How Much Evaluation is Enough?

Marc Collins, Itron, Inc., Toronto, Ontario

ABSTRACT

Unlike a power plant that can have its output precisely metered, the energy efficiency resource requires "output" estimation that can vary in accuracy quite dramatically depending on the vagaries of what, how, where and when the impacts are measured. Naturally, governing bodies want accurate assessments, but not to spend more than is warranted to evaluate efficiency savings. This is especially true for national-scale endeavors. A systematic, stakeholder-based and transparent uncertainty assessment process that provides reasoned analysis inputs necessary to answer the question how much evaluation is sufficient for the intended audience(s) of the results is proposed. Although more time consuming and involved than a technocratic approach, it produces richer results and results that can be used as inputs for any actual evaluation planning that might ensue.

INTRODUCTION

The Environmental Protection Agency's (EPA) Clean Power Plan is expected to generate growth in the utilization of energy efficiency (EE) programs, particularly in jurisdictions where there is less history of these programs being deployed. Guidance is needed to determine the recommended and/or required evaluation, measurement and verification (EM&V) practices and standards that states should use to evaluate energy program savings so that savings estimates can be compared from one state to the next and so that an acceptable level of confidence can be ascribed to the air emissions calculations that fall-out from the savings estimates.

One of the primary goals of guidance is to make it easier for states to assess what level of EM&V effort (in effort and dollars) is needed to evaluate the impacts of their proposed EE and demand response (DR) programs (referred to collectively as EE programs for the rest of this paper), needed to report the results back to EPA, should they decide to use EE programs as a means to comply with the requirements of the Clean Power Plan. Although many analysts believe that EE programs and projects will be among the least expensive of the compliance options for some states to meet their carbon emission rate target (in lbs./MWh), there is some concern and uncertainty about the costs of evaluating these programs. Any EPA guidance should help states better estimate the total cost of operating and evaluating EE efficiency programs, so they can be properly compared to other major options such as fuel switching or developing more non-carbon sources of electricity generation.

The problem with estimating, "how much evaluation is enough," is that the appropriate amount of evaluation does not maintain a linear relationship with any readily available and convenient indicators like the amount of money spent on the EE programs or the number of customers served or the size of the savings target or the size of the initial estimate of the savings to be produced. Although summary-level statistics suggest the percentage of program expenditure dedicated to EM&V is usually in the 2 to 6% range, these calculations are both inconsistent in their methodologies and are likely to be describing quite disparate EE portfolios. A major purpose of

the EPA's desire to provide guidance in the first place is that the jurisdictions most in need are precisely the ones that lack the longer-term experience of building and operating (and regulating) an EE program portfolio. We shouldn't expect the current savings, costs and risks profile of a California or Massachusetts to translate directly to a state developing EE programs for the first time. There is a huge spectrum of conditions and experience, and of course, even California and Massachusetts have major differences between them.

The amount of time and resources needed for evaluation to be net useful, to add value to the overall enterprise of operating an EE program portfolio, varies according to the nature of the portfolio itself. This point is self-evident, but sometimes lost in the quest to define appropriate resources for evaluation. EM&V properly deployed is a management decision-making tool that can be applied to many aspects of an EE program portfolio, from investigating the smallest detail of one individual measure in one program in a portfolio all the way up to providing portfolio-wide performance measurement results needed as reliable inputs into an integrated resource plan. It can be focused on auditing past performance, or predicting future performance. Accuracy expectations vary all over the map, depending on many factors. Assessing long-term market transformation could be a primary goal, or measuring the immediate impact on utility customer service ratings. There are as many combinations and permutations of evaluation goals and needs as there are programs and portfolios.

Generally, evaluation spans from the earliest formative research to the latest after-the-fact summative audits. It is this potential blend of different types and purposes for energy program evaluation that makes planning for EM&V studies so challenging. Add-in that many of the categories are inter-linked—that is, a study in an earlier year may form part of the basis for a conclusion or calculation in a later year, or particular results in one category (e.g., impact) are needed as inputs in another (e.g., cost-effectiveness)—and something of a witches' brew ensues. Figure 1 is a simplified depiction of the six major types of studies, that could and often do connect to one another, and generally flow in time starting with potential and feasibility studies and ending with market effects evaluation. The cycle or parts of the cycle repeat.

