9. Evaluating the Impact of Labeling and Standards-Setting Programs

Guidebook Prescriptions for Evaluating the Impact of Labels and Standards

1. To ensure efficient program design and data collection, begin the evaluation process as soon as you decide to establish a labeling or standards-setting program.
2. Before conducting the evaluation, make sure all the key stakeholders understand the objectives of the evaluation and the resources that are available and necessary for conducting the evaluation.
3. To minimize costs, try to leverage existing sources so that data-collection efforts can focus on primary data. Allocate some of the evaluation budget to up-front costs.
4. Establish a national appliance database, and develop a baseline ("market characterization") representing the appliances that are currently being promoted on the market.
5. At regular intervals, evaluate both the program implementation process and the program impact on energy consumption, emissions, energy bills, and the appliance market.
6. Use a diverse group of data-collection methods rather than relying on just one method.
7. Evaluate the impacts on all key stakeholders, including consumers, manufacturers, retailers, and policy makers.
8. Focus on how the evaluation findings will be used in: a) refining appliance labels and standards, b) improving the implementation of the labeling and standards program, c) supporting other energy programs and policies, d) forecasting energy use, e) conducting strategic planning, and f) carrying out regulatory proceedings.

9.1 Why Evaluation Is a Must and Not a Luxury

Traditionally, energy-efficiency programs have received only a fraction of the attention and resources allocated to the energy supply side. There are many possible causes for this, but perhaps the most important is the relative invisibility of energy-efficiency impacts compared to the easily observable impact of adding new energy supply.

9.1.1 Making the Case

Unfortunately, a major barrier to the implementation and expansion of energy standards and labeling programs in many developing countries is policy makers' lack of confidence in the effectiveness of

labeling and standards. This lack of confidence can in large part be addressed by the presentation of clear evaluation results. Evaluations are needed to "prove" program impacts. As standards and labeling programs are increasingly implemented in developing countries, evaluation is expected to play a critical role in enhancing these programs' effectiveness and in convincing policy makers to adopt these measures through a gradual chain reaction.

If energy-efficiency policies and programs are to take their proper place, their benefits need to be clear, measurable, verifiable, and transparent. Quantifiable benefits are especially important for justifying that adequate funding and resources be allocated to a program. Many labeling and standards programs in developing countries have received seed money from donor agencies; however, this outside support cannot be expected to continue indefinitely and does not form a basis for sustainable program planning. Therefore, over the long term, a case needs to be made for support of standards and labeling programs by national sources.

Properly carried out, program evaluations quantify impacts and benefits in concrete terms, which can be the main evidence of the need to support the programs. Measuring impacts can justify allocation of resources to the program and demonstrate the need for funding that is sufficient to make the program effective. Policy makers will find evaluation results useful during internal discussions about governmental resource allocation in which they may be asked to prove that a program is generating sufficient savings. An evaluation can be designed with almost any level of resources to meet prioritized needs of time, cost, or accuracy.

9.1.2 Assessing the Program

In addition to justifying program funding, evaluations serve a second, equally important function: they assess the efficiency and effectiveness of the program process, revealing potential weaknesses in program implementation so these problems can be corrected. In the long run, this helps guarantee and enhance the program impacts. For example, an early evaluation of the European Union's appliance labeling program showed that the label was not being applied correctly by a large number of retailers, which allowed corrective action to be taken.

9.1.3 State of the Art

Unfortunately, there has been very little post-implementation evaluation of appliance standards and labeling programs although this situation is beginning to change. In the U.S., most impact assessments of efficiency standards have taken place just prior to adoption of new efficiency standards, based on forecasted information about product shipments and customer use (Nadel 1997). These evaluations rarely use field measurements, nor do they attempt to systematically examine what would have happened if standards had not been adopted (Meier 1997; Nadel 1997).

One evaluation of the U.S. federal energy-efficiency standards for residential appliances used a spreadsheet accounting method that tracked shipments of a given product in each year (along with average

annual energy use or energy efficiency of a given product sold in each year), created a base-case scenario that assumed no standards were or will be implemented, and then compared various scenarios with standards to the base case (Meyers et al. 2002).

Many past evaluations of appliance-labeling programs have focused on consumer awareness of the label but have not explicitly linked the label to actual behavior (i.e., to the efficiency of the appliances purchased and to the most likely purchase if there had been no label). However, some evaluations of appliance-labeling programs do include data on actual sales and behavior. Examples include evaluations of the European labeling program (Beslay 1999; Schiellerup and Winward 1999; Waide 1997, 1998; Winward et al. 1998) and the labeling programs in Australia (Harrington and Wilkenfeld 1997), Denmark (Karbo et al. 2002), Thailand (Agra Monenco, Inc. 2000a, 2000b), the state of Vermont (Rosenberg 2003), and the U.S. (du Pont 1998a 1998b; Thorne and Egan 2002).

Whether estimated or measured, the impacts of standards and labeling programs have been dramatic. For example, the E.U. energy label has been an undeniable success in terms of its market transformation impact. As described in insert: *Comprehensive Evaluation of the E.U.'s Labeling Program*, market evaluations have shown a clear and strong evolution of the market toward higher efficiency products since the introduction of the E.U. label. Much of the credit for the label's success must be attributed to its design.

A series of evaluations has also shown that the Thai energy label has been effective (du Pont 1998a 1998b; Agra Monenco, Inc. 2000a, 2000b; Singh and Mulholland 2000.) Insert: *Evaluation of the Thai Labeling Program Using Manufacturer and Consumer Surveys* on pages 234–236 describes an evaluation of the Thai labeling program for refrigerators and air conditioners, which helped to solidify support among Thai policy makers for continuation and expansion of the program. In addition, the evaluation gave credibility to the program results, and Thailand is now known as a regional leader in energy labeling and is an example for policy makers designing programs in Southeast and South Asia.

Future evaluations of labeling and standards-setting programs are likely to be more comprehensive than has been the case so far because labeling and standards programs are designed to be market-transformation strategies (e.g., see Barbagallo and Ledyard 1998; Hagler Bailly 1996, 1998; HBRS 1995; Hewitt et al. 1998; Pacific Energy Associates 1998; Vine et al. 2003; Xenergy 1998).

9.1.4 Planning

Because appliance and equipment efficiency levels are incredibly dynamic and can change very quickly, evaluations are essential for planning the subsequent program steps. For example, categorical energy labels (which are the dominant form of label internationally) require regular evaluations of market impact to determine whether the top efficiency classes are becoming saturated. If this is found to be the case, then the label can be judged a success; however, it most likely also means that it is time to reclassify the efficiency grades upwards so the label can continue to have an impact on the market.

Comprehensive Evaluation of the E.U.'s Labeling Program

The E.U. introduced framework legislation for mandatory energy labeling in 1992 and has since issued product-specific energy-labeling directives for refrigerators and freezers, clothes washers, clothes dryers, combined clothes washers and dryers, dishwashers, household lamps, ovens, and room air conditioners.

