Current Practice in Evaluating Outcomes of Publicly Funded Research, Technology and Development Programs: An Overview (DRAFT)

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

Presented at

NSF 2014 Industry/University Cooperative Research Centers Evaluators Meeting

June 5, 2014

Gretchen Jordan, RTD TIG Co-Chair

Outline

- Purpose, scope, approach
- National guidance
- Evaluation purpose and practice
- Common/comparable logic models, indicators
- Selected evaluation design and methods
- Recommendations to improve practice and learning

Purpose, Scope

- The purpose of this paper is engage RTD evaluators and program managers in a dialogue about a common RTD evaluation language and practice.
- The end goal is a consensus and broader implementation of RTD evaluation that is more useful for learning what works in what context both within and across publicly funded RTD programs.
- This is needed because of the complexity and diversity in RTD programs.
- Our scope is broad but not comprehensive:
  - Publicly funded, Program level
  - All aspects: research, technology, development and deployment
  - Including innovation, defined as a new product, process or organizational practice that is entering the "market"
  - Looking at program contribution to outcomes before, during and after (life cycle)

Our Approach

- General government requirements and guidance on managing and measuring for results began in the U.S. in 1993, continue to evolve.
- There have been several expert studies in the past two decades that provide guidance.
- Practitioners have gained expertise, but this has not resulted in standardized practice or a textbook.
- Within the American Evaluation Association (AEA) the RTD topical interest group began to address this gap in 2012.

Congressional and OMB guidance

- **GPRA Modernization Act of 2010 (GPRAMA)** places a heightened emphasis on priority-setting, cross-organizational collaboration, and the use and analysis of goals and measurement to improve outcomes.
- **OMB Circular A-11**
  - Establishes a federal performance management framework and timeline to implement the requirements of GPRAMA;
  - Depicts & recognizes the role of evaluation;
  - but not an evaluation framework.
- OMB favors experimental or quasi-experimental methods, which "can produce rigorous evidence about program effectiveness. Qualitative evidence can complement that."
- OMB/OSTP annual memos endorse these for R&D programs

Guidance in the AEA "Evaluation Roadmap for Effective Government"

- Emphasizes that "[T]here 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."
- Describes 17 recommendations for federal agencies in the areas of
  - Scope and coverage,
  - Management,
  - Quality and Independence, and
  - Transparency

Relationship to AEA Evaluation Roadmap

While we agree with all of these recommendations, we have singled out two of them to expand upon for RTD programs:

1. Build into each new program and major policy initiative an appropriate evaluation framework to guide the program or initiative throughout its life.
2. 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.

Studies on How to Respond to Requirements – National Academies of Sciences

- Recommendation 1: ... research programs should be described in strategic and performance plans and evaluated in performance reports.
- Recommendation 2: Measurement use needs to recognize what can and cannot be measured. Misuse of measurement can lead to strongly negative results;
- Recommendation 3: ...use expert review to assess the quality, relevance, leadership of research. ...develop guidance [on] expert review processes.
- Recommendation 4: Both research and mission agencies should describe in their … goal of … adequate human resources…
- Recommendation 5: A formal process should be established to identify and coordinate areas of research that are supported by multiple agencies.
- Recommendation 6: The science and engineering community can and should play an important role in GPRA implementation.

Committee on Science, Engineering, and Public Policy. (1999). Evaluating Federal Research Programs: Research and the Government Performance and Results Act.

Studies on How to Respond to Requirements – Government Accountability Office (2012)

- Research program outcomes can be evaluated. This is most likely to be against a standard, including external standards or objectives set by the program.
- More complex demonstrations of outcomes such as comparison pre- and post-action or against a control group were not considered "best suited for" research programs (see Table 5.1).
- Methods suggested for assessing basic research quality and performance include bibliometric analysis of citations and patents and use of expert judgment.
- Effectiveness of an applied research program can be measured by whether it met its goal to improve the quality, precision, efficiency of tools/ processes.

US Government Accountability Office . *Designing Evaluations: 2012 Revision.* GAO-12-208G

Current Practice from RTD TIG review and experience in these areas follows.

- Purposes of RTD outcome evaluation
  - To monitor process and progress
  - To look forward
  - To look back
  - To look for early progress and lessons
- Plans for RTD monitoring and evaluation
- RTD evaluation frameworks guide planning

Why conduct RTD outcome evaluation? Questions decision makers ask

Program Stage	Question Simply Stated	Evaluation "Criteria"	Questions Posed by Government Leaders
Planning	What will the program do, when, why?	Evaluation plan exists	Planned end outcomes? Strategies for achieving these? What part within assessment timeframe?
	Are we doing the right things?	Relevance	How are planned outcomes aligned with organization's goal(s)? What is the program's "critical link" with these outcomes?
Early/Mid Implementa-tion	Are we doing it the right way?	Efficiency Quality Performance (early)	Progress toward this "critical link" with outcomes?
Mid/End of Implementa-tion	What are the outcomes and impact?	Effectiveness Performance Value For Money Sustainability	What are the intermediate outcomes, end outcomes, contribution to these outcomes?
Investment Decisions	What do we do next?	Evaluation findings used	How has evaluation been used to discontinue, redirect, or modify activities and likely results?