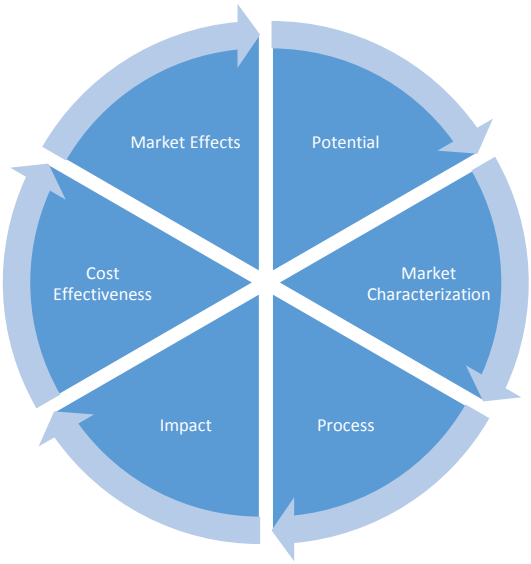

Figure 1. Major Evaluation Study Types

Striking a balance between energy program evaluation scope, rigor and cost is difficult

given this plethora of evaluation goals and priorities, which sometimes conflict or confound each other. In contrast, charitable foundations have a very simple equation to guide evaluation expenditure: does increased donor and potential donor confidence gained from a positive evaluation result in a net gain in donations after paying-out for the evaluation? If so, the evaluation was "enough." If a public lottery fraud squad recaptures more than their cost in misappropriated winnings, it was "enough."

In addition to trying to assess, in some objective manner, what is appropriate, the EPA is also faced with the existing wide range of EM&V experience and practices across all 50 states. Those jurisdictions with long histories of program and evaluation activity are more likely to be reluctant to alter course, particularly if there is not clear evidence that new guidance is a superior approach to what they are already doing.

SYSTEMATIC ASSESSMENT OF UNCERTAINTY

In an ideal world, a convenient and all-encompassing decision tree or mathematical model would be deployed to assess uncertainty. This oracle-like algorithm would produce a scored result advising what, when and where to evaluate. Unfortunately, it is not that simple. Given the multiple purposes EM&V serves and the wide array of situational factors that combine to affect what could be considered an appropriate quantity and quality of evaluation, the even-handed way to answer the question, "how much is enough," in a replicable manner across a series of jurisdictions, is via a systematic assessment process, built-up program-by-program, if necessary.

Systematic uncertainty assessment is best if it combines policy, science (facts) and various stakeholder perspectives (values). Any design of an assessment mechanism must recognize the social reality that values (of various parties involved) can affect interpretation of facts, which in turn can be used in various ways to meet, or subvert, policies. The EPA presumably wants its policies respected and its targets achieved, but not all actors involved in ultimately achieving the goals will agree to the same extent about the merits of the policies. Some may have their own reasons to try to achieve the same end-results, or not. This political reality exists in the energy sector due to its pointed environmental and economic impacts—and suggests that broader stakeholder input is likely a key to the accurate assessment of uncertainty.

Just as a planned power plant may not get successfully sited due to community opposition, planned EE programs may be executed more slowly or poorly compared to plan, or even not at all. EM&V is the feedback mechanism to report this progress, but by recognizing that some risks are predictable to some extent (i.e., not completely random), it can be designed expressly to help manage those risks rather than just act as an *ex post* testimony of the results. Effective stakeholder input into the uncertainty assessment process can successfully marry the social values elements—let's call those for simplicity's sake, behaviors—with the engineering facts, figures and projections. One does not need to look far in current society for examples where behavior, perhaps stemming from beliefs, overrides facts and science.

Uncertainty assessment should contribute to addressing two key decision-making criteria[1]:

1. Is it Time to Act? What, generally, needs to be evaluated and when?
2. Where to Focus Attention? Which programs, measures, markets should be the highest priority for evaluation (uncertainty reduction and knowledge gain)?

What follows is a proposed multi-stage, transparent and systematic uncertainty assessment

[1] Adapted from: Fischhoff, B., and A. Davis. 2014. "Communicating scientific uncertainty." *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111:13664

process designed to elucidate the comparative value of pursuing evaluation research to a greater or lesser extent and among competing ends.

1. Is it Time to Act?

The question, "is it time to act," can be informed by a fairly simple uncertainty inventory that plots knowledge of consequences versus knowledge of probability. Uncertainty is the product of those two and the product to which evaluation can be applied for net benefit.

Knowledge of consequences can range from poor to good, but may have inconveniently variegated subcomponents. For example, there may be a good degree of certainty that if a particular efficient technology is installed and utilized, cost-effective savings to the end-user will occur. However, there may be much less certainty about which end-users would adopt the new technology at various levels of an incentive offering. What is the likelihood that overall program and societal cost-effectiveness will be achieved if most end-users demanded higher incentive levels (to participate)? A cost-test model exists that could predict quite precisely the benefit-cost impacts of various proportions of customers choosing or requiring various levels of incentive payments. The likely inability to accurately predict which proportions will actually occur is the source of poor knowledge of probability. The generic matrix is illustrated below in Figure 2.