Evaluation of the labeling scheme has monitored retailer, distributor, and manufacturer compliance with the legislation and assessed impacts on energy use, energy efficiency, CO_2 emissions, and cost trends. Because the energy label for refrigeration appliances (refrigerators, freezers, and their combinations) was the first to be introduced, this category has received the most attention to date. Two years after the implementation of the labeling program for refrigerators, the European Commission launched a study to assess legislative compliance and program implementation issues and a set of successive studies to assess quantitative sales-weighted energy efficiency, energy, and emissions trends. The implementation/compliance study involved the following steps:

- surveys of representatives to the European Commission's Energy Labeling Committee, 10 retail outlets in each member state, 16 mail-order catalogs in eight member states, and numerous customers, to assess compliance, learn about consumer attitudes and responses, and discover any legal and governmental issues that may have arisen in each country
- independent tests in consumer association laboratories across the E.U. to evaluate the accuracy of manufacturer product-performance declarations
- interviews with manufacturers and retailers to assess their attitudes and responses and discover any concerns that may have arisen

The successive quantitative studies evaluated the sales-weighted efficiency trends of refrigeration appliances, clothes washers, washer-dryers, and household lamps sold in the E.U. up to 1998 and compared these trends to the pre-labeling levels (e.g., circa 1992 for refrigerators). Although these studies examined the impact of labels, several interlocking policies, of which labeling was one, were in effect during this period, including pending minimum energy performance standards (MEPS) and/or voluntary agreements (depending on the appliance) as well as various national and regional incentive programs. Yearly data on the sales volume and average retail prices of individual appliances were purchased on a country-by-country basis from established market research agencies. These data were then matched to separate databases containing model-by-model information on the technical characteristics of the appliances, including all aspects needed to evaluate energy consumption and efficiency. The quantitative assessment found that the sales-weighted efficiency of refrigeration appliances improved by 17.6% from 1992 to 1998. Furthermore, this detailed evaluation provided clear evidence of the distinct impact of the energy label as opposed to other E.U. policy measures such as MEPS. The data on the distribution of refrigerator sales by energy efficiency index (Figure 5-3 in Chapter 5) demonstrate that the categorical label design has not only stimulated consumer demand for higher-efficiency products but has also moved manufacturers to develop products targeting specific higher efficiency thresholds both in advance of (i.e., in anticipation of) and in response to heightened consumer demand (Waide 1998, 1999). This demonstrates the clear value of using a categorical efficiency scale with higher efficiency thresholds that challenge manufacturers to develop more efficient products.

This kind of detailed evaluation based on matching between technical and sales databases

has not been repeated since; however, market research companies have continued to collate data on sales by label class; from these data, it is possible to make a less refined evaluation of impacts than was made in the original evaluation. Average energy efficiency is estimated to have improved by 37% for refrigerators and freezers (Figure 2-5 in Chapter 2 shows the shift to higher efficiency label categories (Waide 2004; GfK 2003)), 21% for clothes-washers, and 35% for dishwashers since the introduction of labeling, at average rates of 4.0%, 3.7% or 6.5% per annum, respectively.

The use of a common efficiency scale and format for all labeled products is also reported to have aided comprehension and "brand" recognition levels, the latter of which are said to be very high. Regrettably, data are not available on the impact of the decision to add the A+ and A++ classes for the refrigerator label, but the small amount of information available suggests that consumers would have found a regrading of the existing A to G scale easier to comprehend.

Compared to a static-efficiency base-case scenario (assuming average efficiency frozen after 1999 at late-1999 levels), it has been estimated that the improvements in refrigeration appliance efficiency for the 25-year period ending in 2020 will be: 398 TWh of energy savings, 56 billion of avoided electricity bills , and 237 megatonnes of avoided CO_2 emissions. These figures are based on the assumption that declared energy consumption equals actual consumption, which is supported by some regionally specific end-use metering studies. The accuracy of the consumption numbers for individual models has sometimes been questioned based on concern that the results of these studies may not be applicable to the entire E.U. The uncertainty results primarily from assuming that energy consumption under standard test conditions is representative of energy consumption in consumers' homes. Also, the frozen-efficiency base-case scenario overstates the savings because efficiency would have increased to some degree in the absence of labels as a result of MEPS, other programs, and uninduced technological innovation.

The compliance/implementation assessment found that implementation of the legislation varied considerably among member states. Both Germany and Italy implemented the legislation within their borders only in 1998 and 1999, respectively, after receiving formal warnings from the European Commission. Retailer compliance was low, with only an average of 56% of refrigeration appliances on display across the E.U. in the summer of 1997 being correctly labeled and considerable variation among member states. There have been some ad hoc follow-up surveys since, and the level of compliance has generally been found to be much higher although some problems still remain. The implementation/compliance study also compared manufacturers' self-declared performance levels with those recorded by independent testing agencies such as those operated by consumer associations. A wide divergence was found because of deviations from the testing protocol by both manufacturers and independent agencies, with efficiency levels declared by consumer associations and manufacturers differing by up to four labeling classes with an average of one class. Since the 1997 analysis and following repeated efforts by member states, the commission, and industry associations to improve the accuracy of the self-declared values, the degree of discrepancy between manufacturer and third-party product energy-performance declarations is said to have diminished although there has been no comprehensive survey to assess the situation. The compliance/implementation assessment also found that the stated impact of the label on consumer purchasing patterns was substantial, from 4% (Greece) to 56% (Denmark), and was strongly related to the level of compliance.

Sources: Boardman 1997; Winward et al. 1998; Waide 1998, 2004; GfK 2003.

A common fault in program design is to postpone working on the evaluation until some years after the program has been implemented, which makes it impossible to confirm the state of the market before the program was implemented. A pre-program market assessment establishes the reference baseline efficiency trend against which impacts can be assessed. To allow for a pre-program assessment, it is essential to begin planning for the evaluation process when the labeling and standards program is being initiated. Early planning allows for effective design of the evaluation program, efficient collection of data, and adequate opportunity to make key stakeholders aware of the importance of the evaluation so that they will likely feel receptive to its findings. As noted below in Step 1, many of the data needed for evaluations are actually an integral part of effective program implementation. This chapter describes the types of activities involved in the evaluation of labeling and standards programs and provides a few examples of how labeling programs have been evaluated. (See insert: *Evaluation of the Thai Labeling Program Using Manufacturer and Customer Surveys.*)

It is important for policy makers to distinguish among different techniques for assessing program effectiveness. Three of the major approaches are described briefly below:

- Process evaluations examine all aspects of the mechanics and operation of a program, including applications, procedures, dissemination, awareness, etc. Process evaluations are usually primarily qualitative in nature, but they also have quantitative elements.
- Impact evaluations address the magnitude and timing of a program impact, such as equipment sales, electricity saved, and amount of pollution reduction attributable to the program. A comprehensive program evaluation will usually entail both a process and impact evaluation.
- Program monitoring is a technique for regularly assessing progress of activities and results against project targets. This useful tool is not discussed in this chapter but is briefly summarized in the insert: *Program Evaluation Differs from Program Monitoring* on page 237.

Figure 9-1 shows the four steps necessary for evaluating labeling and standards-setting programs.

