The Purpose, Practice, and Profession of DSM Evaluation: Current Trends, Future Challenges

Martin Kushler, Michigan Public Service Commission
Ken Keating, Bonneville Power Authority
Jeff Schlegel, Wisconsin Energy Conservation Corporation
Edward Vine, Lawrence Berkeley Laboratory

KABOOM! The sound you hear is that of the field of energy evaluation exploding onto the scene. Over the past two decades, this field of endeavor has grown from virtually nothing to a thriving area of applied research and evaluation. In a time of almost mind-boggling expansion and evolution, the intent of this paper is to step back for a moment, take stock of where we are, and offer some thoughts on where we are headed.

The paper itself is organized into three main components: (1) the "Purposes" (functions and arenas where evaluation is being applied); (2) the "Practice" (key challenges pertaining to the implementation of evaluation/measurement methodologies); and (3) the "Profession" (issues facing the practitioners of evaluation as the profession evolves). In each section, some assessment of the current state of affairs is offered, followed by some prognostications regarding future directions and challenges.

The reader should be advised that this is not an exhaustive review, nor is it a "nuts and bolts" methodologies paper written for evaluators only. (Some other fine sources for that type of content are identified in the text.) Rather, the purpose of this paper is to identify some key issues and challenges facing the discipline, and to spur constructive dialogue regarding their solution -- a dialogue to be joined by both evaluators and those who use evaluation information.

Introduction

> "Making predictions is hard, especially when they're about the future." - Yogi Berra

Writing an assessment of the "future directions" of evaluation, performance measurement, and behavior with respect to energy efficiency in buildings is certainly a daunting task. This is so not just because it involves predicting the future (which, as Yogi suggests, is always difficult), but also because those topic areas encompass such a wide range of timely and important issues. Despite the fact that an enormous amount of ground has been covered in these areas over the past 15 years, we find ourselves today with far more questions than answers. Clearly, there is still much to learn, not only about how to evaluate specific programs, but also about such fundamental questions as: "How do buildings use energy?"; "How do technologies perform in the field?"; and "How can decision makers be influenced to adopt and correctly use efficient technologies?". Furthermore, each of those larger issues inspire dozens of specific "researchable questions" (many of which could themselves be the topic of an entire paper).

On the other hand, the importance of the subject area demands that we not retreat from the task. The United States (and indeed the world) is facing an economic and environmental imperative to improve the efficiency with which energy is used. Evaluation stands to play an indispensable role in meeting that challenge, performing such key functions as: (1) identifying promising technologies and intervention approaches; (2) evaluating and documenting the effectiveness of measures and programs which are implemented; and (3) suggesting practical and cost-effective ways in which such implementation could be improved. It is probably no exaggeration to say that the quality of the contribution by energy researchers and evaluators over the next decade could have a major influence on the quality of life on this planet!

Given the breadth, difficulty, and importance of the subject matter at hand, it is hard for the authors not to be humble. This suggests the appropriateness of a caveat to the reader that this paper cannot, and does not, cover all subjects and concerns that could rightfully be included in such a document. Constraints of space and time not only necessitated a selection of some issues among many, but also an abbreviated treatment of those issues that were selected. We trust that the reader will accept these limitations and be sympathetic to the primary purposes of this paper: to focus attention on some key issues and trends in the evaluation field, to spur debate, and to facilitate constructive responses to the challenges we face.

Format

Before proceeding to the topical content, it may be helpful to address two items pertaining to the format of this paper. The first item relates to defining the focus. In order to establish some workable parameters around this somewhat amorphous subject area, this paper is built upon the following operational assumption: "performance measurement" and "behavior" are elements incorporated within the overall umbrella of "evaluation". This is by no means intended to imply second-order status to the former topics, but seemed to be the most logical mechanism by which we could avoid the need to write three separate papers within one paper. The practical significance of this assumption is principally that the term "evaluation" is often used to broadly represent activities that encompass performance measurement and behavioral research.

The second item relates to defining the structure. This paper is composed of three main segments: (1) a consideration of the functions and arenas of application of evaluation (the "Purposes"); (2) an analysis of key issues involved in the implementation of evaluation methodologies (the "Practice"); and (3) some discussion of the circumstances of the practitioners of energy evaluation (the "Profession").

Under "Purpose," the paper begins with a brief sketch of the historical application of energy program evaluation, leading to a discussion of the current audiences and purposes of evaluation. This sets the stage for an assessment of some likely developments in the role and application of DSM evaluation over the remainder of this decade.

In the section on evaluation "Practice," a brief overview is first provided of the range of methodological approaches currently being employed in DSM evaluation. Ten quick predictions are offered concerning future developments in the application of those methods. This material provides a foundation from which the paper proceeds to identify and discuss in more detail seven key evaluation issues challenging DSM evaluators.

Finally, the paper concludes with a section addressing the "Profession" of DSM evaluation, where some emerging issues confronting this rapidly growing profession are briefly addressed.

Purpose of Evaluation: Users and Functions

Background

Energy program evaluation is several years older than the phrase, "Demand Side Management." Based on an analytic approach that was developed in the 1930's for agricultural programs and expanded during the 1960's to cover social programs, program evaluation was adapted to energy programs in the late 1970's. The early efforts were directed at Carter-era Federal programs, such as the State Energy Conservation Program (SECP) and the Energy Extension Service. These beginning efforts tended to be ad hoc evaluations, which, over time have become more organized and inclusive (see Hirst 1989, for a good historical perspective).

Early utility conservation programs were sufficiently tangential to core concerns of utilities that engineering paradigms, so dominant in utilities, did not end up directing the earliest program evaluation efforts. Instead, the evaluations were often delegated to utility marketing departments -- not surprising given the lack of consideration of conservation as a "real" resource -- where social science evaluation paradigms were better established. Yet, to this day, there has been a tension in utility evaluations between engineering approaches that seemed natural for the measurement of efficiency changes, and the social science paradigm that is better suited for measuring the organizational and behavioral aspects of program implementation. Since measures have to be delivered by a program in order to save a utility energy, and operated by consumers, we have recognized that we cannot look at the effects of measures in isolation.

The original purpose of program evaluation was to see if a program effect could be detected. Simply seeking to determine if the early programs produced statistically significant results was considered a challenge when the programs were emphasizing information and education. As conservation programs became more substantial, offering incentives for physical improvements in efficiency, the evaluation emphasis became measuring the size of the effect, and estimating its apparent cost-effectiveness. Today, with large amounts of utility investment and DSM seen as part of a utility's resource portfolio, the evaluation emphasis has become developing more precise and reliable estimates of the exact size and timing of the resource and capturing the full costs.

The field of DSM evaluation has evolved over the last 15 years to meet, and, indeed, to establish, its current purposes. There were several important developments. Perhaps the first was the shift in focus from considering survey results as data for measuring energy impacts to the use of billing data.

It didn't take long for the practitioners to move from before-and-after studies of program participants to quasi-experimental designs with comparison groups (optimistically called "control groups") to allow attribution of program effects.

By the late 1980's some utilities were beginning to consider evaluation as an integral part of good program management, and process evaluations began to be accepted as crucial for understanding the results of programs.

The need to measure coincident peak and load shape impacts resulted in more specialized metering, and brought load researchers into evaluation. Now, program evaluation has grown in scope and size, and planners, engineers, and interested stakeholders have joined the dialogue with evaluators.

As important as the early government interest was in evaluation, the field only dramatically expanded when major utilities got involved. Their interest in evaluation first evolved as they recognized that DSM programs were important customer service vehicles, and they needed to know if they were providing good service. With the recognition of DSM as an energy resource alternative, there developed an interest in impact measurement. However, it was the decision of many state regulatory bodies to grant shareholder incentives based on measured program results that has provided evaluation with its current high profile. [As of March 1992, 21 states were giving this type of incentive (Conservation Law Foundation 1992). Many of these incentive mechanisms are described by Kelly-Cochrane et al. (1991).]

The goals and uses of evaluation have grown with the size of the DSM budgets and the importance of the results. Yet it is difficult to define the current situation and how it relates to the future. In some states, evaluators are still trying to establish the need for evaluations; in other states, decision makers are making requests of evaluators for precision and breadth which are not consistently achievable. Because the sophistication of DSM evaluation will follow the growth in importance of the DSM resource, the current state of the art or science of evaluation can be represented by the most sophisticated needs which it is currently fulfilling. The current situation is one in which DSM is regarded as a resource, and one in which the results of the evaluations have real monetary impacts on utilities. Without these drivers, the odds are great that evaluation would have remained a tangential effort in most states, if it was considered at all.

Audiences

Although the extent of implementation varies among utilities, the primary users of evaluation are currently utility program managers, program planners, forecasters, regulators, intervenors/collaborative parties, and utility management. In addition, however, the public is becoming more interested as large amounts of ratepayer money are being invested, both in programs and in shareholder incentives. Government, too, is becoming an interested party again, as the US DOE has increased attention to evaluation and the EPA has become responsible for providing pollution credits for successful, and evaluated, DSM programs (leading to the attempt to develop national evaluation/measurement guidelines by Lawrence Berkeley Laboratory). It is these diverse audiences that will guide the role and purposes of evaluation in the future, not the evaluators.

Current Purposes of Evaluation

Currently, evaluation can be seen to have a few core purposes: description and characterization; measurement; and optimization of the programs.

Description and characterization involve: the operation of a program; the market reached and the market that remains; the interaction of the DSM measures with behavior; the DSM resource that remains to be captured; and why the program results occurred as they did.

Measurement includes: measurement of energy savings attributable to the program; measurement of demand impacts, including coincident peak load reductions; measurement of utility and societal costs; and measurement of persistence of savings.

Evaluations are also expected to provide the basis for optimizing programs. They do this by identifying: bottlenecks in program operation; problems in program goals, especially if they are not shared throughout the utility; the things in programs that worked well; barriers to participation; barriers to persistence of savings; and measures that may not be performing as well as expected.

Future Challenges

The near-term future purposes and practices of evaluation will be driven by the need to stretch to meet additional needs, and to meet current ones better. These challenges are set by the new audiences for evaluation. They are asking for sophistication and accuracy and, at the same time, the audiences (managers, implementors, and regulators) are seeking transparent, less expensive, and more standardized evaluations. Although these are seemingly contradictory directions, they are each important trends in the demands being placed on evaluators.

Ironically, evaluators have to some extent been "hoisted upon their own t-test." After years of trying to sell the usefulness of evaluation and its scientific nature, evaluators may have oversold the field. Now regulators are expecting rocket science in the study of phenomena very dependent on the behavior of multitudes of people. As discussed later (in the section on "Evaluation In Practice"), the difficulties of establishing a baseline, or what would have happened without the program, is partly an art. Nevertheless, energy program evaluation, with the exuberance of a new area of study, will be trying to adapt to meet the demands of its audiences and the challenges of the future. In terms of the uses of evaluation, some of these key challenges include: (1) adapting to new areas of application; (2) developing the flexibility to achieve congruence with planning; and (3) accomplishing greater integration with other disciplines.