Why conduct RTD outcome evaluation? Monitoring is a foundation

- Also referred to as performance measurement
- In the longer-term, systematic collection, analysis and feedback of data about the program will provide important data for evaluation
- These functions can provide early progress metrics useful in assessing if the programs are meeting targets and on track to achieve results.
- For example, monitoring may identify the achievement of key technical goals

Why conduct RTD outcome evaluation? Look forward (Prospective evaluation)

- Question of great interest to RTD decision makers –where to invest next
- Methods in use
  - Foresight
  - Expert judgment
  - Stakeholder group techniques
  - Technology roadmaps
  - Stage gate analysis

Why conduct RTD outcome evaluation? Look back (Retrospective)

- Looking back many years is needed to see social/economic impacts of science
- Often do not have data on specifics of the mechanisms or pathways to end outcomes, so cannot inform program improvement
- Since each situation is unique and dynamic, the past cannot be used to inform future investment.
- Methods used for retrospective evaluations:
  - Economic studies (cost benefit, etc.)
  - Mixed method case studies

Why conduct RTD outcome evaluation? Look for early progress and lessons

- There is a growing demand on the part of managers of publicly funded programs for the inclusion of formative analysis in impact evaluation in order to have timely information to inform future program decisions and policies.
- RTD managers can collect and analyze data on outputs and early outcomes that are recognized by key stakeholders as likely precursors to intermediate and longer term outcomes

Evaluation plans

- Intended to organize the evaluation activities according to a logical framework for the program to be evaluated.
- Include program's goals and resources , responsibilities, approaches, metrics/indicators, data requirements, and data collection, analysis and reporting mechanisms
- Some programs need to retrofit a performance management framework to existing operations and to develop plans that systematically guide all aspects of evaluation.
- This has necessitated overlapping efforts in real time to plan, monitor, implement evaluations studies, and build supporting databases.

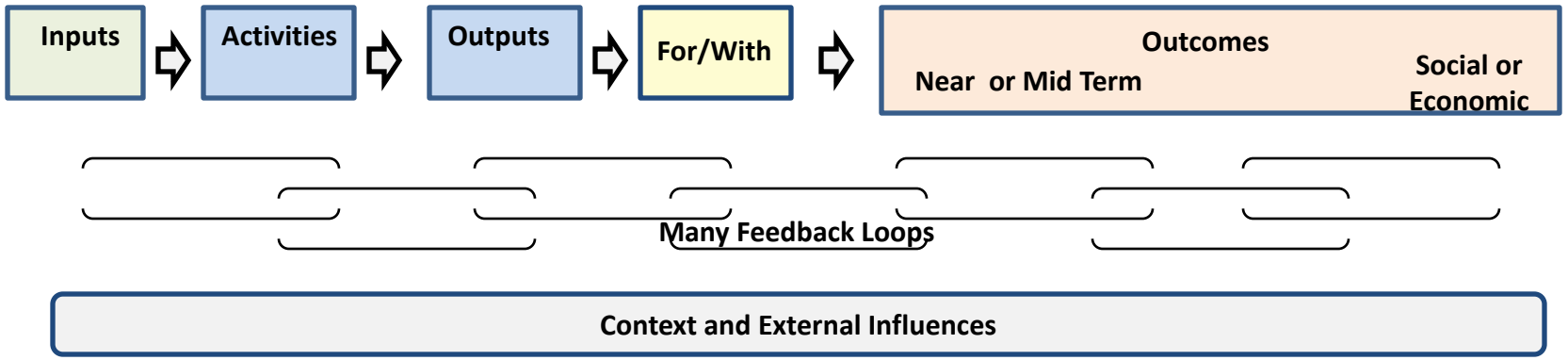
Evaluation plans – example
Advanced Technology Program (ATP)

- Immediately in 1991 established a comprehensive monitoring and evaluation plan
- Enabling factors were a supportive director; a budget for evaluation; a specific Congressional mandate to report outcomes, expert advice; and an internal staff charged with making it happen.
- At the outset, each funded project was analyzed to identify its key technical goals and metrics, and other dimensions of progress; monitored throughout the funding period and 5-years after completion.
- Pioneered benefit-cost evaluation of technology portfolios, compiled extensive databases to serve its evaluation needs, and conducted studies of impact and those aimed at improving an understanding of program dynamics.