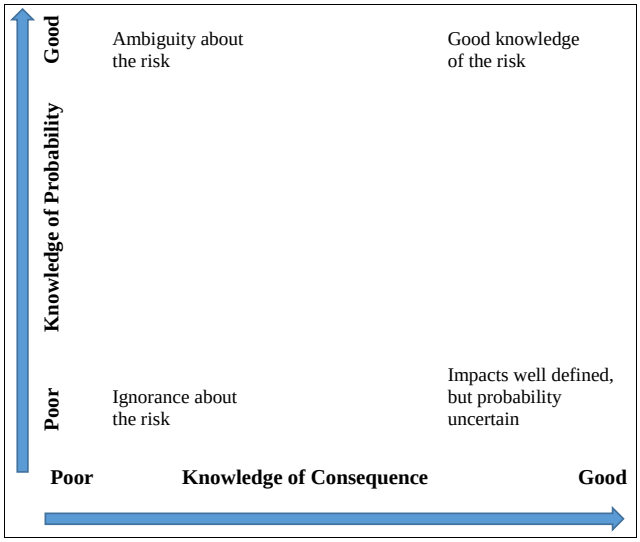

Figure 2. Basic Uncertainty Inventory Matrix[2]

As a first order of assessment and prioritization, measures, programs, or even portfolios could be inventoried using this basic uncertainty matrix. Even when programs already exist, relatively simple analysis of uncertainty contained them is not too difficult. Program managers usually know about data tracking, savings calculations and reporting issues with their own programs, even if relatively minor. Evaluators can quickly scan for signs of trouble. These or any other factors that could cause final results to vary from a plan can be plotted fairly quickly.

To help organize the delving into these sorts of issues, there are several types of uncertainty that should be distinguished in the basic screening process:

[2] Matrix and uncertainty types adapted from Willows, R.I. and Connell, R.K. (Eds.). (2003). *Climate adaptation: Risk, uncertainty and decision-making.* UKCIP Technical Report. UKCIP, Oxford.

Natural and social variability. Weather is the most obvious natural variable that can impact a significant sub-set of measures in energy programs. Although savings are often weather normalized for accounting purposes, use of savings for real world purposes, such as demand response-derived load reductions in an electricity system distribution-constrained zone, requires a taking into account the actual weather and living with the consequences. Rapidly evolving markets are another source of variability that, if not paid sufficient attention, could undermine evaluation efforts. Market adoption patterns sometimes move faster than programs and certainly faster than *ex post* evaluation delivered months or even years later. Any form of context that mimics "shifting sands" could fit into the definition of variability that contributes significantly to uncertainty.

Data uncertainty. Even when an evaluation process has lots of data, or the seeming ability to get lots of data, uncertainty is rarely eliminated. Various types of measurement error, some systematic that result in bias and some based on incomplete data or insufficient resolution (in time span or frequency) can confound many conclusions. The ability to extrapolate conclusively is often less certain than desired. There may exist five years of consumption data for all households in a region—perhaps more than enough for a study's purpose—but if its granularity is monthly versus daily or hourly or 5-minute interval will dramatically affect the certainty of analysis or could even eliminate the ability to answer certain research questions.

Knowledge uncertainty. Management decision-making requires knowledge, but evaluation studies are often designed to unearth knowledge, or may do so by happenstance. Therefore, there is a "chicken versus the egg" conundrum when trying to assess risk related to how designed processes actually work in the field or how dependencies among components of a program work. There could be risk related to the functioning of markets themselves, interactions that could affect behavior, or prices, or, of the individual performance of some elements of a program. When assessing uncertainty, appropriate degrees of ignorance should be acknowledged, particularly when trying to predict future states. Of course responsible parties may perceive acknowledging ignorance as a serious threat due to, among many reasons, because program expenditure approvals are often premised on optimistic depictions of certainty.

Model uncertainty. Assessing engineering uncertainty forms the core of many energy program impact evaluations and also provides one of the best examples of institutionalized model uncertainty. The International Performance Measurement and Verification Protocol (IPMVP) is one of the few almost universally-accepted underpinnings for impact evaluation in the industry. The reference to IPMVP for engineering review of projects is often interpreted as if it represents a minimum quality threshold or a "good" standard. In fact, it is itself a model/protocol that recognizes the inherent model uncertainty in engineering review and attempts to redress the potential vagaries of that uncertainty by modelling decision-making around optimizing uncertainty reduction. In other words, the IPMVP is a procedural model, honed over many long years of effort (but that still contains some uncertainty), that assists in prioritizing options related to measuring project-level engineering savings that inherently contain enough uncertainty that wise people thought the need to devise an IPMVP in the first place! In addition, other forms of model uncertainty pervade the whole range of evaluation types, from cost-effectiveness models to market models to human behavior models and so on.

