerated innovation or if it had broadened or deepened a research effort. There are three conditions that are requisite to establishing that a program intervention has caused part or all of an observed change:

- There is a logical explanation as to why the intervention can be expected to have led to the observed change;
- There is a plausible time sequence whereby the investment and subsequent actions occurred before the observed change, the latter being relative to an appropriately established baseline; and
- There is compelling evidence that the program intervention is the partial or full cause of the observed change after competing explanations are taken into account (i.e., rival explanations are eliminated as causes of the change).

Three conditions are requisite to establishing that a program intervention has caused part or all of an observed change.

The first condition for establishing attribution is addressed by examining the logic of the design of the program intervention within the context of the challenge or problem to be solved. The second condition regarding the time sequence of action followed by observed change is assessed through analysis that compares the current condition with a before-program-intervention baseline. The baseline also provides a means for measuring the change itself. Meeting the first and second conditions provides evidence, but not proof, of cause and effect. The third condition, namely elimination of rival explanations of effect (i.e., did the program intervention cause the observed changes to occur or did something else cause the change?), is necessary to provide more solid evidence of cause and effect. Meeting all three of these conditions, if feasible, is considered a best-practice test for determining attribution of outcome/impact to a program intervention.

In certain areas of research, such as medical and agricultural research, it may be possible to conduct controlled experimentation using control groups to determine if a public program intervention caused an isolated outcome. Barring random experimental design, an alternative may be to conduct the evaluation using a quasi-experimental (that is, not random) design that is sufficiently robust to measure the outcome of a public program intervention while avoiding the bias of self-selection into the program.

When experimental design or robust quasi-experimental design is possible, the evaluation documents observed outcomes. It then assesses the program's attribution to those outcomes by comparing the group that received the program intervention with a control group that did not. If the only difference between the groups is receiving/not receiving the program intervention, the difference in their outcomes can be directly attributed to the program intervention.

There are, however, a number of reasons why problems often arise in attempting to apply experimental and quasi-experimental design in RTD evaluations. The sampling of participants and non-participants in an evaluation may not be truly random and hence the groups may not be comparable. In particular, there tends to be a self-selection bias in terms of who seeks to participate in publicly-funded RTD programs and those who do not. Further, there may be a bias in the process of selecting participants. Populations of both participants and non-participants may be too small in areas of emerging technologies, especially during an early period of development, to produce groups of sufficient size to support random sampling and to meet statistical tests of significance. Additional issues may also arise that can compromise study objectivity when data used to assess outcomes in the two groups are obtained by subjective methods such as self-reporting, interview, and survey. Specifically, either or both groups may have reason to misreport results and there may be unwillingness among non-participants to engage with evaluators in providing data because they see no value to it. Finally, research and the application of research are by their nature very uncertain with many phases and actors; understanding and replicating that same uncertainty in an experimental or quasi-experimental design is often not feasible.

There are a few examples of quasi-experimental evaluation studies that have successfully developed comparison groups for use in RTD evaluation using econometric or statistical techniques to rule out confounding variables. In some cases, comparison groups were drawn from program-compiled data and in other cases this was done through other databases. Feldman and Kelley (2001) analyzed the effect of the Advanced Technology Program (ATP) on firm ability to attract additional funding by comparing a sample of ATP recipients of awards with a comparison group of non-recipient/near winners. The evaluators used multivariate regression and Tobit estimators to adjust for other differences in the two groups that may have influenced their comparative ability to attract funding. The use of regression-discontinuity designs, where awardees' outcomes were compared with the outcomes of research conducted by those whose proposals were in the fundable range but who did not receive funding, or of propensity-score matching to create synthetic matched groups of participants and nonparticipants, are methods that have also been used in RTD evaluation for the creation of comparison groups.

When a non-experimental design is used, it is necessary to assess program attribution using mixed methods that generally include a counterfactual approach. The counterfactual approach is used in either of two modes of application. One mode of application is to query program participants about what they actually did versus what they think they would have done had the program not existed. In this case, the actual data is empirically and objectively based while the counterfactual data is hypothetical and subjectively based. The other mode of application is to obtain actual data empirically and to query experts about what they think would have happened differently in the absence of the research program intervention. Asking participants the counterfactual question of what they otherwise would have done, or asking experts what otherwise would have happened, allows a comparison to isolate the part of the outcome that is attributable to the program investment. In this regard, it resembles experimental and quasi-experimental design approaches. However, instead of using objectively derived data, the non-experimental counterfactual approach relies on the generation of subjectively-derived hypothetical data for comparison. Furthermore, it assumes that participants or experts are able to reliably express estimates for the counterfactual scenario.

Contribution analysis considers other potential explanations for a change and then tests the relative role of each of these as part of a larger "causal package."

5.3. Consider Contribution Analysis

While attribution analysis (or additionality) is critical for establishing causality and identifying whether a program has caused all or part of an observed change, a newer approach called "contribution analysis" plays a complementary role. It could be argued that some RTD programs do a form of contribution analysis for attribution analysis, but in the U.S. few are aware of contribution analysis as a formal methodology. Contribution analysis was developed with the intention of capturing information for program improvement as well as program additionality. This approach, which is currently used most often in European and Canadian evaluation practice, examines context, mechanisms, and outcomes to see what worked under what circumstances (Mayne & Stern, 2013). The central purposes of contribution analysis are to confirm whether a program is working as intended and to identify areas for potential improvement. However, contribution analysis can also be used to make an estimate of how a program has affected the outcome. Contribution analysis is useful for RTD programs because it helps isolate the signal associated with the program in question, a requirement in quasi-experimental approaches. In many RTD domains, investigators receive multiple lines of funding, and often multiple funders engage in parallel in programmatic activity.

Contribution analysis uses qualitative methods to address each of the three conditions of additionality outlined in Section 5.2. For the first condition, contribution analysis begins by examining the logical coherence of a program's theory of change to identify whether there is a logical explanation as to why the investment can be expected to have led to the observed outcome. If a program's logic is not sound and expected outcomes are unlikely to follow from the program's activities, processes, or funding levels, then it is unlikely that the program itself contributed to the observed effects. Second, to assess whether there is a plausible time sequence whereby the investment occurred and the observed change followed, contribution analysis identifies whether results were achieved, when they were achieved relative to the program's lifespan, and whether it is reasonable that the program might have caused the intended effects. For the third condition, namely to address whether there is compelling evidence that the investment/actions are the partial or full cause of the change when competing explanations are taken into account, contribution analysis considers other potential explanations for a change and then tests the relative role of each of these as part of a larger "causal package."

In non-experimental designs, contribution analysis provides a mechanism to ask "what factors contributed to an observed result?" and "what was the relative importance of the program compared with competing explanations?" Both of these questions may be difficult to assess using quantitative methods. Interviews with experts may be one mechanism for collecting qualitative data for the purpose of identifying the relative importance of a program compared with other factors or to identify a program's specific role. Case studies are another qualitative method that could be used to delve into a particular facet of the program logic to assess its role in causality.

5.4 Using Mixed Evaluation Methods

Over the past several decades, as government agencies and evaluation practitioners have undertaken more RTD evaluations, it has been increasingly recognized that a mixed evaluation methods approach, both quantitative and qualitative, is often best. This is because a combination of analytical approaches allows each to make up for deficiencies of the other. In spite of this recognition, however, it is not unusual to see a practitioner or commissioner of evaluation use a single method.

The application of mixed methods also allows questions to be answered from different perspectives. Very importantly, mixed methods not only compensate for the deficiencies in each of the various individual methods, but it also provides an in-depth analysis of the long-term scientific, institutional, and societal outcomes that research partnerships generally are intended to achieve. Mixed methods may include both quantitative and qualitative approaches.

Quantitative methods allow results to be given as numeric measures, often as statistics, and may support further statistical and other numerical analysis that strengthen evidence. Quantitative evaluation methods include but are not limited to:

- Statistical analysis;
- Econometric analysis;
- Benefit-cost analysis;
- Impact assessment methods;
- Bibliometrics and patent analysis;
- Benchmarking;
- Social network analysis;
- Cost-index methods;
- Monitoring using indicator metrics; and
- Various scoring and rating systems.

Examples of quantitative evaluation data include: the number of patents issued and citation index values: a country's ranking relative to others; cost savings resulting from an improved product; density of change metrics for a social network; growth in the size of a customer base over time; the percent of energy supplied by renewable sources; unemployment and wage rates; and rates of return on investment.

Qualitative methods for evaluation of RTD programs include but are not limited to:

- Peer review and expert judgment;
- Site visit reports;
- Descriptions of behavior;
- Focus groups; and
- Case studies.

These methods complement and amplify quantitative data, often increasing the understanding of research findings and aiding the communication of results by providing descriptive detail and illustrative stories. Qualitative results may also assist in the interpretation of the results, integration across research findings, provision of new perspectives, and formulation of hypotheses for further testing and analysis.

Combining statistical findings with case studies, for example, provides a richer and more compelling body of evidence than that achieved through either numbers or stories alone. Providing findings from the application of several quantitative and qualitative methods, especially findings that build on each other, may serve to strengthen evidence, provide alternative perspectives and explanations, and increase confidence in the evidence. For example, showing that a public research program produced publications and patents that were heavily cited by patents of companies that commercialized a subject technology might strengthen statistical survey results of experts who reported that the public research program was important to commercial product innovation. As another example, a case study describing a collaborative effort might amplify numerical data on partnership formation and changes in the measures of density produced by social network analysis.

As concluded by Creswell and Plano Clark (2011), “mixed methods provide a bridge across the sometimes adversarial divide between quantitative and qualitative researchers.” Use of combined methods provides evaluation studies of all scales with a breadth and depth of data that more effectively and efficiently answers evaluation questions and measures programmatic success.

5.5 Valuing Economic and Other Societal Outcomes

The outcomes of federal RTD programs or portfolios of programs on the nation’s well-being may be reflected as knowledge, health and safety, economic, environmental and other societal outcomes. These outcomes can be measured by a variety of metrics that show economic and other societal change.

Economic impact evaluations of federal RTD programs generally compare the resulting economic benefits (outcomes expressed in dollars with a constant purchasing power) against associated investment costs (also expressed in constant dollars) to calculate any of a group of economic performance measures. These measures usually include net present value benefits, benefit-to-cost ratio, and internal rate of return on investment. Economic outcomes may also be expressed in terms of changes in national income, the rate of economic growth, employment, productivity and related measures.

Federal expenditures on RTD may also result in other long-term outcomes that lend themselves to valuation in economic terms as readily as resource effects. These ‘other’ outcomes have traditionally included such effects as changes in the knowledge base, quality of life, sense of wellbeing, environmental quality, health, longevity, safety, security of infrastructure, aesthetics, and other effects not easily valued in dollars. These outcomes have traditionally been expressed in physical units or described qualitatively and presented together with economic results.

Recent advances in federal evaluation, however, have extended dollar valuation to ‘other’ outcomes. For example, the U.S. Environmental Protection Agency (EPA) has developed two models to estimate the mortality and morbidity outcomes of reduced air pollution not only in terms of health and death incident rates, but also in terms of dollars of healthcare costs. These models include BenMap: Environmental Benefits Mapping and Analysis Program (Abt Associates Inc., 2012) and the Co-Benefits Risk Assessment (COBRA) Screening Model (U.S. EPA, 2013).

Similarly, economists and health, safety, and environmental policy analysts have developed alternative approaches over the years for valuing life, life years, states of health, and greenhouse gases (GHG). These approaches have been used by public agencies to evaluate the economic benefits of their programs and technologies that affect disease, illness, safety, security, and the environment. As an example, an Interagency Working Group on Social Cost of Carbon (2010, 2013) developed a range of social cost of carbon (SCC) values to estimate the monetized social value of reducing GHG emissions.

Another example is the evaluation framework that was developed by the Office of Energy Efficiency and Renewable Energy at the U.S. Department of Energy (DOE/EERE) for retrospective impact evaluation; this framework brings together multiple impacts across a portfolio of technologies, projects, or programs (Ruegg, O'Connor, & Loomis, 2013).

5.6 Evaluation Synthesis and Aggregation

Evaluation synthesis brings together a group of existing evaluation studies and, after screening the studies for relevance, quality and strength of evidence, organizes the data in order to answer evaluation questions with the assembled data (GAO, 1992). The synthesis refers to the organizing of data from the group of existing studies as well as the processes for extracting and using the results to answer new questions. The evaluation questions often include those of broader scope and pertaining to larger portfolios and related policy. These questions may focus on overall effectiveness, identifying which areas have been found to work better or worse than others; the comparison of programs and portfolios of programs; or specific program/portfolio features. The utility of evaluation synthesis lies in its ability to look across studies to point to features of an intervention that matter most and that are not otherwise visible through a single study approach. However, this utility requires that conflicts in the findings be resolved.

The utility of evaluation synthesis lies in its ability to look across studies to point to features of an intervention that matter most and that are not otherwise visible through a single study approach. However, this utility requires that conflicts in the findings be resolved.

Evaluation synthesis is typically initiated by a question that may be addressed through a synthesis of past studies. Once the question(s) is clearly stated, the next step is to assemble documentation from past studies that may to some extent have addressed the current question. This documentation generally consists of journal articles, research reports, and databases. To conduct the evaluation, researchers need to:

- Develop and implement a document/database review strategy;

- Develop a synthesis model for organizing the existing evaluation results for use in addressing the new question;
- Conduct the synthesis analysis; and
- Present the new evidence and conclusions.

The analysis may include identification of persisting gaps in knowledge that call for further targeted evaluation studies or new policy experiments to supplement the synthesized findings from existing studies.

An example of synthesized evaluation is provided by the former Advanced Technology Program (ATP) of the National Institute of Standards and Technology. The ATP systematically developed case studies for all completed projects using a standard set of progress metrics. Periodically, the ATP had evaluators synthesize across these case studies to develop statistical profiles for completed projects using a database of progress metrics from the individual case studies.

Evaluation synthesis is also exemplified by an impact evaluation completed for the French National Institute for Agronomic Research (INRA) (Joly et al., 2013). In order to assess the specific contribution of INRA, the evaluation used contextual and process analysis to identify and analyze mechanisms that generate various dimensions of impact. The synthesis was possible because three tools were standard across 30 matched case studies from five research divisions within INRA. These tools were (i) a chronology showing time frame, main events, and turning points; (ii) an Impact Pathway (similar to a logic model) showing productive intermediaries/ interactions and contextual factors; and (iii) an impact vector using a radar chart of impact dimensions. Meta cases were completed for the three areas associated with the tools (e.g., genomic breeding) and these meta cases identified the production of actionable knowledge and the lag between research and impact with intermediary results. The meta cases also examined the roles of each case of success, organizing them on two dimensions: (i) structural role (i.e., upstream research consortium or downstream intermediaries/regulation and (ii) anticipatory role (i.e., exploring new options or insuring existing options).The plotting of the successes into the associated two-by-two matrix was informative for program managers and also illustrates the utility of evaluation synthesis for learning.

The U.S. Department of Energy’s Energy Efficiency and Renewable Energy Office is currently in the process of aggregating results across multiple impact studies. The aggregation is possible because the impact analyses of the various program portfolios were performed according to a standard methodology.

6. A Generic Framework for RTD Programs with Examples

The ***third area of recommendations*** in this paper is for the RTD community to move toward the utilization of agreed upon evaluation frameworks in order to learn from the synthesis of findings across evaluations.

A generic high-level logic model and menu of indicators (a logical framework) has been developed for this paper and is described below. It builds on theories and frameworks for new product development, diffusion of innovation, and technology and knowledge transfer (Tassey 2007; Rogers, 2003; Reed & Jordan, 2007). It also takes into consideration various discussions on national innovation systems, the associated roles for government, and on how best to evaluate science and technology programs (Arnold, 2004; Ruegg & Feller 2003; Feller, Gamota & Valdez, 2003). Nevertheless, a grand theory or research plan that connects all of these is yet to emerge and learning about these factors may be impeded by current RTD program evaluation practices. In particular, a review by Autio concluded that current RTD program evaluation practice often remains limited to expert reviews or single case studies that lack the structure and characterization of circumstances required for aggregation or comparison across studies (Autio, 2014).

The generic framework described herein is intended to set the stage for a dialogue about a common language and frameworks for publicly-funded federal RTD program monitoring and evaluation. The generic framework demonstrates the diversity of RTD programs and their outcomes while suggesting a common framework for collecting and analyzing data that would allow synthesis and aggregation across that diversity. A glossary has also been provided to further assist the development of a common language through the use common terminology and clarity in meaning (see Appendix B). These tools, namely the generic framework and the glossary, were developed to assist in organizing the thinking about RTD program outcomes and context. In addition, these tools were developed to assist research program managers and evaluation practitioners in planning evaluations of a broad range of short, intermediate, and longer term outcomes.

The generic framework demonstrates the diversity of RTD programs and their outcomes while suggesting a common framework for collecting and analyzing data that would allow synthesis and aggregation across that diversity.

6.1 Key Variables in Diverse Outcomes and Contexts of Federal RTD Programs

The context for any RTD program is the innovation ecosystem. For systems level evaluation, this context is best characterized by three levels: the micro, characterized by team or organizational issues; the meso, which reflects RTD sectors; and the macro, described by national rules and objectives (Arnold, 2004). The three levels are shown in Figure 3. Of these three, the meso level is of particular importance because program outcomes/impacts differ by sectors and the sectors themselves differ in the amount of investments made in each type of RTD, the rates of technical change, and the ease of adoption. Mission, policy, and programmatic decisions also tend to be sector-specific. Bottlenecks can be spotted more easily at the meso level due to its connection with both the macro and micro level (Jordan, Hage, & Mote, 2008).