Key to the plan is an evaluation framework

- Understood in the context of performance management and answering specific evaluation questions
- Components
  - logic model including context
  - indicators
  - evaluation design and methods
- A logic model is a plausible and sensible model of how the program will work under certain environmental conditions to solve identified problems.
- It should reflect a theory of change/program theory.

A Logical Framework



Indicators for						
Inputs	Activities	Outputs	Interactions	Near term Outcomes	Mid term Outcomes	Social or Economic Outcomes
Characteristics of likely differentiating factors; External influences on achievement						
Micro			Meso/Sector			Macro

The Need For Common Frameworks and Practice and Comparable Studies

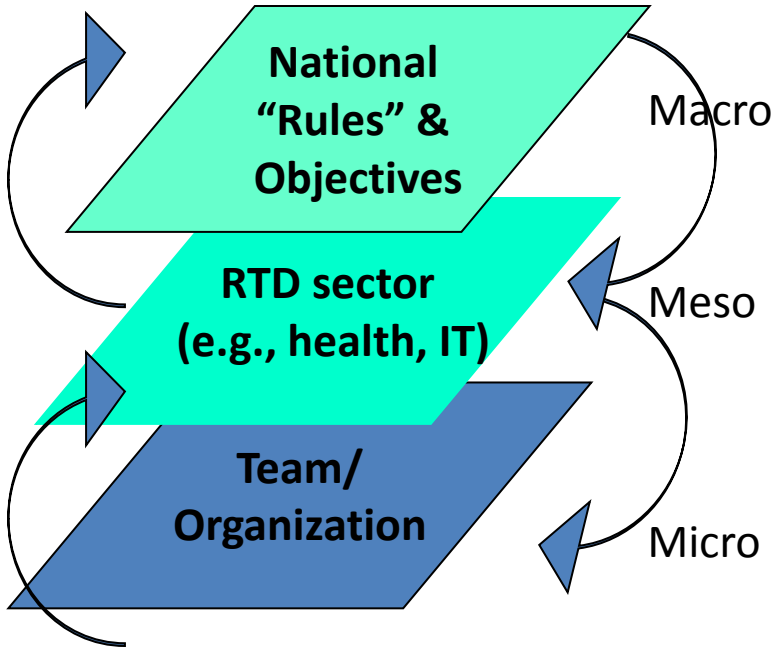
- Ideally there would be sufficient data and theory to enable policy makers to better target interventions, even to the point of comparing the cost, size and speed of pay off among alternatives.
- To build data and theory for the innovation process and system, there will need to be multiple studies and synthesis across those.
- Synthesis is easier if studies use similar terminology, good research design, and make clear the full context in which an intervention occurs.
- Only if we move toward common language, indicators, and notation of context can we aggregate, synthesize, compare.

A Proposed Generic Framework – With Context To Describe the Diversity in RTD Programs