Adapting to New Applications. One of the challenging aspects to the field of evaluation has been the way that the questions always seem to stay ahead of the techniques for answering them. The users of evaluation are already starting to ask new questions that are difficult to answer. These include measuring the impact of DSM programs on the transmission and distribution needs of the utility. This is not a question like, "what general value can we give to DSM for reduced line losses or reduced maintenance?" Rather, the issue is site-specific effects on the carrying capacity of a particular feeder as a result of the reliable impacts of specific DSM programs or measures. In 1992, it was the highest rated new evaluation priority for the DSM Taskforce of the Electric Power Research Institute.

Another area in which we are only seeing the first questions is on the persistence of DSM measures. Persistence is a key to the reliability of the DSM resource necessary to displace or defer future generating plants, and persistence, as expressed as effective measure life, is essential to the cost-effectiveness of many measures. Without measuring and verifying persistence, of both energy savings and coincident peak impacts, the real cost of the resource can be dramatically understated. Without this type of emphasis, DSM planners may be forced to face the skepticism of some utility planners who have been heard to say: "First you defined it as the lowest cost resource, and therefore you can pay any price to get it." (The persistence issue will be addressed in greater detail later in this paper.)

A third new area is developing for evaluators. This is called "verification." As summarized in Table 1, verification differs from traditional evaluation in several ways. The standards which determine how verification is done are much more contractual than they are academic. There is a strong emphasis on transparency, rather than sophistication of methods. There is a need for closure (time to get paid) even if a more precise answer could come from more time and effort. In addition, verification requires the type of trust that comes from collaborative or joint oversight of the process. The evaluator is no longer the judge of how to do the evaluation, but someone who carries out

Table 1. Some Differences Between Evaluation and Verification

	Evaluation	Verification
Statistical Significance	Essential	Mean or Median
Standards	Academic	Contractual
A Priori Rules	Change and Experiment	Change by Mutual Consent
Time Lines	Tend to Fill All Available Time (plus)	Timely; Restricted
Complexity	The More the Better	Trust is Based on Transparency
Change of Results	Everything is Subject to Change with New Information	Need for Closure

a pre-approved plan--and sometimes a party to the negotiation of that plan.

One common area of application for verification is in connection with measurement of program results to satisfy regulators who grant shareholder incentives. The key assets of timeliness and transparency are very important in this situation. (A similar circumstance is likely to apply to the Environmental Protection Agency, which will grant clean air emissions credits to utilities.)

Another emerging area for verification activity is in regard to utility contracts with third party DSM providers. Among the diversity of ways in which the DSM resource can be delivered is through third party implementors, whether not-for-profit, or more usually, for profit. These third parties are increasingly demanding the right to compete with utilities for the right to develop and deliver the DSM resource, just as third parties are beginning to achieve the right to compete with the utility to build and operate generation facilities. A key concept is that they claim to only want to be paid for the actual savings achieved. They present a double challenge to evaluators: they are demanding that evaluators rigorously evaluate utility programs so that the utility is not credited for savings with less evidence for results than they are; and they present methods for contractually measuring the savings for their own projects that someone in the utility must agree will do the job--called verification plans.

Evaluators, with their understanding of the complexity of measurement issues from their experience in trying to measure DSM impacts, are the likely people to review the verification plans. This role will become increasingly important as utilities with personnel constraints, in terms of numbers and expertise, and under pressure to open up DSM resource acquisition to competition, will allow third parties to implement their programs more frequently in the future. Yet some prior contracts of this type have turned into disappointments for utilities, because non-evaluators sometimes make the decisions on the quality and acceptability of the verification. It can be said that some third party firms made a profit not because they were very good at implementing DSM, but because they were good at writing verification plans that were unusually generous to themselves.

Whether in negotiating third party contracts or working to implement an evaluation plan approved by utility regulators, evaluators will want to set up a verification scheme that ensures an objective count, a "fair count," and care must be taken to get independent measurement. They will have the responsibility to maintain the integrity and credibility of the verifications. While working for utilities, they must be the conscience of the utility. This role will grow for evaluators, and they will have to make some adjustments in their way of thinking.

Congruence with Planning. Although verification would require more structure, and contractually determined output will be quite specific, there is a also a need for a more flexible aspect to evaluation. In addition to producing designs and products that are complete, the evaluation function in a utility may involve more ad hoc research to support the planning function in a utility. Planners depend heavily on assumptions about market behavior, measure performance, pre-existing conditions, and the state of current practice in new construction. They need to know how valid each of their assumptions are in order to improve their planning. In a way, a full blown evaluation is overkill; bottom line results are not informative. Planners need to know specifics. With the concentration of resources and skills in evaluation groups, they have been increasingly expected to fill a basic research role within utilities.

This expanded responsibility includes standard measurement issues such as savings and costs, and clearly can expand into transmission and distribution savings and persistence of savings, all of which are assumed in the models. Evaluations will become less focused on output, results and reports, and more interested in model inputs; that is, rather than simply saying that the savings weren't what was expected by the planning models for showerheads or residential lighting, evaluations will be expected to say that the pre-existing conditions assumed in the resource planning models were incorrect. The pre-existing flow for the average shower may have been 3 gallons per minute, not 4 gallons per minute as assumed; or the residential light was on for only two hours a day averaged over 5 bulbs per home, instead of 4 hours per day per bulb. This type of research has a broader scope than traditional evaluations, but the results will not only help explain the why's of what happened, help improve the program, and eliminate ineffective measures, but can also be directly translated into the planning models.

Integration With Other Disciplines. The recent growth in the importance of evaluation has paralleled the recognition of DSM as a resource and the provision of regulatory incentives. This has put pressure on utilities to re-direct staff resources toward evaluation. It is common for market researchers, planners, load researchers, and others within utilities to find themselves on an evaluation team. Together with the needs, identified above, for coordination with program design and implementation through responsibilities for validating assumptions and dealing with verification, evaluators are no longer an isolated group. They are central to the design and implementation of DSM.

As resources have been concentrated in evaluation, evaluators are beginning to serve a role as coordinators of basic program research. They not only can coordinate some R&D within the utility, but with outside parties as well. For example, through cooperative metering projects, joint evaluations, and research into current practice for motors and new commercial construction, evaluators in New England have been pooling their resources for the last few years. In the Pacific Northwest, public and investor-owned utilities have joined together to develop evaluation methods for new commercial construction programs. In California, the California Conservation Inventory Group (composed of utilities, state government, and various interested parties) is conducting cooperative research on a variety of issue areas. A similar group (the Wisconsin Center for Demand-Side Research) exists in Wisconsin. Through their shared research interests, and with sensitivity to the cost of evaluations, evaluators have begun to work collaboratively. (See Misuriello and Hopkins 1992, for a recent report identifying a number of topic areas with good potential for cooperative research.)

To fully adapt to this new role, evaluators have at least one additional challenge -- they need to learn to better communicate their results to these other interested parties and the wider audiences for their work outside the utility. Evaluators can no longer assume that they are always writing for other evaluators. Buzz words, which help simplify concepts for evaluators in the field, serve to confuse others and can appear to be adding an artificial level of complexity. More than ever before, with so much riding on the evaluations, evaluators need to take the time to clarify, simplify, and interpret results for all relevant audiences.

Evaluation in Practice: Challenges and Opportunities

Introduction

In addition to the expanding purposes and areas of application discussed above, many important methodological and implementation issues also face the evaluation field. DSM evaluation is a profession that has not yet matured. While the art, science, and practice of evaluation have all improved substantially in the last few years, there are still many challenges to address.

Perhaps the greatest challenge is presented by the fact that, at its core, evaluation is attempting to measure the unmeasurable. Impact evaluations compare what happened to program participants to what would have happened to the participants if the program had not existed (Figure 1). While the changes in consumption and demand for

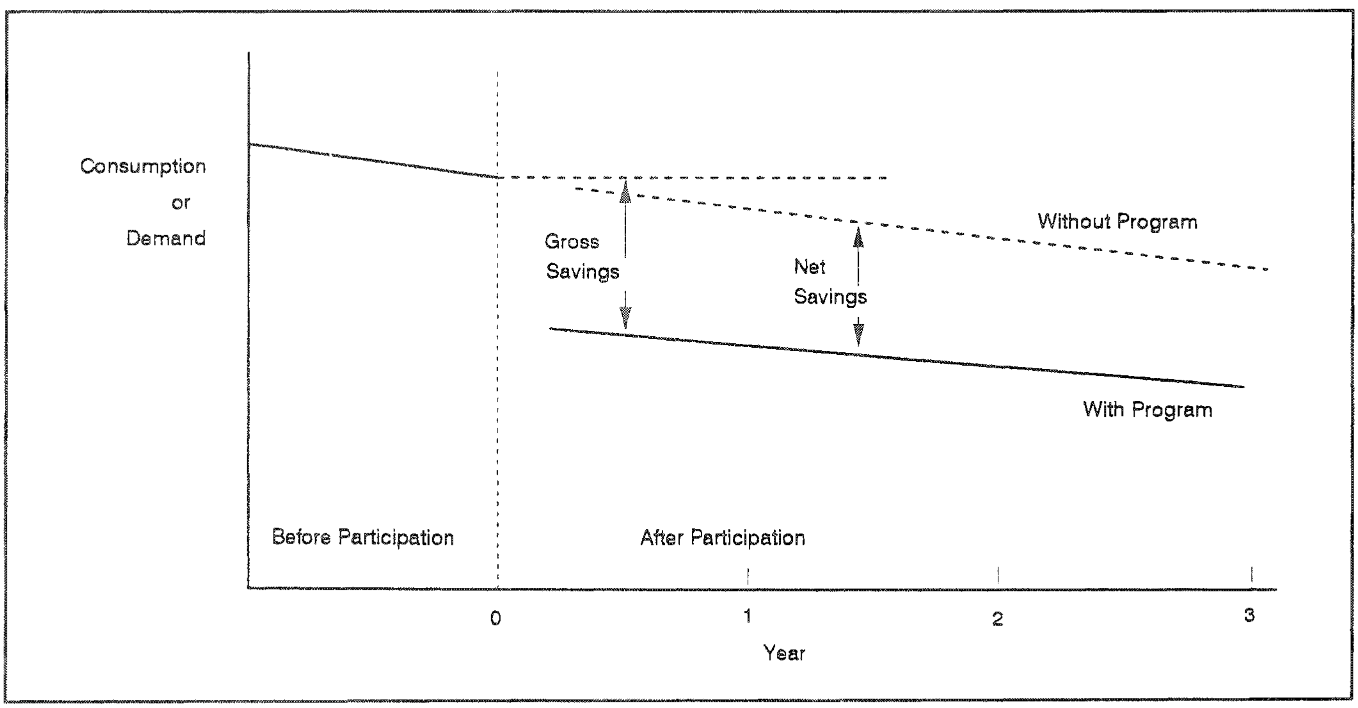

Figure 1. Illustration of Gross and Net Savings (Hirst and Reed 1991)

participants (gross savings) can be measured directly, the savings attributable to the utility program (net savings) can only be inferred by comparing measured changes for participants to an estimate of a baseline describing what would have happened to participants without the program.[1] The estimation of accurate and appropriate baselines is a fundamental challenge facing evaluators today. (This challenge is a component of several of the key issues discussed later in this paper.)