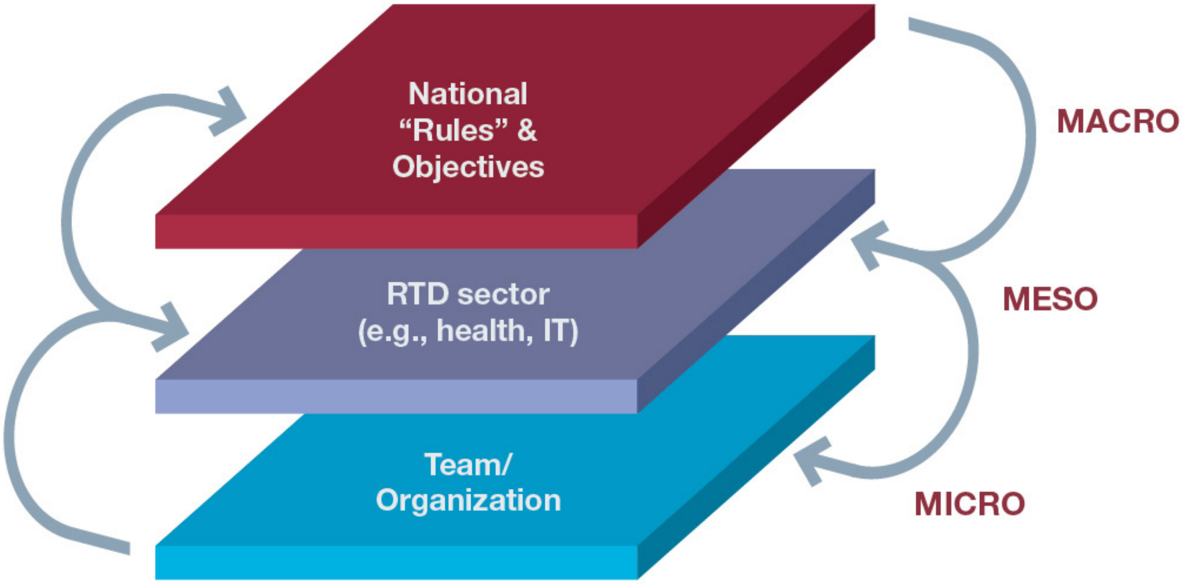


Figure 3. Systems Evaluation for RTD Includes Three Levels: Micro, Meso, and Macro

The discussion that follows focuses on key sources of variation in RTD programs, program outcomes and contextual influences on those. These need to be considered for RTD program evaluation design at the three different levels of the system.

Micro Level – Resources (Team and Organization)

RTD programs differ substantially in terms of resources (i.e., inputs). This can range from research being done by a single individual on a small budget to big science or technology development completed by large teams with large budgets and everything in between. The organizations in which the RTD programs take place also differ by size, organizational goals (e.g., teaching universities, federal mission laboratories, etc.), infrastructure (e.g., shared equipment), risk tolerance, and managerial strategies and styles.

Micro Level – Nature of the Research Problem

RTD programs encompass a wide array of activities that include basic research, applied research, and technology development research. The nonlinear model of RTD recognizes that research is an iterative process that also encompasses the research components that are needed during manufacturing and commercialization (utilization)[6] (Kline & Rosenberg, 1985). These different types of research are:

- Basic research: research that generally aims to improve the understanding of a phenomenon;
- Early-on applied research: research conducted with the goal of establishing proof of concept;
- Development research: research focused on the development and validation of new or improved products, processes, practices or policies;
- Manufacturing research: research that determines how to manufacture or assemble the new product or process with the desired qualities and cost; and
- Commercialization research: research with the goal of understanding what meets a market need or how to stimulate adoption of a new product.

The stage of research may not always be clear because the iterative RTD process does not move smoothly from basic research to applied research to development and production/ commercialization. For example, elements of basic or applied research are often needed after the development stage has been reached.

Within each type of research, the degree of radicalness of the objectives varies within a spectrum ranging from small incremental change from the current state of the art to new-to-the-world objectives. Another spectrum relates to the scope of the problem tackled, extending from quite a narrow scope to one that is quite broad or systemic (Jordan, Hage, &Mote, 2008). By taking these different spectrums into account, it becomes apparent that a program of research that works toward radical change in an entire system would have dramatically different objectives and desired outcomes than one that seeks incremental change in a narrow area.

Meso Level – Interactions

Regardless of the type of research, collaborations and interactions with next stage users are critical, as are indicators of potential or actual influence (shown as “For/With” in Figure 1). This is because the translation of RTD program activities into economic or social outcomes/impacts requires that the associated research outputs and short term outcomes be taken up by a variety of stakeholders. It is the actions taken by these stakeholders, or the support that they

[6] Nonlinearity of that being evaluated suggests the need for non-linear evaluations.

give to the actions of others, that achieves the translation of outputs and short term outcomes into mid-term and longer-term outcomes within a given context. These processes and the associated contexts have been described as the system or ecosystem for innovation.

The next stage users (i.e., the target audience for the outputs) for RTD programs could be very diverse. For example, users may include the research community itself because researchers may draw from the knowledge pool or utilize the new research methods or facilities. Users may also consist of industry members who take over the development and commercialization of prototypes demonstrated by federal RTD programs or who utilize government-developed standards or generic technologies. Government policy makers are another potential user group, with the research findings being taken up and applied by them in regulations or government programs such as education or environmental protection. It is also possible that research findings are taken up directly by advocacy groups, the media, and the public, thereby affecting debate, attitudes, and behaviors.

Regardless of the type of research, collaborations and interactions with next stage users are critical, as are indicators of potential or actual influence.

Other important factors are the diversity and continuity of the interactions with the stakeholders throughout the research program, including the degree of integration achieved (Jordan 2013). The term “connectedness” captures both the interaction and the level of integration of the parties involved in terms of knowledge sets and goals. For example, research could involve multidisciplinary, interdisciplinary, or transdisciplinary research teams. The latter are teams with multiple functions that have the downstream users of the research represented within the team. Interactions may also be inter-sectoral (e.g., partnerships among multiple levels of government). Diversity is also reflected by the different mechanisms used for interactions, including: joint planning; co-funding; co-location of researchers; establishing language or curriculum that bridges disciplines; and assigning responsibility for bridging various interests and knowledge sets (i.e., having individuals or organizations who act as intermediaries or boundary spanners between different fields or functions).

Meso Level – Nature of the Application and Ease of Adoption

There is considerable diversity in the breadth of applications of federal RTD programs and it includes both private and public value (Bozeman and Sarewitz 2011). A given research or technology platform may be deployed in multiple application areas to achieve extremely different outcomes, with the expectations and planning for these outcomes being conditioned by the various issues and challenges in each application area. For example, a basic research finding might be applied in a treatment for cancer, a new defense capability, a more energy efficient motor, or as a way to solve world hunger. . Another example is that nanotechnology may be deployed in applications as diverse as energy materials and food packaging. Considerable diversity can also occur when different research investments have similar goals within a given sector of the economy or when focused on a given problem. This diversity reflects that different research investments may entail vastly different subject matter, activities, and outcomes.

The diversity of RTD program outcomes relates to the variable length of time required for program outcomes to be realized. In part, the experimental nature of research itself often results in prolonged and unexpected time lapses between RTD activities and social or economic outcomes. The variety of speeds at which different types of new products and organizational innovations are adopted and move into wider use also affect the length of time required for the emergence of program outcomes. For example, a time lapse of fifteen years or more between basic research and the outcome of an associated product in the market is not unusual. In contrast, research results on how to get an existing product or practice more widely adopted could see results much more quickly. Longer lengths of time may be necessary to appreciate the impact of radical research advances if radical changes in supporting technology, distribution infrastructure, or user behavior and skill are also required. Similarly, advances that require higher capital investment and that have higher risks associated with making system changeovers (e.g., a new chemical process implemented on an industrial scale) typically have longer time lapses before adoption than, for example, new software. Finally, variances in the absorptive capacity and resources in different economic sectors influence the ease and speed of adoption of a new product or practice.

Macro Context
Dissimilarity in the macro context can result in different outcomes among programs or within the same program over time. One such factor in the macro context is the level of economic activity in the region or country because of its potential influence on market size and resource availability, particularly the availability of capital and capabilities. Other factors in the macro context are institutional rules (e.g., RTD tax credits), intellectual property law, restrictions or incentives to coordinate, immigration, and banking policies. Social and cultural norms may influence acceptance or rejection of certain areas of research (e.g., embryonic stem cell research) and certain behaviors (e.g., entrepreneurial activity) and are therefore part of the macro context.

6.2 A Generic High-Level Logic Model for RTD Programs

Figure 4 provides a high-level logic model that depicts the inputs, activities, strategies, and target audiences for achieving a diversity of RTD end goals. In the model, science outcomes have been purposefully separated from the application of that science and the program's long term goals such as social and economic outcomes. The separation from the program’s long term goals emphasizes that while much basic research has the end in mind, the outcomes and long term goals to which the RTD program contributes typically do not occur during the time that the RTD outputs or early outcomes are under the direct influence of the RTD program.

There were two reasons for separating science outcomes from the application of that science. First, depending on the program's context, its primary goals could be scientific, applications-oriented, or both. The assessment of research progress in the service of scientific outcomes should therefore be identified separately from the application of research knowledge to technology development, informing policy decisions, or generating other impacts that help to achieve an agency's strategic goals. Second, given the importance placed in the United States on agency-level performance planning, it is useful to integrate evaluation activities with agencies' performance measurement efforts. The inclusion of both science-oriented and societal-oriented outcome measures in the high-level logic model, even though separated, is intended to facilitate agencies' in matching their evaluation activities to their performance objectives. For example, mission agencies typically demonstrate progress toward societal-oriented outcomes while research agencies usually demonstrate progress toward scientific outcomes.

The logic model in Figure 4 has an application and progress stage before sector, social and economic outcomes (i.e., end outcomes) for technology and development activities that draw on science outcomes. Many intermediate outcomes can be anticipated to occur, and do occur, during the application and progress stage. This includes outcomes such as the career paths of graduate students, products moving through different stages of development, and consumers moving from awareness to decision to adoption. Of note, however, is that program planning and evaluation often overlook this detail on intermediate outcomes and hence the "magic in the middle" is missed.

Feedback loops between the elements in Figure 4 indicate the iterative nature of the relationships. Although several additional feedback loops could have been depicted in the diagram, including a loop to emphasis that resources and activities should be planned based on the desired social or economic outcomes, this was not done so as to keep the diagram to the fundamentals.

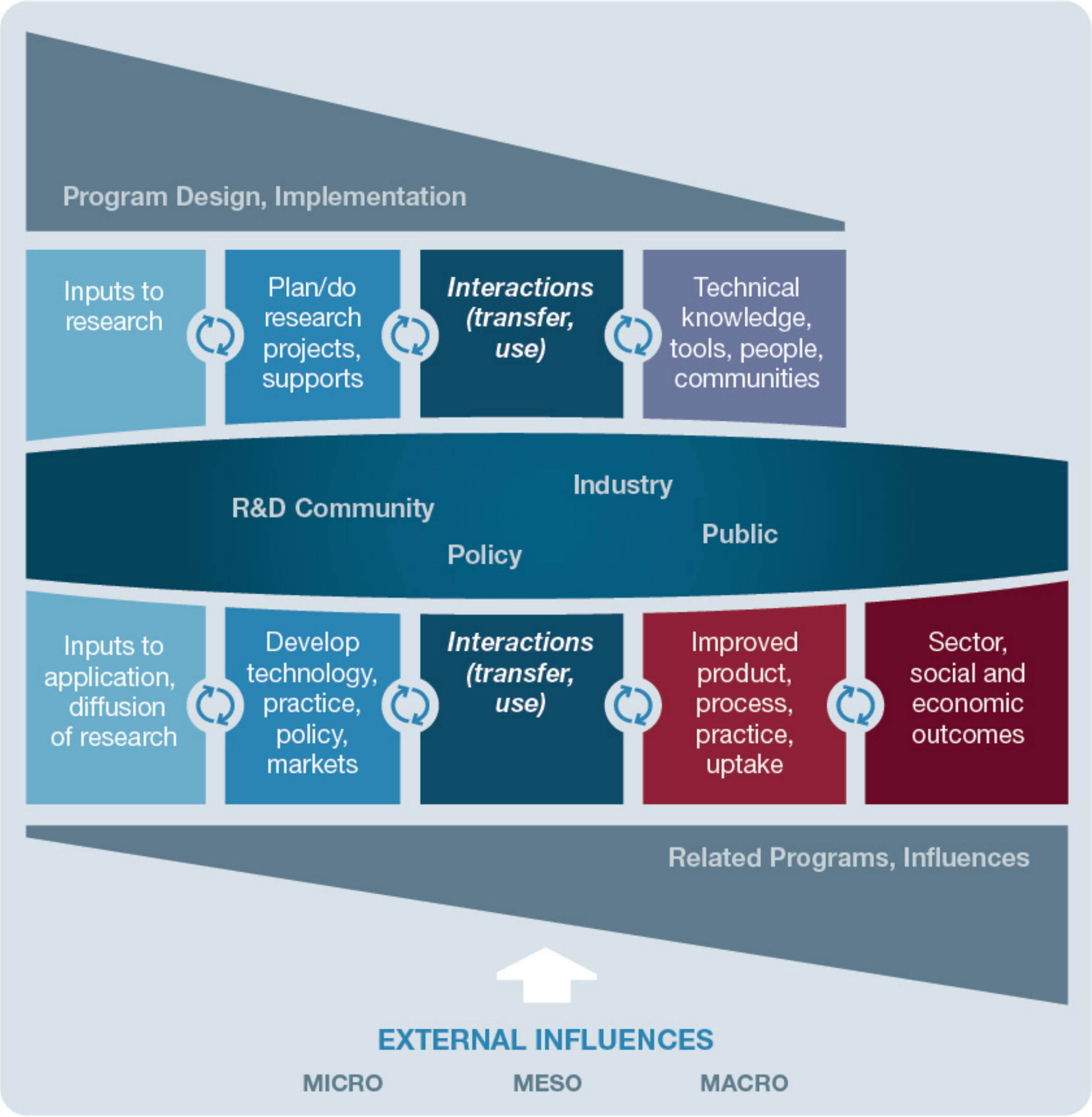

Program Design, Implementation
Inputs to research
Plan/do research projects, supports
Interactions (transfer, use)
Technical knowledge, tools, people, communities
R&D Community
Industry
Policy
Public
Inputs to application, diffusion of research
Develop technology, practice, policy, markets
Interactions (transfer, use)
Improved product, process, practice, uptake
Sector, social and economic outcomes
Related Programs, Influences
EXTERNAL INFLUENCES
MICRO
MESO
MACRO

Figure 4. A High-Level Logic Model for RTD Programs

Interactions are highlighted between the science and the application of that science because these are almost always accomplished by different actors. Although logic models do not always separate out this For (transfer) and With (partnership) element, this paper highlights how these interactions are an important area of measurement for RTD programs. As borrowed from other frameworks (Reed & Jordan, 2007; CAHS, 2009), four pathways to application and to long-term goals such as socio-economic outcomes have also been included in the high-level logic model, namely the RTD community, government policy, industry, and the public.

The top left of the generic logic model is the essential step of program design and implementation. Quite simply, program outcomes are not likely to be achieved if program design and implementation are not done thoughtfully and correctly. This includes program planning, developing selection criteria and processes, planning for evaluation, and program management. At the bottom right of the logic model, related programs and influences are highlighted to emphasize that a RTD program never is the sole reason intermediate outcomes and longer term goals are achieved. Lastly, the three levels of influence external to the program (micro, meso, and macro) are shown at the bottom of the figure. Levels of influence were included in the model to emphasis that every program operates within a larger innovation ecosystem that consists of factors that drive and constrain program success. These factors need to be considered during program and performance planning and during any assessment of program performance.

The program design and evaluation concepts illustrated in Figure 4 require only minor modification to meet requirements in OMB Circular A-11 for federal RTD programs (OMB, 2013, see 3.1 Requirements). For example, a more detailed logic model using the outline of Figure 4 would illustrate the RTD program's strategic objectives, long-term goals, and the outcomes to which it contributes. It would also depict the program's annual performance goals including indicators, targets, and timeframe to define the intended level of performance to be achieved during the year in which the annual performance goal is to be realized. While various types of indicators (e.g., outcome, output, customer service, process, efficiency) may be used as performance indicators, OMB Circular A-11 encourages agencies to use outcome indicators as performance indicators where practical.

Using the generic high-level logic model (see Figure 4) and example outcomes (see Table 2), a federal RTD program could develop a program-specific logic model that describes its end goals and the strategies for achieving these. This model can then be used to identify an evaluation framework with relevant evaluation questions as well as progress and outcome indicators. Once data is compiled for these indicators and periodic in-depth evaluations are conducted, the RTD program would be in a position to answer the questions frequently asked by government leaders, including funders (see Table 1).

Using the generic high-level logic model and example outcomes, a federal RTD program could develop a program-specific logic model.

To illustrate this, four examples that highlight different types of programs and outcome pathways are summarized in Table 3 and discussed in some detail in Appendix A.

6.3 A Framework of Frameworks

A set of more detailed generic logic models and frameworks could be developed to provide further guidance for the RTD evaluation community on how to plan and conduct outcome evaluations using similar program theories and indicators for similar programs. Although beyond the scope of this paper, enough is known about a number of commonly used RTD efforts that generic logical frameworks could be developed for each of these within a categorization scheme that drills down from the very summary level of Figure 4 to more detailed levels. Such a set, this framework of frameworks, could cover outcomes and pathways to outcomes for various sectors (e.g., health, energy, etc.). The set could show detail for pathways to outcomes for combinations of characteristics, such as type of research (e.g., applied) and types of context (e.g., favorable RTD networks already exist; technical, business and government infrastructure supports adoption of the new product; pent up demand; etc.). As well, some frameworks could show detail on commonly used interaction mechanisms such as strategic clinical networks in health research, Engineering Research Centers, or collaborations such as Sematech.

