Evaluation for What?

- The purpose of evaluation is to improve the agency's ability to further its objectives.
- Hence: start with objectives, not with metrics

Objectives

- DOE, like any public agency, has multiple objectives.
- It's important to be clear about these multiple objectives, because in many cases different metrics will be needed to assess progress towards different objectives

Multiple Objectives (illustrative)

- Energy/environmental: reduction of climate impacts of energy production and use
- Economic: reduced operating costs and/or commercial risks of energy technologies
- New Knowledge: new understanding of physical relationships or knowledge of properties of materials or technologies
- Capability: enhancement of the technical capabilities of the workforce
- Social: reduction of inequality across regions and social groups
- ….

Challenges in Evaluation

1. How to measure the complex, partially intangible progress towards objectives (“metrics”)
2. Difficulty of causally linking program actions to progress (“causality”)
3. Long, variable and/or unknown pathways to progress (“lags”)

Thinking about Metrics

- Some policy objectives are inherently intangible and/or hard to observe
- A “proxy” or “indicator’ metric is an observable quantity that we believe is correlated with the underlying unobservable objective.
- Sometimes, policy works by inducing changes that are not goals in and of themselves, but are desired as steps on a known pathway to the goal. Progress on such “intermediate outcomes” is useful evidence of potential eventual progress on the underlying goal.

Different kinds of metrics

Category	Examples	
	Objective	**Metric**
Direct measure	Reduced climate impact	Gms C02 per unit of output
	Reduced energy cost	$ per kwh
Proxy or indicator	New knowledge	patents; publications
		Expert assessments
	Public engagement in science	Semantic analysis of social media
Intermediate outcome	Reduced climate impact	Increase in Technology Readiness Levels ("TRL")
		Private investment in low-carbon technologies
	Public engagement in science	Development and use of educational materials

What makes a good metric?

- A high signal/noise ratio
- Errors that are unbiased and uncorrelated with other phenomena of interest
- Stability over time and across settings in the relationship between the proxy and the underlying concept
- Low susceptibility to manipulation

“Innovation” as a policy objective

- Innovation offers particular challenges as a policy objective, because even before we think about measurement, there is conceptual ambiguity regarding what we are looking for:
  - Advancement of knowledge
  - ‘Number of …’ (new products, new firms)
  - Change in context-specific performance, e.g. speed or capacity
  - Increase in consumers’ plus producers’ surplus
- These are not different measures of the same concept; they are different characterizations of the underlying objective

Beware the danger of “surrogation”

- Once created, a metric can take on a life of its own, become an end in itself rather than a noisy indicator of some underlying objective
- GDP was invented by economists as an indicator of the overall ability of the economy to meet people’s wants. Now politics treats it as a goal.
- Choi, J., Hecht, G. W., & Tayler, W. B. (2013). Strategy selection, surrogation, and strategic performance measurement systems. *Journal of Accounting Research*, *51*(1), 105-133.

Causality

- Evaluation is, at a fundamental level, always a comparison.
- We want to know the effect of a program or policy on the state or behavior of people or firms or systems affected by a program (the “treated group”)
- We can never get meaningful evaluation by studying only the treated group.
- To evaluate the effect of the program or policy, we must compare the state or behavior of the treated group to what that group would look like had it not been “treated.”

The ‘But for’

- The conceptual ideal for the ‘but for’ comparison is the exact same people/firms/system as the treated group, in a hypothetical world in which they never encountered the program being evaluated.
- We can’t study that, so we try to approximate it with some other group that is chosen in such a way that we can expect it to be similar to the treated group (the “control” group).

Finding a control group

- The ‘gold standard’ for identifying a control group is the Randomized Control Trial (“RCT)
- Unsuccessful applicants to a program are often a good control.
- Comparisons across subgroups of the treated can be informative
- Other ‘natural experiments’ often arise, in which accidental program features or events introduce elements of randomization to the treatment process

Lags between treatment and outcome

- There is often an unknown lag period between interventions to advance science and technology and the realization of desired outcomes.
- Evaluation must then focus on the generation of intermediate outcomes that can be expected to increase the likelihood of realizing ultimate outcomes.
- For this to meaningful, there should be an explicit model of the process that links the intermediate outcomes to the desired ultimate outcomes.

General observations

- The “fat tail” problem
- The “Hawthorne effect”
- Validation of metrics
- Semantic analysis, the internet and big data

The perfect should not be allowed to be the enemy of the good. Some knowledge of how policies work is better than no knowledge.

THANKS FOR LISTENTING!

QUESTIONS WELCOME

adam.jaffe@motu.org.nz