Still, many useful tools have been developed for pursuing evaluation objectives. The next section focuses on the current practice of DSM evaluation. Common evaluation approaches and methods are summarized, and current trends and future directions in approaches are discussed. This is followed by a discussion of seven key issues facing evaluators and decision makers, including: (1) the role of behavior in evaluation, (2) net and gross energy savings, (3) estimating coincident peak load savings and load shape impacts, (4) persistence of savings, (5) limits to measurement, (6) dealing with uncertainty, and (7) the role of process evaluation.

Current Methodological Approaches

There are many approaches to evaluating DSM programs. No one approach is perfect or preferred for all evaluation tasks. The choice of which evaluation approach to use depends on the type of program, the evaluation goals and objectives, the specific research questions to be addressed by the evaluation, the needs and desires of decision makers regarding the accuracy and timeliness of results, and the resources available to conduct the evaluation. Process, market, and impact evaluations may (and often do) use different approaches and methods. The selection of an impact evaluation approach also depends on whether the evaluation is measuring the effect of a program or the performance of technologies, and whether net or gross impacts are being evaluated.

The most common impact evaluation approaches can be categorized as engineering methods, billing/statistical analysis, metering, surveys, and site visits. Process evaluations use analysis of program information, interviews, focus groups, and surveys. These evaluation approaches are summarized below. The approaches are discussed in greater detail in a handbook produced by Oak Ridge National Laboratories (Hirst and Reed 1991), and in a two-volume report produced by the Electric Power Research Institute (Violette et al. 1991).

Note that these evaluation approaches occur within a process that combines experimental designs (or quasi-experimental designs), data collection methods or sources, sampling methods, and analysis or estimation methods. For example, a billing analysis may use a pre/post design with a comparison group (quasi-experimental design), customer billing data (data source), a random sample of participants and non-participants (sampling method), and mean-per-unit estimation (analysis method). It is important to think of evaluation approaches this way when designing an evaluation to ensure that all methodological options are explored, choices are not limited to what is most common or most familiar, and the most appropriate methods (or combinations of methods) are selected.

Engineering Methods. In this method, program impacts are estimated based on generally-accepted engineering calculations and models. The engineering estimates may be developed using simple algorithms or detailed computer models. Engineering estimates are conceptually straightforward, relatively quick and easy to develop, replicable (if the same methods and assumptions are used), and inexpensive. Perhaps their greatest advantage is that they can be developed in a timely manner. Unfortunately, engineering estimates are often inaccurate (Nadel and Keating 1991), mainly because they are based on unvalidated assumptions and do not properly account for customer behavior. Not only are engineering estimates often inaccurate, but they also tend to be biased towards overestimating savings. While detailed simulation models can provide greater accuracy than simple calculations, the models are time-consuming and more expensive to use, and the degree of accuracy achieved depends on the skills and experience of the modeler.

More accurate engineering estimates (from calculations or simulation models) generally result when site-specific information is used for inputs rather than default data or assumptions. However, since many engineering models were intended for building design, some do not accept measured data easily. Further, there is no current agreement among practitioners on how to calibrate or tune engineering models using measured data.

Engineering methods are most useful for impact accounting, evaluating new construction programs, estimating coincident peak load savings using measured energy savings and load shape information, and developing estimates for use in statistical models.

Billing/Statistical Analysis. Billing/statistical analysis approaches include many methods from simple pre/post billing analysis to complex regression models. Billing analysis methods estimate program impacts based on energy consumption data contained in the customer billing system. Billing analysis generally compares consumption data before and after participation in the program for a sample of participants (treatment group) and non-participants (comparison group). Many billing analysis methods normalize consumption data to account for changes due to weather, such as in the PRISM model (Fels 1986), and other factors such as business modifications, energy prices, or general economic conditions.

Some billing/statistical analyses use multivariate regression techniques, often in two-stage or pooled time-series cross-sectional models. These regression models can be used to: (1) increase precision by explaining some of the variability in consumption and savings due to other factors such as weather, economic factors, or customer attitudes; (2) increase precision and inform engineering estimates by using engineering estimates as independent variables in the regression equation; (3) adjust for free riders and self-selection bias; (4) estimate savings by measure type; and (5) explore other associations between savings and other variables.

Conditional Demand Analysis is one type of regression model that uses data on appliances and existing end-uses to estimate consumption or savings (either one can be estimated depending on the model used) (Parti and Parti 1980, Parti and Rogers 1991). Conditional Demand Analysis models total consumption as the sum of the consumption of each appliance or end-use. As such, it generally uses customer and appliance survey data as inputs.

Billing/statistical analysis is most useful for estimating net and gross energy savings from the program. Its main advantage is that it captures both the technological and behavioral effects of the program on customers' consumption. The main disadvantage of billing analysis is that it captures all changes and variability in customers' consumption, some of which are due to other factors besides the program. Therefore billing analysis is limited to programs where the impacts due to the program are large enough that they can be discerned from changes in consumption due to other factors (many DSM programs fit this category). While some billing analyses models are designed to control for changes in consumption due to other factors, these models are often limited by the ability to obtain data (e.g., on business modifications, customer attitudes and actions, etc.)--especially for comparison group facilities. Because billing analysis is relatively inexpensive, it is often done even if other methods are being employed.

Metering. Metering approaches use direct measurements of energy consumption, demand, and other parameters to estimate energy savings, coincident peak load reductions, and load shape impacts. There are several types of metering methods, including end-use metering, whole-facility metering, short-duration metering, and spot metering.

End-use metering uses multi-channel metering equipment that records data at frequent intervals to measure the energy consumption of specific end-uses or devices. End-use metering has the major advantage of providing an accurate, high-resolution measurement focused on the end-use (or end-uses) of interest. In addition, since end-use metered data is differentiated by time period, it can be used to estimate coincident peak load reductions and load shape impacts. The main disadvantage of end-use metering is its cost, generally ranging from $2,000 to over $20,000 per facility, which limits the number of sites that can be metered (Michaels, Hoffman, and Schon 1991; Eto et al. 1990). Also, there are practical limits to end-use metering such as customer approval and timely collection of pre-participation data, since metering equipment must be installed before data collection can begin.

End-use metering is most useful for evaluation of gross energy and coincident peak load savings (rather than net savings, since end-use metering is rarely installed in non-participating facilities), technology evaluation (where one is interested in the gross change in consumption due to a particular measure), providing data for statistical analyses and combination approaches, addressing specific research issues such as estimating interactive effects, and joint utility projects where the relatively high costs can be spread across several utilities.

Three improvements would make end-use metering more appropriate on a wider scale: (1) reduction in per-site costs due to improvements in metering equipment or metering procedures and processes (e.g., installation, data collection and processing, analysis, etc.); (2) less intrusive meters that would ease installation and reduce customer objections, as well as reduce potential testing bias; and (3) procedures for identifying participating customers early so that meters could be installed and sufficient pre-participation data could be collected. In addition, end-use metering would need to be used in non-participating facilities before it gains wide acceptance as a method for estimating net impacts. Advances in metering technology (particularly advances in electronics and signal processing) and procedures are underway to meet these challenges. These advances will be applied to all types of metering: end-use, whole-facility, short duration, and spot metering.

Whole-facility metering uses high-resolution metering to measure the energy consumption of an entire facility, rather than the consumption of an individual end-use or device. Evaluation activities sometimes can be coordinated with load research activities to take advantage of the whole-facility data provided by load research recorders.

The main advantages of whole-facility metering are that it collects time-differentiated data that can be used to estimate coincident peak load savings, is less expensive than end-use metering ($500 to $3,000 per facility), and can be installed without the knowledge or consent of the customer (Michaels, Hoffman, and Schon 1991). Whole-facility metering can be used to estimate net program impacts if installed in both participating and non-participating facilities, since the meter captures all of the changes in consumption due to the program (including technological and behavioral effects). It has the same disadvantage as billing analysis in that the data includes all changes in consumption or demand, not just those due to the program, thereby limiting whole-facility metering to larger-impact programs. Also, whole-facility metering does not focus directly on the measure or end-use of interest, though end-use impacts can be estimated statistically if they are a large portion of the total load.

Spot and short-duration metering measure changes in key values either instantaneously or over short time periods. These types of metering can be very useful for measuring wattage, changes in wattage on affected circuits, operating hours, and frequency of use. These data can be used in engineering methods and statistical analyses. Some short-duration meters can provide time-differentiated measurements that can be used to estimate coincident peak load reductions and demand impacts. The main advantages of spot and short-duration metering are that they directly measure parameters of interest, are relatively inexpensive, are easy to use and install, and can be coordinated with other evaluation activities such as site visits and surveys.

Surveys and Site Visits. Surveys are data collection techniques used to collect information from customers (program participants and non-participants), trade allies, program staff, and other individuals involved in the program. These methods are often used in process evaluations where program design, planning, administration and management, marketing, implementation, and delivery are assessed. Mail, telephone, or in-person surveys can be conducted.

Surveys can also be used in impact evaluation. Surveys can be used to assess customer attitudes and actions, collect data on factors that influence customers' decisions to participate in a program, estimate schedules and operating hours, and collect data on changes that effect consumption (e.g., occupancy, usage, facility size, business expansion/contraction, additional measure installation, etc.).

Site visits are used to gather information on-site via facility inspections and customer surveys. Information from site visits can be used for many purposes, including verifying and improving tracking system data, informing engineering estimates, supporting billing/statistical analyses, improving program design, observing program staff and contractors, assessing program comprehensiveness, evaluating the effectiveness of technical aspects of the program, and determining measure persistence. Site visits can also be combined with other data collection methods such as spot metering.

Combination Approaches. The combination of multiple methods in an evaluation is a major trend in impact evaluation. The goal of many of the combination approaches is to increase precision (or alternately, reduce sample size for a given level of precision) by explaining some of the variability in consumption or savings through the use of available information. Using multiple methods and data sources allows an evaluation to use available information in an effective and cost-efficient manner.

There are several approaches for using multiple methods--three major approaches are summarized here (Violette 1991).

Triangulation involves comparing the results of more than one method used to estimate program impacts. The results of each method are compared to the results of the other methods to gain insight into the quality and reliability of each estimate. Triangulation is sometimes used to compare different estimates of free rider ratios from customer surveys, trade ally surveys, and participation models.