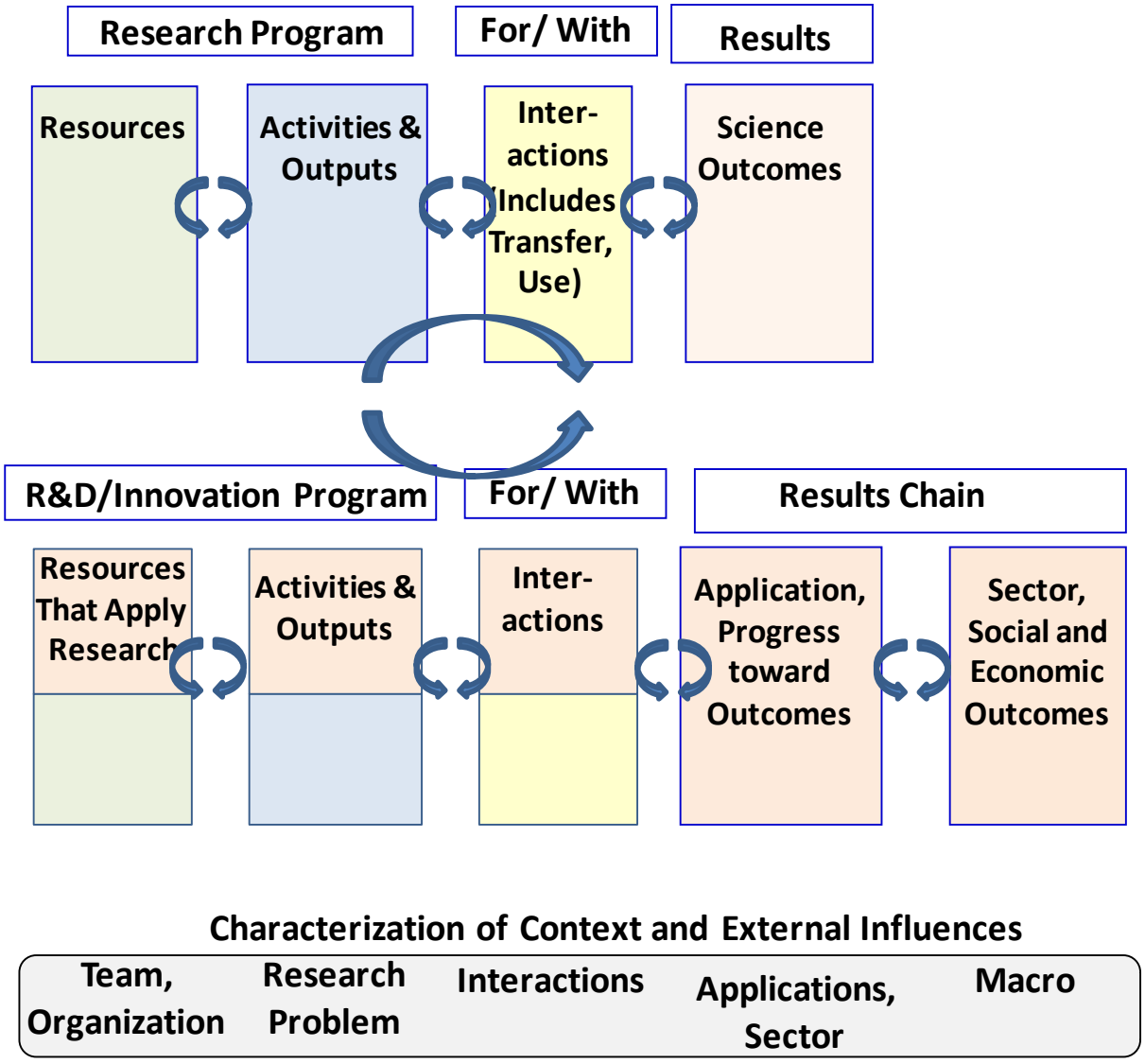
- Separates science outcomes from application and end outcomes.
  - to distinguish science questions from impact and policy questions
  - end outcomes of current work will typically not occur during the time these are under the direct influence of the program.
  - important to measure dissemination and take up or potential take up
- Technology and development activities may or may not draw on science outcomes. For any new innovation there is an "application and progress" stage before end outcomes.
  - Many intermediate outcomes occur and can be anticipated
  - Often detail on intermediate outcomes is left out of planning and evaluation.

Program Context Is the Innovation System. Context Must Characterize 3 Levels for Systems Evaluation.

- Program outcomes/ impacts differ by sectors because sectors differ in
  - Amount of investment for types of RTD
  - Rates of technical change
  - Ease of adoption
- Mission, policy and programmatic decisions are often sector specific
- Bottlenecks can be spotted more easily here
- Meso/sector level connects macro with micro



A Proposed Generic Logic Model and Context To Outline the Diversity in RTD Programs



We would need a framework of frameworks to describe major archetypes

- Outcomes and pathways to outcomes for various sectors (e.g., health, energy)
- Detail for pathways to outcomes for combinations of characteristics, e.g.,
  - Applied research in area where RTD networks already exist, technical, business and government infrastructure supports adoption of new product, which is an expressed need of consumers
  - The opposite of that
- Detail on commonly used mechanisms such as strategic clinical networks in health research, Engineering Research Centers, or collaborations such as Sematech

A Menu of Indicators For the Generic Logic Model

- Each element is described by the listing of indicators for types of outcomes different RTD programs are aimed at delivering.
- This results in a menu of many outcomes of RTD that can be measured, depending on
  - the type of RTD and its desired objectives,
  - target audiences for the application of the RTD, and
  - timing of the evaluation relative to the time passed since the activities took place.
- The list, while not comprehensive, reflects outcomes identified in numerous evaluation frameworks and literature reviews.

Logical Framework of Indicator Categories - 1

Resources for Research

- Funds for research, for research support
- Knowledge of researchers, technologists
- Team quality, organization
- Tools, techniques available
- Research environment

Activities/Outputs

ACTIVITIES

- Plan
- Investigate
- Prove concept
- Prototype

OUTPUTS

- Ideas/Knowledge advances (Excellence, Novelty, Publications, tech reports)
- New research tools, techniques
- People trained
- Preparation for transition to application

Interactions

CONNECTEDNESS

- With other scientists (pre-development)
- Across functions with developers, manufacturers, marketing
- Inter-sectoral
- With intermediaries
- With potential application users

LEVEL OF INTEGRATION (co-located, boundary spanners, etc.)