Of all of the types of uncertainty that need to be taken into account for a basic inventory, model uncertainty may be the most important to delve into. Any "model" has inherent structural characteristics that often become institutionalized—and not always for the better. It is possible over time to "forget" key building blocks of the construct and thereby risk misunderstanding results or misusing data and analysis in ways never intended by the model builders.

The California Standard Practice Manual[3] cost-effectiveness tests illustrate uncertainty issues related to models quite well. These tests are, like the IPVMP, almost universally accepted as valid and are used (at least one of the five) across the EE industry and by almost every jurisdiction in North America as the benchmark for benefit-cost reporting. This universality has a tremendous advantage in that a huge swath of "model selection uncertainty" is removed from the equation because everyone chose to use the same model.

However, model input values are a source of model uncertainty. In this cost test example, some input information is likely to be accurate (e.g., total incentives paid to customers), while other inputs are less or much less so (e.g., total program expenditure including what specifically is included in that definition, or, incremental cost to customers for their efficiency investment/project).

Model parameters include such items as avoided cost assumptions, which should vary from region to region based on the nature of the local energy supply mix and which resources are utilized on the margin. However, even if there is high certainty related to the supply mix and marginal unit composition (which is unlikely), there is often uncertainty related to the avoided cost estimates themselves, even though they form such an important basis for the test results. In fact, it is highly likely that any avoided cost assumption matrix is itself the product of a complex model with its own array of model uncertainty factors. Other example model parameters inside the cost tests include the discount rate used for normalization of the costs and benefits into today's money (which rate should be used for which test), or, the valuation of capacity benefits that can change dramatically over time, so picking a static value (for purposes of extrapolation) is difficult.

Model outputs do not contain the same degree of uncertainty inherent in model inputs and parameters, but because past outputs/results are often used as primary data sources for the current round of uncertainty assessment, they still need to be handled with care. To avoid uncertainty inadvertently propagating through to the next cycle of analysis, some reality-checking should occur. For example, a past series of cost-test results, all calculated in a consistent manner with consistent assumptions, could be subjected to sensitivity analysis. By testing alternate model inputs and parameters, perhaps garnered from stakeholder perspectives, a sensitivity analysis could reveal broader aspects of uncertainty that may be otherwise hidden or ignored in the existing cost-test regime results that appeared to be relatively consistent.

2. Where to Focus Attention?

The basic inventory of uncertainty that falls-out of the review described in the previous section (a two-dimensional matrix with four uncertainty types) encapsulates the amount and degree of uncertainty necessary to decide, generally, whether and where evaluation action is warranted. This section describes the next level of assessment needed to determine which evaluation activities should take precedence—under the assumption that doing "everything" is usually not a viable option. A reasonably objective rating and prioritization process is needed to produce a set of evaluation activities that arguably provide more value in terms of reducing uncertainty than they cost to perform. As mentioned earlier, this is not possible to calculate using a simple equation, but can emerge from the product of a facilitated multi-perspective review, using a common framework of assessment. It is probably too involved and complex for most jurisdictions to undertake as a

[3] Danforth, C., Weiss, D., Woychik, E., and California Public Utilities Commission. 1983. *California Standard Practice for Cost-Benefit Analysis of Conservation and Load Management Programs: Joint Staff Report*. California Public Utilities Commission. Updated documents: http://www.cpuc.ca.gov/PUC/energy/Energy+Efficiency/EM+and+V/

first step, so it is proposed here to flow after the basic inventory step.

It should also be noted that expending effort on a detailed, multi-party uncertainty assessment exercise is not only useful for answering the initial scoping questions related to, "how much evaluation is enough?" The majority of the thought process dedicated towards identifying risks, pinpointing and cataloguing the various types of uncertainty, assessing their weight, considering trade-offs, debating importance, and so on, are all precisely the desired inputs needed for a good quality evaluation planning process. In other words, any actual evaluation planning later would be shortened and enhanced by utilizing the product of this preliminary process. It essentially works to shift detailed uncertainty analysis sooner than it usually occurs and shifts it from the purview of evaluation contractors and their direct clients onto a broader stage.

This more detailed uncertainty assessment process, illustrated below in Figure 3, can be thought of in terms of a series of concentric activities that culminate in a collection of prioritized and comprehensively understood research options. The outermost circle represents the basic uncertainty inventory described above in Section 1. It provides the starting point and foundation for the more thorough framing of risks and uncertainty here in Section 2. The next step involves developing uncertainty indicators that allow for a shared and consistent assessment of uncertainty and risks. That is followed by a cross-check against the current knowledge base, designed to further inform the indicators as there may be data points or even trends already formed related to some indicators. Lastly, the product of the entire exercise is a collection of uncertainty and risk assessments relevant to the jurisdictional boundaries and scope of the quest. This could be a statewide assessment for EPA compliance purposes, or a narrower, utility-specific or even a program-specific focus.