Leveraging uses data from one method as inputs to another method. Using the information as inputs may help explain some of the variance of savings, thereby increasing precision. Two examples of leveraging information are summarized below.

Stratified ratio estimation is an approach that combines engineering estimates with other evaluation approaches (such as metering or billing analyses) to obtain greater precision through the use of two sampling and estimation techniques--stratification and ratio estimation (Northeast Utilities 1991). Both of these techniques increase precision by using available information (e.g., consumption/ft^2 or engineering estimates) to explain some of the variation in consumption or savings for the population. Ratio estimation generally obtains a more precise estimate of a population quantity (such as average energy savings) by calculating the ratio between the measured savings and the engineering estimates of savings for a sample, and using this ratio to adjust the engineering estimates of the entire population.

Statistically Adjusted Engineering (SAE) estimates is an approach that combines engineering estimates with statistical analysis using billing or metering data. SAE can be used to estimate the accuracy of engineering estimates by using the engineering estimates in a regression equation rather than dummy variables. SAE approaches can also be used to disaggregate whole facility metered data into individual end-uses. Both of these applications may be appropriate for DSM evaluation.

Bayesian methods use a systematic approach to update and adapt prior information based on additional information from a new evaluation (Violette 1991, Violette et al. 1991). Prior information on savings could come from engineering estimates, previous studies, or expert judgement and be used for inputs to a Bayesian estimation framework.

Process Evaluation Methods. Surveys, focus groups, interviews, site visits, and the analysis of program information are used in process evaluations to assess program design, planning, administration and management, marketing, implementation, and program delivery. These methods are also used to explore attitudes and actions of customers and trade allies, and for determining customer satisfaction with the program.

Process evaluation methods are used to collect information from individuals involved in the program such as customers (program participants and non-participants), trade allies, and program staff. The information can be collected in-person, by telephone, or by mail.

Some of the important advances in process evaluation are not in the basic methods themselves, but in how the methods are applied. For example, technical process evaluations use site visits and surveys to assess the technical aspects of a program, such as measure selection procedures, the level of comprehensiveness, quality of program delivery and measure installation, and the abilities of program staff and contractors. Technical process evaluations are also conducted as part of impact evaluations to explore why savings were achieved (or not achieved). Often buildings with high or low savings (or savings that vary widely from expectations) are identified for follow-up site visits. This information is used to identify additional opportunities for savings and to improve the technical components of the program.

One major trend in evaluation that will only increase in the future is the integration of impact and process evaluations. Two examples of this integration are the use of site visits and surveys in technical process evaluations, and the use of customer surveys to collect data for both process and impact evaluations.

The Future. Ten key predictions can be made about the future of evaluation approaches. First, many evaluations will use multiple methods in combination approaches to enhance cost-effectiveness, increase precision, and reduce uncertainty. Second, the use of metering will increase, especially as costs drop and as less intrusive equipment becomes available. In the near future, spot, short-duration, and whole-facility metering will be used much more frequently. Third, more data on customers, including customer opinions, attitudes, actions, and decisions will be used in impact evaluation. Fourth, simple before/after billing analysis will become routine. Fifth, billing data will be used more commonly with information on customer decisions and actions in multivariate statistical models to develop more accurate estimates of net program impacts by developing better estimates of the baseline (what would have happened in the absence of the program). Sixth, site visits will be used much more frequently to collect and verify data, assess the effectiveness of technical aspects of the program, and identify why savings were achieved (or not achieved). Seventh, evaluation will be integrated with other customer analysis activities such as market research and load research. Eighth, techniques and experience from other disciplines will make contributions to improving DSM evaluation approaches. Behavioral science and market research experience may make the largest contributions. Ninth, more effective methods will be developed to conduct multi-year studies and evaluate the persistence of savings. And tenth, process and impact evaluation approaches and methods will become more coordinated and integrated.

Key Evaluation Issues

The evaluation community is faced with a number of critical issues that are currently being addressed and some that will need to be addressed in the near future (Table 2). While there are many important issues listed in Table 2 (and other issues could surely be added to this list), space limitations caused us to focus on seven issues that we believe are of particular importance. These issues are: (1) the role of behavior in evaluation, (2) net and gross energy savings, (3) estimating coincident peak load savings and load shape impacts, (4) persistence of savings, (5) limits to measurement, (6) dealing with uncertainty, and (7) the role of process evaluation.

Table 2. Examples of Key Issues in Program Evaluation

Role of behavior in evaluation
Net and gross savings
Estimating coincident peak load savings and load shape impacts
Persistence of savings
Limits to measurement
Dealing with uncertainty
The role of process evaluation
Integration of impact evaluation and process evaluation
Maximizing precision versus minimizing bias
Assessing market transformation
Predicted versus measured savings
Avoiding lost opportunities and cream skimming
Quality assurance, confirmation, and validation
Timeliness of information and feedback
Presentation of results--clarity, honesty, and objectivity
Verification versus evaluation of program savings
Definition of key DSM program evaluation terms
Comparability of results (across programs, utility service territories, states, and countries)
Generalizing results from metered subsamples to larger populations
Incorporation of environmental externalities
R&D needs for measuring technology performance

Importance of Behavior in Evaluation. Behavioral variables play an important role in both impact and process evaluation. Some reasons for this include: (1) programs are conducted by individuals within organizations that have their own internal dynamics that may affect overall savings of a program (e.g., the relationship between top management and middle and lower management, the capability and inclination of field staff regarding the use of new software tools in selecting energy efficiency measures and targeting customers, etc.) (Vine et al. 1992); (2) the adoption and diffusion of energy efficiency measures is dependent on key behavioral factors (e.g., customer aversion to risk, sociodemographics of participants, maturity of the marketplace) (Braun and Williams 1979; Claxton et al. 1981; Sharma 1983); (3) the performance of energy efficiency measures is not only dependent on the technical performance of a particular measure, but also, and perhaps more importantly, on how the measure is used and maintained (Vine 1992); and (4) the modeling of energy use in buildings relies on the use of key variables representing important behavioral elements (e.g., thermostat settings and appliance/equipment usage) (Vine et al. 1982; Gladhart and Weihl 1990).

In the very early years of energy program evaluation, there was a fair amount of social science research applied to the behavioral aspects of energy efficiency (e.g., see Stern and Gardner 1981; Yates and Aronson 1983). Since the mid 1980s, however, (and particularly since utility DSM incentives have become popular) there has been a heavy emphasis on impact evaluation and technical measurement and engineering methodologies. Although some have articulated the need to integrate behavioral research into energy evaluation (e.g., Lutzenhiser 1990) and some very interesting behavioral research has continued (witness the papers in the Human Dimensions panel from the 1990 ACEEE conference), most emphasis has tended to center on the technical/engineering aspects.

Increasingly, however, the realization is growing that it is necessary to integrate important behavioral variables into impact evaluation methodologies. For example, in a recent comparison of engineering estimates of gross savings to impact evaluations of net savings, Nadel and Keating (1991) found large percentage differences between the two estimates in the following types of programs: residential retrofits; residential appliances and equipment; residential new construction; and commercial retrofit. Where engineering estimates and impact evaluations differed, a number of common explanations emerged, including the following: (1) wrong assumptions in the engineering estimates (e.g., over-estimating lighting system operating hours); (2) "takeback" effects for some measures (air-conditioner rebates in moderate climates, low-flow showerheads, and residential compact fluorescents) were not modeled in the engineering estimates; (3) quality control problems during measure installation, commissioning, and maintenance; and (4) greater than expected adoption of conservation measures by nonparticipating customers (which will lower the net savings attributable to a program).

At this point, it is being further recognized that social science (behavioral) research questions are central to a number of critical evaluation issues. Some examples include the following:

1. How does an energy efficiency program affect people's behavior? (Important for calculating net energy savings, self-selection bias, free riders and free drivers, cream skimming, and the rebound effect);

2. How do occupants/consumers use (operate) energy efficiency measures? (Important for determining baselines and persistence of savings, and for comparing measured and estimated savings);

3. What is the experience and expertise of building managers? (Important for determining baseline and persistence of savings, and for comparing measured and estimated savings);

4. How reliable is the information supplied by consumers? (Important for determining baselines and comparing measured and estimated savings);

5. How heterogeneous is behavior? (Important for designing samples and determining baselines).

Although behavioral research still remains relatively underemphasized in the DSM evaluation arena, the situation is changing.

The Future. Social science researchers should be poised to expect a renaissance in behavioral research. As increasingly sophisticated modeling techniques are developed and perfected, accurate and reliable behavioral inputs will be of increasing value. Even more importantly, as the flood of impact evaluations which have been initiated in the last couple years begin to produce results--results which will undoubtedly be discrepant from predictions/expectations--there will be a clamor to provide information to "explain" those results. Research regarding the "human dimension" will be a crucial aspect of the necessary response.

Net and Gross Savings. Gross savings are the changes in annual energy use and peak demand experienced by participants in the utility's program. In contrast, net savings are typically the savings that can be directly attributed to a utility program. In other words, net savings are the difference between gross savings and the change in consumption and demand that participants would have achieved had the program not existed. Nonprogram savings (or energy increases) reflect customer responses to changes in electricity and fossil-fuel prices, changes in economic activity or personal income, introduction of new electricity-using technologies, and other nonprogram factors. Estimation of nonprogram savings typically requires a comparison group to approximate what program participants would have done without the program. Nonparticipating customers that are eligible for the program are often used as a comparison; sometimes, preparticipation data for participants are used as a comparison (Hirst and Reed 1991).

Because the purpose of an impact evaluation is to measure changes in energy use (for both participants and nonparticipants), a baseline must be identified from which a change can be measured or estimated. Determining this baseline is a critical step, but fraught with difficulties. In fact, many of the estimation problems associated with impact evaluation are related to the selection of an appropriate baseline. Three key problems associated with this selection are: self-selection bias, free riders, and free drivers.

Self-Selection Bias and Free Riders. Self-selection bias occurs when program participation is voluntary: systematic differences may occur between program nonparticipants, who either chose not to participate or were unaware of the program, and participants. Free riders are defined as those participants in a conservation program that would have installed energy conservation measures even if there had been no program. Customer surveys of different types of utility programs report free rider fractions ranging from 20 to 80%. Utilities need to minimize the effect of free riders if they do not wish to pay for some DSM actions that would have been installed without the program. Furthermore, utilities need to measure this effect if they wish to distinguish between gross savings and net savings. Because free riders represent a cost to the program but offer no direct benefits in return, DSM programs that are highly successful from a gross savings perspective may prove to be less attractive when free riders are considered. Accordingly, the accurate measurement of free ridership is a significant issue for evaluators and program managers of DSM programs.