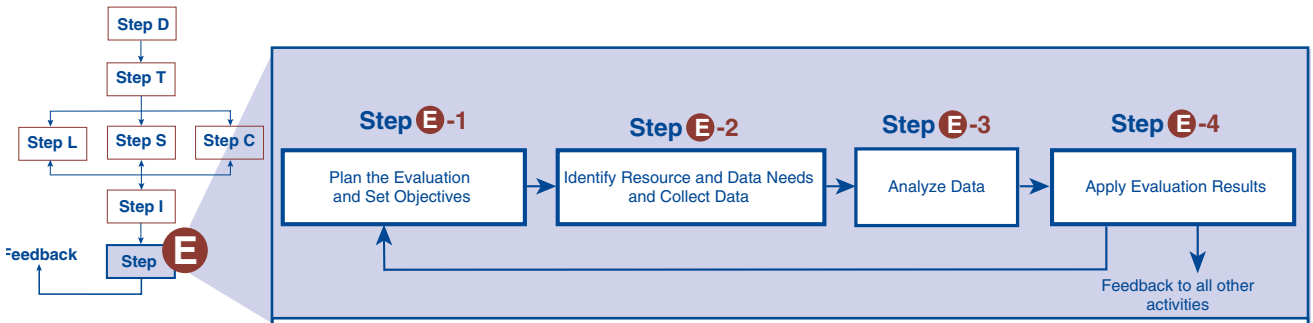


Figure 9-1 Major steps in evaluating a labeling or standards-setting program

Evaluation of the Thai Labeling Program Using Manufacturer and Consumer Surveys

In early 1994, the Electricity Generating Authority of Thailand (EGAT) approached the five Thai manufacturers of household refrigerators and quickly gained their cooperation for a voluntary energy-labeling program. The efficiency scale on the label ranges from 1 to 5, with 3 as the average and 5 as the most efficient. A selection of the models in this size range was tested during fall 1994 to establish the average efficiency level. Models that fell within 10% of the mean were rated at 3; models that were 10 to 25% more efficient than the mean were rated at 4; and models that were more than 25% more efficient than the mean were rated at 5.

A similar labeling program for air conditioners began in early 1996. Negotiations with air-conditioner manufacturers were more difficult than those with refrigeration manufacturers because of the diverse and fragmented nature of the Thai air-conditioner industry, which consists of 200 manufacturers, many of which are small, local assembly operations. Most Thai air conditioners are produced by the 15 largest firms. Unlike in the refrigerator market where efficiency levels were relatively similar among manufacturers, the Thai air-conditioner market has a trimodal distribution: low-cost, low-efficiency, locally produced models; higher cost, moderate-efficiency, locally produced models; and high-end, high-efficiency models dominated by imports. The air-conditioner manufacturers chose to place energy labels only on the most efficient units, those with a rating of 5. Thus, consumers were faced with a choice between buying a unit with a label (i.e., a rating of 5) or a unit with no label (i.e., an invisible rating of 4, 3, or worse).

In 1999, the Thai demand-side management (DSM) office commissioned a comprehensive evaluation of its energy-labeling programs. The evaluation had three major components:

- a process evaluation, to gather qualitative data about the behavior and attitudes of consumers and manufacturers and their reactions to the program
- a market evaluation, to assess the impact of the program on manufacturer decisions and market penetrations
- an impact evaluation, to assess the program's effect in terms of energy and demand savings

The study was carried out using two primary data collection techniques:

- a manufacturer survey, which entailed development of a detailed survey questionnaire that was administered through in-person interviews with marketing and production personnel at 50 manufacturing and distribution firms
- a detailed, five-page residential survey that was administered by a team of 18 surveyors to 2,000 households in Bangkok and in three upcountry cities in Thailand

The evaluation found a high level of awareness of the label among Thai consumers. Non-participants (consumers who purchased a refrigerator or an air conditioner without a label) indicated that they did not buy a labeled refrigerator for the following reasons:

- they were not aware of energy-efficient refrigerators
- labeled units were not available where they purchased the unit
- the salesperson recommended a non-labeled unit

The evaluation yielded the following findings specific to the air-conditioner program:

- participants tended to have higher incomes than non-participants
- testing and labeling had a high degree of credibility among consumers
- the zero-interest loan program offered by EGAT for air conditioners had a very low participation rate because of lack of

Continued on next page

Evaluation of the Thai Labeling Program Using Manufacturer and Consumer Surveys (continued)

support by retailers and the perception that the process was complicated and involved intensive paperwork.

The manufacturers of both refrigerators and air conditioners reported that they were highly satisfied with the program. For air conditioners, however, the retailers were not satisfied; only 29% of the Green Shops (stores that participated in EGAT's no-interest loan offer for models rated 4 and 5) surveyed felt that the marketing campaign by EGAT was adequate. A number of the manufacturers suggested that the program could be improved by improving the speed and accuracy of the testing process. They also recommended that EGAT consider targeting promotional and educational campaigns at increasing the interest and ability of salespeople to market the higher-efficiency models.

The impact evaluation was based on direct metering of air conditioners and refrigerators in several hundred homes. The metered savings were combined with data from the surveys of residential households and manufacturers and with program data on the size and efficiency of models, to estimate the energy and demand savings attributable to the program. The table below summarizes the savings for the Thai energy-labeling programs.

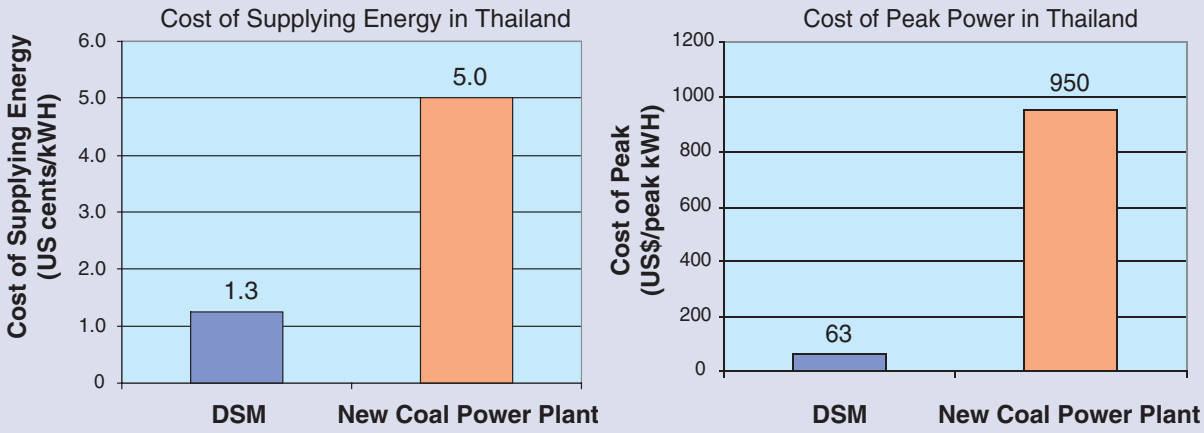
Since this evaluation, additional data have been collected and analyzed through 2000. Based on EGAT's final evaluation of the results of the DSM programs (including a thin-tube fluorescent lamp program in addition to the refrigerator and air-conditioner labeling program). The following figures show the favorable cost of saved energy and cost of avoided peak, which has increased the confidence of Thai policy makers in the program's benefits.

Source: Agra Monenco, Inc. 2000a, 2000b; Phumaraphand 2001.

Summary of Evaluated Savings from Thailand's Energy-Labeling Programs

	Number of labels	Energy Saving (GWh/yr)	Demand Savings (MW)		Benefit-Cost Ratio		
			Ave.	At peak	Customer Resource Cost	Utility Resource Cost	Total Resource Costs*
Refrigerators	3,698,117	235	80	14.0	2.2	9.8	2.8
Air conditioners	395,488,171	173	176	17.8	1.4	5.2	0.67

** The Total Resource Cost (TRC) is lower than anticipated because few residential air conditioners are running during the new afternoon system peak (14:00 - 17:00 hours), and because all differences in the price of efficient and standard units were assumed to be due to differences in the energy efficiency of the unit.*



Some of these steps are interactive and, as noted above, the conceptualization of them should be incorporated into an evaluation research plan early in the process of designing and implementing programs.

The remainder of this chapter discusses the above evaluation steps in detail.