Logical Framework of Indicator Categories - 2

Outcomes		
Near Term	Mid Term	Long term

SCIENCE OUTCOMES

- Addition to knowledge base
- Citations, awards, leadership
- Affect organization, integration of knowledge
- Addition to science infrastructure (tools,
- Facilities, People)

R&D/Innovation/APPLICATION OUTCOMES

- Interactions with science, other R&D, Innovation entities
- Fund changes to technical Infrastructure
- New platforms for RTD, technical standards
- Industry funds further research, development
- New products, processes, organizational models
- Governments use in policy, programmatic decisions
- Public groups use in decisions, advocacy
- New skills, behaviors, attitudes, conditions

ADOPTION / INFRASTRUCTRE OUTCOMES

- Business/Organizations/Government and Public groups support or develop production, delivery systems, adoption by end user

VALUE OF THOSE APPLICATIONS:

Economic

- general
- business
- other sectors

Social

- health
- environment
- security
- other

Logical Framework of Indicator Categories - Context

Macro

- Availability of Capital
- Availability of Capabilities
- Ease of coordination

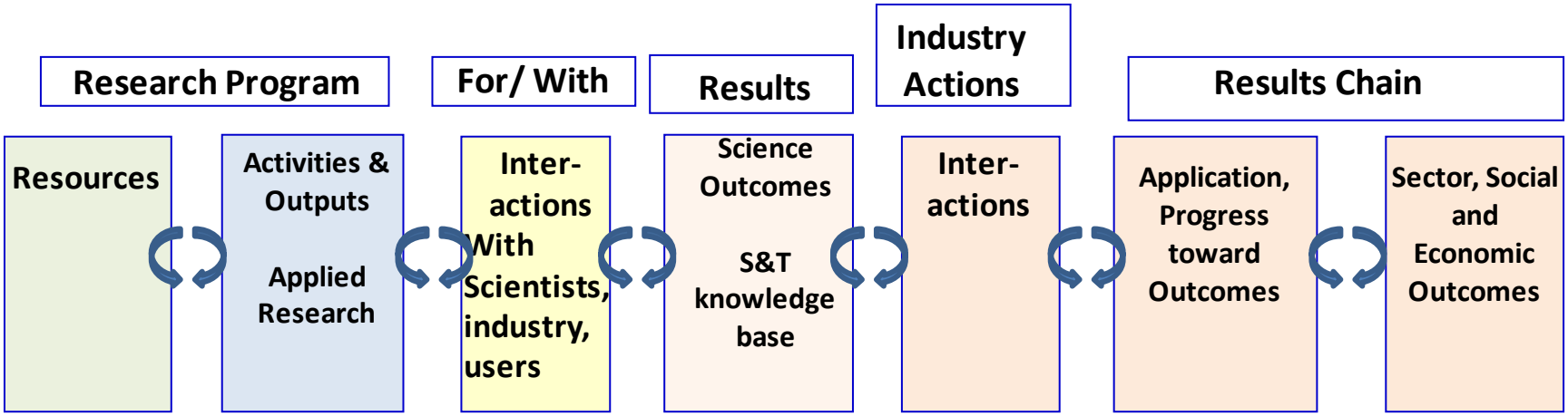
Meso/Sector

Characteristics of Interactions:	Nature of the application of research:
a. diversity	a. Breadth
b. continuity	b. Timing
c. mechanism used	c. Radicalness of change for application
	d. Sector speed for technical change
	e. Sector absorptive capacity, resources

Characteristics of the team (size, diversity, organizational/manage-ment, readiness, etc.)	Nature of the research problem a. research type b. radicalness c. scope

Micro

Example: U.S. DOE Wind R&D Linkages with Commercial Wind Generation



Indicators

$ spent by year Topics funded	Publications, technical reports. 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.	Scale up of DOE prototype turbines, etc. System Integration. New innovations	Cost per MW reduced. Improved system reliability, durability. Market growth. Spillovers.	Wind power capacity (MW). Fossil fuel, pollution avoided.

Design, Methods

Historical Tracing study that included: Publication and patent analysis; interviews with researchers, program managers, firms, and technology and market experts; network analysis

Ruegg and Thomas, Linkages from DOE's Wind Energy Program, 1999.