If free ridership is viewed as a subset of self-selection bias, then the most appropriate procedures for addressing self-selection may not allow the analyst to separately identify the impacts of free riders independent of the broader self-selection bias (Violette et al. 1991). The pros and cons of various data analysis methods for estimating the impact of DSM programs on energy consumption with regard to these issues are discussed by Violette et al. (1991), and the various options for measuring free riders (surveys, marketplace data, comparison groups, and statistical models) are discussed by Saxonis (1991). This is not just a theoretical problem: regulators in some states are requiring the specific estimation of free-rider effects.

The measurement of free riders can be enhanced with survey instruments, data from market studies and comparison groups, and modeling techniques. Surveys can be improved by targeting questions to address (1) free ridership, (2) net program impacts, and (3) types of programs (e.g., by type of appliance or market delivery mechanism). In addition to developing a positive relationship with manufacturers and distributors, market data can be obtained by requiring dealers to provide sales data as a prerequisite for participating in a program. Monitoring activity of a nonparticipating region with similar characteristics to the test market is a practical and efficient option for analyzing free riders. Finally, a number of modeling techniques are available for evaluating self-selection bias and free riders (see Violette et al. 1991), but few studies have used these new techniques.

Finally, program designs can be improved to reduce the impact of free ridership: (1) require preinstallation inspections to ensure that potential participants do not have the efficient equipment promoted in the program; (2) encourage the use of very efficient (advanced technology) equipment; and (3) target customer groups with, historically, low free rider rates (e.g., low-income households and small commercial and industrial customers).

Free Drivers. A free driver contributes to the goals of the program (e.g., reduced energy consumption) but is not formally a program participant (Saxonis 1991). A free driver is affected by the program either through a conscious awareness of the program or because of program-induced changes in the marketplace (e.g., a customer who purchases a product that qualifies for a rebate but does not claim the rebate, or a builder who constructs a home to program standards but does not choose to participate in the program).

Free drivers are more likely to be a significant problem for programs that have been in existence for several years and have achieved high participation levels. Research on free drivers is limited, but three approaches are available to address this issue (Saxonis 1991). First, use a historical baseline from the early years of the program. Second, use survey methods to determine whether nonparticipants have changed their energy use as a result of the program and to determine (via trade ally interviewing) if the market for the DSM actions promoted by the program has moved in such a way that nonparticipants are impacted. And third, use a community (or set of communities) outside the area in which the program is offered, and compare the distribution of efficiencies for program-sponsored measures in the participating area with those in the comparison area.

The Future. In the future, evaluation activities will focus on the following issues. First, as program evaluations continue to rely on measured data, evaluators will be asked to combine engineering and measured data for calculating gross and net energy savings. Most likely, future program savings will reflect "hybrids" of engineering and measured data. Second, more attention will be devoted to the calculation of net energy savings (i.e., savings caused by the program), leading to the allocation of more resources to social science (behavioral) research. Third, obtaining a good control group and good baseline data and calculating free riders and free drivers will continue to plague evaluators and will lead to the use of new methods and new applications of existing methods. Fourth, as more program evaluations are conducted, comparing and transferring results and methodologies across programs, utility service territories, states, and countries, and over time, will need to be facilitated. The national database on energy efficiency programs (DEEP), focusing on measured results from program evaluations, will assist the evaluation community in this matter (Vine 1992c). Fifth, clearer definitions of what is included in "net savings" or "net benefits" will be needed. In particular, the question of how (or whether) to account for the benefits due to market transformation resulting from utility DSM programs needs to be addressed. Finally, the inherent difficulty of the task of precisely measuring concepts such as self-selection bias, free riders, and free drivers will force evaluators to be more modest in their claims and their audiences to be more realistic in their expectations.

Estimating Coincident Peak Load Savings and Load Shape Impacts. Most impact evaluations have concentrated on estimating energy savings. More attention needs to be paid to estimating coincident peak load savings (kW or peak day therms), especially where the benefits of DSM programs include substantial avoided capacity. In addition, information on the impact of the DSM programs and technologies on load shapes is needed to determine the value of the savings (where time-differentiated impacts are important), and to develop more effective and better targeted programs. This information is also important for resource planning.

In general, estimating coincident peak load savings is more difficult and costly than estimating energy savings, especially where billing analysis is used to estimate energy savings. Two types of information are needed to estimate coincident peak savings: (1) change in connected load, resulting in a need to know both pre- and post-installation connected load; and (2) the coincidence of the load with the time of system peak (either peak hour for electric utilities, or peak day for gas utilities). For some dual-peaking utilities, savings must be estimated for both the summer and winter peaks. Measurement approaches generally rely on some form of metering to collect the necessary data. The techniques employed to measure coincident peak load reduction can also be used to estimate other types of demand savings, such as load shape impacts or changes in monthly customer demand.

Several approaches for evaluating coincident peak load savings, load shape impacts, and customer demand savings are summarized below. These approaches are listed roughly in the order of increasing cost, and of increasing accuracy.

1. <u>Engineering methods</u>. Engineering methods are the most common and least costly method used -- but also the least accurate. The accuracy of engineering estimates can be improved by using better load shape data for the particular utility system, by calibrating the estimates using energy savings data, and by developing better engineering methods based on measured data.

2. <u>Site visits and short-duration, time-differentiated metering</u>. These methods can collect more accurate data for use as inputs to engineering methods. For example, site visits can be used to collect more accurate data on operating schedules, time-differentiated metering (light loggers) can measure operating hours and coincidence with system peak, and spot metering can measure the connected load (or change in connected load). To be most accurate, these measurements should be made both before and after installation.

3. <u>Whole-facility metering</u>. Whole-facility metering can measure coincident peak load savings and load shape impacts at the facility level. If installed in comparison facilities (e.g., for future participants or as part of load research) it can also provide an estimate of net peak savings. Statistically Adjusted Engineering (SAE) can be used with load shape data to statistically disaggregate whole-facility data into individual end-uses where the change is large enough and is not confounded by collinearity or other problems.

4. <u>End-use metering</u>. End-use metering can measure coincident peak load savings and load shape impacts at the end-use level. To be most useful, the metering equipment should be installed before participation in the program to ensure the collection of pre-participation data--this can be a barrier in many programs. Metered end-use data for non-participants will also be necessary to estimate net peak load savings.

The Future. As DSM becomes more of a resource, and overall program impacts increase, accurate estimates of coincident peak load savings and load shape impacts will become very important. All of the approaches summarized above will be used more frequently in the near future--the question is how can they be applied in each program most effectively and efficiently. Metering methods will have a large role, and spot, short-duration, and whole-facility metering will make substantial contributions, especially in smaller facilities or in smaller programs where funding for evaluation is lower. As less expensive and less intrusive metering equipment becomes available due to advances in electronics, and as more effective data collection and analysis procedures are developed, metering approaches will become routine.

Improved engineering methods and better load shape data will also be needed in the near future. While there should be (and needs to be) improvement in the accuracy of the data, some have suggested that engineering estimates and measured data will actually converge over time. Since it is unlikely that estimates from engineering methods and measurement approaches will converge to the degree desired (even if they are routinely calibrated), and since technologies, programs, and customer attitudes and actions are always changing, some form of measured data will be necessary well into the future.

Persistence of Savings. The persistence of energy and demand savings is an important issue to many stakeholders: e.g., building owners, architects and engineers, utility program managers and evaluators, regulators, utility shareholders, resource planners, forecasters, researchers, and consumers. For instance, resource planners need to know if the energy saved through energy efficiency programs will offset generating resources (i.e., is it reliable and durable?), and utility shareholders need to know if financial incentives (based on measured energy savings) received from their utility's investment in energy efficiency measures will continue. The rewarding of financial incentives for utility investment in energy efficiency measures makes it even more imperative to use credible, defensible measure-life data in cost-effectiveness testing, DSM planning, and program impact evaluation.

In a recent study of research opportunities to improve DSM impact estimates, the persistence of energy savings was noted as probably the single, largest, unanswered question in demand-side management (Misuriello and Hopkins 1992). Until recently, persistence was assumed to be relatively constant, and most analyses of persistence relied on engineering estimates of measure life and judgement (with adequate justification). For example, most planners assumed that knowing the physical life of a measure installed was sufficient to determine persistence: i.e., first-year savings continued for the life of the measure (e.g., 20 years). Recently, this assumption has been challenged as the issue of persistence has gained more prominence in the evaluation of energy efficiency programs. In fact, the limited empirical research conducted so far (see below) raises questions about the validity of using manufacturer's claims for physical measure lives as a basis for projecting persistence.

Two dimensions of persistence exist: measure persistence and program persistence (Keating 1991). The former focuses on measure lifetime and operation, while the latter emphasizes total and net program impacts (Gordon et al. 1988; Vine 1992). The two dimensions are interrelated in that program persistence includes measure persistence variables as well as other factors. Key parameters affecting the different dimensions of persistence are noted in Table 3 (based on Misuriello and Hopkins 1992).

Table 3. Parameters Influencing the Persistence of Energy Savings

Measure Persistence

- Technical lifetime
- Measure installation
- Measure performance or efficiency decay
- Measure operation (behavior)
- Measure maintenance, repair, commissioning
- Measure failure
- Measure removal
- Changes in the building stock (renovations, remodels, alterations, additions)
- Occupancy changes (turnover)

Program Persistence

- Rebound (snap-back, take-back) effects
- Surge effect (additional measures added by customer after initial program participation)
- Replacement effect (replacing efficiency measures with less or more efficient measures)
- Energy use by control group

Persistence Studies. Until recently, there have been few published studies on the topic of persistence of energy savings. The study of persistence is in its infancy because (1) true impact evaluation research on DSM programs has recently begun; (2) few programs of the duration necessary to provide the data required to investigate persistence have been evaluated; and (3) until recently, persistence has not been focused upon as an important issue. Readers are encouraged to read summaries of some of the key persistence studies in Vine (1992), Keating (1991), SRC (1992b), Cohen et al. (1991), and Maketi (1992). Also, although not focused specifically on persistence of energy savings, the Family Energy Project at Michigan State University conducted several phases of behavioral monitoring during the early and mid 1980's with the objective of identifying patterns in energy behavior and factors influencing those patterns (Weihl and Gladhart 1990). The methods, approach, and results of this project should be useful for researchers studying the persistence of energy savings (and energy-related behavioral research in general). The key highlights of recent persistence research are noted below.

(1) Reliance on technical or average service lifetime may overestimate DSM measure savings, particularly in the commercial sector where renovations and remodels occur frequently and where removal or deactivation of energy conservation measures occurs often (Petersen 1990; Hickman and Steele 1991; Skumatz et al. 1991). In addition, certain building types appear to be more susceptible to frequent remodeling and turnover: e.g., office, retail, restaurant, and warehouse sectors.

(2) Studies on measure life in the residential sector have focused on removal rates of particular measures: typically, measures with the highest removal rates are low-flow showerheads, compact fluorescent bulbs, and door weatherstrips (SRC 1992a).