6.4 A Menu of Indicators Associated With the Generic Logic Model

Each element in the generic high-level logic model (see Figure 4) can be further described by the types of outcomes different RTD programs are aimed at delivering given the: type of RTD; the desired objectives of the RTD program; the target audience(s) for the application of the research; and the timing of the evaluation relative to the time passed since the activities took place. The list in Table 2, while not comprehensive, reflects several diverse outcomes identified in numerous evaluation frameworks and literature reviews on current RTD and innovation indicators.

Table 2. Examples of Indicators and Outcomes Across the Scope of RTD Programs

Program Design, Implementation:

- Efficiency, effectiveness of planning, implementing, evaluating; Stakeholder involvement
- Robustness of program partnerships, other delivery infrastructure
- Progress in required areas (e.g., e-government)

Contextual Influences:

- Characteristics of researchers (team size, diversity)
- Nature of RTD problem (type, scope, radicalness)
- Characteristics of interactions (continuity, diversity, etc.)
- Nature of research application (breadth, depth, timing, radicalness of change; sector absorptive capacity)
- Characteristics of macro environment (availability of capital, capabilities; ease of coordination)

Inputs/Resources for Research:

- Expenditures on research
- Expenditures on research support activities, such as database development, research planning and priority setting
- Depth, breadth of knowledge base and skill set of researchers and technologists, teams, organizations
- Capabilities of research equipment, facilities, methods that are available
- Vitality of the research environment (management, organizational rules, etc.)

Activities (the Research Process) and Outputs:

- Plan, select, fund, researchers, research projects, programs
- Quality, relevance, novelty, of selected researchers, projects, programs
- New knowledge advances (publications, technical challenges overcome)
- Quality and volume of other outputs (grants made, projects completed, number of reports, people trained, etc.)

Interactions (Includes Transfer and Use):

- Research collaborations, partnerships formed; preparation for transition to application
- Dissemination, exchange of research outputs (publications, inclusion in curricula, etc.)
- Industry engagement, co-funding, follow on funding for the research
- Public engagement, awareness of outputs (participation, media mentions)

Science Near-Term Outcomes:

- Citations of publications; patent applications, patents
- Awards, recognition, professional positions
- Expansion of Knowledge base in terms of technical leadership and absorptive capacity
- Advances in research/technical infrastructure (new research tools, scientific user facilities, testing facilities)
- People educated in RTD area and research methods
- Linkages/communities of practice/networks
- Technical base (technology standards, research tools, databases, models, generic technologies)
- Commercialization/utilization support base (manufacturing extension programs, supportive codes, etc.)

More RTD or RTD Diffusion Activities, Outputs and Interactions:

- Public funds expended for these RTD or Diffusion programs; Leveraged investments by private sector
- Translational or cross-functional teams; Presence of intermediary organizations
- Technical milestones achieved, prototypes built/scaled up, additions technical knowledge and infrastructure
- Dissemination, exchange of knowledge; consultation; citation
- Additions to diffusion/adoption infrastructure (capabilities, delivery, etc.)

Application of Research, Progress toward Outcomes:

- New technology development advances (movement through stages, functionality)
- Product commercialized; policy /practice implemented; attitude or behavior changed
- New "technology" commercialization/diffusion advances (supply chain develops, adoption of new process technology)

For each of the above:

- Utilization/influence, sustainability of influence on decisions, behavior, physical or financial factors

Sector, Social and Economic Outcomes/Impacts:

- Modeled monetized benefits
- Health status
- Security, safety measure
- Sustainability measure
- Income levels
- Jobs
- Benefit to cost ratio
- Quality of life
- Environmental quality
- Production levels
- Cost savings
- Competitiveness

Related Programs and Major Influencers:

- Date of formal handoffs to or take up from partners, others
- Chronological account of who else did what, when

Table 3. Tailoring Generic Logic and Indicators to Program Activities and Applications

Program Activity	Program Goal	Indicators of Success	Examples
Fundamental research/discovery	Influence R&D community and follow on research	Publications/citations to those publications, research collaborations	Example 1: National Science Foundation (NSF) Human and Social Dynamics Program
Applied research/”use-inspired” basic research	Inform government regulatory policy	Research results cited in new guidelines/ standards/regulations; wholly new approaches to addressing policy problem developed	Example 2: Environmental Protection Agency (EPA), Clean Air Act
Applied research/technology development	Further development, commercialization by industry	Intellectual property protected (e.g., patents), citations to patents, technology licensing to industry, partnerships formed, technologies commercialized	Example 3: Department of Energy (DOE) Wind Energy Technology Development
Diffusion/dissemination	Healthcare industry and public aware of, utilize lower cost delivery of care	Data systems in place, awareness, change in attitudes, delivery of care, cost-effectiveness of care	Example 4: Innovation in Health Care Delivery (notional)

7. Conclusion

This paper was developed to engage the audience in a dialogue about current RTD evaluation practice, how it has progressed, and how these practices might be further improved. The ultimate goal is to contribute to a consensus and broader implementation of a common evaluation language and practice within and across publicly-funded RTD programs. To achieve this, we have provided the larger context and guidance on RTD evaluation planning and implementation based on extensive review of the literature, practical experience, and the advice of expert reviewers. This context and guidance includes a newly developed generic high-level RTD logic model with accompanying output and outcome indicators; guidance on designing, monitoring, and evaluating outputs and outcomes of publicly-funded RTD programs; and a variety of examples from different types of RTD programs at different stages of implementation.

The discussion and examples contained in this paper support the following key recommendations:

Recommendation #1: Build into each new program and major policy initiative an appropriate evaluation framework to guide the program or initiative throughout its life.

- Evaluation should be undertaken because evaluation is a valuable management tool at all stages of the program life cycle;
- Evaluations should be planned using a logical framework that reflects the nature of RTD in a meaningful way; and
- Decision makers' questions may call for both retrospective and prospective evaluation, and for evaluation of outputs and early outcomes that are linked to longer term outcomes.

Recommendation #2: More needs to be done to develop appropriate methods for designing programs and policies, improving programs, and assessing program effectiveness.

- More can be done to use or insist on the use of the robust set of methods that exists for evaluating RTD outcomes;
- Evaluation methods for demonstrating program outcomes should be chosen based upon the evaluation purpose and specific questions being answered and the context;
- Mixed methods are usually best, especially when outcomes of interest go beyond advancing knowledge to include social or economic outcomes, where neither expert judgment nor bibliometrics are sufficient; and
- There are options for assessing attribution, although it is recognized that experimental design is seldom an option and contribution to a causal package is more useful.

Recommendation #3: The RTD community should move toward the utilization of agreed upon evaluation frameworks tailored to the RTD program type and context in order to learn from synthesis of findings across evaluations.

- There needs to be continued movement toward a common language and common evaluation frameworks by type of RTD program and context, with common questions, outcomes, indicators, and characterization of context; and
- Methods need to be further developed and used in relation to evaluation synthesis and the research designs, data collection, and analysis that support it.

The Research, Technology & Development Evaluation Topical Interest Group of the American Evaluation Association invites comments and suggestions on this paper and welcomes additions to the materials in the Appendices. We also welcome ideas on how to engage the RTD evaluation community in further dialogue with the intention to do as the title says, namely to improve current practice in evaluating outcomes of publicly-funded RTD programs.

References

Abramowitz M. (1956).Resource and output trends in the United States since 1870. *American Economic Review*, 46(2):5-23.

Abt Associates Inc. (2012).*BenMAP: Environmental benefits mapping and analysis program: User's Manual.* U.S. Environmental Protection Agency. Retrieved from: http://www.epa.gov/air/benmap/models/BenMAPManualOct2012.pdf

America COMPETES Act. "Public Law 110-69." *Washington, DC: GPO. Section* 1105 (2007).

American Evaluation Association (AEA). (2013). *An Evaluation Roadmap for a More Effective Government*: AEA. Retrieved from: http://www.eval.org/d/do/472

Arnold E. (2004). Evaluating research and innovation policy: A systems world needs systems evaluations. *Research Evaluation*, *13*(1), 3-17

Autio E. (2014). *Innovations from Big Science: Enhancing Big Science Impact Agenda.* UK Department for Business Innovation and Skills. Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/288481/bis-14-618-innovation-from-big-science-enhancing-big-science-impact-agenda.pdf

Baumol WJ. (1986).Productivity growth, convergence, and welfare: What the long-run data show. *American Economic Review*, 76(5):1072.

Bozeman B, & Sarewitz D. (2011). Public value mapping and science policy evaluation. *Minerva,* 49(1):1-23.

COSEPUP. Committee on Science, Engineering, and Public Policy. (2008). *Evaluating research efficiency in the U.S. environmental protection agency.* Committee on Evaluating the Efficiency of Research and Development Programs at the U.S. Environmental Protection Agency, National Research Council. Board on Environmental Studies and Toxicology. Washington, D.C.: National Academies Press. Retrieved from: http://www.nap.edu/catalog.php?record_id=12150

COSEPUP. Committee on Science, Engineering, and Public Policy, National Academy of Sciences, National Academy of Engineering and Institute of Medicine. (1999). *Evaluating federal research programs: research and the Government Performance and Results Act*. Washington, D.C.: National Academy Press. Retrieved from: http://www.nap.edu/openbook.php?record_id=6416

Cooper RG, Edgett SJ, & Kleinschmidt EJ. (2002). Optimizing the stage-gate process: What best-practice companies do-I. *Research Technology Management*, 45(5):21-27.

Creswell JW, & Plano Clark VL. (2011*). Designing and conducting mixed methods research. Los Angeles, CA: SAGE.*

EPA. (2013). *User's manual for the Co-Benefits Risk Assessment (COBRA) screening model.* Retrieved from: http://epa.gov/statelocalclimate/documents/pdf/cobra-2.61-user-manual-july-2013.pdf

Excellence in Research for Australia (ERA) Retrieved from: http://www.arc.gov.au/era/

Feldman MP, & Kelley MR. (2006). The ex ante assessment of knowledge spillovers: Government R&D policy, economic incentives and private firm behavior. *Research Policy*, 35(10):1509-1521.

Feller I, Gamota G, &Valdez W. (2003) Developing science indicators for basic science offices within mission agencies. *Research Evaluation,* 12(1):71-79.

Feller I, & Stern PC (Eds.) Committee on Assessing Behavioral and Social Science Research on Aging, National Research Council. (2007). *A strategy for assessing science: ehavioral and social research on aging*. Washington, DC: National Academies Press (US). Retrieved from: http://www.nap.edu/openbook.php?record_id=11788

Funnell SC, & Rogers PJ. (2011). *Purposeful program theory: effective use of theories of change and logic models*. San Francisco, CA: Jossey-Bass.

GAO. Government Accountability Office. (1992). *The Evaluation Synthesis*, GA/PEMD-10.1.2, Washington, DC.

GAO. Government Accountability Office. (2012). *Designing Evaluations: 2012 Revision.* Report No.: GAO-12-208G

GAO. Government Accountability Office. (2013). Managing for Results: 2013 Federal Managers Survey on Organizational Performance and Management Issues. Section 6—Survey of Organizational Performance and Management Issues.. (Report No. GAO-13-519SP). Government-wide survey results. Retrieved from: http://www.gao.gov/special.pubs/gao-13-519sp/results.htm#question_109.

GPRA. (1993). The Government Performance and Results Act (GPRA) (P.L. 103-62). Retrieved from: http://www.whitehouse.gov/omb/mgmt-gpra/gplaw2m

GPRMA. (2010). US Congress. *Public Law 11-352: GPRA Modernization Act of 2010*. Retrieved from: http://www.gpo.gov/fdsys/pkg/PLAW-111publ352/pdf/PLAW-111publ352.pdf

Guthrie S, Wamae W, Diepeveen S, Wooding S, Grant J. (2013). *Measuring research: A guide to research evaluation frameworks and tools*. Santa Monica, CA: RAND Corporation. Retrieved from: http://www.rand.org/pubs/monographs/MG1217.html

Hall BH, Mairesse J, & Mohnen P. (2009) *Measuring the Returns to R&D*. No. w15622. National Bureau of Economic Research. Retrieved from: http://www.nber.org/papers/w15622

Interagency Working Group on Social Cost of Carbon, United States Government. (2010).*Technical support document: Social cost of carbon for regulatory impact analysis under Executive Order 12866.* Retrieved from: http://www.epa.gov/oms/climate/regulations/scc-tsd.pdf

Interagency Working Group on Social Cost of Carbon, United States Government. (2013).*Technical Support Document: Technical update of the social cost of carbon for regulatory impact analysis- under Executive Order 12866*. Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/inforeg/social_cost_of_carbon_for_ria_2013_update.pdf

Joly, PB, Colinet L, Gaunand A, Lemarie S, Laredo P, Matt M. (2013). Designing and implementing a new approach for the ex-post impact of research - A return of experience from the ASIRPA project. Retrieved from: *www.fteval.at/upload/Joly_session_1.pdf*

Jordan G. (2010). A Theory-Based Logic Model for Innovation Policy and Evaluation. *Research Evaluation*, 19(4): 263-274.

Jordan GB. (2013). A logical framework for evaluating the outcomes of team science: prepared for the Workshop on Institutional and Organizational Supports forTeam Science, National Research Council. Retrieved from: *http://sites.nationalacademies.org/DBASSE/BBCSS/DBASSE_085357*

Jordan, GB., Hage J, & Mote J. (2008). A theories-based systemic framework for evaluating diverse portfolios of scientific work, part 1: Micro and meso indicators. In C.L.S. Coryn & Michael Scriven (Eds.), *Reforming the evaluation of research. New Directions for Evaluation, 118,* 7–24.

Jordan G, Mote J, Ruegg R, Choi T, & Becker-Dippmann A. (2014). *A framework for evaluating R&D impacts and supply chain dynamics early in a product life cycle: Looking inside the black box of innovation*, prepared for the U.S. Department of Energy. Retrieved from: http://www1.eere.energy.gov/analysis/pdfs/evaluating_rd_impacts_supply_chain_dynamics.pdf

Kline SJ, & Rosenberg N. (1986). An overview of innovation. In R Landau & N Rosenberg (Eds.), *The positive sum strategy: Harnessing technology for economic growth* (275-306). Washington,DC: National Academy Press. Retrieved from: http://www.nap.edu/openbook.php?record_id=612&page=275

Link AN, & Vonortas NS. (2013). *Handbook on the theory and practice of program evaluation.* Cheltenham: Edward Elgar.

Mankins, JC. (1995). *Technology* Readiness Levels: A white paper. National Aeronautics and Space Administration (NASA). Retrieved from: http://www.hq.nasa.gov/office/codeq/trl/trl.pdf

Marjanovic S, Hanney S, & Wooding S. (2009). *A historical reflection on research evaluation studies, their recurrent themes and challenges*. Santa Monica, CA: RAND Corporation. Retrieved from: http://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR789.pdf.

Martin BR, & Tang P. (2007) *The benefits from publicly funded research*. (SPRU Electronic Working Paper Series. Paper No. 61,). University of Sussex, Science Policy Research Unit. Retrieved from: http://www.sussex.ac.uk/spru/documents/sewp161.pdf

Mayne J, & Stern E. (2013). *Impact evaluation of natural resource management research* programs: a broader view. (ACIAR Impact Assessment Series Report No. 84). Canberra: Australian Centre for International Agricultural Research. Retrieved from: http://aciar.gov.au/files/ias84.pdf

Morris ZS, Wooding S, & Grant J. (2011). The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*, 104(12):510-520.

National Academy of Sciences. (2000). *Experiments in international benchmarking of U.S. research fields.* Washington, DC: National Academy Press. http://www.ncbi.nlm.nih.gov/books/NBK26380/

National Academy of Sciences. (2004). *Research priorities for airborne particulate matter: IV. continuing research progress.* Committee on Research Priorities for Airborne Particulate Matter, & National Research Council. Washington, DC: The National Academies Press. Retrieved from: http://www.nap.edu/openbook.php?record_id=10957

National Academy of Sciences. (2007). *Framework for the review of research programs of the National Institute for Occupational Safety and Health.* National Research Council Committee for the Review of NIOSH Research Programs. Retrieved from:

http://www.cdc.gov/niosh/nas/pdfs/Framework081007.pdf [See also *NIOSH Research Programs: The National Academies Evaluation of NIOSH Programs*: Centers for Disease Control and Prevention (CDC). Retrieved from: http://www.cdc.gov/niosh/nas/]

National Research Council. (2008). *Evaluating Research Efficiency in the U.S. Environmental Protection Agency***.** Washington, DC: The National Academies Press. Retrieved from:

http://www.nap.edu/catalog.php?record_id=12150

National Science and Technology Council. (1996). Assessing Fundamental Science: a report from the subcommittee on Research committee on Fundamental Science. Washington, DC: National Science Foundation. Retrieved from: http://www.nsf.gov/statistics/ostp/assess/

National Science and Technology Council (NSTC). (2007). *A Multiyear Federal Research Plan for Particulate Matter Within the Context of the NRC's Committee on Research Priorities for Airborne Particulate Matter's Report IV.* Air Quality Research Subcommittee of the Committee on Environment and Natural Resources. Retrieved from: http://www.esrl.noaa.gov/csd/AQRS/reports/pmplan.pdf

OMB. Office of Management and Budget. (2003*). Memorandum for the Heads of Executive Departments and Agencies, FY 2005 Interagency Research and Development Priorities*, June 5, 2003. Retrieved from : http://www.whitehouse.gov/sites/default/files/omb/memoranda/m03-15.pdf

OMB. Office of Management and Budget and Office of Information and Regulatory Affairs. (2012). *2012 Report to Congress on the Benefits and Costs of Federal Regulations and Unfunded Mandates on State, Local, and Tribal Entities*. Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/inforeg/2012_cb/2012_cost_benefit_report.pdf

OMB. Office of Management and Budget. (2013). *Fiscal Year2014 Analytical Perspectives: Budget of the U.S. Government*. [Chapters 8 and 9 on Evaluation]. http://www.whitehouse.gov/sites/default/files/omb/budget/fy2014/assets/spec.pdf

OMB. Office of Management and Budget. (2013). *OMB Circular A-11: Preparation, submission, and execution of the budget*. Washington, D.C.: Executive Office of the President, Office of Management and Budget. Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/assets/a11_current_year/a11_2013.pdf

OMB. Office of Management and Budget. (2014). *OMB Circular A-11 Part 6:* Preparation and submission of strategic plans, annual performance plans, and annual performance reports.