9.2 Step E-1: Plan the Evaluation and Set Objectives

Evaluation should be thought of as an integrated part of the overall data collection and management process of a standards and labeling program. Evaluations are built on data. The data used for evaluations—e.g., market size and shares, trends, drivers, breakdowns—are also an integrated part of the overall program design and implementation When planning an evaluation, policy makers should realize that the data collected will become part of an overall data set related to the program and thus part of an overall data-collection and management effort.

9.2.1 Evaluating Labeling vs. Evaluating Standards Programs

An impact evaluation of labels and standards will examine efficiency and capacity improvements, macro energy and environmental impacts, the range of models and features on the market, the costs and benefits to different groups (e.g., consumers, industry, retailers, society), and manufacturer competition. For both labeling and standards-setting programs, it is important to evaluate the program's process as well as its energy and economic impacts. For appliance standards, an evaluation should focus on manufacturers' decisions and changes in the efficiency of models sold in the marketplace and on the effectiveness of

Program Evaluation Differs from Program Monitoring

The terminology for monitoring and for evaluation is similar. The main difference is that monitoring is a part of the project implementation cycle, during which the project activities and results are measured against benchmarks or objectives set out in a logical framework approach (commonly called a logframe, or LFA, see Saldanha and Whittle 1998) or in a project agreement. Evaluations are carried out at a discrete point in time, usually upon completion or midway through a project (sometimes called a mid-term review) whereas monitoring is an ongoing process that should be carried out at regular intervals throughout the project. The results of monitoring are often used as input to the evaluation process. The frequency of monitoring may vary from project to project, but the aim should be to monitor at least annually.

Monitoring identifies day-to-day problems during implementation of the program and examines whether the past and planned activities will realistically achieve the planned results. Monitoring's main purpose is to track activities, identify weaknesses, and serve as an "early warning system" that allows for timely intervention if a project is not functioning well. Increasingly, international development assistance agencies are using monitoring to ensure the efficiency, effectiveness, and sustainability of their projects. Monitoring can also be used specifically within the energy sector as an essential tool for project management and quality control.

Source: Danish Energy Management A/S 2000

compliance procedures. The evaluation of a labeling program should include all of the above but should also assess the sales and purchase process to determine the impact of labeling on retailer and consumer decisions. An evaluation of a labeling program involves both quantitative and qualitative research to understand the process of consumer decision making and the actions of multiple stakeholders involved in the manufacture, sale, and distribution of appliances. Finally, labeling programs affect behavior over a longer period and their impacts are often more subtle than the impacts of standards because standards take effect in a step function on a particular date and can be fully verified over a reasonably short timeframe.

9.2.2 The Objectives of Evaluation

An evaluation can focus on a program's process and/or its impact on energy use and demand, the environment, and other areas that affect people and the economy. The best evaluations should have both process and impact components.

Process Evaluation

Process evaluation is an important tool for assessing program impacts as well as for improving program design, acquiring more participants, and increasing cost-effective energy savings generated by the program. A successful example of process evaluation, mentioned previously, showed that the E.U. label was not being applied correctly by a large number of retailers, which allowed corrective action to

be taken. In contrast, the U.S. label was evaluated many years after it was first introduced and the evaluation found that the label was widely misunderstood (e.g., a large proportion of U.S. consumers mistook operating cost information for operating savings), and no corrective action was taken (du Pont 1998a).

A process evaluation measures how well a program is functioning and is often qualitative. Although policy makers sometimes assume that a program is functioning smoothly and therefore may not see the need for or value of this type of evaluation, process elements are critical to the implementation and success of a program. This is especially true because program success usually depends on a number of separate activities all functioning as designed. If one element doesn't function as planned, the program may either fail or have a significantly reduced impact. For example, the success of an energy-labeling program requires, at a minimum, the following:

- correct labels for the designated products to be supplied with the product
- product purchasers to see and use the labels when making purchase decisions
- the information on the labels to be accurate
- the information to be correctly interpreted by purchasers
- a significant number of purchasers to be motivated by the label to consider purchasing more efficient equipment
- the market to be able to respond by supplying more efficient equipment

Process evaluation elements include:

- assessing consumer priorities
- tracking consumer awareness
- monitoring correct display of labels in retail showrooms
- measuring administrative efficiency (e.g. registration times)
- checking and verifying manufacturer claims (maintaining program credibility)

Impact Evaluation

An impact evaluation determines the energy and environmental impacts of a labeling program. Impact data can also be used to determine cost effectiveness. Impact evaluations can also assist in stock modeling and end-use (bottom-up) forecasting of future trends. Impact evaluation elements include:

- determining the influence of the label on purchase decisions
- tracking sales-weighted efficiency trends
- determining energy and demand savings

Impacts can be very difficult to determine accurately, especially for a labeling program. One of the fundamental problems is that, once a program such as energy labeling has been in place for some time, it becomes increasingly difficult and hypothetical to determine a "base case" against which to compare the program impact.

Evaluation Issues

Both process and impact evaluations should be performed regularly during the life of a labeling and standards program, and especially during initial implementation. Evaluation frequency will depend upon the type of program and technology; as a general rule of thumb, an evaluation should be conducted every two to three years at a minimum. Waiting longer between evaluations can lead to neglect and result in a stagnant or ineffective program.

Process and impact evaluations of labels and standards can be conducted based on either "resource-acquisition" or "market-transformation" objectives. From a resource-acquisition perspective, the primary objective of evaluation is to calculate energy and demand savings and greenhouse gas (GHG) emissions reductions (i.e., the reduced need to purchase energy from a power plant) and the associated cost of acquiring these resources during the first few years of the implementation of a standard.

From a market transformation perspective, the primary objective of evaluation is to see whether sustainable changes in the marketplace have resulted from labels and standards programs. For example, although a labeling program may take longer to implement and its energy-saving impacts may be seen over a longer period of time than is the case for standards, changes in attitudes and purchasing behavior can be evaluated during the first year of a labeling program. Program designers with the goal of market transformation are increasingly relying on theories with hypotheses about how the program might affect market players (Theory Evaluation or Logic Models). Program designers with this perspective benefit from evaluations that test their hypotheses through interviews and tracking of market indicators, which can then be translated into impacts. In addition, there are theories of how a market will evolve so that private actors might shift toward promoting more efficient products in the absence of a program. A theory-based approach, similar to a process evaluation, would test many of the hypotheses presented in this chapter such as: "most/some/all consumers will use labels as part of their purchase decisions" or "labels will encourage manufacturers to improve the energy performance of their products."

An appliance-labeling program influences the activities of many market players, including consumers, retailers, and manufacturers. Figure 9-2 shows how the various actors interact and affect the purchase environment, and, ultimately, the purchase decision of the consumer. Evaluators initially focus on "leading indicators": changes in the attitudes and behavior of market players), which can be measured in shorter periods of time than "lagging indicators": energy savings, appliance sales, and GHG emissions reductions.