(3) Most of the persistence studies that have examined energy savings have focused on residential weatherization programs (both low-income and standard income) in the Pacific Northwest. The limited information on energy savings from these studies has shown that DSM program participants have not tended to increase their energy use over time; however, it is also true that the control group of non-participants have tended to lower their energy use over time (Sumi and Coates 1988 and 1989; Ecker 1991). As a result, the difference in energy use between the participant group and non-participant group narrows and net savings is reduced. Nevertheless, preliminary results indicate that the potential for the durability of net program savings is very good.

The Future. For the future, the outlook is somewhat more encouraging for obtaining a more comprehensive analysis of persistence. The California Conservation Inventory Group (CCIG) is conducting a statewide scoping effort on the persistence of energy savings. The members of the CCIG include the California Energy Commission, the California Public Utility Commission, each of the State's major electrical and gas utilities, the Natural Resources Defense Council, Lawrence Berkeley Laboratory, and the California Institute for Energy Efficiency. The purpose of CCIG is to collect and analyze statewide data on the potential to reduce energy use through conservation.

Recognizing that persistence is a statewide problem, funds from the utilities are being used to address this critical issue. The research project will: (1) identify current research in the area of effective measure life and other persistence issues; (2) evaluate the feasibility of utilizing existing utility conservation program information to estimate effective conservation measure life; and (3) develop a panel group evaluation design to follow a group of current program participants over a number of years. The New York State Energy Office and the Bonneville Power Administration (BPA) are also planning a persistence study in the residential and commercial sectors. Finally, the New England Electric System (NEES), Northeast Utilities, and BPA are planning to study measure life in the commercial sector.

Several strategies are available for ensuring the persistence of energy savings and will be (and should be) the focus of future efforts in this area: (1) measurement and verification plans; (2) technology performance tools; (3) operations and maintenance; (4) building commissioning; (5) training and education; (6) program design; and (7) cooperative evaluation projects (Vine 1992). In this paper, we highlight the first two strategies.

Measurement and verification plans developed by utilities can be an important tool to examine the persistence of energy savings. For example, the Central Maine Power Company has developed a measurement and verification plan that requires measurement of their energy efficiency programs one year after measure installation (for years 1991, 1992, and 1993) and a second measurement in 1997 (Haeri 1991).

In January 1992, the Massachusetts Department of Public Utilities ordered the Cambridge Electric Light Company and the Commonwealth Electric Company to monitor savings annually until a consistent savings level (within 10%) is maintained for at least two consecutive years in order to establish confidence that significant changes in savings will not occur within the intervals (every three years) proposed by the utilities (Hastie 1992).

New and/or improved technologies are also needed for measuring persistence. A recent report identified viable research and development opportunities that could improve capabilities to determine the energy use and demand reductions achieved through DSM programs and measures (Misuriello and Hopkins 1992). Table 4 identifies these recommendations, most of which are applicable to the study of persistence. Because of the magnitude of the effort, coordinated and cooperative research is needed.

Finally, as all these new (and essentially descriptive) persistence studies begin to identify the durability of savings as a serious problem plaguing DSM programs (which they very likely will), DSM implementors will respond. As a result, a new applied area of persistence research will emerge: evaluating the effectiveness of DSM program modifications created to enhance program and measure persistence.

Limits to Measurement and Evaluation. In today's environment of increased demand-side activity, the demands on measurement and evaluation approaches have often exceeded their capabilities. In some cases, the

Table 4. R&D Recommendations for Improving the Analysis of Persistence

Priority 1 Opportunities
Fill Critical Steps to Meet Basic Evaluation Requirements

Statewide DSM persistence study
Impact methods using whole-building data
Short-term measurement techniques
Integration of statistical, engineering, and behavioral models
Value engineering study to reduce monitoring costs
Specialty "test kits" for DSM field measurement
Field test methods for HVAC measures
Methods to evaluate low-impact measures
Methods to evaluate low-frequency measures

Priority 2 Opportunities
New or Advanced Methods to Replace Less Effective Ones

Guidelines to calibrate simulation models with measured data
Literature guide to experimental design
Statewide baseline performance data compilation
Reduce multicollinearity through advanced sampling techniques
Expert system applications for field monitoring projects
Site measurement plan "recipe book"
Intra-building sampling techniques

Priority 3 Opportunities
Enhancements to Existing Methods

Protocols for data collection project planning
Improvements for DSM administrative tracking systems
Feasibility study on self-metering appliances
Resolve issues with interactive and secondary effects

Source: Misuriello and Hopkins (1992).

capabilities of impact evaluation approaches have been overpromised, misrepresented, or misunderstood. Many regulators and utility managers have unrealistic expectations regarding the performance or cost of evaluation approaches.

There are technical, economic, and practical limits to measurement. The main factors that limit evaluation are summarized below.

First, the net effects of demand-side programs can never be directly measured, only estimated. DSM impacts are determined by two factors: the outcome that would have occurred in the absence of the program, and the outcome that actually occurred with the program in place. The latter can be measured, but the former can only be inferred, using experimental designs and statistical methods widely accepted by evaluators.

Second, the determination of net impacts of demand-side activities relies largely on social science approaches. The impacts of demand-side programs are more than the technological impacts of the measures installed in the programs--changes due to the behavior of customers and trade allies must also be assessed. While the technological impacts of a demand-side measure can be measured directly, the behavioral and market impacts of technologies and programs are much more difficult to assess. For example, the flowrates of showerheads can be measured and the resulting energy savings estimated using a simple engineering formula. However, this method does not consider behavioral factors such as changes in water temperature setpoint, or changes in the length or frequency of showers.

Third, DSM programs live in a dynamic environment where many things are changing over time. With the increasing amount of DSM activity, and associated market transformation, it may become even more difficult to determine the effects of a particular DSM program. For example, many customers participate in more than one program, sometimes making it difficult to identify true participants and nonparticipants.

Fourth, there are limits to the commitment to and understanding of evaluation on the part of decision makers. Decision makers are not always certain of the value of evaluation, and therefore do not support it consistently. In addition, as evaluation issues and the approaches used to address those issues have become more complex, the abilities of decision makers to understand the issues and approaches has diminished.

Fifth, there are limits to the resources available for evaluation. It is not clear whether utilities, commissions, and the public are willing to bear the large costs associated with truly rigorous impact assessment.

Sixth, there are practical limits to imposing measurement requirements on program delivery systems and customers. Program forms, processes, and data collection procedures are designed primarily for delivering program services, not for evaluation. In addition, customers generally prefer minimal contact, and evaluation often increases customer contact.

The Future. While these limits exist, evaluators should not become paralyzed by them. There are methods available to address many evaluation questions adequately, and new methods (or new applications of existing methods) are being developed to help address some of the technical limits.

However, many of these limits will remain no matter how sophisticated the evaluation approaches become. In the near-term future, there is some risk of friction between evaluators and evaluation consumers because evaluators may have occasionally over-sold the current capabilities of the field. In order to avoid misunderstandings and disappointments in the future, both parties bear some responsibilities. Evaluators need to be forthright about what can and cannot be achieved, and they should identify and discuss the limitations of their research. Consumers of evaluation results need to be well informed about the strengths and limitations of evaluation approaches, and realistic expectations should be maintained. In addition, decision makers must be prepared to devote the necessary resources and time for evaluators to address their key issues and questions.

Dealing With Uncertainty. Utilities, regulators, and intervenors are putting strong pressure on evaluators to ensure that DSM savings estimates are accurate and reliable. Much of this pressure comes from advocates for supply-side options (including those within the utilities), and from ratepayer groups (especially large industrial intervenors) who are concerned about rate impacts and want to make sure that the savings from DSM are real, not imagined (Hughes 1991).

Evaluators and decision makers have tended to focus almost entirely on statistical uncertainty resulting from sampling error--but this is only one kind of uncertainty.[2] The other two main types of uncertainty in DSM evaluation are bias and unmeasured assumptions (e.g., discount rates, measure lifetimes, etc.). While the degree of statistical uncertainty due to sampling error can be quantified using standard statistical techniques, the uncertainty due to bias or unmeasured assumptions is not included in the estimate of statistical uncertainty. Appropriate experimental and quasi-experimental designs can be used to minimize bias. The uncertainty due to unmeasured variables can be explored using sensitivity analyses or modeled using probabalistic models.

In the last two years, much effort has been spent on increasing precision in impact evaluation. Sometimes this has resulted in a statistically precise estimate of program impacts that was biased because the sample was not representative of the population, or because the comparison group did not consist of a group of customers who were reasonably comparable to the treatment group. Evaluation efforts need to be balanced more appropriately between reducing statistical uncertainty (increasing precision) and minimizing bias. This is especially important now because DSM is increasingly seen as an important resource, and because many utility performance incentives are based on a single number representing savings or net benefits. If there is a large amount of bias in the estimate of savings, the incentive amount provided will be wrong. Even worse, supply side resources will be built (and not needed) or not built (and needed).

There has also been much debate about statistical uncertainty in the last few years. Some have argued that very precise results (i.e., +/- 10% precision at a 90% confidence level) are necessary, while at the other end of the spectrum others have stated that evaluation cannot produce findings that meet that level of statistical precision and that less precise results (i.e., +/- 20 or 30% precision at an 80% confidence level) are good enough.

In general, it would be helpful for evaluators to make the understanding of statistical uncertainty (confidence and precision) more intuitive and transparent. One way to achieve this is to use a high confidence level (e.g., 90%) consistently and use statistical precision (the confidence interval) to represent how uncertain the result is. The greater the statistical uncertainty, the wider the confidence interval. Changing confidence levels arbitrarily risks reducing trust and further complicating the issue.

Evaluators should plan on attaining a level of precision that is appropriate for the decisions that the findings will impact. Different questions need different levels of accuracy [e.g., are savings there at all, is the program cost-effective (a threshold issue), resource planning (how much savings), etc.]. In many cases this will be between +/- 20% and +/- 30% precision (at a 90% confidence level). For some programs +/- 30% precision will be the best that can be attained with current evaluation approaches and funding levels.

Clearly, in order for statistical precision to improve, evaluation techniques need to be developed that help explain variability due to other effects besides the program. Even with such improved techniques, +/- 15 or 20% precision may be the best that can be achieved without substantial increases in sample size (and costs).

The Future. Uncertainty is a normal part of utility regulation and of life in general--especially when we are trying to predict what will happen in the future. Evaluators need to be concerned with uncertainty (it should not be ignored), but not to the point that it paralyzes them or the decision makers. Since the many users of evaluation results have varying levels of expertise (especially when it comes to statistics), evaluators need to address uncertainty explicitly, consistently, and in a manner that builds trust and supports intuitive understanding.

Each of the three main types of uncertainty (statistical uncertainty, bias, and unmeasured assumptions) need to be considered when planning an evaluation. Any potential threats to validity need to be explored, and techniques should be used to minimize their impact on the evaluation. The effect of any substantial uncertainty remaining should be explored using sensitivity analyses and other techniques to document: (1) how the remaining uncertainty affects the results of the evaluation; and (2) the implications for future decisions (e.g., resource planning).