Washington, D.C.: Executive Office of the President, Office of Management and Budget. Retrieved from: http://www.whitehouse.gov//sites/default/files/omb/assets/about_omb/Overview_of_strategic_plans.pdf

OMB and OSTP. (2010).Offices of Management and Budget and Science and Technology Policy. Memorandum for the heads of executive departments and agencies, Subject: Science and Technology Priorities for the FY 2012 Budget. Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-30.pdf

OMB and OSTP. (2013).Offices of Management and Budget and Science and Technology Policy. Memorandum for the heads of executive departments and agencies, Subject: Science and Technology Priorities for the FY 2015 Budget. Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-16.pdf

OSTP. Office of Science and Technology Policy Interagency Working Group. (2008). The Science of Science Policy: A federal research roadmap. Report on the Science of Science Policy to the Subcommittee on Social, Behavioral and Economic Sciences, Committee on Science, National Science and Technology Council, Office of Science and Technology Policy. Retrieved from: http://www.whitehouse.gov/files/documents/ostp/NSTC%20Reports/39924_PDF%20Proof.pdf

Reed JH, &Jordan, G. (2007). Using systems theory and logic models to define integrated outcomes and performance measures in multi-program settings. *Research Evaluation,* 16(3): 169-181.

Reed, JH., & Jordan, G. (2007). Impact Evaluation Framework for Technology Deployment Programs, U.S. DOE. Retrieved from: http://www1.eere.energy.gov/analysis/pdfs/impact_framework_tech_deploy_2007_main.pdf

Reimsbach-Kounatze C. OECD (2015). The Proliferation of Big Data and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis, OECD Digital Economy Papers No. 245. Retrieved from: http://www.oecd-ilibrary.org/science-and-technology/the-proliferation-of-big-data-and-implications-for-official-statistics-and-statistical-agencies_5js7t9wqzvg8-en.

Research Excellence Framework. (REF). Retrieved from: http://www.ref.ac.uk/

Rogers, E. (2003). *Diffusion of Innovations Fifth Edition*, New York: New York Free Press.

Rogers J, Youtie J, & Kay L. (2012). Program-level assessment of research centers: Contribution of Nanoscale Science and Engineering Centers to US Nanotechnology National Initiative goals. *Research Evaluation, 21(5)*:369-380.

Romer PM. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5):71-102.

Ruegg R, & Jordan G. (2007). *Overview of evaluation methods for R&D programs.* U.S. Department of Energy. Retrieved from: http://www1.eere.energy.gov/analysis/pdfs/evaluation_methods_r_and_d.pdf

Ruegg R, & Feller I. (2003). *A toolkit for evaluating public R&D investment: Models, methods, and findings from ATP's first decade.* Gaithersburg, MD: Economic Assessment Office.

Advanced Technology Program, National Institute of Standards and Technology. Retrieved from: http://www.atp.nist.gov/eao/gcr03-857/contents.htm

Ruegg R, O'Connor A, Lomis R. (2014). *Evaluating realized impacts of DOE/EERE R&D programs: Standard Impact Evaluation Method.* Report No.: DOE/EE-1025. Retrieved from: http://www1.eere.energy.gov/analysis/pdfs/evaluating_realized_rd_mpacts_9-22-14.pdf

Solow RM. (1957).Technical change and the aggregate production function. *The Review of Economics and Statistics*, 39(3):312-320.

STAR METRICS: National Academy of Sciences. Retrieved from: http://sites.nationalacademies.org/PGA/fdp/PGA_057189

Tassey, G. (2007). *The technology imperative*. Cheltenham: Edward Elgar Publishing.

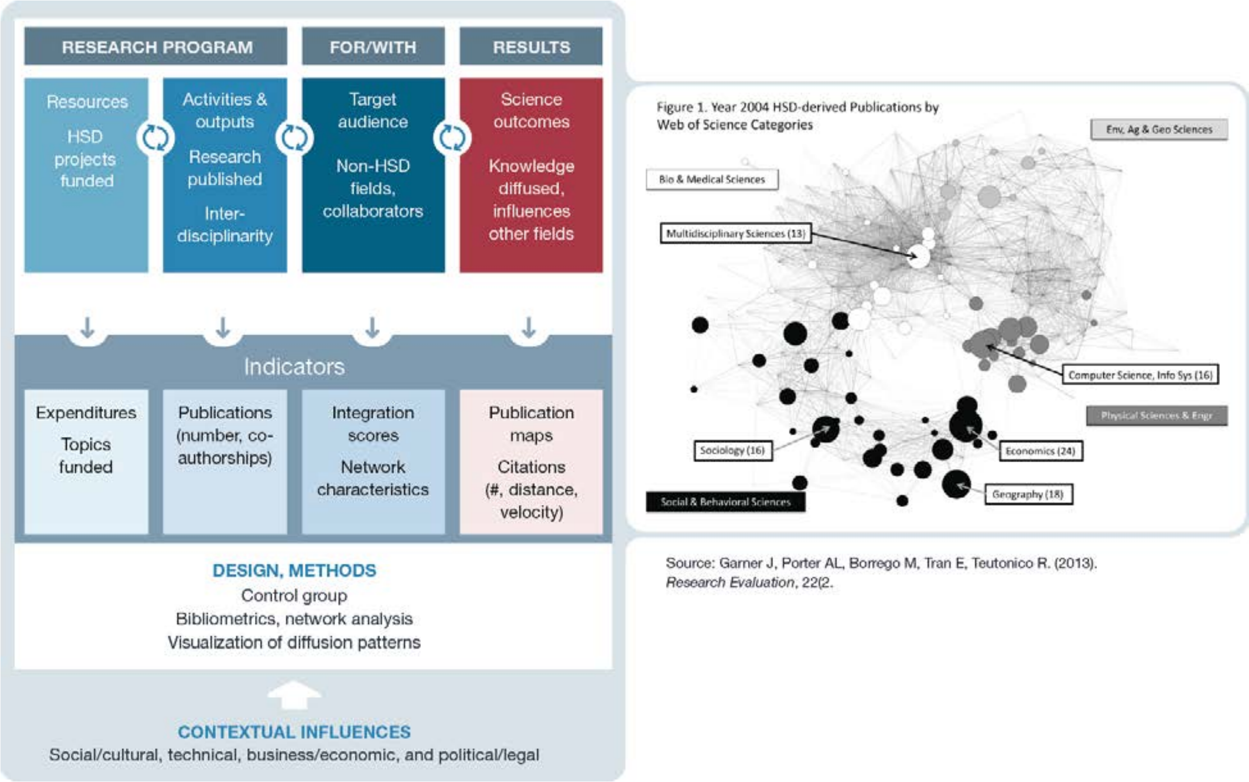
APPENDIX A. Examples of Applications of the RTD Logical Framework

The generic high-level logic model and menu of indicators must be customized to reflect the unique nature of the RTD program and evaluation context. Four distinct examples are provided below that illustrate such customizations. The examples use a common format for explanatory purposes including a high-level diagram, accompanying set of indicators, and list of evaluation methods. In actual use, this format works only for a summary of an evaluation framework. The evaluation framework itself would consist of a more complex logic diagram that is accompanied by a document that explains the program and program logic/theory. The document would also contain tables of indicators with methods for collecting data and analyzing these indicators.

EXAMPLE 1. The Human and Social Dynamics Basic Research Program

The first example was drawn from a study of the National Science Foundation (NSF) Human and Social Dynamics (HSD) Program (Garner et al. 2013). More detailed information about the evaluation and its findings is provided in section F-1 of Appendix F.

In brief, the primary goal of the HSD program was to advance knowledge to other fields of science. To do so, the program intended researchers from multiple disciplines to collaborate in the conduct of research. The results of the program are scientific and aimed at the research/university community. Because of the multidisciplinary nature of the research, indicators of success included measures of whether the research itself was interdisciplinary (e.g., the integration score) as well as whether results were influential in multiple research communities across disciplinary boundaries (see Figure A-1). Given this focus, the evaluation made heavy use of bibliometric techniques that relied on publications and citations to those publications while also recognizing the limitations of that approach.


RESEARCH PROGRAM
FOR/WITH
RESULTS
Resources
HSD projects funded
Activities & outputs
Research published
Inter-disciplinarity
Target audience
Non-HSD fields, collaborators
Science outcomes
Knowledge diffused, influences other fields
Indicators
Expenditures
Topics funded
Publications (number, co-authorships)
Integration scores
Network characteristics
Publication maps
Citations (#, distance, velocity)
DESIGN, METHODS
Control group
Bibliometrics, network analysis
Visualization of diffusion patterns
CONTEXTUAL INFLUENCES
Social/cultural, technical, business/economic, and political/legal
Figure 1. Year 2004 HSD-derived Publications by Web of Science Categories
Env, Ag & Geo Sciences
Bio & Medical Sciences
Multidisciplinary Sciences (13)
Computer Science, Info Sys (16)
Physical Sciences & Engr
Sociology (16)
Economics (24)
Geography (18)
Social & Behavioral Sciences

Source: Garner J, Porter AL, Borrego M, Tran E, Teutonico R. (2013). *Research Evaluation*, 22(2.

Figure A-1. A Logical Framework for a Basic Science Program

Reference

Garner J, Porter AL, Borrego M, Tran E, Teutonico R. (2013).Facilitating social and natural science cross-disciplinarity: Assessing the human and social dynamics program. *Research Evaluation*, 22(2):134-144.

EXAMPLE 2. Research to Inform Regulatory Policy

The second example is of a RTD program specifically designed to inform regulatory policy. Requirements in the Clean Air Act (CAA) link research, assessment of scientific knowledge, science-policy decisions such as protective health standards, and evaluation (Pahl et al., 2008)**.** When enacted in 1970, the CAA required the U.S. Environmental Protection Agency (EPA) to establish national ambient air quality standards (NAAQS) for six air pollutants including airborne particulate matter (PM) (CAA, 1991). Referred to as *criteria* air pollutants, these six pollutants are created by numerous anthropogenic and natural sources and are harmful to public health and the environment.

The CAA requires that the EPA Administrator promulgate ambient air quality standards that are based on these criteria and requisite to protect the public health with an adequate margin of safety. The 1977 amendments to the CAA added a requirement to evaluate and, if appropriate, revise existing criteria for these pollutants every five years to: reflect advances in scientific knowledge on the effects of the pollutant on public health and welfare; and recommend to the EPA Administrator any new national ambient air quality standards and revisions of existing criteria and standards that may be appropriate. In adding these 1977 evaluation provisions, Congress recognized the need for evaluation to inform decisions by EPA and by the executive, legislative, and judicial branches of government (U.S. Congress House Committee on Interstate and Foreign Commerce, 1977).

By documenting new research knowledge, the use of the latest scientific knowledge, and critical judgments about causality related to NAAQS decisions, an integrated science assessment (ISA) serves to document the impact of the research program (EPA, 2009). Each ISA identifies the current state of knowledge regarding the relationship of a NAAQS pollutant(s) and human morbidity and mortality. Analysis of successive ISAs identifies where knowledge has improved and uncertainties have been reduced. Moreover, the content of ISAs can be analyzed to identify which new findings – stemming from particular research publications – represent the critical new scientific knowledge that has been gained. Tying the publications back to federally-sponsored particulate matter research serves as an indicator of the success of participating agencies' research programs (National Science and Technology Council, 2007).[7]

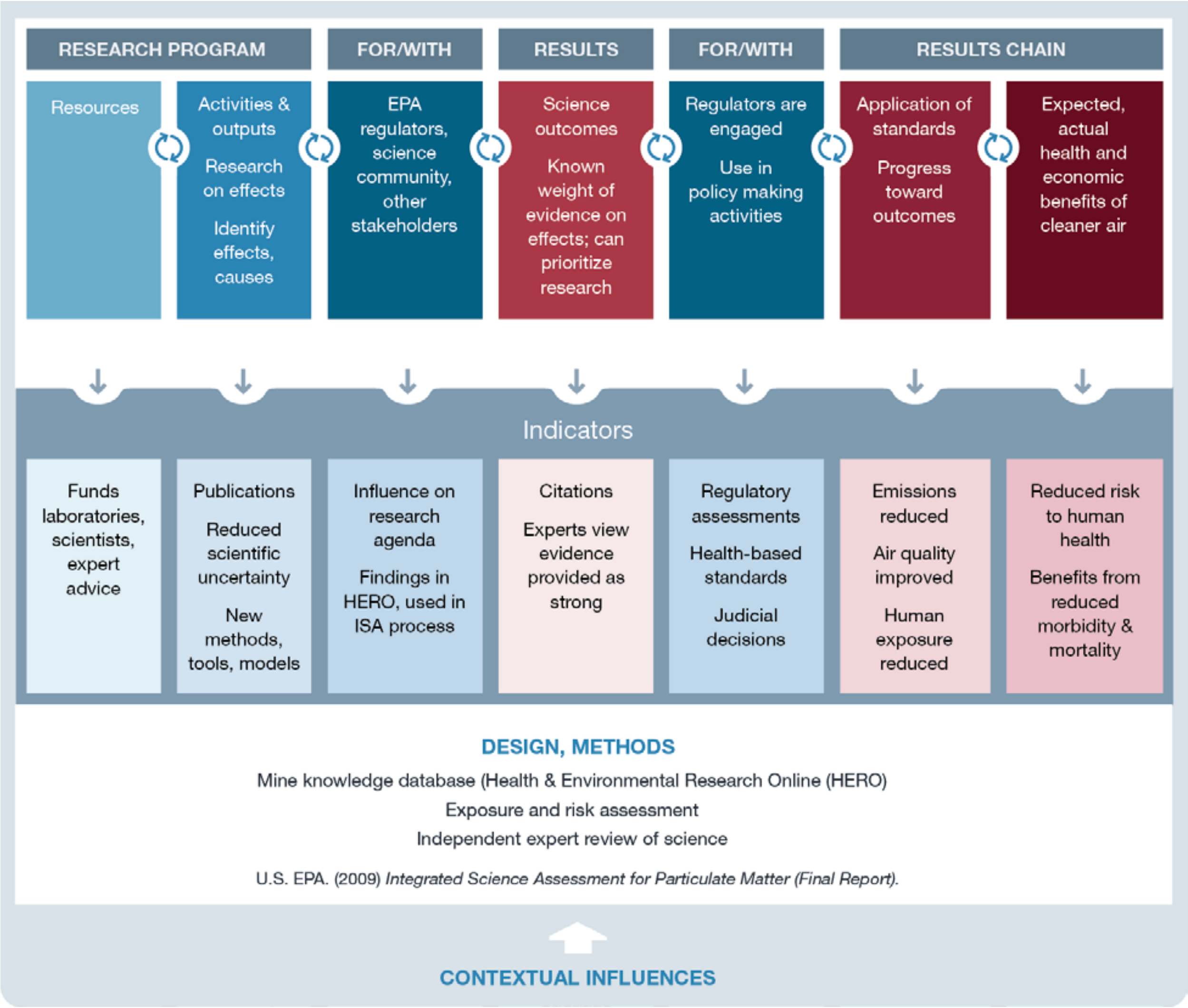
As EPA and federal partners develop multiyear research plans (e.g., Clear Air Research Multi-Year Plan 2008 Committee on Research Priorities for Airborne Particulate Matter, 2004) that link research themes to proposals for project-level plans, early indicators of the value of such research includes scientists' knowledge of key research questions that need to be investigated to help reduce uncertainty and further improve the knowledge base. This knowledge, which is available from successive ISA's, helps scientists from many federal research programs develop new knowledge related to causality, thereby contributing also to scientific outcomes. In 2010, EPA created a publicly accessible and transparent database known as *Health and Environmental Research Online* (HERO, www.hero.epa.gov) to document the use of the latest research on particulate matter to inform scientific judgments about causality for the Particulate Matter (PM) National Ambient Air Quality Standards (NAAQS) (e.g., judgments about the indicator, averaging time, level, and statistical form of the NAAQS). Thus, citation and use of research studies in the HERO database are indicators of scientific outcomes (see Figure A-2).

Intermediate-term outcomes occur as protective health standards are updated, resulting in reduced emissions, improved in air quality, and reduced human exposure. The ISA and HERO also document the use of science to respond to policy questions. Federal particulate matter research publications include quantified long-term health impacts related to the NAAQS.

[7] Such an approach is most likely be useful for the assessment of preclinical, clinical, or epidemiologic research that is intended to directly tie changes in air pollution to changes in morbidity and mortality.

Specifically, US cities with improvements in PM air quality demonstrate corresponding improvements in health, as measured by life expectancy (Pope, Ezzati, & Dockery , 2009). Intermediate outcomes (e.g., the promulgation of new standards based on the scientific evidence) and long-term outcomes (e.g., changes in human health response and risks; decreased social costs associated with air pollution) are ultimate indicators of the value of research supported by EPA and its federal partners.