Example: NSF Human and Social Dynamics Program

Research Program	For/ With	Near Term Results

Resources	Activities & Outputs	Target Audience	Science Outcomes
HSD projects funded	Research published. Interdisci-plinarity	Non-HSD fields, collabor-ators	Knowledge diffused, Influences other fields

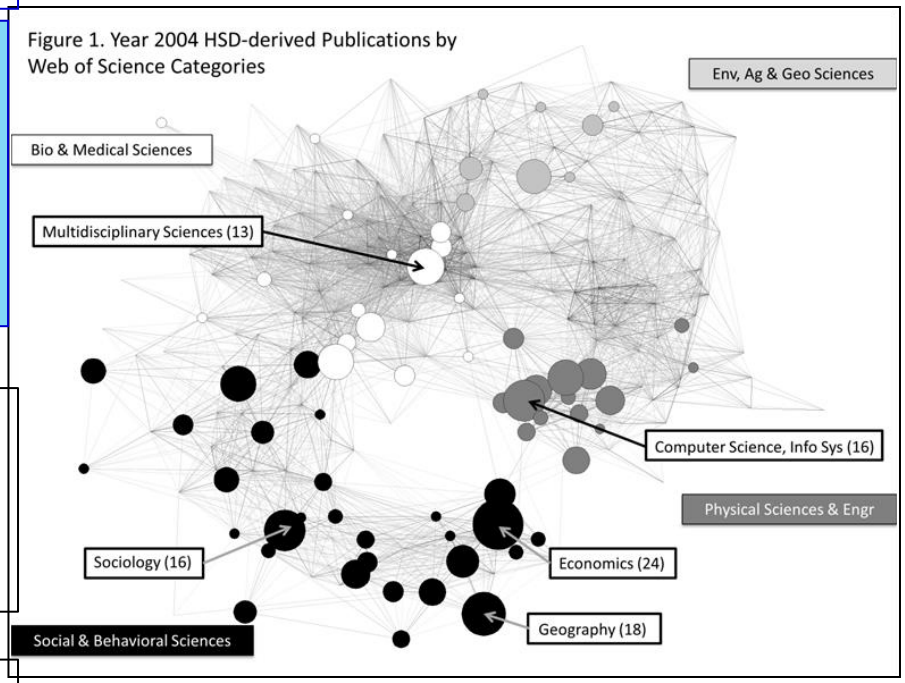
Indicators

Expendi-tures Topics funded	Publications -number -co-authorship	Integration scores Network characteris-tics	Publication maps Citations - #, distance -velocity

Design, Methods

Control group
Bibliometrics, Network analysis
Visualization of diffusion patterns

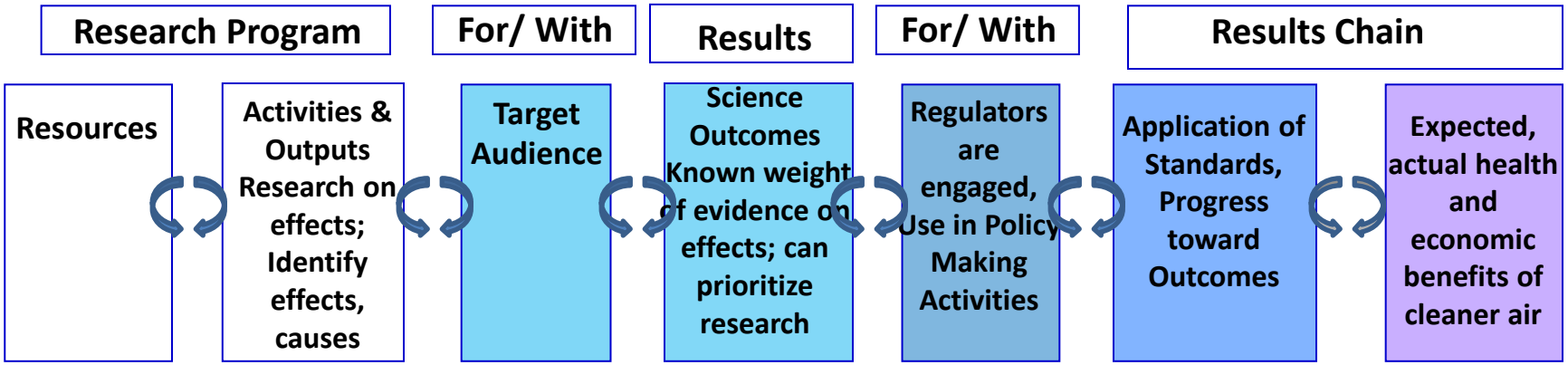
Contextual Influences: Social/Cultural, Technical, Business/Economic, and Political/Legal)



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

Program-Level Research to Inform EPA National Ambient Air Quality Standards

(for example, Particulate Matter NAAQS under the Clean Air Act)



Indicators

1. Resources 2. Laboratories 3. Scientists 4. Advice from independent experts (e.g., NRC)	1. Partnerships 2. Number & sequence of projects funded 3. PI workshops 4. Research publications 5. New methods, tools, models, etc.	1. Science community 2. Stakeholder organizations 3. Public databases – Health & Environmental Research On-line	1. Integrated Science Assessment 2. Science assessments by stakeholder organizations 3. Review by CASAC	1. Policy assessment 2. Regulatory impact assessment 3. Health-based standards 4. Judicial decisions	1. Emissions reduced 2. Air quality improved 3. Human exposure reduced 4. Dose to target organs reduced	1. Reduced risk to human health 2. Improved health in susceptible subpopulations 3. Benefits from reduced morbidity & mortality