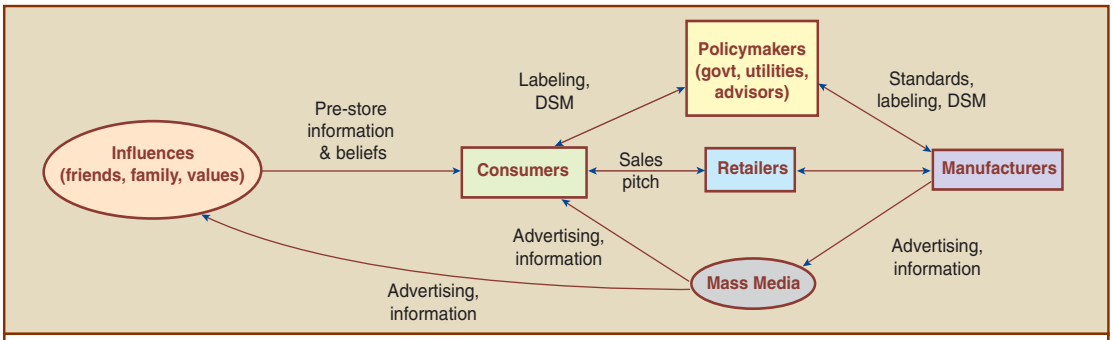


Figure 9-2 The appliance purchase environment

Planning the evaluation of a program with such complex interaction among stakeholders can be a challenge.

9.3 Step E-2: Identify Resource and Data Needs and Collect Data

The costs of evaluation and the types of data needed vary depending on a number of factors, as described in the subsections below.

9.3.1 Resources Needed for Evaluation

The cost of evaluating labeling and standards programs varies depending on a number of factors, such as the type of evaluation (process, impact), the quantity and type of available data, and whether energy savings are calculated by engineering estimates based on data from manufacturers or textbooks, and/or by end-use metering of a sample of products. Most comprehensive evaluations rely on the collection of survey, sales, billing, and end-use data. The use of end-use monitoring equipment to measure energy consumption for specific appliances will increase the cost of evaluation, as will the purchase of commercially available market research data on sales of different models. Although most evaluation costs arise after a program has been implemented, some of the evaluation budget should be allocated for up-front costs when the labeling and standards-setting programs are being discussed and the evaluation research plan is being developed.

9.3.2 Data Needed for Evaluation

The *type of data* needed for evaluation will also vary depending on a number of factors, such as the type of evaluation (process, impact), the quantity and type of existing data (versus data that must be collected, i.e., primary data collection), and whether measured data are needed.

Many types of data are useful for evaluating the impact of labeling and standards-setting programs, and many methods are available for collecting these data. The data requirements for labeling programs are similar to those for standards-setting programs in many but not all ways. For example, label impact

Table 9-1 Evaluation Data: Type and Sources

Labeling and standards-setting program evaluation uses a variety of data from a variety of sources.

Data Type	Main Data Sources
Customer and retailer knowledge, awareness, understanding, and decision making	• Surveys of customers and retailers and in-depth interviews
Availability of products	• Sales data from manufacturers, trade associations, or government • Surveys of manufacturers and retailers
Prices for efficient products	• Surveys of customers, retailers, and manufacturers
Market penetration	• Sales data from manufacturers, trade associations, or government • Surveys of participant and non-participant customers • Surveys of suppliers
Energy use	• Manufacturer data • Independent laboratory data • Engineering specifications • Metered end-use data
GHG emissions	• Reported emissions factors • Utility dispatch model data

evaluations are likely to rely more heavily on consumer surveys than would evaluations of standards programs although some assessment of individual consumer attitudes is useful in standards-setting evaluations as well. Much of the necessary data may already be available at the time the program is being designed. However, impact evaluation becomes especially important if inadequate research went into the design of the label initially. Whenever possible, secondary data sources (e.g., industry, commercial, and government reports) should be analyzed first because these are the most cost-effective sources of information. Once these sources are used, primary data collection should begin, based on interviews and surveys and focusing first on the most important data needs for the country in question. Table 9-1 gives information on the types of data needed and how they should be collected.

A caution is in order. Definitive data to support assessment of the impact of labeling and standards programs is, at best, difficult to obtain. Understanding of true consumer purchase behavior requires a carefully constructed research protocol, and ad hoc research is not likely to provide the necessary information. Consumers' verbal endorsements of the value of an attribute of an appliance or label may not coincide with their financial decision. Manufacturing costs and mark-up rates throughout the distribution chain are generally not available. Market share and consumer purchase choices are also influenced by many factors unrelated to relative energy efficiency. The amount of time and resources appropriate for evaluation are often greater than initially anticipated and budgeted.

A first step in evaluation is to collect model-specific data for establishing a national appliance database. This database will contain information on the models that are manufactured and their annual sales,

prices, and technology characteristics. The database can be used to monitor national appliance-efficiency trends. When energy use is analyzed, utility bill data or end-use metered energy data should be collected (sometimes, the change in energy use for an appliance is too small to be reflected in a utility bill, hence the need for end-use metering).

When energy savings are projected, particularly near the beginning of a labeling or standards-setting program, data are typically collected on equipment energy-use trends under standard test conditions and then linked to sales and retirement data in a stock model to project past and future impacts. As a complementary activity, end-use metering data are collected to: a) calibrate the energy-use data based on engineering estimates used in the stock model; b) establish the accuracy and failings of the test procedure; c) enable corrections to be developed for energy data measured under the test procedure (e.g., a factor of 0.85 was applied for energy-labeling purposes to U.S. freezer energy-consumption results recorded under standard test conditions); and d) establish other avenues for energy savings (such as advice to consumers that is informed by data on their energy consumption).

9.3.3 Types of Data

Other types of data needed include the attitudes and behavior of key market players and characteristics of the market (e.g., number of manufacturers and retailers, percent of appliances in stock that are energy efficient). Finally, it is important to note that it is always possible to carry out some level of evaluation, no matter how crude the data sources and how limited the resources. Evaluators should not be discouraged if they cannot gather data of the highest quality; compromises in accuracy can be made to limit cost without making an evaluation useless.

9.3.4 Data-Collection Methods

As noted earlier, it is very important to collect data in the beginning of designing and implementing standards and labeling programs. Whenever possible, cooperative agreements with industry should be encouraged for the purpose of gathering data on sales and efficiency levels. Sales data can be obtained from surveys of manufacturers, retailers, and/or contractors. Products in stores can be inspected visually to assess compliance with labeling programs and to collect information on stocking practices (sometimes this is done by a "mystery shopper" who visits stores unannounced and unidentified). Appliances can be tested in laboratories to measure energy use and assess the accuracy of labels. Finally, interviews with consumers, retailers, manufacturers, and contractors often play a central role in assessing the extent of market transformation.

9.4 Step E - 3: Analyze Data

A comprehensive analysis is needed to evaluate resource acquisition and market transformation. For a simplified model of analyzing an appliance labeling program, see Figure 9-3. Although this type of analysis has usually been focused on labeling programs, it can also be used to evaluate standards programs.

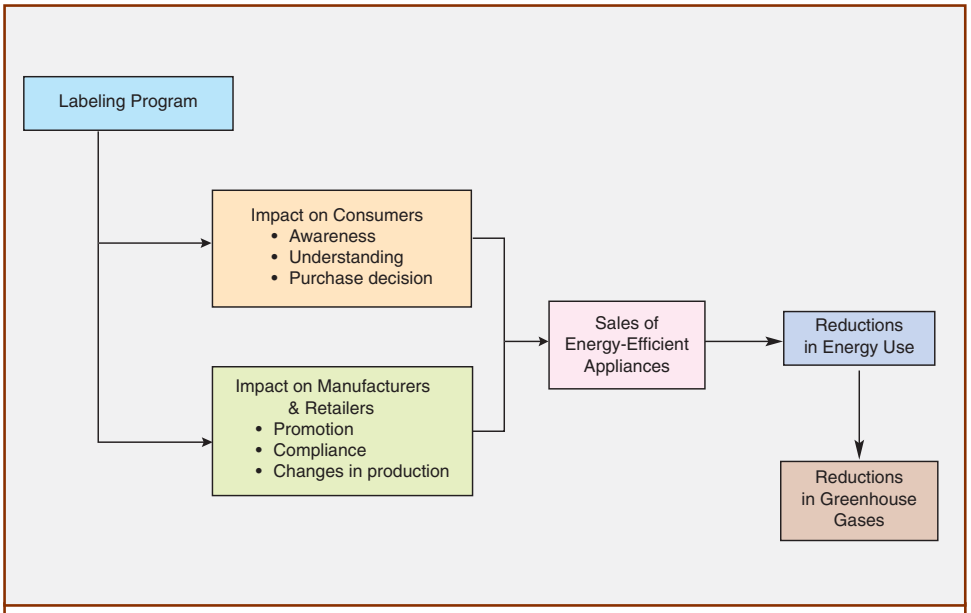


Figure 9-3 Simplified example of analyzing appliance-labeling programs

A simplified procedure is available to evaluate resource acquistion and market transformation.