While quantitative techniques such as citation analysis may suffice for identification of whether federally-supported science has contributed to changes in knowledge, expert judgment would play a vital role in identifying the true extent to which research, relative to (and in conjunction with) other efforts, are important in providing the scientific backdrop to regulatory decisions and downstream impacts.


RESEARCH PROGRAM
FOR/WITH
RESULTS
FOR/WITH
RESULTS CHAIN
Resources
Activities & outputs
Research on effects
Identify effects, causes
EPA regulators, science community, other stakeholders
Science outcomes
Known weight of evidence on effects; can prioritize research
Regulators are engaged
Use in policy making activities
Application of standards
Progress toward outcomes
Expected, actual health and economic benefits of cleaner air
Indicators
Funds laboratories, scientists, expert advice
Publications
Reduced scientific uncertainty
New methods, tools, models
Influence on research agenda
Findings in HERO, used in ISA process
Citations
Experts view evidence provided as strong
Regulatory assessments
Health-based standards
Judicial decisions
Emissions reduced
Air quality improved
Human exposure reduced
Reduced risk to human health
Benefits from reduced morbidity & mortality
DESIGN, METHODS
Mine knowledge database (Health & Environmental Research Online (HERO)
Exposure and risk assessment
Independent expert review of science
CONTEXTUAL INFLUENCES

U.S. EPA. (2009) *Integrated Science Assessment for Particulate Matter (Final Report).*

Figure A-2. A Logical Framework for Research and Science Judgments that Inform Protective Health Standards

References

Clean Air Act 108 (1991). Air Quality Criteria and Control Techniques, 109, National Ambient Air Quality Standards, 110, Implementation Plans 7408-7410.

EPA. (2008). Clean Air Research Multi-Year Plan: 2008 - 2012, EPA 620/R-08/001, June 2008. Retrieved from Health and Environmental Research Online. Retrieved from: www.hero.epa.gov

EPA. (2009). *Integrated Science Assessment for Particulate Matter (Final Report).* Washington, DC: Report No.: EPA/600/R-08/139F. Retrieved from: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=216546

EPA, (2009). Review of Integrated Science Assessment for Particulate Matter (Second External Review Draft, July 2009). Clean Air Act Scientific Advisory Committee. November 24, letter to EPA Administrator Jackson from Chair of the Clean Air Scientific Advisory Committee. Retrieved from: http://yosemite.epa.gov/sab/sabproduct.nsf/151B1F83B023145585257678006836B9/$File/EPA-CASAC-10-001-unsigned.pdf

Pahl D, Wilson W, Evans R, Kowalski L, Vickery J, and Costa D. (2008). Research, policy, and evaluation: systematic interaction informs air quality decisions. *Research Evaluation,* 17(4):251-263.

Pope CA 3rd, Ezzati M, & Dockery DW. (2009). Fine-particulate air pollution and life expectancy in the United States. *New England Journal of Medicine*, 360(4):376-386. Retrieved from: http://www.nejm.org/doi/full/10.1056/NEJMsa0805646

U.S.C. House Committee on Interstate and Foreign Commerce. (1977). *Clean Air Act Amendments of 1977.* Washington, D.C.: U.S. Government Printing Office. House Report No. 95-294 to accompany H.R. 6161 (95th Congress, 1st session)

EXAMPLE 3. Department of Energy Wind Energy Technology Development

A technology program carried out by the U.S. Department of Energy (DOE) to develop cost-effective renewable power generation from wind energy serves as the third example (Ruegg & Thomas, 2009). In this example, the question addressed by evaluation was: what evidence is there linking research outputs of DOE's Wind Energy Program to key innovations in commercial wind power generation for both utility-scale and distributed-use power markets?

Pre- and post-program conditions in system costs, performance, and power generation by wind turbines in the U.S. were documented. Yearly DOE wind program investment costs and research awards to industry and universities were recorded. Paths of knowledge flow were documented through multiple evaluation techniques, including bibliometrics, patent citation analysis, analysis of databases, simple depiction of the network, and interviews with industry and government experts (see Figure A-3).

The study produced substantial quantitative and qualitative evidence linking DOE's Wind Energy Program to key innovations in commercial power generation for both utility-scale and distributed-use power markets.

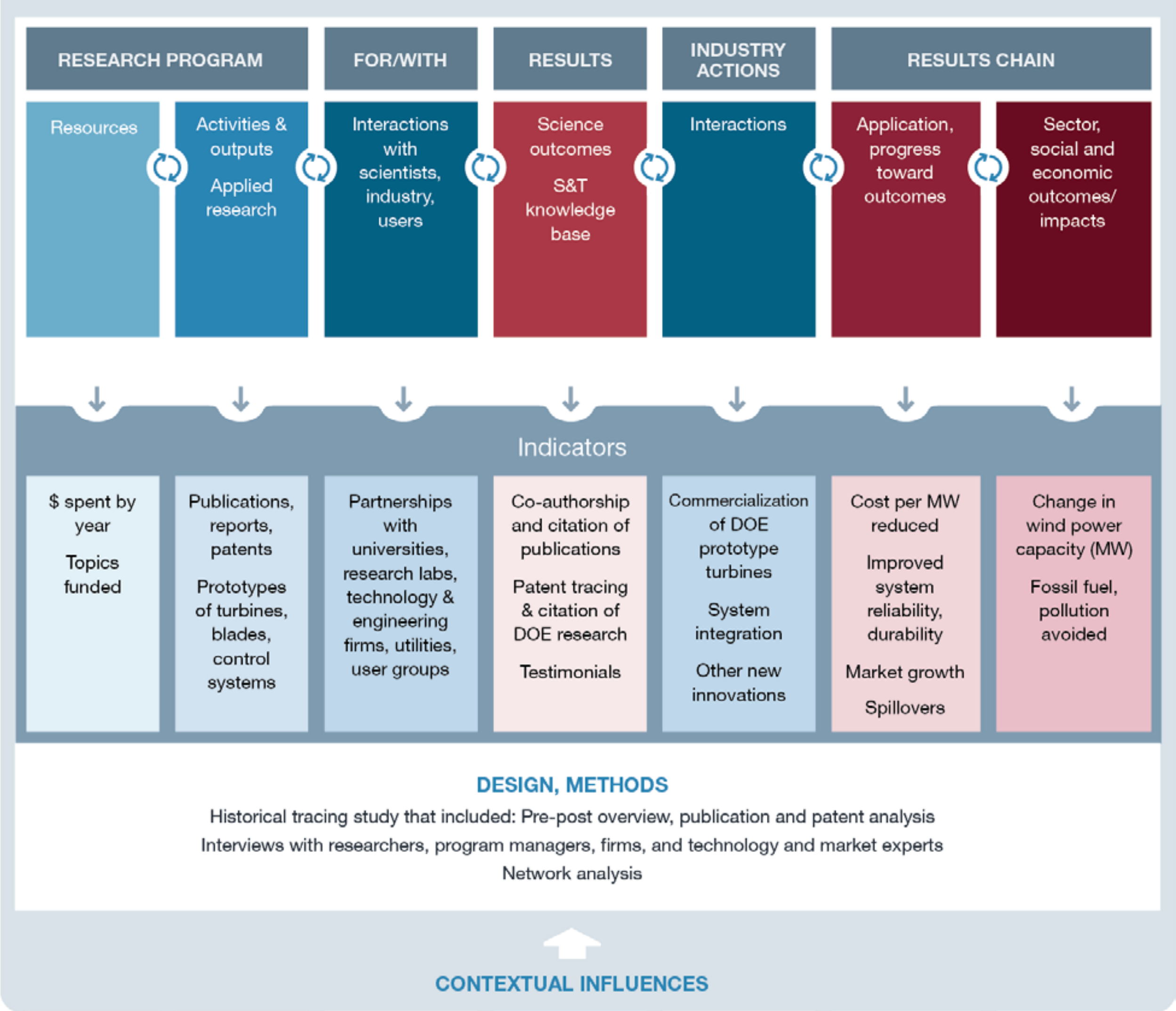

RESEARCH PROGRAM
FOR/WITH
RESULTS
INDUSTRY ACTIONS
RESULTS CHAIN
Resources
Activities & outputs
Applied research
Interactions with scientists, industry, users
Science outcomes
S&T knowledge base
Interactions
Application, progress toward outcomes
Sector, social and economic outcomes/ impacts
Indicators
$ spent by year
Topics funded
Publications, reports, patents
Prototypes of turbines, blades, control systems
Partnerships with universities, research labs, technology & engineering firms, utilities, user groups
Co-authorship and citation of publications
Patent tracing & citation of DOE research
Testimonials
Commercialization of DOE prototype turbines
System integration
Other new innovations
Cost per MW reduced
Improved system reliability, durability
Market growth
Spillovers
Change in wind power capacity (MW)
Fossil fuel, pollution avoided
DESIGN, METHODS
Historical tracing study that included: Pre-post overview, publication and patent analysis
Interviews with researchers, program managers, firms, and technology and market experts
Network analysis
CONTEXTUAL INFLUENCES

Figure A-3. A Logical Framework for R&D Linkages with Commercial Wind Generation

Reference

Ruegg, R,& Thomas, P. (2009). *Linkages from DOE's Wind Energy Program R&D to Commercial Renewable Power Generation*. U.S. Department of Energy, Energy Efficiency & Renewable

Energy. Retrieved from: http://www1.eere.energy.gov/analysis/pdfs/wind_energy_r_and_d_linkages.pdf

EXAMPLE 4. Innovation in Health Care Delivery

This notional example, based on no specific evaluation study, illustrates a health services delivery research program (see Figure A-4). The program is intended to mine patient medical records (suitably de-identified as required to protect privacy) to identify successful innovations in health care delivery that serve to improve health outcomes at lower cost. The program's primary activity is to build mechanisms for exchanging data across diverse health care delivery systems and patient populations. This would create integrated databases of patient records that allow for large-scale data mining efforts. For the program to be a success, a strong network needs to be developed that consists of participants who agree to provide data of comparable depth and quality to allow for the unified database to be built and maintained. Therefore, initial indicators of success lie in the:

- Development of partnerships;
- Agreement among partners on the goals of the activity;
- Agreement upon a standard format for medical records that is both attainable by the partners and sufficiently detailed with respect to patient backgrounds, medical activities, costs, and patient outcomes to allow for analysis; and
- Development of the database itself, with participants.

Data underlying these indicators would be collected from interviews with participants. The interviews would also be used to gather information from a process perspective to capture challenges that need to be overcome and areas for future improvement.

A critical assumption in the program is that mining of the integrated dataset would identify innovations in healthcare delivery that are lower-cost or health-improving. Additionally, the goals of the program are contingent on the translation of those innovations into procedural changes and the subsequent adoption of those changes across the network of participants (and beyond). The program, therefore, needs to translate initial findings into practical recommendations that physicians can adopt; disseminate those findings to the physicians in (and beyond) the network; and provide information to patients explaining the rationale for any change from the accepted standard of care that justifies the new, recommended treatment procedures. Especially if the program is publicly-funded, another dissemination goal would be to catalyze randomized controlled trials of the innovation(s) relative to the accepted standard of care to provide stronger evidence for the innovations, thereby speeding acceptance throughout the health care delivery system. Indicators of the success of dissemination efforts lie in whether:

- Stakeholders (physicians, patients, and insurers) know of the findings and recommendations;
- Stakeholders have accepted the findings and recommendations as legitimate;
- Stakeholders have considered whether to implement the findings;
- Stakeholders have been accepted and internalized the findings.

Data underlying these indicators would be collected from stakeholder surveys. A pre-post design, in which baseline knowledge and attitudes would be collected even before innovations are identified, would be useful so as to be able to attribute results to the program itself. If more resources were available for evaluation purposes, the surveys could be extended to nonparticipants (e.g., physicians in health systems outside the ambit of the program or residents of different localities, etc.) to control for underlying changes in attitudes and knowledge unrelated to the program itself.

Ultimate objectives for the program lie in reducing the cost of medical care (e.g., eliminating unnecessary tests or preventable hospitalizations) and reducing patient morbidity and mortality. If successful, the program would change the behavior both of physicians and the patients they serve.

Initial indicators of success would come from within the health systems themselves; the database not only serves as a research tool to identify the innovations but also as a mechanism for identifying whether participating physicians have changed practices. It also serves for assessing the cost and health status implications of those changes. To determine whether the innovations have been disseminated beyond the program participants themselves, initial indicators would be whether randomized clinical trials based on those innovations are fielded and successful. Publication of the results of such trials in major medical journals (and analysis of editorials accompanying the results of those trials) would identify whether the medical community considers these innovations to be substantial. Changes in medical treatment guidelines that recommend the innovations as the new standard of care could be monitored and collected as well. Effects on cost and health status at a national scale may be difficult to identify, especially in the case of a fragmented national health program, as neither surveys nor national-level health indicators may be sufficiently precise to observe the effects of such change.

As in all of the examples, the context in which the research program is fielded matters greatly. The barriers to launching such a program in the context of a unified national, single-payer health care delivery system would need to be reduced. A standard format for medical records and an organized set of medical stakeholders would already exist. Moreover, the incentives within the system would be strong to launch such an effort as the national government, which sets health care prices and monitors expenditures, would have both the overall incentive to reduce costs and the bureaucratic means to ensure that best practices identified would be disseminated across the system. In a fragmented system (such as that exists within the United States), more upfront effort likely would be required to forge the underlying partnerships and create the data system, especially if it were necessary to work across groups of stakeholders (e.g., multiple health care delivery networks in a locality or region, multiple insurance companies). Moreover, it might be more difficult to align the incentives of the stakeholders in disseminating and using best practices; insurance companies and physicians, for example, might interpret data generated by the research differently.

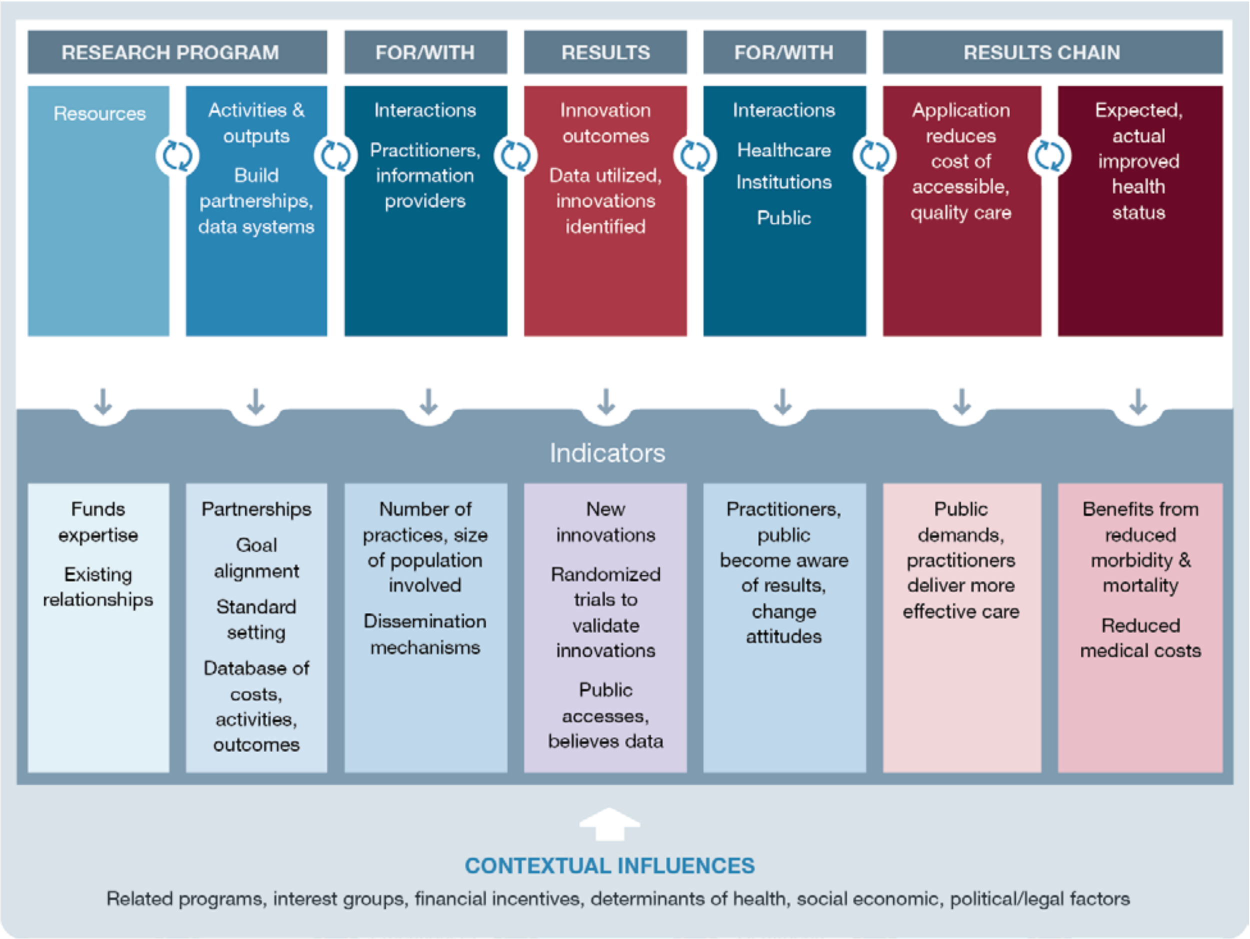

RESEARCH PROGRAM
FOR/WITH
RESULTS
FOR/WITH
RESULTS CHAIN
Resources
Activities & outputs
Build partnerships, data systems
Interactions
Practitioners, information providers
Innovation outcomes
Data utilized, innovations identified
Interactions
Healthcare Institutions
Public
Application reduces cost of accessible, quality care
Expected, actual improved health status
Indicators
Funds expertise
Existing relationships
Partnerships
Goal alignment
Standard setting
Database of costs, activities, outcomes
Number of practices, size of population involved
Dissemination mechanisms
New innovations
Randomized trials to validate innovations
Public accesses, believes data
Practitioners, public become aware of results, change attitudes
Public demands, practitioners deliver more effective care
Benefits from reduced morbidity & mortality
Reduced medical costs
CONTEXTUAL INFLUENCES
Related programs, interest groups, financial incentives, determinants of health, social economic, political/legal factors