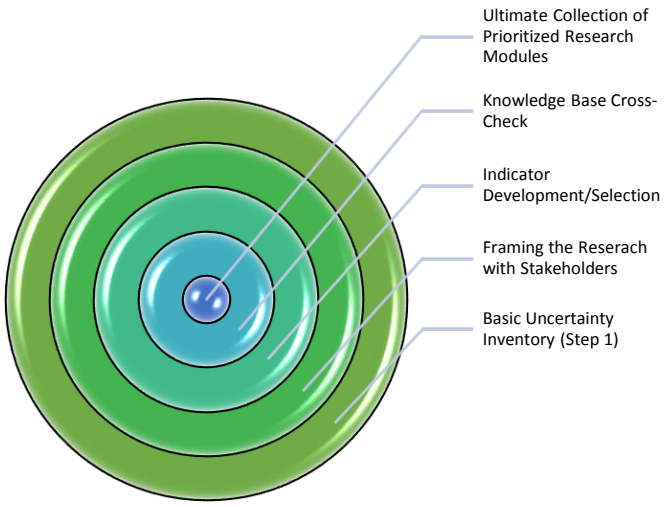

Figure 3. Specifying and Further Defining Uncertainty[4]

A. Framing the Research with Stakeholders

Fundamentally, this step is a further detailed review of the inventory of uncertainty and uncertainty factors that emerged from the first step. Here, explicit decisions about which contextual factors to include or exclude are required. Given this core function, it should be apparent why

[4] Process adapted from Janssen, P., A. Petersen, J. van der Sluijs, and J. Riseby. 2005. "A guidance for assessing and communicating uncertainties." *Water, Science & Technology* 52 (No. 6): 125-131.

involving the full range of stakeholders, with their various perspectives, would be useful at this stage.

Are there multiple needs for evaluation results? For example, if an integrated resource planning (IRP) process is the driving force behind a portfolio of programs and the IRP could even affect program selection within the portfolio, IRP planners should be stakeholders at the table identifying their particular definition(s) of uncertainty. Greenhouse gas (GHG) reduction/carbon-related concerns may be paramount—as in our EPA scenario. If so, the level of tolerable uncertainty may still be low, but likely not as low or precisely the same as for IRP purposes. A utility rate-setting process may depend on evaluation results. These types of accounting exercises value accuracy, but probably to a lesser extent than the first two examples, because they operate on some form of revenue reconciliation process—so an overage or underage one year simply gets adjusted in a following year. Only if results gets significantly out of whack in any particular year does it present a serious accounting reputation problem. Political savings targets (for a state) might be the driving force for DSM programs. The world of politics can make strange bedfellows in that it could actually be politically advantageous to have less certain results if the overall political strategy itself happens to be precarious (for whatever variety of reasons). Basic ratepayer/ stakeholder reassurance of value-for-money could be a major driver. What investment return on the programs is the utility getting? What are the community/regional benefits?

Systematically reviewing the likely differing perspectives related to uncertainty and risk is not overly complicated, but is often subverted by parties' attempts to interpret or prioritize others' values related to the shared facts at hand. Instead, it should be determined if there are different stakeholders associated with each/any relevant driving force (some examples listed above) and, if so, make sure they are involved in the assessment process. Are there other needs (e.g., not represented at the table) that should be taken into account? If so, do they require another stakeholder representative, or is the need definition sufficiently unambiguous and unanimous?

How much interconnection of various priority needs exists? For example, a political savings target may include components of energy, capacity and GHG reductions simultaneously. One element may trump others, but this could be open to some interpretation. Needs should not be approached in isolation unless it is unanimously agreed they are not inter-related.

Document what specific "answers" each end user (or stakeholder participant) needs from the evaluation. Determine in which research questions these needs were described and what framing was chosen. Since evaluation may not have occurred yet, there may be no formal research questions available. If not, the most opportune time to craft them is at this juncture. When attempting to develop consensus around relevant research questions, which will not always be achieved, at least attempt to ensure the logical framing for the questions is unanimously accepted.

There may be situations where fundamental misunderstandings are revealed, such as a relationship where electricity savings trigger higher natural gas use. Some conundrums or disagreements could be policy based. Justifications related to any differences in perspective should be documented, including discussion of potential consequences with regard to the merit and scope of pursuing various evaluation options.