9.4.1 Baseline

It is critical for an evaluation to establish a realistic and credible baseline, that is, a description of what would have happened to energy use if labels and/or standards had not been implemented. Determining a baseline is inherently problematic because it requires answering the hypothetical question "what would have happened in the absence of labels and/or standards?" To accurately evaluate energy savings, it is necessary to analyze energy use of a sample of households/facilities before and after the installation of an energy-efficient product. For example, energy use might be measured for a full year before the installation of the efficient appliance and then for several years after the installation. Some types of appliances may not require a full year of monitoring, however. If loads and operating conditions are constant over time, short-term (e.g., one-week) measurements may be sufficient to estimate equipment performance and efficiency. These data would then be used for calibrating engineering estimates that could generally be applied to the population of energy-efficient products. Frequently, load research data are available for establishing product baselines (see Section 9.3.4).

Market characterization studies are also necessary for developing a baseline of existing technologies and practices. These studies provide detailed data on end users (consumers), including estimates of market size, analyses of decision making, identification of market segments, and analysis of market share by market event (retrofit, renovation, remodeling, replacement). Market characterization studies also provide detailed data on the supply side—manufacturers, retailers, and contractors (e.g., designers and installers)—including information on relationships among supply-side actors; development of market segments; business models of each entity; and the nature of distribution channels, stocking/selling practices, and trade-ally reactions to labeling programs.

Baseline development is often highly contentious and, at best, a good guess of what might have been. In many cases, it is as important to quantify the level of efficiency improvement from before the time of the program startup in order to demonstrate that progress is continuing. Finally, it is important to note that the baseline issue is important for all types of energy-impact evaluations and is a crucial element of the assessments conducted to determine carbon savings from energy-efficiency and renewable energy projects under international carbon-trading provisions and agreements (see Section 9.4.6 below, Vine and Sathaye 1999, Kartha et al. 2004).

9.4.2 Impacts on Consumers

A key point in evaluating the effect of labeling programs on consumers is the degree to which the label's presence affects consumer purchasing decisions in favor of more efficient appliances. In addition to observing actual consumer purchasing and sales trends, consumer evaluations should also focus on consumers' level of awareness and understanding of energy and on the factors that affect their purchases of energy-efficient appliances. Specific types of questions to address in this type of evaluation include:

- What is the level of awareness, among buyers and potential buyers, of the energy label, related product materials, retailer advice, and advertising?
- What is the relative level of importance of various consumer purchase criteria—such as brand, price, perceived durability, product features, size, color, energy use, environmental factors – in the consumer's appliance purchase decision?
- What is the relative level of importance given to the energy label, related product materials, retailer advice, and advertising in the buyer's choice of appliance?
- How well does the customer understand the label, related product materials, and advertising?
- What is the customer's perception of the usefulness of the label, related product materials, retailer advice, and advertising?
- What sorts of changes do consumers propose to the label, related product materials, and advertising to make each more effective?
- What is the importance of energy or fuel efficiency in the buyer's choice of the appliance? How does this relate to other customer purchase priorities?
- How does the customer use the appliance?
- What are the life-cycle cost impacts, accounting for possible changes in the price of the equipment, operating expenses, and installation or maintenance expenses?

Socio-economic data can also be analyzed to help understand the effectiveness of labeling and standards-setting programs for different socio-cultural situations: e.g., low-income households versus high-income households, recent purchasers versus the general public. Market segmentation can be used to develop education, information, and advertising programs that complement labeling and standards-setting programs. For example, program material can be translated into different languages, and program providers

can use residents of targeted communities to educate local populations about the benefits of energy-efficient equipment.

There is an array of econometric and statistical models for analyzing the contributions of many factors to program impacts on consumers. These are generally considered to be advanced evaluation tools and range widely in cost depending on many characteristics, especially their level of accuracy; however, it will often suffice to use simple tools and methods (see Vine and Sathaye 1999).

9.4.3 Impacts on Manufacturers and Retailers

Evaluators assess the impact of labeling and standards programs on appliance manufacturers and/or retailers by examining the following:

- consolidation of competition
- impact on features, product utility, and consumer choice
- impact on manufacturing jobs
- impact on private-sector advertising in support of labeling programs
- impact on sales (and market share)
- compliance with the programs
- promotion of labels to retailers (e.g., direct promotion, print advertising, in-house product presentations and training, trade fairs, product catalogs, help desks)
- direct promotion to consumers (by both manufacturers and retailers)
- direct and indirect costs to manufacturers (increased cost of production, research and development efforts to improve appliance efficiency, distribution of labels, promotion and support of labeling programs)
- changes in the production process to manufacture more efficient models
- impacts similar to those affecting consumers (see Section 9.4.2)
- placement of energy labels on appliances in retail outlets

9.4.4 Program Compliance, Enforcement, Training, and Education

Once appliance labels and standard-setting programs have been implemented, it is important to regularly monitor whether program requirements are met, enforcement measures are taken where there is non-compliance, retailers and distributors are trained in explaining the program to consumers, and consumers understand the meaning of the label and/or standard.

For example, in many labeling and standards-setting programs, it is the manufacturers' responsibility to ensure that the information they supply is correct. Often, there is no automatic system of independent

testing. Occasionally, third-party testing agencies are used. In the U.S. and Canada, manufacturers test their own products in certified test laboratories and report the results on the label. In principle, such a system can work well because any manufacturer can challenge the veracity of a competing manufacturer's claim. This system of self-certification and challenges is used in the U.S. and is generally thought to provide acceptable compliance.

In Europe and Australia, the practice depends on the product concerned, but, for most household appliances, the test laboratory does not have to be certified. Manufacturers are responsible for the accuracy of their claims and are at risk if they use a non-certified laboratory and a control agency subsequently fails the product. Under E.U. legislation, it is the responsibility of the member states to ensure that E.U. law is enforced in their states (Waide 1997). In the past, some serious inaccuracies in energy consumption reporting have been identified for refrigerators, freezers, and clothes washers in the E.U. This indicates, as described in more detail in Chapter 4, that it is necessary to compare manufacturer-reported energy consumption to test results from a third-party laboratory as well as to monitor energy use in the field (although end-use metering in the field does not take place under standard test conditions, it gives information on the relevance of the laboratory studies) to determine whether the appliance rating and label should be changed (e.g., see Meier 1997; and Winward et al. 1998).

A labeling program also depends on retailers' efforts to make sure that labels are attached to appliances for consumers to read. Thus, it is imperative for evaluators to assess retailers' compliance with the program (see Winward et al. 1998). Australia has developed a model "Check Testing Program" for evaluating a sample of models in the market to monitor whether labels are applied and whether the test results reported on the label are accurate (Grubert 2001).