Design, Methods

1. Publicly accessible database of new scientific knowledge – Health & Environmental Research On-line (HERO)
2. Integrated Science Assessment
3. Exposure assessment
4. Risk assessment
5. Independent expert review of science questions, publications, scientific progress, new scientific knowledge – Clean Air Scientific Advisory Committee (CASAC)

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

Evaluation Design and Methods Making Use of a Framework and Context

- Synthesis with standardized case studies
- Interim impact study with 4 tiers
- Attribution/contribution analysis

Evaluation Synthesis

- Takes existing studies, and based on the quality of the study and strength of evidence, uses findings as a database of what is known at that time.
- Helps answer policy questions that no single study could answer because a single study cannot be large enough in scope.
- After conflicts in findings can be resolved, looking across studies points to
  - features of an intervention that matter most, that are not visible in a single study.
  - which may be background variables, or research design, or stability across groups.
- Can show where there are gaps in knowledge that call for further targeted evaluation studies or new policy experiments.

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

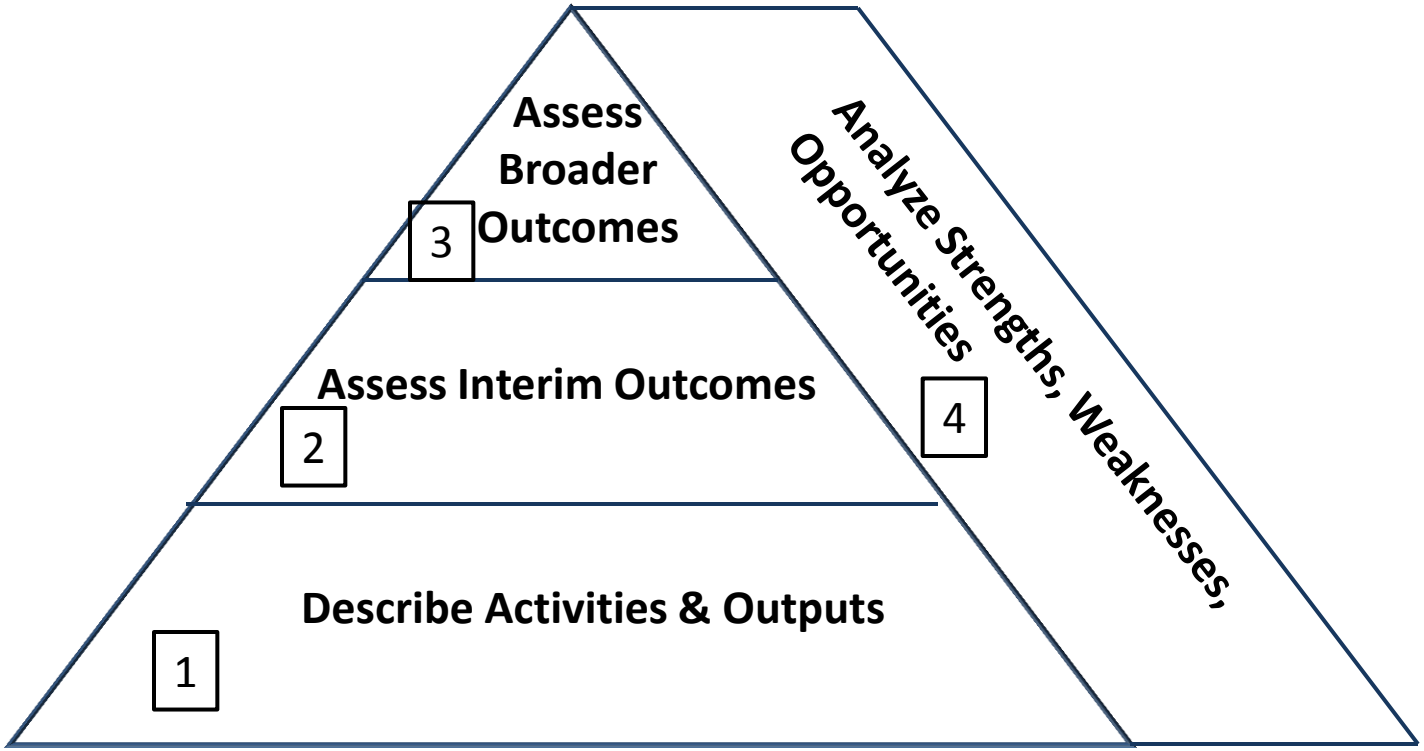
For Example, Standardized Case Studies

- Standardized case studies share a common framework and characterize key aspects of a program and its context, so study data can be aggregated and hypotheses tested with that data combined data
- Example that built on Research Value Mapping approach: French National for Institute for Agronomic Research (INRA)
  - Contextual and processual analysis to identify and analyze mechanisms that generate various dimensions of impact, in order to determine the specific contribution of INRA
  - Standardization overcomes some of the limitations of case studies while retaining the benefits of thick descriptions and stories.