Transparent documentation of any potentially relevant aspects of stakeholder needs that deliberately will not be addressed in the research questions is also critical to avoid future disappointment or conflicts around questions that are not being answered. It is helpful to include any reasons why they are not dealt with and the known consequences of not dealing with them. To avoid potentially confounding or conflict-ridden discussions, it is helpful to pose this simple question: would evaluation results and conclusions have been any different if the missing aspects

had been included? Although this is phrased in the past tense, the binary yes/no answer to this question is usually knowable in advance. In some cases, it may need to be addressed in an actual, yet-to-be-developed, evaluation plan, or, may not be fully known until an evaluation is performed.

A critical element of the framing exercise is to examine what is the role of evaluation in the ongoing program design process. Are results likely to be used as *ad hoc* advice or to help evaluate an existing program or portfolio (or savings target) in the context of a continuous improvement strategy? Is the evaluation research meant to elucidate future policy options or to raise awareness about impending problems, or focus on the current and past only? It must be asked if the research is intended to identify or expound upon possible solutions to problems or just report on the fundamental performance indicators. Is one of the purposes of the results to provide counter-evidence for a hearing or process? For any of these purposes, a transparent indication of controversies and any known plurality in views is helpful to fully assess uncertainty that could be redressed by evaluation activities.

Although probably already inherent in every party's views, document what, if anything, has been said about the issue or indicator in the past? Dissecting this history and various (shared) interpretations of it, rather than allowing it to remain unspoken, and potentially used for divisive purposes, is helpful. The use of previous knowledge of (conclusive and inconclusive) evaluation results is a major element useful to plan an evaluation scope. For instance, a utility could have worked informally with its largest customers on power factor and energy efficiency matters for many years in advance of the introduction of formal DSM programs. The baseline is therefore not one of "no attention paid to efficiency" or necessarily low efficiency performance. Evidence of what was actually happening prior to the program introduction might be an eye-opener that causes a change to the scope and nature of future evaluation efforts—possibly because some helpful data already exists, or, perhaps due to a more variegated existing landscape than anticipated. It is useful to document what added value and meaning can be taken from the present understanding of the context and how it is believed to affect what needs to be examined in future evaluations.

The implicit assumption in this entire section is that stakeholder participation aids in the uncertainty assessment process by including more and broader perspectives, particularly related to values and the value-laden interpretation of facts. Involvement in this framing process is also excellent preparation for any eventual stakeholder advisory role in the evaluation work itself.

B. Indicator Development and Selection

Indicators play a very important role in data- and/or model-based studies to highlight important aspects of the needs that require evaluation to address. Developing and then understanding performance indicators for DSM measures, programs or portfolios serves two purposes: it allows for tangible definitions, or boundaries, of uncertainty and it potentially confirms some common agreement around DSM performance indicators themselves. The latter assists with stakeholder consensus building.

As an example, a popular type of residential consumer behavioral program could be proposed. The region's IRP planners may be interested in how behavior change will affect afternoon peaking in the summer months (specific kW reductions during specific hours). The program vendor may be interested in demonstrating that the savings from this program are less costly than other possible program options ($/kWh) to the utility. The utility's customer service department may be interested in whether incremental calls to the call centre are complaints versus questions as its indicator of choice. Participating customers may be looking for straight bill savings. The DSM portfolio manager may be most interested in understanding whether the

behavior promoted in the program is sustained and whether it has the potential to become a new baseline in the near-term future (savings replicability and persistence).

No matter what the examples, it is necessary to map the main indicators of interest, across the various constituencies, used to measure performance, so as to then determine how these relate to each other and to uncertainty. The process can involve first plotting conceivable alternatives (as in the example above) and a discussion of their implications and shortcomings. The group substantiates an ultimate choice of indicators, including the identified shortcomings and any potential controversies (e.g., the program design may suffer from various selection biases, the Hawthorne effect or other problems[5]). Stakeholders involved in the indicators selection process can help identify how much support there is among programs and evaluation professionals in the sector and within society (including decision-makers and politicians) for the use of the selected indicators. The collective group can and should also identify what could lead to a lack of support for any of the particular indicators. If the group is particularly well-functioning, they could plan to deal with a scenario of low initial support or a future withdrawal of support.

There is benefit to various parties' preferred indicators being accompanied by a common understanding of any inherent shortcomings and potential controversies. For example, if support for DSM programs was suddenly in jeopardy or being withdrawn, the group should be able to identify performance indicators and evaluation metrics that might address decision-maker or politician concerns—such as the levelized cost of energy (LCOE). Relevant metrics that facilitate comparisons with other resources could avoid getting too far into arcane details or avoid debating metrics that may have negative policy overtones for one group or another. Stakeholders should also be aware of the uncertainty associated with each of the indicators, including those being used to justify opposing positions.