Figure A-4. A Logical Framework for Innovation in Healthcare Delivery to Reduce Costs

APPENDIX B. Glossary with References

Terms and Definitions

Activities: Actions taken or work performed through which inputs, such as funds, technical assistance and other types of resources are mobilized to produce specific outputs. (OECD-DAC)

Applied research: Original investigation undertaken in order to acquire new knowledge. It is, however, directed primarily towards a specific practical aim or objective. (OECD-Frascati)

Attribution: The assertion that certain events or conditions were, to some extent, caused or influenced by other events or conditions. This means a reasonable [causal] connection can be made between a specific outcome and the actions and outputs of a government policy, program, or initiative. (EPA)

Baseline study: An analysis describing the situation prior to a [program], against which progress can be assessed or comparisons made. (OECD-DAC)

Basic research: Experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundation of phenomena and observable facts, without any particular application or use in view. (OECD-Frascati)

Beneficiaries: The individuals, groups, or organizations, whether targeted or not, that benefit, directly or indirectly, from the [program]. (OECD-DAC)

Contribution analysis: explores attribution through assessing the contribution a program is making to observed results. It sets out to verify the theory of change behind a program and, at the same time, takes into consideration other influencing factors. [This] provides reasonable evidence about the contribution being made by the program. (Mayne, 2008)

Cost-benefit and cost-effectiveness analyses: These analyses compare a program’s outputs or outcomes with the costs (resources expended) to produce them. When applied to existing programs, they are also considered a form of program evaluation. Cost-effectiveness analysis assesses the cost of meeting a single goal or objective and can be used to identify the least costly alternative for meeting that goal. Cost-benefit analysis aims to identify all relevant costs and benefits, usually expressed in dollar terms. (GAO, 2011)

Evidence: Information that increases the probability of the truthfulness or accuracy of a proposition. Examples of evidence may include but are not limited to, performance measurement, research studies, program evaluation, statistical data series, and data analytics. Evidence can be quantitative or qualitative and has varied degrees of reliability. The credible use of evidence in decision-making requires an understanding of what conclusions can be drawn from the information, and equally important, what conclusions cannot be drawn from it. (OMB)

Ex-ante (evaluation): An evaluation that is performed before implantation of an intervention (prospectively). (OECD-DAC)

Ex-post (evaluation): Evaluation of an intervention after it has been completed (retrospective). It may be undertaken directly after or long after completion. The intention is to identify the factors of success or failure, to assess the sustainability of results and impacts, and to draw conclusions that may inform other interventions. (OECD-DAC)

Experimental development: Systematic work, drawing on existing knowledge gained from research and/or practical experience, which is directed to producing new materials, products or devices, to installing new processes, systems and services, or to improving substantially those already produced or installed. (OECD-Frascati)

For whom/with whom: Program partners and the target audience the program is trying to influence. Referred to in the paper as "Interactions" and by some as "Reach."

Formative evaluation: Evaluation intended to improve performance, most often conducted during the implementation phase of projects or programs. Formative evaluations may also be conducted for other reasons such as compliance, legal requirements, or as part of a larger evaluation initiative. (OECD-DAC)

Goal(s): The higher-order objective to which a development intervention is intended to contribute. (OECD-DAC) See also **"Objective(s)"**

Impact: This paper uses the terms 'outcome' and 'impact' interchangeably recognizing the importance of effects from early progress to ultimate outcomes and differences in timing for those ultimate outcomes. Some define the term more narrowly. Positive and negative, primary and secondary long-term effects produced by an intervention, directly or indirectly, intended or unintended. (OECD-DAC)

Impact/Outcome evaluation: This form of evaluation assesses the extent to which a program achieves its outcome-oriented objectives (GAO, 2011). See Outcome evaluation.

Impact evaluation (defined as net effect): Impact evaluation is a form of outcome evaluation that assesses the net effect of a program by comparing program outcomes with an estimate of what would have happened in the absence of the program. This form of evaluation is employed when external factors are known to influence the program's outcomes, in order to isolate the program's contribution to achievement of its objectives. (GAO, 2011)

Indicator*: A variable that measures a phenomenon of interest to the evaluator. The phenomenon can be an input, an output, an outcome, a characteristic, or an attribute. (World Bank)

*Note: [An indicator can be either] a quantitative or qualitative factor or variable that provides a simple and reliable means to measure achievement, to reflect the changes connected to an intervention, or to help assess the performance of a development actor. (OECD-DAC)

Innovation: The implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations […] A common feature of an innovation is that it must have been implemented. A new or improved product is implemented when it is introduced on the market. New processes, marketing methods, or organizational methods are implemented when they are brought into actual use in the firm's operations. (OECD/Eurostat)

Input: Inputs include the labor (the range of skills, expertise and knowledge of employees), capital assets (including land and buildings, motor vehicles and computer networks), financial assets, and intangible assets (such as intellectual property which are used in delivering outputs). (OECD, 2009)

Knowledge translation: A dynamic and iterative process that includes synthesis, dissemination, exchange and ethically-sound application of knowledge to improve the health of Canadians, provide more effective health services and products and strengthen the health care system. (CIHR)

Note: It should be noted that this definition holds true for the application of knowledge for practical purposes outside of health too.

Logic model: A diagram and text that describes and illustrates the logical (causal) relationships among program elements and the problem to be solved, thus defining measurements of success. (EPA)

Monitoring: A systematic process of collecting and recording information on the progress and direction of ongoing actions, generated mainly for management purposes. (ETAN Expert Working Group)

Measure or metric: See "**Indicator**." Note: In the U.S. the term measure is used more often than "variable", and a metric is the unit of measurement for that measure. While "metric" and "indicator" are often used interchangeably, "indicator" conveys the notion that it only partially captures the measure.

Objective(s): Specific results or effects of a program's activities that must be achieved in pursuing the program's ultimate goals. (EPA)

Outcome: Changes or benefits resulting from activities and outputs. Short-term outcomes produce changes in learning, knowledge, attitude, skills or understanding. Intermediate outcomes generate changes in behavior, practice or decisions. Long-term outcomes produce changes in condition. (EPA) See also **"Impact"**

Outcome evaluation: This form of evaluation assesses the extent to which a program achieves its outcome-oriented objectives. It focuses on outputs and outcomes (including unintended effects) to judge program effectiveness but may also assess program process to understand how outcomes are produced. (GAO, 2011)

Output: The products or results of the process. These might include, for example, how many people a project has affected, their ages and ethnic groups or the number of meetings held and the ways in which the findings of the project are disseminated. (WHO)

Performance assessment: Includes both performance measurement and program evaluation. (GAO, 2011)

Performance management: The systematic process of monitoring the achievements of program activities; collecting and analyzing performance information to track progress toward planned results; using performance information and evaluations to influence decision-making and resource allocation; and communicating results to advance organizational learning and communicate results to stakeholders. (USAID)

Performance measurement: Performance measurement is the ongoing monitoring and reporting of program accomplishments, particularly progress toward pre-established goals. It is typically conducted by program or agency management. Performance measures may address the type or level of program activities conducted (process), the direct products and services delivered by a program (outputs), or the results of those products and services (outcomes). (GAO, 2011)

Program: A “program” may be any activity, project, function, or policy that has an identifiable purpose or set of objectives. (GAO, 2011) Note: We use program to be a broad set of activities.

Program evaluation: individual systematic studies conducted periodically or on an ad hoc basis to assess how well a program is working. They are often conducted by experts external to the program, inside or outside the agency, as well as by program managers. A program evaluation typically examines achievement of program objectives in the context of other aspects of program performance or in the context in which it occurs. Four main types can be identified, all of which use measures of program performance, along with other information, to learn the benefits of a program or how to improve it. (GAO, 2011)

Process evaluation: This form of evaluation assesses the extent to which a program is operating as it was intended. It typically assesses program activities' conformance to statutory and regulatory requirements, program design, and professional standards or customer expectations. (GAO, 2011)

Program theory: An explicit theory or model of how an intervention contributes to a set of specific outcomes through a series of intermediate results. The theory can have two components: a theory of change about the central mechanisms by which change(outcomes) comes about for individuals, groups, and communities and a theory of action about how the program is constructed to activate the theory of change (Funnell &Rogers, 2011)

Reach: The beneficiaries and other stakeholders of a [program]. (OECD-DAC)
See also "**Beneficiaries**.**"**

Research and experimental development (R&D): Creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications. R&D covers three activities: "**Basic research**", "**Applied research**", and "**Experimental development**." (OECD Frascati)

Stakeholders: Agencies, organizations, groups, or individuals who have a direct or indirect interest in the [program] or its evaluation. (OECD-DAC)

Target group: The specific individuals or organizations for whose benefit the [program] is undertake. (OECD-DAC)

References

Environmental Protection Agency (EPA). (2007). *Program evaluation glossary*. Office of the Administrator, Office of Policy, Office of Strategic Environmental Management, Evaluation Support Division. Retrieved from: http://ofmpub.epa.gov/sor_internet/registry/termreg/searchandretrieve/glossariesandkeywordlists/search.do?details=&glossaryName=Program%20Evaluation%20Glossary

European Technology Assessment Network (ETAN) Expert Working Group. (1999). *Report to the European Commission, Options and Limits for Assessing the Socio-Economic Impact of European RTD Programmes*. Retrieved from: ftp://ftp.cordis.europa.eu/pub/etan/docs/master-impact.pdf

Funnell S, & Rogers P. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: Jossey-Bass.

Government Accountability Office (GAO). (2011). *Performance measurement and evaluation: Definitions and relationships*. (GAO-11-646SP). Retrieved from: http://www.gao.gov/products/GAO-11-646SP

Mayne J. (2008).Contribution analysis: An approach to exploring cause and effect. *ILAC Brief Number* 16.Retrieved from: http://www.cgiarilac.org/files/ILAC_Brief16_Contribution_Analysis_0.pdf

Office of Management and Budget (OMB). (2013). *Circular A-11, Strategic plans, annual performance plans, performance reviews, and annual program performance reports, 200.21 Definitions*. Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/performance/a-11_part-6_2013.pdf

Organization for Economic Co-operation and Development , Development Assistance Committee (OECD-DAC). (2002). *Glossary of Key Terms in Evaluation and Results Based Management*. Retrieved from: http://www.oecd.org/development/peer-reviews/2754804.pdf

Organization for Economic Co-operation and Development (OECD). (2002)*. Frascati Manual: proposed standard practice for surveys on research and experimental development* (6th ed.). Paris, France. Retrieved from: www.oecd.org/sti/frascatimanual

Organization for Economic Co-operation and Development (OECD). (2009). *Innovation in Firms: A Microeconomic Perspective*. Paris, France.

Organization for Economic Co-operation and Development, Development Assistance Committee (OECD-DAC). (2002). *Glossary of Key Terms in Evaluation and Results Based Management*. Retrieved from: http://www.oecd.org/development/peer-reviews/2754804.pdf Other languages available: Chinese, Italian, German, Russian, Spanish, etc.

Organization for Economic Co-operation and Development & Eurostat (OECD/Eurostat). (2005). *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data* (3rd ed.). Paris, France. Retrieved from: http://www.oecd.org/innovation/inno/oslomanualguidelinesforcollectingandinterpretinginnovationdata3rdedition.htm

RAND Europe. (2013). *Measuring research: A guide to research evaluation frameworks and tools*. Retrieved from: http://www.rand.org/pubs/monographs/MG1217.html

United States Agency for International Development (USAID). (2011). *USAID Evaluation policy*. Retrieved from: http://www.usaid.gov/sites/default/files/documents/1868/USAIDEvaluationPolicy.pdf

World Health Organization (WHO). 2013. *Health impact assessment: Glossary of terms used*. Retrieved from: http://www.who.int/hia/about/glos/en/index.html

APPENDIX C. Logic Model Resources

On Line Training

Centers for Disease Control and Prevention. (2010). *Logic Model*. Retrieved from: http://www.cdc.gov/nccdphp/dnpao/hwi/programdesign/logic_model.htm

Environmental Protection Agency.(n.d.).Pictureing your program: An introduction to logic modeling web-based online training course. Retrieved from: http://www.epa.gov/evaluate/lm-training/index.htmTaylor-Powell E., & Henert E. (2008). *Developing a Logic Model: Teaching and Training Guide*. Retrieved from: http://www.uwex.edu/ces/pdande/evaluation/pdf/lmguidecomplete.pdf

W. K. Kellogg Foundation. (2005). *Logic Model Development Guide*. Retrieved from: http://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide

Books

Frechtling J. (2007). *Logic Modeling Methods in Program Evaluation*. San Francisco: Jossey-Bass.

Funnell S, & Rogers P. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: Jossey-Bass.
Knowlton LW, & Phillips, CC. (2013). *The Logic model guidebook: Better strategies for great results (2nd Ed.)*. Thousand Oaks, CA: Sage

Mayne J, & Stern E. (2013). *Impact evaluation of natural resource management research* programs: a broader view. (ACIAR Impact Assessment Series Report No. 84). Canberra: Australian centre for International Agricultural Research. . Retrieved from: http://aciar.gov.au/files/ias84.pdf

Articles

Funnell S. (2000). Developing and using a program theory matrix for program evaluation and performance monitoring." In P. Rogers, T. Hacsi, A. Petrosino, and. T. Huebner (eds.), *Program Theory in Evaluation: Challenges and Opportunities*. New Directions for Evaluation, no. 87. San Francisco: Jossey-Bass.

McLaughlin J A, and Jordan GB. (1999). Logic Models: A tool for telling your performance story. *Evaluation and Program Planning*, *22*(1):65–72.

McLaughlin J A, and Jordan GB. (2004). Logic models. In Wholey JS, Hatry, HP, & Newcomer, KE (Eds.), *Handbook of Practical Program Evaluation* (2nd ed.). San Francisco, CA: Jossey-Bass.

Montague S, & Porteous N L. (2012). *The case for including reach as a key element of program theory. Evaluation and Program Planning, 36(1):177-183.*

APPENDIX D. Partial Listing of Specific Indicators by Outcome Area and Evaluation Question

Table D-1. Indicators of Knowledge Advance, measured by Published Research

Policy question	Indicator/Metric	Data source
How much quality research has a program produced?	Number of peer-reviewed publications	Publications lists
	Number of peer-reviewed publications per unit of funding	Publications lists
	Number of peer-reviewed publications per unit of time	Publications lists
	Journal impact factor-weighted number of peer-reviewed publications	Publications lists
	Number of conference presentations	Publications lists
	Number of white papers/non-refereed reports	Publications lists
How has the published research affected the research community?	Total citations to peer-reviewed publications in peer-reviewed journals	Publications lists
	Normalized citations to publications relative to field average, relative to journals in which publications appear	Publications lists
	Citation velocity of peer-reviewed publications (normalized to field, or not)	Publications lists
	Citations in grey literature (e.g., via Google Scholar)	Publications lists
	Analysis of communities/journals in which citations appear (e.g., using map of science)	Publications lists
	Comparison of communities/journals in which publications appear and communities/journals in which citations appear (e.g., using map of science)	Publications lists
	Publications identified by expert review as seminal/key (e.g., high-impact research, transformative research, germinating new fields)	Publications lists
	Publications introducing terms, theorems, approaches that become widely used in the research community	Publications lists
How has the published research affected industry?	Citations to publications in patent applications	USPTO database
	Citations to publications in invention disclosures	University data
	Publications identified by expert review as seminal/key to developing new products/processes/services/industry lines of research	Publications lists

Table D-2. Indicators of Research Collaborations

Policy question	Indicator/Metric	Data source
How have the research collaborations of participants changed subsequent to the program?	Density of research network as measured through publications	Publications lists
	Centrality of program participants in field-level research networks as measured through publications	Publications lists
	Role of program in creating new collaborations among participants	Interviews/surveys
	Role of program in creating new collaborations between participants and others	Interviews/surveys
	Role of program in enhancing existing collaborations among participants	Interviews/surveys
	Role of program in enhancing existing collaborations between participants and others	Interviews/surveys
	New collaborations between researchers from different backgrounds/disciplines	Interviews/surveys
	New collaborations between researchers from different universities	Interviews/surveys
	New collaborations between researchers from different countries	Interviews/surveys
	New collaborations between researchers from different sectors	Interviews/surveys
How do program participants integrate knowledge differently in their research subsequent to program participation?	Change in research cited by program participants in their publications (e.g., integration score)	Publications lists
	Value of program in changing sources of knowledge/disciplinary focus of research	Interviews/surveys
	Increase in number of collaborators	Interviews/surveys or publication list
	Increase in diversity of collaborations based on discipline of collaborations	Interviews/surveys or publication list
	Increase in diversity of collaborations based on university/geographic location of collaborators	Interviews/surveys or publication list
	Increase in diversity of collaborations based on sector	Interviews/surveys or publication list
Are collaborative publications more effective than previous research?	Increase in average citation rate of authored publications	Publications lists
	Increase in journal impact factor of authored publications	Publications lists
	Increase in publication rate (publications per unit of time)	Publications lists
	Increase in publication rate (publications per unit of funding)	Publications lists