When reporting the findings of an evaluation, it would be helpful for evaluators to report on all three types of uncertainty, and discuss how each type of uncertainty contributes to the overall uncertainty of the result. Ranges should be used to represent the "real" uncertainty of the result rather than a single number (if feasible).

Note that while the debate about uncertainty has focused on the demand-side, there is substantial uncertainty on the supply side (e.g., future fuel prices, cost of capital, regulatory risk, environmental costs, very few operational combined cycle plants, etc.). The focus in the future needs to shift more to the overall uncertainty (including cost) of delivering energy services to customers, rather than solely on DSM savings versus generating plant output.

The Role of Process Evaluation. Process evaluations examine the operation of a program to see how well a program is being implemented, and they can provide (1) prompt feedback to help improve program design and implementation, and (2) important insights that can be used to interpret findings from impact evaluations. Process evaluations typically focus on program goals, history, and activities, and are often based on interviews with utility program staff, program managers, participants, and trade allies. The rewarding of financial incentives for utility investment in energy efficiency measures has elevated impact evaluation to a higher level of importance than before (which had already been considered by many to be more important than process evaluation). Unfortunately, process evaluation has been relegated to a second order of importance when, in fact, process evaluation can serve many important objectives which can complement (not just supplement) impact evaluation.

The goal of process evaluation is program optimization through: improvement in implementation efficiency, assessment of market segments and targeting of specific segments, improvement in quality of measure installation, identification of program-design issues, interim accounting of program progress through reviews of the program database, and examination of special issues (e.g., measure life and program comprehensiveness) (Bronfman and Peters 1991).

The Future. As DSM-program performance becomes more closely tied to integrated resource planning and incentive regulation, process evaluation will assume more importance (Bronfman and Peters 1991). First, because many programs are designed and implemented quickly, early process evaluations will be critical to fine-tune programs, assess the market potential, and identify market segments not reached. Second, because process evaluations are in the field early and because impact evaluations often do not produce results for a year or two, process evaluations will be instrumental in producing early reports on program effects and effectiveness: e.g., assistance in designing the program database, assessing the reliability of program data, and periodic reporting of implementation data (e.g., the number of installations, estimated and actual costs, and projected savings). And third, the results of process evaluations will need to be reported in the context of other evaluation findings, so that the results of different evaluation components will provide a complete, integrated evaluation.

Finally, there is an increasingly likely scenario whereby process evaluations, long relegated to second order status in the eyes of many, may ironically emerge as a savior of DSM. The past few years have seen a surge of hastily implemented DSM programs and associated regulatory incentive schemes, closely followed by mandated impact evaluations. As the results of these evaluations roll in, and the achieved savings (in all likelihood) fall short of expectations, there will be a time of great turmoil. In some quarters there may be a backlash against DSM. The preferred response, however, will be a resurgence of interest in process evaluation as a means to better understand and improve DSM program performance.

The Profession

The past decade and a half has been a time of phenomenal growth and change for the profession of energy program evaluation. As Eric Hirst (1989) colorfully observed, an energy evaluation conference in the mid 1970s could have been "held in a phone booth". Indeed, it could be argued that even a decade ago there was no "profession" to speak of, just a collection of a relatively few individuals that had happened to evaluate some particular programs.

Today there are hundreds, if not thousands of individuals employed at some facet of DSM program evaluation. At least two large professional associations have developed which include DSM evaluation, and there are three major national conferences devoted to DSM research and evaluation. The profession of DSM evaluation has virtually exploded onto the scene. At this point in the evolutionary process, it may be wise to pause and take stock of the current state, and future prospects, of this booming profession.

In an effort to shed some light on future directions, this paper will focus on three key issues: the growth and development of DSM evaluation as a "profession"; the increasing shortage of trained and experienced personnel; and the potentially worrisome trends in the institutional location and perspective of DSM evaluation professionals.

Development As A Profession

If the profession of DSM evaluation was born at the beginning of the 1980's, the current period could be described as its adolescence. This is a period of some tumult and confusion, with rapid growth and voracious appetite -- but some lack of coordination and maturity.

Noteworthy positive developments have included the recognition, at least, that it is time to begin to develop a coherent identity. Steps in that direction have included the formation of professional associations: e.g., the Association of Demand Side Management Professionals (ADSMP) and the Demand Side Management Society (DSMS) of the Association of Energy Engineers; as well as the establishment of several very successful ongoing professional conferences: e.g., the ACEEE Summer Studies, the International Energy Program Evaluation Conferences (the "Chicago" Conferences), and the National Demand-Side Management Conferences. The growth in the size of the profession is illustrated by the mushrooming attendance figures at these conferences.

Like most professions, DSM evaluation has also begun to grapple with issues of professional conduct, such as standards and ethics. As of yet, however, there is little resolution. There have been several flirtations with the concept of establishing specific standards for evaluation practice, but the consensus seems to be that while some general "guidelines" would be helpful, it is premature to think in terms of accounting-type methodological "standards" for DSM evaluation. The most recent effort in this area is currently being undertaken by a committee at ADSMP (Peters 1992). (See also the recent Chicago conference paper by Buller, Quigley, and Miller 1991, for some useful discussion on this issue.)

With the higher financial stakes involved in utility DSM programs (resulting from the increasing implementation of regulatory incentives), the issue of ethics is beginning to take more prominence. The concern usually tends to be that there will be pressure on evaluators (typically hired by the utility) to produce positive results (e.g., see Hartnett and Kelleher 1991). However, this potential for influence is by no means unidirectional. Despite the progressive pronouncements of a few innovative industry leaders, many utilities are implementing DSM under coercion of one type or another -- and some may see their long term interests as being served by a negative DSM evaluation result.

The Future. As more parties become involved in the oversight of DSM programs (e.g., through collaboratives) -- particularly parties with little technical knowledge or experience -- there will be an increasing demand for basic standards of acceptable evaluation methodology, including simpler and more understandable methods (e.g., the "transparent" methods discussed by Keating 1991).

Similarly, the growing financial stakes surrounding DSM evaluation will likely lead to a heightened concern regarding ethics and credibility (e.g., Hall 1992) -- especially if there is a major scandal (which is quite possible given the frenetic pace and somewhat haphazard manner in which incentives are being promulgated). Although many areas of utility operation have large financial stakes (e.g., load forecasting, fuel purchase contracts, etc.), DSM evaluation is newer; more visible; more controversial; tends to attract the attention of more parties; and is much more likely to be conducted by outside contractors (who have a tendency to be regarded by some as "hired guns").

In the boom market which DSM has experienced, the profession has had the luxury of a kind of free-for-all "frontier" mentality. That is likely to change, both due to demands from the users of evaluation as well as to the natural maturing process of the profession. That will tend to make evaluation work more stable and controlled -- but probably less fun.

Personnel Shortage

Anyone working in this area is likely acutely aware of the shortage of trained and experienced people in the DSM evaluation field. While this is good news for those selling their talents, it does create some major problems. These range from the inability to find personnel for a desired evaluation project, to the seemingly constant threat of losing your own good staff to "greener pastures".

Recent research (Hall and Skelton 1991) indicates high likelihood for continuing robust growth in utility DSM activities. That, coupled with increasing governmental attention to environmental issues related to energy, suggests that one cannot expect much relief from the "talent gap" anytime soon.

The Future. On the positive side, efforts are underway (through the auspices of ADSMP) to operate a "training institute" for DSM professionals. That should help. Some additional creative longer-term solutions might include mechanisms such as the establishment of student groups at universities, the development of academic curricula, and the use of sponsored internships.

However, in the meantime, there are some ominous signs on the horizon. Serious questions are being raised about whether the evaluation infrastructure can keep pace with the skyrocketing demands for its services (Schlegel, Prahl, and Kushler 1991). Many of the leading consulting firms, having very successfully marketed their services, are becoming stretched to the breaking point in terms of being able to deliver the evaluations already planned. Down-side risks from this range from delays and/or quality dilution to an epidemic of coronary attacks (e.g., note the career stress issue addressed in Dethman 1992). Longer-term effects on the profession include pressure toward routinization of evaluation functions so that they are (a) faster and (b) able to be performed by staff with less training. In sum, this is a problem with no quick solution in sight.

Institutional Perspective

A final issue for discussion, which intersects with each of the preceding two issues, is the striking extent of migration of energy evaluation professionals from the public to the private sector that has occurred over the past decade. This is not surprising given the convergence of three factors: (1) the dramatic reductions in the federal government role in energy conservation during the Reagan administration; (2) the rise of Demand-Side Management in the utility sector; and (3) the more recent budgetary woes of the states.

All of these factors have resulted in a major shift in the labor market (resulting from a natural tendency to "go where the jobs are"). The extent of this transformation is illustrated by the analysis of attendees at two of the major energy evaluation conferences during this time period (see Table 5).

The shift is most visible in the data from the Chicago conference, where the combined categories of state/local government, regulatory and federal (non-lab) attendees went from 47% in 1984 to 14% in 1990; while the combined categories of utility sector and private consultants went from 25% to 67% over the same time period. The trend was similar, although less pronounced, at the ACEEE Summer Studies.

While being careful not to cast aspersions upon anyone's career choices, there are some important reasons why the collective profession of DSM evaluation ought to be concerned.

First is the issue of credibility. Evaluation, of all fields, has a premium value on maintaining an image of objectivity (i.e., being unbiased, independent and credible). To oversimplify: if all or most of the top professionals in the field work for utilities and/or consulting firms that do most of their work for utilities, there is a serious risk of a perceived credibility problem among the other interested parties.

Second is the issue of perspective. Talented people focusing through the lens of a particular interest perspective (i.e., the utility sector), however conscientious and honorable, are going to attend to certain concerns and needs and neglect others (e.g., societal/governmental/academic objectives). Again, simply put, there are many legitimate needs (e.g., regulation, R&D, codes and standards, government programs) that suffer if they cannot attract and retain the talented professionals necessary to serve those needs. In the near term, due to the pace of utility DSM activity, this shortage is particularly acute in the regulatory sector.