C. Knowledge Base Cross-Check

The penultimate step in the process towards building the collection of prioritized research modules is to assess the adequacy of the knowledge base that is available for any proposed evaluation work or uncertainty assessment. By examining and cataloguing what is already known, uncertainty can be viewed in its most relevant context. All of the earlier work done to frame issues and formulate indicators gets applied here, but at a more granular level. This cross-check acts like a filtering process based on the group's best assessment of the current knowledge base.

The cross-check filtering criteria definitely include quality. Each indicator is likely to have quality criteria relevant to answering the research questions associated with that particular indicator. For example, a past program's project savings realization rate (the difference between expected/reported savings and evaluated savings) may have a certain accuracy and reliability (90% ± 10%) associated with it. But the plausibility of that result may be called into question due to whether or not the savings were weighted by strata or some other methodological detail. They may be called into question due to a reputational concern—about the utility or the evaluation contractor. The estimates may or may not have been based on a strict interpretation of the IPVMP, suggesting professional or scientific support, or lack thereof, for the results. These results may have been the third in an annual series, possibly suggesting robustness if the results were fairly consistent.

In addition to possessing quality criteria, existing evaluation results may be the subject of policy-relevant controversy. Controversies within the evaluation and regulatory arenas, as well as

[5] Fischhoff, B. and A. Davis. 2014. "Communicating scientific uncertainty." *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111:13667

from individuals or organizations, could be used to play-up some uncertainty issues for policy purposes. The development of the Department of Energy's Uniform Methods Protocol[6] (UMP) is arguably an attempt to reduce controversy by employing standard evaluation and analysis methods, where practicable, that have been screened and approved by a wide peer group. Controversies can act as an impediment to obtaining satisfactory answers to research question by limiting the availability and/or quality of expertise, empirical data, theoretical underpinning and model development. For example, the UMP process provided a "safe" forum for those in the industry to collectively discuss how to improve and standardize some evaluation processes. Outside such a forum, raising questions, even about the identical issues tackled in UMP, could be frowned upon due to professional etiquette because all professions tend to lapse into some degree of institutionalization and acceptance of the *status quo*. The knowledge base cross-check is designed to mimic the UMP philosophy of openness and transparency for the greater goal of overall system improvement. As in the other sections, documentation of how any deficiencies and limitations are best addressed, either during proposed evaluation studies or after their completion, is best accomplished with peer review from stakeholders providing determination whether controversies have been adequately dealt with.

D. Ultimate Collection of Prioritized Research Modules

The final crystallization of all of the input from all of the prior assessment layers is to document the implications of any deficiencies and/or limitations for the scope, quality, and acceptance of the findings of any proposed evaluation work (i.e., the attempt to fill gaps in knowledge and reduce uncertainty). Before attempting to determine an overall scope and range of detailed, time-consuming and often resource-intensive evaluation work, it is wise to create a series (collection) of building blocks (modules) that can be prioritized. The modules represent areas of risk and uncertainty that could benefit from some or another extent of evaluation. These need to be graded and sorted to answer the ultimate question, "how much evaluation is enough?" The sum of these modules, each filtered and documented in a systematic and transparent process, then represents the total reasonable scope for evaluation that can claim to usefully contribute to management decision-making and risk management (through uncertainty reduction) in a manner that can be understood, warts and all, by the full range of stakeholders.

While each module may vary considerably in its own scope and particular nature, so will the breadth of the sum of the collection of the modules. There is no need for consistency or conformity. Each jurisdiction and scenario has a unique uncertainty context and needs, with the core question of what evaluation would be most appropriate to mitigate it being the consistent element. The proposed systematic approach is designed to ensure that the right questions are asked by the right people before arriving at that conclusion. Analysis should occur to an extent that proximal budgeting and high-level resource planning is possible. This information is, not coincidentally, needed later for actual evaluation planning and budgeting.

There are common questions that should be answered and documented for each potential module (uncertainty area or issue). Much of the response required at this stage may be harvested from prior steps, but should be consolidated into one consistent overview to aid in scoring and prioritizing. A real-world example of a template follows at the end of the paper in Figure 4. To aid in understanding what is involved in filling-out that "form," what follows is a recitation of the necessary Questions or issue Statements with a corresponding Response framework.

Q: Are decision stakes high or is there a lack of consensus about policy goals or the

[6] http://energy.gov/eere/about-us/ump-home

significant role of value-laden uncertainties? R: Yes, could even affect selection of indicators; Yes, could affect interpretation of conclusions; Yes, could have implications for the socio-political arena; Could trigger stakeholder-level debate; Could generate professional-level or methodological debate; No, not controversial enough to affect evaluation plans.