Table D-3. Indicators of Other Research Outputs

Policy question	Indicator/Metric	Data source
What types of other knowledge outputs are associated with the program?	Number and type of data systems/software developed	Annual reports/program-collected data (e.g., survey)
	Number and type of resources/research tools developed	Annual reports/program-collected data (e.g., survey)
	Number and type of research databases developed	Annual reports/program-collected data (e.g., survey)
	Number and type of samples developed	Annual reports/program-collected data (e.g., survey)
How have these research outputs been used by the research community?	Number (and distribution e.g., by sector) of downloads of data/software	Google analytics/registrations
	Number (and distribution e.g., by sector) of users of knowledge outputs	Surveys/interviews
	Value of knowledge outputs to users	Surveys/interviews

Table D-4. Indicators of Intellectual Property Protected

Policy question	Indicator/Metric	Data source
How much protected IP has been developed by the program?	Number of invention disclosures	University data/annual reports
	Number of patent applications	University data/annual reports; USPTO citations
	Number of patents received	University data/annual reports; USPTO citations
	Patent applications/patents received per unit of program funding	University data/annual reports; USPTO citations
	Patent applications/patents received per unit of time	University data/annual reports; USPTO citations
	Number of trademarks, copyrights, etc. received	University data/annual reports; USPTO citations
	Number of trademarks, copyrights, etc. per unit of program funding	University data/annual reports; USPTO citations
	Number of trademarkers, copyrights, etc. per unit of program time	University data/annual reports; USPTO citations
Has that IP been disseminated to industry?	Number of licenses granted	University data/annual reports; interviews
Is that IP of value?	Number of citations of patents (normalized by patent class)	USPTO
	Patent citation velocity (normalized by patent class)	USPTO
	Value of IP to company that licensed it	Interviews/expert judgment

Table D-5. Indicators for Formal Training

Policy question	Indicator/Metric	Data source
How many individuals were trained by/through the program?	Number of high school students or undergraduates participating in research	Annual reports
	Number of graduate students participating in research	Annual reports
	Number of graduate students whose theses/degrees were funded (in whole or in part) by the research	Annual reports
	Number of postdoctoral researchers/fellows participating in research	Annual reports
	Number of Master's theses supported (in whole or in part)	Annual reports
	Number of PhD theses supported (in whole or in part)	Annual reports
How diverse were the individuals trained	Number of men vs. number of women	Annual reports
	Number/percentage from underrepresented groups (e.g., Hispanic or Latino, African American, Native Americans)	Annual reports
	Number/percentage with disabilities (physical, learning/developmental)	Annual reports
	Number/percentage who were from low-income backgrounds/first generation college	Annual reports
	Number/percentage who were veterans	Annual reports
	Number/percentage from rural populations	Annual reports
What are the next steps of trainees?	Number of undergraduates continued in STEM careers immediately subsequent to completing their degrees	Alumni records/alumni interviews or surveys
	Number of undergraduates continued to graduate training upon completing their degrees	Alumni records/alumni interviews or surveys
	Number of graduate students continued to further graduate training (e.g., MS to PhD) after completing their degrees	Alumni records/alumni interviews or surveys
	Number of graduate students continued to postdoctoral research positions	Alumni records/alumni interviews or surveys
	Number of graduate students continued to faculty positions	Alumni records/alumni interviews or surveys
	Number of postdoctoral researchers continued to additional postdoctoral training	Alumni records/alumni interviews or surveys
	Number of postdoctoral researchers continued to faculty positions	Alumni records/alumni interviews or surveys
	Number of graduate students/postdoctoral researchers continued to positions in government laboratories/FFRDCs	Alumni records/alumni interviews or surveys
	Number of graduate students and postdoctoral researchers continued to STEM careers	Alumni records/alumni interviews or surveys
How do long-term career trajectories of trainees evolve?	Number/percentage of trainees who remain in research in field over the course of time period studied	Alumni records/alumni interviews or surveys
	Percentage of trainees eventually receiving tenure (compared with similar others)	CV analysis
	Time to tenure of trainees (compared with similar others)	CV analysis
	Number/percentage of trainees who remain in STEM over the course of time period studied	Alumni records/alumni interviews or surveys
	Percentage of trainees (compared with similar others) engaged in interdisciplinary research	Alumni records/alumni interviews or surveys
	Percentage of trainees (compared with similar others) engaged in research with industry/government/FFRDC researchers	Alumni records/alumni interviews or surveys

Table D-5. Indicators for Formal Training (cont'd)

Policy question	Indicator/Metric	Data source
How do long-term career trajectories of trainees evolve? (cont'd)	Percentage of trainees (compared with similar others) serving in government advisory roles (e.g., on science advisory boards)	Alumni records/alumni interviews or surveys
	Percentage of trainees (compared with similar others) serving as reviewers for government grant programs	Alumni records/alumni interviews or surveys
	Awards won by former trainees (compared with similar others)	Alumni records/alumni interviews or surveys
	Future grants/grants histories of former trainees (compared with similar others)	Government grants databases

Table D-6. Indicators of Informal Training, Education, and Outreach

Policy question	Indicator/Metric	Data source
What skills were gained by trainees?	Types of skills gained/learned/developed by trainees	Interviews/surveys
	Value of skills gained to trainees	Interviews/surveys
	Value of mentorship to trainees	Interviews/surveys
	Other craft skills/tacit knowledge learned by trainees	Interviews/surveys
Did the researchers integrate research and education?	Number of courses/course modules created based on research	Interviews/surveys
	Number of courses/course modules updated or enhanced based on research	Interviews/surveys
	Number of online educational resources created and disseminated	Interviews/surveys
	Number of students reached by these educational products	Interviews/surveys
	Number of outreach events conducted (e.g., K-12 classroom visits, public lectures)	Interviews/surveys
	Number of participants in these events	Interviews/surveys
	Diversity of participants in these events	Interviews/surveys
	Inclusion of/featuring of research results in informal science products (e.g., museum exhibits)	Interviews/surveys
	Descriptions of research results in lay publications/media releases	Interviews/surveys

Table D-7. Indicators of Other Effects on Investigators' Careers

Policy question	**Indicator/Metric**	**Data source**
Has the program assisted investigators in advancing their careers?	Number of transitions to tenured positions by program participants	CV analysis; interviews
	Reduced time to tenure (as compared with similar faculty not involved in program)	CV analysis; interviews
	Number of promotions (e.g., associate professor to full professor)	CV analysis; interviews
	Reduced time to promotion (as compared with similar faculty not involved in program)	CV analysis; interviews
	Number of faculty members changing universities for 'better' positions	CV analysis; interviews
	Role of program participation in career enhancement	Key stakeholder interviews

Table D-8. Effect on Research Field

Policy question	Indicator/Metric	Data source
Has the program enhanced the leadership role of participants in the research field?	Recognition of participants as leaders in the relevant research community	Interviews with key stakeholders
	Awards granted to participants	Interviews with key stakeholders
	Participants begin to serve as journal editors	Interviews with key stakeholders
	Participants begin to serve as conference/workshop organizers	Interviews with key stakeholders
	Participants asked to serve on review panels or advisory boards relevant to program	Interviews with key stakeholders
	Participants begin to serve in leadership roles in relevant professional societies	Interviews with key stakeholders
Has the program led the research community to organize in different ways?	Number of workshops held on topic of program	Interviews with key stakeholders
	Value/impact of workshops	Interviews with key stakeholders
	New strategic planning documents or research strategies developed	Interviews with key stakeholders
	New research consortia organized	Interviews with key stakeholders
	New networks of institutions form to conduct research	Interviews with key stakeholders
Has the program led to the creation of new fields or subfields?	Number of new journals created in field	Interviews with key stakeholders
	Role of participants (if any) in launching new journals	Interviews with key stakeholders
	Number of new sections of professional societies created	Interviews with key stakeholders
	Role of participants (if any) in launching new professional society sections	Interviews with key stakeholders
	Change in the structure of research collaborations across the entire research field	Interviews with key stakeholders
Has the program led to the creation of new solicitations/grant mechanisms by science funders?	New solicitations on topic of research catalyzed by program	Interviews with key stakeholders
	Funding level of those solicitations	Analysis of government funding databases
	Content of grants awarded in response to those solicitations	Analysis of government funding databases

Table D-9. Effect on Participating Institutions/Universities

Policy question	Indicator/Metric	Data source
What effect does the program have on the universities where research is conducted?	Number of new pieces of research equipment purchased	Interviews with key stakeholders
	Use of research equipment by investigators who are not program participants	Interviews with key stakeholders
	Number of new faculty hired	Interviews with key stakeholders
	Number of new courses/course modules developed	Interviews with key stakeholders
	Number of new degree programs developed	Interviews with key stakeholders
	Changes to university or departmental policies and procedures because of program	Interviews with key stakeholders

Table D-10. Indicators of Public Impact

Policy question	Indicator/Metric	Data source
Has research associated with the program influenced stakeholder behavior?	Change in stakeholder knowledge of result of research	Interviews/surveys
	Change in stakeholder attitudes regarding research results	Interviews/surveys
	Change in stakeholder behavior based on research results	Interviews/surveys
Has research associated with the program changed public policy?	New legislation based on or influenced by research results	Interviews with key stakeholders
	New regulations based on or influenced by research results	Interviews with key stakeholders
	Changes to existing regulations based on or influenced by research results	Interviews with key stakeholders
	New guidelines/guidance documents based on or influenced by research results	Interviews with key stakeholders
	Changes to existing guidelines/guidance documents based on or influenced by research results	Interviews with key stakeholders

Table D-11. Indicators of Leveraged Funding

Policy question	Indicator/Metric	Data source
Have program participants received follow-on research funding?	Number of new grants received by researchers in topic	Analysis of grants databases
	Funding level of new grants	Analysis of grants databases
	Increased diversity of funding agencies from whom grants have been received	Analysis of grants databases
Have program participants (and their universities) received funding for IP developed during the program?	Revenue from licenses of IP generated by program	University data/annual reports; interviews
Has IP generated by the program had commercial impact?	Number of new technologies/processes/products	University data/annual reports; interviews
	Number of spin-off companies formed to commercialize protected IP	University data/annual reports; interviews
	Number of jobs created based on commercialization of technologies	Key stakeholder interviews
	Revenue to companies based on commercialization of technologies	Key stakeholder interviews
	Tax revenue generated based on commercialization of technologies	Key stakeholder interviews

APPENDIX E: Background on U.S. Federal Evaluation Requirements

Introduction

In the US, many relevant reports and articles have been published since enactment of the Government Performance Results Act of 1993 (GPRA). Additional publications were stimulated by enactment of the GPRA Modernization Act of 2010 (GPRAMA).

In preparing this background information, the authors have focused on the actual requirements communicated to the heads of federal agencies. These requirements represent the directions from the Executive Office of the President—which is responsible for directing the implementation of GPRAMA by all executive agencies. In general, these requirements are documented in OMB Circular A-11; in particular, many are presented in *Part Six—Preparation and Submission of Strategic Plans, Annual Performance Plans, Performance Reviews, and Annual Program Performance Reports*.

To supplement these requirements, the authors present relevant observations from the US Government Accountability Office (GAO) and the Congressional Research Service (CRS).

Government Performance and Results – 1993 and 2010 Requirements

It is important to recognize that GPRA is one component of a statutory framework that Congress put in place to (a) establish and maintain internal systems and controls that identify and address major performance and management challenges, and (b) identify areas at greatest risk for fraud, waste, abuse, and mismanagement. For example, the Federal Managers' Financial Integrity Act of 1982 (FMFIA) was enacted to require ongoing evaluations and reports of the adequacy of the systems of internal accounting and controls for federal agencies. FMFIA requires the General Accounting Office (GAO) to issue standards for internal control in government. Supplementing the GAO standards, Office of Management and Budget (OMB) Circular A-123 identifies specific requirements for assessing and reporting on controls, monitoring major performance and management challenges, and identifying and assessing systemic risks of fraud, waste, abuse, and mismanagement.

When GPRA was enacted in 1993, its provisions were intended to (a) help federal program managers address their performance and management needs and (b) help Congress address its policy, oversight, and budgeting responsibilities. GPRA required executive agencies to complete strategic plans in which they define their missions, establish results-oriented goals, and identify the strategies that will be needed to achieve those goals. GPRA required agencies to consult with Congress and solicit the input of others as they develop these plans. Through this strategic planning requirement, GPRA required federal agencies to reassess their missions and long-term goals periodically—as well as the strategies and resources they will need to achieve their goals.

The term *resources* was defined to include the operational processes; skills and technology; and human, capital, information, and other resources needed to achieve agency goals.

GPRA also required executive agencies to prepare annual performance plans that describe goals for the upcoming fiscal year that are aligned with their long-term strategic goals. These performance plans include results-oriented annual goals linked to the program activities displayed in budget presentations as well as indicators the agency will use to measure performance against the results-oriented goals.

GPRA also required agencies to measure performance toward the achievement of their goals in the annual performance plan and report annually on their progress in program performance reports. These reports were intended to provide important information to agency managers, policymakers, and the public on what each agency accomplished with the resources it was given—as well as information about any unmet goals and the actions needed to meet them in the future.

The Office of Management and Budget (OMB) has an important role in the management of federal government performance as well as GPRA implementation. For example, part of OMB's overall mission is to ensure that agency plans and reports are consistent with the President's budget and administration policies.

GPRA represents significant progress to address the performance and management challenges inherent in the vast and complex missions of the federal government and its agencies. However, this progress is tempered by a number of limitations. For example, GPRA did not address:

- Systematic approaches needed to respond to government-wide challenges shared by multiple federal agencies;
- Development of a government-wide inventory of federal programs – including a common practical definition of program that extends beyond a budgetary context and information about each program's outcomes;
- Challenges faced by federal managers to measure the performance and outcomes that result from federal resources invested in research and development programs; and
- The preparation, use, or methods for presenting results of program evaluations.

Federal agencies developed their first strategic plans four years after GPRA was enacted (in fiscal year 1997) and updated these plans every 3 years since then. Thus, the enactment of GPRAMA was based on an assessment of nearly thirteen years of federal experience with GPRA and on a number of GAO reports to Congress on various aspects the implementation process and its challenges.

GPRAMA was signed by President Obama on January 4, 2011. It expanded the federal government's performance management framework, retaining and amplifying some aspects of

GPRA while also addressing some of its weaknesses. GPRAMA retained requirements for strategic planning, performance planning and performance reporting on progress to achieve their missions albeit with new names and some additional requirements.

GPRAMA places an increased emphasis on agency priority-setting, government-wide cross-organizational priority-setting and collaboration to achieve shared goals, the use and analysis of goals and measurement to improve outcomes, engaging leaders in performance improvement, and on easily-accessible and timely information about these topics. For example, GPRAMA adds agency-level requirements for priority goals and quarterly progress reviews for these goals. GPRAMA adds executive-branch-wide requirements – e.g., for federal priority goals, for federal government performance plans, for an OMB performance website, and for an OMB-led inventory of all federal programs. GPRAMA requires additional government-wide reporting of key performance information on a quarterly basis.

GPRAMA requires new government-wide organizations and officials – for example, chief operating officers, performance improvement officers, and a performance improvement council – to create and manage the new processes and products. GPRAMA expands the roles and responsibilities of OMB to help manage and communicate about the new processes, products, organizations, and results. The scope and complexity of these new requirements are illustrated in a table in OMB Circular A-11 that begins at section 210.6 and extends through section 210.14.

GPRAMA requires that, within a year of its enactment, the Director of the Office of Personnel Management, in consultation with the Performance Improvement Council, identify the key skills and competencies needed by federal government personnel to develop goals, evaluate programs, and analyze and use performance information for the purpose of improving Government efficiency and effectiveness. Within two years of its enactment, the Director of the Office of Personnel Management shall work with each agency to incorporate these key skills into training for relevant employees at each agency.

Informed by recommendations from national advisors such as GAO and the congressional Research Service, GPRAMA addresses many of the limitations encountered during the thirteen years of federal experience implementing GPRA. From the program evaluation perspective however, GPRAMA does not clarify or improve on approaches to designing or using results from program evaluations that Congress described in GPRA. This limitation is significant because of GPRAMA's increased emphasis on government-wide and cross-agency coordination to achieve common goals and objectives.

Evaluation is also included in the annual Budget priority memo sent jointly by OMB and the Office of Science and Technology Policy (OSTP):

- 2012 and 2013
    - In accordance with OMB Circular A-11 and the GPRA Modernization Act of 2010, agencies should describe the targeted outcomes of research and development (R&D) programs using meaningful, measurable, quantitative metrics where possible and describe how they plan to evaluate the success of those programs.
- 2010
    - Agencies should develop outcome-oriented goals for their science, technology and innovation activities; establish timelines for evaluating the performance of these activities, and target investments toward high-performing programs in their budget submissions.
    - Agencies should support the development and use of “science and science policy” tools that can improve management of their R&D portfolios and better assess impact of their science, technology and innovation investments.
- 2009
    - Budget submissions should also describe how agencies are strengthening their capacity to rigorously evaluate their programs to determine what has been demonstrated to work and what has not. Budget submissions should show how such assessments allowed agencies to eliminate or reduce funding for less-effective, lower quality, or lower priority programs in 2011, and how they will be applied in the future.
    - Agency submissions should explain how the agency plans to take advantage of today's open innovation model in which the whole chain from research to application does not have to take place within a single lab.