Third is the issue of the ability of the individuals involved to function collectively as a true profession. For example, the major energy evaluation conferences (e.g., Chicago, ACEEE) originally tended to be fairly academic in style. Over time, the orientation has transformed to more of a profession, and now is edging toward a business trade show. This has implications for things like collegiality and information sharing. It's one thing to critique a colleague's paper in an atmosphere of academic and scientific interchange. It's another thing when audience and presenter are direct competitors, and impressions created about a project might have direct economic implications in the future. Similarly, some care is required to keep

Table 5. Institutional Affiliation of Attendees at Two Major Energy Conferences

	ACEEE		Chicago Evaluation	
	1984 (n=290)	1990 (n=599)	1984 (n=215)	1991 (n=429)
Advocate Group	4%	5%	2%	3%
Regulatory	4%	6%	7%	3%
Other State/Local	10%	7%	29%	6%
Federal (non-Lab)	7%	5%	11%	4%
National Labs	17%	12%	5%	4%
Academia	13%	8%	13%	4%
BPA, WAPA, TVA, etc.	7%	5%	1%	4%
International	9%	11%	1%	3%
Utilities	11%	17%	18%	43%
Consultants	5%	17%	7%	24%
Other Misc.	12%	7%	5%	2%

Source: Lists of participants distributed by the respective conferences.

professional papers and presentations from becoming marketing vehicles for particular firms, individuals and/or techniques.

The Future. Until government priorities and budgets turn around, it is difficult to see when the trend toward "privatization" of the energy evaluation function will change. Given that circumstance, the profession will need to explore creative mechanisms for sustaining its credibility (e.g., perhaps a demonstrably independent panel of experts available to resolve disputes and arbitrate between parties in contested cases). Similarly, great vigilance will be required to avoid ethical abuses. (If self-policing proves ineffective, some type of a more formal mechanism may evolve.) Finally, it is hoped that forums like the professional associations, national conferences, and professional journals can be periodic "neutral zones" where collegiality and scientific interchange can continue to prevail over more narrow business interests.

Conclusion

This is a time of great expansion and rapid change in the field of energy evaluation/performance measurement and associated behavioral research. Given that the importance of energy efficiency continues to grow--both through expanding utility DSM activities as well as through a reinvigorated governmental interest due to issues such as environmental concerns--the field should continue to see robust development.

All this activity is not without its problem areas, of course. As this paper has attempted to describe, there are a number of emerging concerns, ranging from the need to make further progress on specific methodological issues (e.g., persistence, net-to-gross savings estimation, etc.) to the necessity for maturation (and some introspection) as a profession.

On the whole, however, these are exciting and challenging times. Fortunately we can take comfort in that old principle of Socrates (or was it Yogi Berra?): "it's better to be too busy than bored."

Acknowledgements

The authors would like to thank Jeff Haberl, Liz Hicks, Bonnie Morrison, Mary Ann Piette, and Rob Pratt for their comments and suggestions, which were very helpful in the writing of this document. Also, special thanks to Melanie Lord for all her help in the typing and preparation of this paper.

Endnotes

1. We use the term "gross savings" rather than "total savings" proposed by Hirst (1991) to describe the change in consumption and demand for participants. We believe gross savings is clearer and that total savings can be misleading and become confused with program savings. Also note that net savings are not always a portion of gross savings as Figure 1 suggests; net savings can exceed gross savings when the general population is increasing consumption or demand.

2. Measurement error is sometimes included in estimates of statistical uncertainty.

References

Claxton, J., C. Anderson, J. Ritchie, and G. McDougall. 1981. "Diffusion of Energy Conservation and Technologies." *Consumers and Energy Conservation.* D. Leonard-Barton, ed. p. 97-107.

Braun, B., and S. Williams. 1979. "Changing Household Energy Uses." *Changing Energy Use Futures*, Volume 2, Pergonan Press, p. 857-864.

Bronfman, B., and J. Peters. 1991. "Process Evaluation of DSM Programs." *Handbook of Evaluation of Utility DSM Programs*, E. Hirst and J. Reed, ed., p. 135-148. Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Buller, S., D. Quigley, and W. Miller. 1991. "How Can We Assure Quality in DSM Program Evaluation?" *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois, p. 456-460.

Conservation Law Foundation. 1992. *CLF 1991 Annual Report.* Conservation Law Foundation, Inc., Boston, Massachusetts.

Cohen, S., C. Goldman, and J. Harris. 1991. *Measured Energy Savings and Economics of Retrofitting Existing Single-Family Homes: An Update of the BECA-B Database.* Lawrence Berkeley Laboratory, Berkeley, California.

Dethman, L. "Refurbishing the Evaluation Toolkit." 1992. *Evaluation Exchange*, 2(1):12-14. Synergic Resources Corporation, Bala Cynwyd, Pennsylvania.

Eto, J., H. Akbari, R. Pratt, and S. Braithwait. 1990. "End-Use Load Shape Data Application, Estimation, and Collection." *Proceedings from the ACEEE 1990 Summer*

Study on Energy Efficiency in Buildings. American Council for an Energy-Efficient Economy, Washington, D.C.

Fels, M. 1986. "PRISM: An Introduction." *Energy and Buildings*. 9(1).

Gladhart, P., and J. Weihl. 1990. "The Effects of Low-Income Weatherization on Internal Temperature, Occupant Comfort, and Household Management Behavior." *Proceedings from the ACEEE 1990 Summer Study on Energy Efficiency in Buildings*, Volume 2, p. 2.43-2.52. American Council for an Energy-Efficient Economy, Washington, D.C.

Gordon, F., M. McRae, M. Rufo, and D. Baylon. 1988. "Use of Commercial Energy Efficiency Measure Service Life Estimates in Program and Resource Planning." *Proceedings of the ACEEE 1988 Summer Study on Energy Efficiency in Buildings*, Volume 3, p. 3.84-3.97. American Council for an Energy-Efficient Economy, Washington, D.C.

Haeri, H. 1991. *Measurement and Verification Plans*. Central Maine Power, Augusta, Maine.

Hall, N. 1992. "Manipulative Evaluation: Producing Results to Achieve Our Organizational Goals." *Evaluation Exchange*, 2(2):8-10. Synergic Resources Corporation, Bala Cynwyd, Pennsylvania.

Hall, N., and L. Skelton. 1991. "Projected Growth of Demand-Side Management Programs, 1991-1995: Results of a Program Growth Survey of Electric Utilities." *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois, p. 132-136.

Hartnett, J., and M. Kelleher. 1991. "The Need for Objective Program Evaluation for Long-Term Demand-Side Management Success." *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois, p. 450-455.

Hastie, S. 1992. "Commission Briefs." *Evaluation Exchange*, 2(1):16. Synergic Resources Corporation, Bala Cynwyd, Pennsylvania.

Hickman, C., and T. Steele. 1991. "Building Site Visits as Supplement to Program Evaluation." *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois, p. 174-180.

Hirst, E. 1989. "Evaluating Energy-Efficiency Programs: A Personal View of Progress and Potential." *Proceedings of the 1989 International Energy Program Evaluation Conference*, Chicago, Illinois, p. 511-523.

Hirst, E., and J. Reed. 1991. *Handbook of Evaluation of Utility DSM Programs*, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Hughes, J. 1991. "What Happens When Utility Financial Incentives Depend on Evaluation Results?" Panel presentation in *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois.

Keating, K. 1991. "An Individual Perspective on Where We've Been, Where We Are, and Where We're Going." Keynote Address, *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois.

Keating, K. 1991. "Persistence of Energy Savings." *Handbook of Evaluation of Utility DSM Programs*, E. Hirst and J. Reed, ed., p. 89-99. Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Kelly-Cochrane, D., M. Kelsey, J. Chamberlin, M. Reid, and J. Brown. 1991. *Utility DSM Shareholder Incentive Study*. Association of Demand-Side Management Professionals, Berkeley, California.

Lutzenhiser, L. 1990. "Explaining Consumption: The Promises and Limitations of Energy and Behavior Research." *Proceedings from the ACEEE 1990 Summer Study on Energy Efficiency in Buildings*, Volume 2, p. 2.101-2.110. American Council for an Energy-Efficient Economy, Washington, D.C.

Misuriello, H., and M. E. Hopkins. 1992. *Research Opportunities to Improve DSM Impact Estimates*. California Institute for Energy Efficiency, Berkeley, California.

Nadel, S., and K. Keating. 1991. "Engineering Estimates vs. Impact Evaluation Results: How Do They Compare and Why?" *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois, p. 24-33.

Northeast Utilities. 1991. *Integrating Information for DSM Evaluation Using the Engineering Calibration Approach*. RLW Analytics, Inc. Cambridge, Massachusetts.

Parti, C., and M. Parti. 1980. "The Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector." *Bell Journal of Econometrics*, Spring 1980, p. 309-321.

Parti, M., E. Rogers, and A. Sebald. 1991. "Conditional Demand Analysis of Commercial and Industrial Customers to Determine the Effects of Demand-side Program." *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois, p. 323-327.

Peters, J. 1992. "Topic Committee Reports: Guidelines For Evaluation Practice." *Strategies*, 3(2):7. Association of Demand-Side Management Professionals.

Petersen, F. 1990. "Remodel and Tenancy Changes: Threats to the Reliability of Commercial Conservation Savings." *Proceedings of the ACEEE 1990 Summer Study on Energy Efficiency in Buildings*, Volume 3, p. 3.165-3.172. American Council for an Energy-Efficient Economy, Washington, D.C.

Schlegel, J., R. Prahl, and M. Kushler. 1991. "Measurement in the Age of Incentives." *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, Illinois, p. 182-190.

Sharma, A. 1983. "Energy Conservation in U.S. Buildings: Solving the High Potential/Low Adoption Paradox From a Behavioral Perspective." *Energy Policy*, 11(2):148-167.

Skumatz, L., K. Lorberau, R. Mole, R. Bordner, and R. Chandler. 1991. *Bonneville Measure Life Study: Effect of Commercial Building Changes on Energy Using Equipment*. Synergic Resources Corporation, Seattle, Washington.

SRC. 1992a. *Dayton P&L Say Yes Program*. Synergic Resources Corporation, Seattle, Washington.

SRC. 1992b. *Effective Measure Life and Other Persistence Issues in DSM Programs: Literature Review*. Synergic Resources Corporation, Oakland, California.

Vine, E. 1992. "DEEP: Database on Energy Efficiency Programs." *Proceedings from the ACEEE 1992 Summer Study on Energy Efficiency in Buildings*, forthcoming.

Vine, E. 1992. *Persistence of Energy Savings: What Do We Know and How Can It Be Ensured?* Lawrence Berkeley Laboratory, Berkeley, California.

Violette, D. 1991. "Analyzing Data." *Handbook of Evaluation of Utility DSM Programs*, E. Hirst and J. Reed, ed., p. 51-72. Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Violette, D., M. Ozog, M. Keneipp, F. Stern, and P. Hanser. 1991. *Impact Evaluation of Demand-Side Management Programs, Volume 1: A Guide to Current Practice, Volume 2: Case Studies and Applications*. Electric Power Research Institute, Palo Alto, California.

Weihl, J., and P. Gladhart. 1990. "Occupant Behavior and Successful Energy Conservation: Findings and Implications of Behavioral Monitoring." *Proceedings of ACEEE 1990 Summer Study on Energy Efficiency in Buildings*, Volume. 2, p. 2.171-2.180. American Council for an Energy-Efficient Economy, Washington, D.C.