Q: What role are uncertainties generally expected to play? R: No significant role; A significant role; A critical issue.

Q: Where are the most important uncertainties expected to be found and what is known about their nature? R: Particular measures; Program delivery elements; Portfolio scope and scale; Technological evolution; Etc.—this list could be vast.

Q: What actions or methods would be required to better characterize the most important uncertainties? R: How feasible are they to execute?; Are available resources adequate?; What specific uncertainty assessment activities need to be carried out?

Q: What role is policy expected to play? R: Uncertainties about existing policy targets are drivers for evaluation work; Evaluation results could reinforce or change policy direction; Evaluation could feed yet-untested policy options; Uncertainties most relevant to policy will be investigated and explicitly communicated; A policy could enable better evaluation and/or reduce uncertainty directly.

Q: What are the main causes of any uncertainty? R: Limited knowledge; Unpredictable variabilities.

S: The performance measurement of indicators selected can be made explicit: In quantitative terms (e.g., sensitivity, ranges, impacts); In qualitative terms

S: The contributions of the main sources of uncertainty: Should be quantified; Can be quantified; Can be qualified/are more appropriately qualified.

S: If unable to quantify, uncertainty will be addressed: By identifying "unknown unknowns"; By identifying resources required to get to the stage of being able to quantify; Not at all.

S: If progress reducing uncertainty related to critical assumption *n* is not achieved: The quality of other assumptions are jeopardized; Robustness of overall conclusions weakened; Little to no impact on other assumptions.

Q: Is selected indicator (too) close to industry standard practice or to a (policy) target: R: Yes; No.

Q: Would a relatively small change in the uncertainty of an estimated indicator may have a significant effect on estimated costs, impacts, or risks: R: Yes; No.

Q: Is there a lack of consensus about the (type of) knowledge required to reduce uncertainty: R: Yes, significant; Yes, minor; No.

S: Major uncertainties exist regarding markets and/or social systems under study (behavioral issues): Yes; No.

S: Research methods likely to be used have predictable uncertainties and limitations associated with them: Yes, requires additional attention (measurements, models, scenarios, expert judgement); Not significant.

3. Using a Template and Conclusion

The exercise of working through an uncertainty assessment template can be accomplished singularly, but as described is better with appropriate stakeholder involvement and if done in a transparent manner. Group facilitation could be carried-out by one of the stakeholders capable of a comprehensive understanding the technical implications of the uncertainty factors, or by a third

party. An evaluator or evaluation contractor is a likely candidate—but only if trusted by a broad spectrum of the stakeholder community. A firm that works with a variety of utilities, regulators, agencies and community-based stakeholders would be ideal.

The use of a Delphi panel-like approach to attempt consensus could be successful given the range of stakeholders at the table. A Delphi panel would require more up-front "homework" by the facilitator, but would pay-off with a more streamlined and focused process itself and likely better quality prioritization ranking.

Advice on "how much is enough evaluation" can avoid criticism for model uncertainty (being the product of a black box) through a systematic, multi-perspective and well-vetted gap analysis that ultimately provides more than enough fodder for scope and cost estimation by experienced energy program evaluation managers. Although more work is required, the product of the process is more likely to garner broad support, and the precise source and cause of any disagreement will be clear to all involved.

UNCERTAINTY MATRIX			Level of uncertainty (deterministic knowledge – total ignorance) (knowing for certain – not even knowing what you do not know)			Nature of uncertainty		Qualification of knowledge base (backing)			Value-ladenness of choices		
Location of uncertainty ⇩			Statistical uncertainty (range + chance)	Scenario uncertainty (range = "what-if" option)	Recognized ignorance	Knowledge-related uncertainty	Variability-related uncertainty	Weak -	Fair 0	Strong +	Small -	Medium 0	Large +
Context	Assumptions on system boundaries plus economic, environmental, technological, social and political context												
Expert judgement		Narrative; storyline; advice											
M o d e l	**Model structure**	Relationships and inclusions											
	Technical model	IPMVP; UMP; Cost tests;											
	Model parameters												
	Model inputs	Quality of input data											
Data	Measurements; consumption and billing information; potential studies; market characterization; process and impact evaluation; cost-effectiveness results; market effects studies												
Outputs	Indicators; statements; broad range of possible evaluation results												

Figure 4. Example of an Uncertainty Assessment Template[7]

[7] Modified and derived from Petersen, A., P. Janssen, J. van der Sluijs, J. Risbey, J. Ravetz, J. Wardekker, and H. Martinson Hughes. 2013. *Guidance for Uncertainty Assessment and Communication 2nd Edition*, (p. 27), PBL Netherlands Environmental Assessment Agency.