References

Brass, CT. (2012). *Changes to the Government Performance and Results Act (GPRA): Overview of the New Framework of Products and Processes*. Congressional Research Service. Retrieved from: https://www.fas.org/sgp/crs/misc/R42379.pdf

GAO. Government Accountability Office. (2013). Managing for Results: 2013 Federal Managers Survey on Organizational Performance and Management Issues. Section 6—Survey of Organizational Performance and Management Issues. (Report No. GAO-13-519SP). Government-wide survey results. Retrieved from: http://www.gao.gov/special.pubs/gao-13-519sp/results.htm#question_109.

OMB. (2013). *OMB Circular A-11: Preparation, submission, and execution of the budget*. Washington, D.C.: Executive Office of the President, Office of Management and Budget.

Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/assets/a11_current_year/a11_2013.pdf

OMB and OSTP. (2013).Offices of Management and Budget and Science and Technology Policy. Memorandum for the heads of executive departments and agencies, Subject: Science and Technology Priorities for the FY 2015 Budget. Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-16.pdf

US Congress. Public Law *103-62: Government Performance and Results Act of 1993*. Retrieved from: http://www.whitehouse.gov/omb/mgmt-gpra/gplaw2m

US Congress. *Public Law 11-352: GPRA Modernization Act of 2010*. Retrieved from: http://www.gpo.gov/fdsys/pkg/PLAW-111publ352/pdf/PLAW-111publ352.pdf

Appendix F. Summary Examples of RTD Evaluations and Measurement and Evaluation Systems

F-1. Assessment of Cross-Disciplinarity, Emergence of Research Communities, and Research Influence

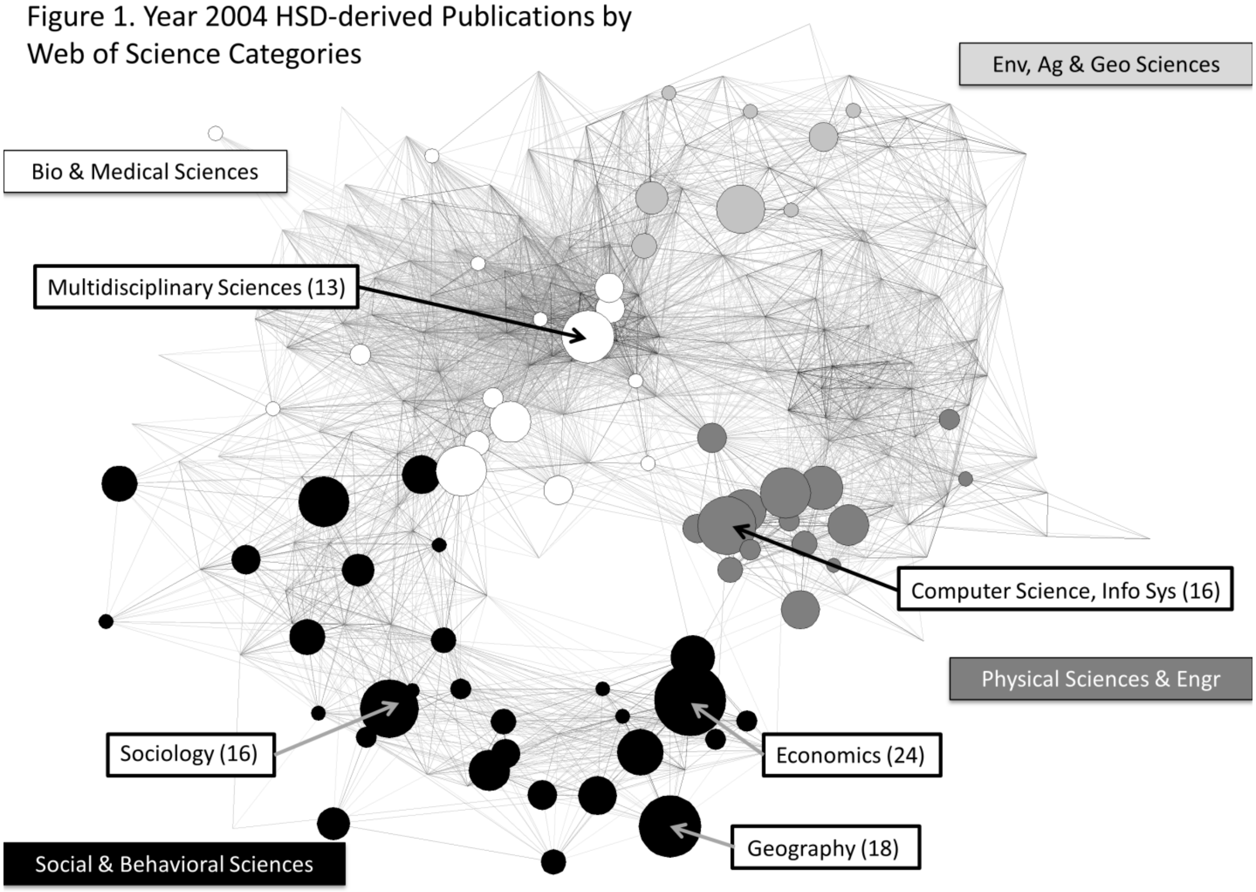
Provided by Alan Porter

An innovative study assessed the cross-disciplinary character and near term outcomes of the research supported by a unique U.S. National Science Foundation program on Human and Social Dynamics ("HSD"). Research that integrates the social and natural sciences is vital to address many societal challenges, yet is difficult to arrange, conduct, disseminate and evaluate. Publication maps and citation distance and velocity measures offer empirical measures of the interdisciplinarity and extent of research influence of the NSF Program. More generally, this research assessment illustrates the possibility of gauging and visualizing research knowledge diffusion patterns, expressly across research fields, associated with a program's research portfolio.

Figure F-1 maps publications deriving from HSD support. The Web of Science (http://thomsonreuters.com/web-of-science) is a leading database that indexes articles published in some 12,000 leading science and social science journals. This widely used database groups those journals in 224 Web of Science Categories (WoSCs). The base map locates those 224 categories based on how frequently their journal articles cite each other (Garner, Porter, Borrego, Tran, & Teutonico, 2013). The WoSCs appear as the endpoints of the black lines showing strength of citation links for the Year 2010 for Web of Science. We then overlay the Year 2004 HSD project publications upon that base map – larger nodes indicating more journal publications. The map shows exceptional diversity -- research publications deriving from this support chiefly pertain to the Social & Behavioral Sciences [lower part of the map], but extend widely into the Bio & Medical Sciences [upper left], Environmental Sciences [upper right], and Physical Sciences & Engineering [lower right] as well. Also identified in the map are the five leading WoSCs in which HSD-supported research was published (e.g., led by 24 in "Economics" journals).

The study compared the HSD project publications to those from comparison projects (also funded by NSF); those are less diverse. It also mapped the papers that *cite* the HSD publications and, separately, mapped those that cite the comparison papers. Importantly, again, the HSD map shows strong engagement by the sciences. This comparison offers evidence that HSD research exerts influence beyond the social and behavioral sciences. The HSD-citing papers overlay map is similar in appearance to the HSD publications map shown here (Garner, Porter, & Newman, 2014).

Figure 1. Year 2004 HSD-derived Publications by Web of Science Categories



In addition to visualizing these data, the study applied several measures to help understand how interdisciplinary the research is. Integration scores, based on the diversity of references cited, indicate that the HSD-derived publications draw upon more diverse knowledge sources than do those of comparable programs. The study did not find notable differences in the frequency with which more integrative (more interdisciplinary) HSD or comparison group papers were cited. Diffusion scores, together with science overlay maps, show that uptake of the HSD publications extends into the natural, as well as social, sciences. Research networking analyses, together with a new composite mapping approach, point toward successful catalysis of a new research community.

The study team was particularly interested in the challenge in tracking the transfer of research knowledge. They experimented with a variety of measures, leading to the advent of two simple, but novel, metrics that we believe offer special potential for research assessment (Rafols, Porter, and Leydesdorff, 2010). One measure – “citation velocity” – calculates how quickly published articles are cited by other journal papers. They found that the 2004 and 2005 HSD-derived papers were generally cited with similar lag times as the comparison group papers.

A second measure – “citation distance” – gauges how far away the citing paper’s journal is from that in which the paper was published. This is based on the WoSCs of each, with distance scaled as in the science overlay mapping (such as Figure 1). The study explored five research questions using these two new measures. To investigate, they focused on 63 heavily cited HSD papers and 63 heavily cited comparison group papers. Those small numbers of papers receive a lot of citations – 4431 for the HSD papers and 5230 for the comparison group in about seven years following publication. Most importantly, they find that HSD publications are cited, on average, by more distant disciplines than are a set of comparison group publications. They also obtained evidence of different citation velocities of papers in different disciplines. Also, papers published in high impact (highly cited) journals tend to get cited faster and in more closely related journals (nearby disciplines).

References

Garner J, Porter AL, Borrego M, Tran E, Teutonico R.. (2013). Facilitating social and natural science cross-disciplinarity: Assessing the human and social dynamics program. *Research Evaluation*, 22(2):134-144.

Garner J, Porter AL, & Newman NC. (2014). Distance and velocity measures: Using citations to determine breadth and speed of research impact. *Scientometrics*, 100:687-703.

Rafols I, Porter AL, & Leydesdorff L. (2010).Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9):1871-1887.

F-2. Evaluating the Regional Experimental Support Centre (RESC) Program

Provided by Liudmila Mikhailova

CRDF Global (formerly U.S. Civilian Research & Development Foundation) is an independent nonprofit organization that promotes international scientific and technical collaboration through grants, technical resources, training and services. One of the many examples of the U.S. federal funded programs to support RTD in the countries of Eurasia was the U.S. Department of State Regional Experimental Support Center (RESC) funded under the Freedom Support Act of 1992 and administered by CRDF Global. The major goal of the RESC program was to increase the capacity of selected research institutions and universities in the former Soviet Union by providing up-to-date, state-of-the-art equipment for use in research and development activities.

A long-term outcome assessment study of the RESC Program was conducted by the CRDF Global Director of Evaluation in 2009-2010. The evaluation methodology utilized multiple methods of quantitative and qualitative data collection and data analysis. A mixed-methods approach with an in-depth case study analysis included site visits to 17 Centers of Excellence in seven countries of Eurasia, observations of equipment in use, document reviews, surveys and in-person interviews with RESC PIs, and face-to-face and focus group interviews with 153 senior and young scientists, engineers, institution administrators and students. The Centers' impacts were assessed against the program goals and objectives and linking inputs with outputs and outcomes.

The RESC Program resulted in a wide range of benefits and had an impact on four major levels: 1) helping to advance scientific and technical knowledge and developing skill sets for utilizing new research methods; 2) equipping the centers of excellence with state-of-the art equipment that contributed to institutional capacity building, the development of excellence in a number of scientific and technical fields, and the advancement of R&D activities in the regions and respective countries; 3) acting as a catalyst for the integration processes of bringing research to university systems and incorporating research activities into the curricula; and 4) helping scientists learn about Western concepts of S&T management and engaging them in commercial research that led to the creation of a more attractive climate for domestic economic activity and direct foreign investment.

The RESC Program contributed significantly to the development of a food, drug and alcohol testing facility in Yerevan, Armenia that facilitates the country's imports and exports; an environmental testing center in Baku, Azerbaijan that encourages responsible development of the country's oil resources, and a nanotechnology center in Nizhny Novgorod, Russia that supports the local automotive industry and contributes to the local potential that helped the city attract a massive direct foreign investment from Intel Corporation.

Other institutional and societal outcomes of RESC included new international S&T collaborations, the engagement of young scientists and former weapon scientists, publications in peer-reviewed journals, patenting, knowledge transfer to academia, and commercial contracts. RESC also contributed to U.S. foreign policy objectives by creating a friendlier environment for domestic economic activity and direct foreign investment.

Another set of benefits were summarized as return on engagement, which resulted in 630 M.S. and Ph.D. students engaged in the RESC research projects and 161 Ph.D. students who defended their doctoral thesis using RESC equipment for their research. More than 300 young scientists worked full-time or part-time in the RESC projects and more than 9,300 students were taught and trained on the RESC equipment through summer practicum, course work or special certification exams that students take using the equipment for analytical tests proficiency.

Reference

Mikhailova L. (2010).*CRDF RESC: Regional experimental support centers program funded by the U.S. Department of State under Freedom Support Act (FSA) - Final Evaluation Report.* CRDF Global. Retrieved from: https://www.crdfglobal.org/docs/default-source/final-evaluation-reports/resc-regional-experimental-support-centers-program-impact-evaluation-report-september-2010.pdf

F-3. Research Assessments: Informing Organizational Decision Making and Evaluating Health Research Impact in Alberta, Canada

Provided by Kathryn Graham and Deanne Langlois-Klassen

Alberta Innovates – Health Solutions (AIHS) is a Canadian-based, publicly-funded provincial health research and innovation funding organization mandated to improve the health and social and economic well-being of Albertans. To better inform organizational decision making and to demonstrate the value for money of its investments, AIHS implemented a standardized *Research to Impact Framework* that it developed through the integration of practice-based evidence and evidence-based practices (Graham et al., 2012).

An evaluation framework published by the Canadian Academy of Health Sciences (CAHS) (2009) was especially relevant for AIHS given its inclusion of wider impacts in assessing returns on investments. From AIHS' perspective, the potential construct validity of the CAHS framework was also advantageous as the framework had been designed to accommodate the Canadian health research context. The CAHS framework also provided an updated logic model that reflected the evidence-based healthcare system and environmental factors associated with health (including health products and services) and individual behaviors. An important addition to the logic model was the identification of system stakeholders who act as the primary agents through which advances in health research can lead to impacts (Jordan, 2011). Finally, the CAHS framework provided a toolbox to assist with the implementation of health research evaluation. The toolbox consisted of a comprehensive set of impact categories (advancing knowledge, capacity building, informing decision making, health impact, and social and economic impacts) across four pillars of research as well as a library of indicators and metrics at different levels of aggregation.

The applicability and feasibility of the CAHS framework within AIHS' local context was determined through a series of retrospective and prospective studies using a mixed methods approach. The retrospective studies aimed to determine if data in existing researcher and administrative program records could be meaningfully classified and analyzed according to the CAHS framework. Conversely, the utility of using the CAHS framework during the early implementation phase of new programs to inform data collection forms, analysis and reporting processes was assessed through prospective studies. During the same period, AIHS implemented practice-based approaches such as logic models and balanced scorecards (Kaplan, & Norton, 1996) to map the necessary processes and pathways between inputs and impacts. These approaches were cascaded across multiple levels (organizational and program) and reflected multiple perspectives (financial, internal, stakeholder, etc.).

AIHS funded investments in research realized a multitude of diverse impacts that ranged from science outcomes to wider impacts. For example, the assessment of AIHS' Independent Investigators program demonstrated results in the following impact categories:

- Capacity building: over 1,000 trainees were supported, infrastructure was built as evidenced by the establishment of 15 laboratories, and investigators leveraged approximately $210 million in additional funding;
- Advancing knowledge: scientific productivity was illustrated through more than 3,900 publications;
- Informing decision making: a number of products, policies, guidelines and services were achieved by investigators either individually or in collaboration with the health system, industry, government, etc.; and
- Wider health and socio economic impacts: 107 patents, five spin-off companies, improvements in health care efficiencies and effectiveness, therapeutics, and diagnostic techniques were generated.

Several important insights were also gained through the implementation of the *Research to Impact Framework:*

- Operationally, it is of paramount importance to align the evaluation framework to key strategy documents and to have it reflect the mission and vision of the organization.
    - For AIHS, this included aligning the *Research to Impact Framework* to the Government of Alberta's Health Research and Innovation Strategy (2010) and to AIHS' mission and vision (Alberta Innovates – Health Solutions, 2013).
- Evaluation frameworks should be used to guide the selection of appropriate indicators and metrics for performance evaluation and monitoring activities.
    - AIHS' experience suggests that this is best informed by mapping the *Research to Impact Framework* to the specific goals and objectives of each program and/or the organization depending on the level of assessment required.
- To improve operational efficiency when using existing information management systems that were not designed on the framework's data architecture, it will be necessary to implement data sets with a minimal number of key performance indicators (KPIs) for each program that key stakeholders consider meaningful.
    - The use of KPIs that key stakeholders consider to be meaningful also encourages the stakeholders' subsequent use of the evaluation results.

- Reporting on the key performance indicators for each program goal can be enriched through the inclusion of impact stories.

This was demonstrated in AIHS by using impact stories to highlight the key program results for stakeholders in a way that is more meaningful and insightful for them.

References

Alberta Innovates - Health Solutions (AIHS). (2013). *Strategic Framework.* AIHS. Retrieved from: http://www.aihealthsolutions.ca/docs/Strategy_FrameworkF.pdf

Canadian Academy of Health Sciences (CAHS), Panel on Return on Investment in Health Research. (2009). *Making an impact: A preferred framework and indicators to measure returns on investment in health research.* Ottawa (ON), Canada: Canadian Academy of Health Sciences (CAHS). Retrieved from: http://www.cahs-acss.ca/wp-content/uploads/2011/09/ROI_FullReport.pdf

Government of Alberta - Alberta Advanced Education and Technology and Alberta Health and Wellness. (2010). *Alberta's Health Research and Innovation Strategy.* Retrieved from: http://eae.alberta.ca/research/initiatives/ahris.aspx

Graham KER, Chorzempa HL, Valentine PA, Magnan J. (2012).Evaluating health research impact: Development and implementation of the Alberta Innovates – Health Solutions impact framework. *Research Evaluation*, 21(5):354-367. Retrieved from: http://rev.oxfordjournals.org/content/21/5/354

Jordan GB. (2011). Do existing logic models for science and technology development programs build a theory of change? American evaluation association conference: Anaheim, California.

Kaplan RS, & Norton DP. (1996). *The balanced scorecard: Translating strategy into action.* Boston, Mass.: Harvard Business School Press.