In sum, evaluation studies can assess current levels of manufacturer compliance and remedial enforcement activity. Evaluators may also examine the use of formal legal processes to impose penalties on persistent rule breakers (see Winward et al. 1998) and may assess the effectiveness of training and education programs as well.

9.4.5 Sales

As noted above, one of the two key "lagging indicators" for evaluation is sales. Market share is also considered a lagging indicator because it is established after the changes that actually cause a difference in purchase habits. Market-share information is critical for the final analysis of a program's effects, but it is often not immediately available during program implementation. Nevertheless, it is possible to evaluate the impact of a labeling program by comparing sales-weighted trends in appliance efficiencies both before and after the introduction of labels. For example, Figure 9-4 shows the sales-weighted, annual-average distribution of dishwashers by energy-label class in the E.U. in 1994, prior to the introduction of energy labeling, and in 2003 some years after labeling. The figure shows that the predominance of purchases shifted from inefficient models (classes D, E, and F) in 1994 to more efficient classes (A, B, and C) in 2003.

Analyses can focus not only on sales but also on changes in prices and technology characteristics (e.g., sizes of appliances). Improvements in appliance energy efficiency are not necessarily related to an increase in the price of the appliances sold. Despite the existence of a strong relationship within the market between average refrigerator price and efficiency, the average refrigerator sold in the E.U. in 2002 was 4 euros less expensive but significantly more efficient than the average sold in 1994 (see Figure 9-5).

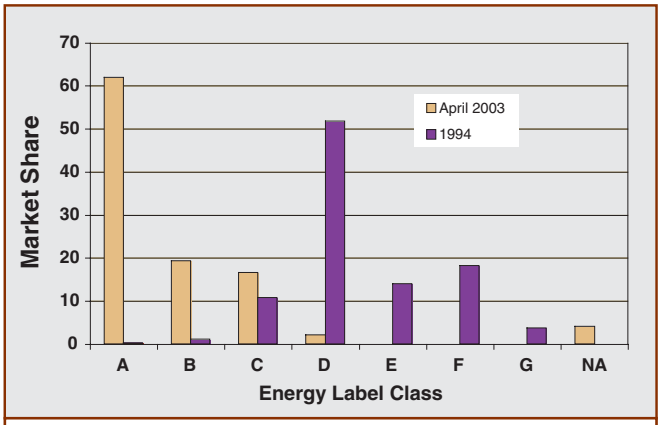


Figure 9-4 Impact of the E.U. dishwasher energy label (dishwasher sales as a function of energy label class from 1994 to 2003)

Impacts can be shown as increases in the sale of efficient products.

9.4.6 Energy Savings and Greenhouse Gas Emissions Reductions

Estimation of reductions in GHG emissions is becoming increasingly important as climate change becomes a driver for many sustainable-energy projects, including energy-efficiency efforts. The main international vehicle has been the Clean Development Mechanism; however, the future of this mechanism is uncertain due to the Kyoto Protocol taking effect February 2005. A number of institutions are being set up for trade in carbon reductions, and several international agencies (including the World Bank, the Dutch government, and others) have begun to actively purchase GHG reductions on a small scale.

Impact analysis can show that efficiency doesn't necessarily increase price.

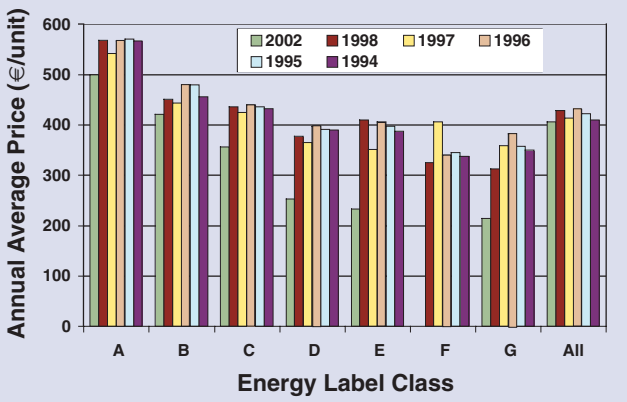


Figure 9-5 Impact of the E.U. refrigerator energy label (E.U. average refrigerator price as a function of energy label class from 1994 to 2002)

At the household or facility level, it is impossible to measure energy savings directly because, to do so, it is necessary to know how much energy would have been used if a specific appliance had not been purchased, which cannot be determined. Nevertheless, any of a number of evaluation methodologies can be used for estimating energy savings, especially for a large sample. These include engineering methods, statistical

models, end-use metering, short-term monitoring, and combinations of these methodologies (Vine and Sathaye 1999).

Changes in market share of energy-efficiency products (sales), for example, can be estimated and multiplied by the amount of unit energy saved (e.g., on average or by type of product). Tracking changes in product and market characteristics over time gives a good initial indication of the type of market shift that takes place in the early stages of labeling or during the lead-up to a new standard coming into force. Detecting trends in consumer preferences toward more efficient products on the market is a more subtle exercise. Here, both sales-weighted trends and changes in consumer sentiment need to be monitored. To maximize the accuracy of the energy savings determined from shifting between any two models, a sample of products can be metered in situ to determine the actual amount of energy used.

At the national level, energy savings can be determined using simple calculations (e.g., spreadsheets) or detailed energy end-use models. The assumptions used in engineering analyses are adjusted to account for real-world data (e.g., actual consumption in the field, fraction of households owning a particular appliance, usage in hours per year) from surveys and end-use monitoring (see McMahon 1997 and Greening et al. 1997).

Once net energy savings have been calculated by subtracting baseline energy use from measured energy use, net GHG emissions reductions can be calculated in one of two ways: average emissions factors can be used, based on utility or non-utility estimates, or emissions factors can be calculated based on specific generation data (Vine and Sathaye 1999). In both methods, emissions factors translate consumption of energy into GHG emissions. Normally, the use of average emission factors is accurate enough for evaluating the impact of energy-efficiency labels and standards. In the rare cases where the other impact analyses are highly sophisticated and regional variations are important, use of plant-specific factors may be warranted.

In contrast to using average emission factors, calculated factors have the advantage that they can be specifically tailored to match, by time of day or season of the year, the characteristics of the activities being implemented. For example, if an appliance-labeling program affects electricity demand at night, then baseload power plants and emissions will probably be affected. Because different fuels are typically used for baseload and peak-capacity plants, baseload emissions reductions will also differ from the average.

The calculations become more complex (and more realistic) if the emission rate of the marginal generating plant is multiplied by the energy saved for each hour of the year instead of multiplying the average emission rate for the entire system (i.e., total emissions divided by total sales) by the total energy saved. For more detailed analysis, the utility's existing system dispatch and expansion plans can be analyzed to determine the generating resources that would be replaced by saved electricity and the emissions associated with these electricity-supply resources.

It is also necessary to determine whether planned energy-efficiency measures would reduce peak demand sufficiently and with enough reliability to defer or eliminate planned capacity expansion. If so, the deferred or replaced baseline source would be the marginal expansion resource. This type of analysis may result in fairly accurate estimates of GHG reductions, but it is more costly than the simpler method and requires expertise in utility-system modeling. In addition, this type of analysis is becoming more difficult in regions where the utility industry is being restructured. In restructured markets, energy may come from multiple energy suppliers either within or outside the utility service area, and the marginal source of power is difficult to forecast.