For Example, Standardized Case Studies – 2 INRA Study

- Tools standard across the studies
  - Chronology: time frame, main events, turning points
  - Impact Pathway: productive intermediaries/interactions, contextual factors
  - Impact Vector: Radar chart of impact dimensions
- 30 cases in five research divisions; meta cases for three (e.g., genomic breeding) identified
  - Production of actionable knowledge
  - Lag between research and impact (possible extrapolation of intermediary results)
  - Structuration role, e.g., upstream research consortium or downstream intermediaries or regulation
  - Anticipatory role, e.g. exploring new options or insuring existing

A Scheme for Combining Interim Outcome Assessment with Formative Questions: Four Tiers of Analysis



Evaluate Tier 1, then 2, then 3, then 4.
Each tier builds on the tier(s) before it.

Source: Jordan, Ruegg, et.al. A Framework for Evaluating R&D Impacts and Supply Chain Dynamics Early in a Product Life Cycle, forthcoming 2014.

Attribution Using Frameworks and Context

- Three conditions required to establish cause and effect:
  - a logical explanation for why the investment can be expected to have led to the observed outcome.
  - a plausible time sequence of the investment occurred and the observed change relative to an appropriate baseline follows.
  - compelling evidence that the investment/actions are the partial or full cause of the change when competing explanations are taken into account.
- Reliable control groups in experimental or quasi-experimental study design is seldom possible for RTD. A sampling of participants and non-participants may not be truly random, groups not comparable.

Contribution analysis – an alternative

- A use of program theory
- Examines the role the program played (plays) in the larger system.
- Shares the credit
- Has the advantage of also informing next steps
- Being used more in Europe and Canada
- Specifically, Contribution Analysis examines *context, mechanisms,* and *outcomes* to see what worked under what circumstances (John Mayne, 2012)

DRAFT AEA RTD GROUP RECOMMENDATIONS

Recommendation 1.
Build an evaluation framework into each new program

- RTD program managers should undertake evaluation because evaluation is a valuable management tool.
- Outcome evaluation would be strengthened if (1) funds were set aside to do RTD program evaluation and build evaluation capacity, and (2) evaluation findings were used within and across programs.
- RTD program managers should plan evaluations using a logical framework and aim analysis at decision makers' questions.
- Questions call for both retrospective and prospective evaluation, and for evaluation of outputs and early outcomes linked to longer term outcomes.
- RTD evaluation needs to move toward a common language for outcomes and indicators, and characterization of context of these.
- Then evaluation studies can be aggregated, compared, and synthesized to build program theory.

Recommendation 2.
Promote the use and further development of appropriate methods

- Outcome evaluation methods can be useful for designing programs and policies, and improving programs, not just for assessing program effectiveness.
- There are multiple methods. Which are chosen depends upon questions being answered and context. Mixed methods are usually best.
- The nature of RTD programs means that the necessary conditions for experimental designs, including random controlled trials, usually do not exist.
- There are methods to be further developed and used, particularly in terms of the building program theory connected to the data collection and analysis.

Summary

- The objective of the AEA RTD interest group is to engage RTD evaluators and program managers in a dialogue about a common RTD evaluation language and practice.
- The end goal is a consensus and broader implementation of RTD evaluation that is more useful for learning what works in what context both within and across publicly funded RTD programs.
- To that end we have proposed a high level generic logical framework (model, indicators, design).
- The paper is still DRAFT. Everything here is a candidate for further discussion.
- We also welcome ideas on how to engage the community.

Acknowledgement

Thanks to volunteers from the RTD TIG

Kathryn Graham	Alberta Innovates - Health Solutions, Canada
Gretchen Jordan	360 Innovation LLC
Liudmila Mikhailova	CRDF Global
Dale Pahl	US EPA
Juan Rogers	Georgia Tech
Rosalie Ruegg	TIA Consulting Inc.
Josh Schnell	Thomson Reuters
Robin Wagner	US NIH
Madeleine Wallace	Windrose Vision LLC
Brian Zuckerman	Science and Technology Policy Institute
Liza Chan	Alberta Innovates - Health Solutions, Canada
Cheryl Oros	Liaison, Evaluation Policy Task Force
Deanne Langlois-Klassen	Alberta Innovates - Health Solutions, Canada

For more discussion or questions, contact

Gretchen Jordan
360 Innovation LLC
gretchen.jordan@comcast.net