9.5 Step E-4: Apply Evaluation Results

Initially, it is important to ensure that policy makers allocate funding for and place priority on developing a framework for evaluation and data collection. Later on, if a technically sound evaluation produces significant results, it is imperative that these results be used, where appropriate, to:

- refine the design, implementation, and evaluation of labeling and standards-setting programs
- support other energy programs and policies
- support accurate forecasting of energy demand for strategic planning
- improve the accuracy of models and analyses in regulatory proceedings

Because the value and amount of load reduction achieved by programs varies by time and location, it is very useful to categorize evaluation results by date, time, and geographical location.

9.5.1 Refining Labeling and Standards Programs

The results from evaluations can be used to improve the design, implementation, and future evaluations of labeling and standards programs. For example, evaluation results can be used to reexamine the accuracy of the inputs used in designing the program. In addition, they can be used to assess whether the programs can (or should) be extended to other appliances that are not currently covered. Ideally, the program designers become the clients of the evaluation department, and the evaluation results feed directly into the next round of program design or improvement.

9.5.2 Supporting Other Energy Programs and Policies

The evaluation of labeling and standards-setting programs can help design appliance rebate programs, appliance standards or negotiated agreements (if none exist), procurement actions, and labeling programs for other appliances. Chapter 10 elaborates on these topics.

9.5.3 Forecasting Energy Use and Strategic Planning

Evaluation results can be used, with caution, to support forecasting and resource planning. In particular, the following elements of an evaluation should be considered before the results are used:

- How representative is the study sample in relation to the population of interest to planners?
- How accurate and precise are the energy and demand impact results?
- Did the evaluation use appropriate control samples?

If comprehensive data on market energy-efficiency trends, sales volumes, and usage patterns are established as part of the evaluation process, these data can be used as inputs to an end-use stock model to make long-range energy consumption and emissions forecasts. This kind of forecasting is useful to guide policy development because it enables the estimated impact of various policy and implementation changes to be simulated in advance.

9.5.4 Using Evaluation Results and Data for Other Regulatory Purposes

Some regulators have standard practice guidelines or manuals on how to conduct cost-effectiveness analysis. Evaluation results are often vital inputs to the cost-benefit tests contained in these manuals. For example, the results from studies on measure retention, technical degradation, and persistence of savings are used for calculating the ongoing costs and benefits used in cost-effectiveness analysis.

9.6 Considering Key Evaluation Issues

This section describes a set of mitigating or potentially confounding issues—free riders, accuracy and uncertainty, and complexity—that may impact or bias the evaluation results and explains how to deal with these issues in the context of the overall evaluation. If resources are available, one must take these factors into account for the evaluation results to be completely credible and defensible.

9.6.1 Free Riders

To evaluate the impacts of standards and labeling programs, one needs to know what customers would do in the absence of these programs. Labeling and standards programs affect only some purchases. Furthermore, some consumers would have purchased the same efficient products even if there had been no program. In an evaluation analysis, these consumers are called “free riders.” The savings associated with free riders are not “additional” to what would occur in the baseline case (Vine and Sathaye 1999). Therefore, free riders should be excluded when estimating savings attributed to the programs. This can be accomplished either by accounting for free riders in the baseline or making a separate adjustment.

For example, if a comparison group's utility bills show an average reduction in energy use of 5% during a given period of time before a label or standard is implemented and then shows a total reduction in energy use of 15% during an equivalent period afterward, it may be reasonable to judge that 5% of the total reduction would have occurred anyway, consistent with the preceding period, and thus to attribute only a 10% reduction in energy use to the standards program (15% total minus the 5% trend that was already occurring and therefore would likely have continued).

Free riders can be evaluated either explicitly or implicitly. The most common method of explicitly estimating free ridership is to ask participants what they would have done in the absence of labeling (this is sometimes referred to as "but for the project" analysis). Based on answers to carefully designed survey questions, participants may be classified as free riders or assigned a free ridership score. As in other surveys, the questionnaire must be carefully worded and interpreted; people's stated preferences and anticipated behavior often differ from their actual preferences and behavior.

It can be especially challenging to evaluate free ridership for labeling programs when other market-transformation programs, such as rebates for efficient appliances, are in place. Because these market-transformation campaigns are specifically designed to create—over time—a situation in which purchasing energy efficient appliances is common practice even in the absence of any program, it is difficult to estimate the increasing rate of efficient purchasing that would result if only the other market transformation programs were in place.

Because estimating the free rider effect is difficult, simple and highly uncertain assumptions are often made about free ridership. If resources are not available for conducting a sophisticated analysis, evaluators may be able to use other sources that implicitly address this issue (e.g., comparing to appliance investment behavior in other regions or in other countries where there are no appliance labeling or standards-setting programs).

9.6.2 Accuracy and Uncertainty

Because of the difficulties and uncertainties in all aspects of estimating energy savings, the degrees of precision and confidence associated with savings measurements should be identified. Ideally, evaluators should estimate and report the precision of their measurements and results in one of three ways:

- quantitatively, by specifying the standard deviation around the mean of an assumed bell-shaped normal distribution
- quantitatively, by providing confidence intervals around mean estimates
- qualitatively, by indicating the general level of precision of the measurement using categories such as "low," "medium," and "high"

9.6.3 Policy and Market Complexity

One of the criteria for examining the success of a market-transformation program is whether observed market changes can be appropriately attributed to the program. Analysis can be conducted more reliably when there is a single type of intervention than when multiple actions (e.g., standards, labeling, procurement, rebates, the phase-out of chlorofluorocarbons, and industrial changes) are occurring simultaneously. It is difficult to distinguish the relative contributions of multiple activities to observed changes in the market. Although logic diagrams and market-influence diagrams are extremely useful tools to structure the analysis, they are generally not powerful enough to handle the evaluation of the complex characteristics of the appliance, equipment, and lighting markets.

In order to reliably claim that observed efficiency improvements were caused by labeling and standards programs, it is necessary to carefully consider and reject other possible explanations for the observed market changes. In particular, the presence of multiple interventions (e.g., changes in energy pricing and metering, financing and incentives, improvements in technology, and regulatory and voluntary programs by government and the private sector—see Chapter 10) may affect the baseline as well as the implementation of labeling and standards programs. An effective external comparison group may help isolate some of these influences. Also, causal modeling may provide a useful approach to making separate attributions to different influences although it is very difficult to create a quantitative model, and manufacturers are often reluctant to make the necessary data available. Quantitative determinations are often difficult to make and can involve substantial costs that may or may not be worthwhile. Venture into this realm of analysis with caution. Sometimes it may be best to simply report the total impact of the program and present the reasons that the program being assessed is a major contributor to that outcome.

Standards and labeling program planners have a strong interest in the evaluation process. Gathering evaluation results by defining objectives, identifying necessary resources, monitoring program performance, and assessing program impacts is a valuable output of a standards and labeling program. The results can be used either to revise an existing program's objectives or as building blocks in establishing a new program. But it is always difficult to measure a program's performance and impact. In some cases, this is the result of lack of data or lack of resources to obtain that data. In others, it may be that the program's direct results are masked by the effects of other complementary, simultaneous programs. Given real-world budget and time constraints, it is difficult to do a "perfect," comprehensive evaluation. However, even paying limited attention to evaluation and making approximate assessments can provide very useful input to program planners and implementers. Simple evaluations, done in a thorough and transparent manner, can serve most of the evaluation needs of a standards and labeling program. For energy-efficiency standards and labeling programs, doing some evaluation is almost always better than doing none.