An Evaluation Framework for a Portfolio of Research, Development & Demonstration Programs

Helen Kim, New York State Energy Research &Development Authority
Larry Pakenas, New York State Energy Research &Development Authority
Rick Ridge, Ridge & Associates
Scott Albert, GDS Associates
Gretchen Jordan, Sandia National Laboratories

Abstract

This paper addresses an approach used to evaluate a portfolio of research, development and demonstration (RD&D) programs administered at the State level. The following project types were identified: (1) research for policy, including environmental research, (2) product development, (3) demonstrations, and (4) pre-deployment activities. In addition, the following performance criteria were identified for use in evaluating projects: (1) knowledge creation, (2) knowledge dissemination, (3) commercialization progress, (4) energy benefits, (5) economic benefits, and (6) environmental benefits. Data were assembled to measure progress of projects according to these criteria and compiled into an "accomplishments packet" consisting of quantitative and qualitative information. The packet was distributed to peer reviewers to assess the significance of the accomplishments. The results of the assessment are discussed. The applicability of the approach to the various types of projects and technologies is also presented.

Introduction

The societal impacts of RD&D programs are difficult to evaluate due to the long-term focus and the influence of multiple unforeseen events. There is also an added difficulty when evaluating an entire portfolio of RD&D projects, as objectives and outcomes vary by project.

A review of the evaluation literature reveals a number of approaches developed specifically for assessing the impact of RD&D programs that include books by Bozeman and Melkers (1993), Link (1996), and Ruegg & Feller (2003). The Composite Performance Rating System (CPRS) described in Ruegg and Feller (2003) seemed most promising for NYSERDA's RD&D programs due to its ease of application and applicability to a portfolio of diverse projects. The CPRS was developed to evaluate the U.S. Department of Commerce's Advanced Technologies Program (ATP). The purpose of the CPRS is to consolidate various dimensions of project success into a single performance index that is a measure of knowledge creation, knowledge dissemination, commercialization success, and future outlook for an individual project. The index ranges from zero to four stars, with four stars indicating a top-rated project (Shipp, Kirtley & McKay, 2004).

In the current study, the CPRS methodology was modified to apply to NYSERDA's broad portfolio of projects. The portfolio of projects consists of six activity types:

- Research for policy: includes projects that result in information for policy makers and the research community, including manufacturers of products. Examples include environmental research, market assessments, technology assessments, and development of business models.
- Product development Stage 1: activities related to product-specific proofs of concept.
- Product development Stage 2: activities related to developing and improving products.
- Product development Stage 3: activities related to product testing.

- Demonstration: demonstration of products that are commercially available to increase awareness and knowledge. Examples include site-specific demonstration of combined heat and power projects.
- Pre-deployment: activities designed to accelerate adoption of commercially available products. Examples include training of photovoltaic systems installers.

The activity types address different barriers and thus have different goals and outcomes. The logic model shows indicators of progress for each activity type. Some of these indicators are quantifiable (number of papers and patents, amount of additional investment) and others are qualitative (changes in behavior, changes in procedures).

Using the logic model, outcomes were categorized into six categories:

1. Knowledge creation
2. Knowledge dissemination
3. Commercialization progress
4. Energy benefits
5. Economic benefits
6. Environmental benefits

Data were collected on these outcomes and compiled into accomplishment packets that were provided to peer reviewers who rated the significance of the accomplishments given program resources and the goals of the program. The rest of this paper will discuss the application of the framework that occurred in two phases. The first phase focused on five individual projects and the second phase focused on two programs consisting of relatively homogeneous projects.

Phase 1

Phase 1 was a pilot study designed to test the applicability and usefulness of the framework. The answers to the following questions were sought:

- Can the six outcomes be meaningfully measured?
- Can the six outcomes be measured equally well across the different activity types?
- What is the cost associated with measuring the outcomes?
- Would external reviewers be able to assess the impact of the projects using the outcome measures?
- Will the ratings be reliable and valid?

To answer these questions, program staff were asked to nominate projects they believed to have significant measurable outcomes and were funded through the Systems Benefits Charge (SBC). The SBC program was initiated in late 1998 and therefore, the pool of completed projects was limited. The purpose of targeting projects with measurable outcomes was to ensure that sufficient data existed to measure the outcomes, recognizing that the selection method would result in projects with above average performance. The six projects selected for assessment represented about 80% of the nominated projects. The six projects were selected by the evaluation team on the basis of obtaining a full range of activities in the logic model. The projects, activity types, funding amounts, and technology area are shown in Table 2. Some projects involved more than one activity type as the project evolved over time.

Table 1. Projects in Phase 1

Project Name (Funding)	Activity Type(s)	Technology Area
21st Century HVAC Research Consortium ($.7 million NYSERDA, $9.4 million total)	Research for Policy	HVAC
Aggregating Distributed Generators ($.5 million NYSERDA, $1.1 million total)	Product Development Stage 2 and Stage 3, Demonstration	Demand Response
Development of Continuous Ambient Particulate Monitor ($.5 NYSERDA, $1.2 million total)	Product Development Stage 2 and Stage 3	Environmental Monitoring
Truck Stop Electrification ($.15 million NYSERDA, $3.3 million total)	Product Development all stages and Demonstration	Transportation
Turnkey Pump and Compressed Air Program ($.4 million NYSERDA, $.7 million total)	Demonstration	Industrial Process
Green Power Marketing*	Pre-Deployment	Power Generation

* This project was later dropped from the evaluation due to time and budget limitations.

Shown in Table 3 are the number of reviewers who were recruited for each project and the reviewers' affiliations.[1] In addition to external reviewers, the project manager for each project also assessed the project.

Table 3. Number of Reviewers Solicited, Number of Participating Reviewers, and Number of Completed Assessments

Project	Number of Reviewers Solicited	Number of Reviewers Recruited	Number of Completed Assessments	Reviewers' Affiliations
21st Century HVAC Research Consortium	6	6	6	- 2 engineers with substantial experience designing and evaluating HVAC equipment. -Heating equipment manufacturer -Member, ASHRAE 90.1 committee - NYSERDA project manager
Aggregating Distributed Generators	7	6	4	-Engineer from national laboratory - Independent system operator staff - NYSERDA project manager
Continuous Ambient Particulate Monitor	5	5	5	- EPA staff - Air quality professionals from state agencies

[1] The peer review effort conducted for this evaluation is based largely on the "*Peer Review Guide*" prepared by the Office of Energy Efficiency and Renewable Energy: Peer Review Task Force (2004).

Project	Number of Reviewers Solicited	Number of Reviewers Recruited	Number of Completed Assessments	Reviewers' Affiliations
				-NYSERDA project manager
Truck Stop Electrification	12	10	10	- National Laboratory staff - Member of National Idling Reduction Plan working group - Editor of National Trucking magazine - Trucking company executive - In-state and out-of-state transportation officials and utility executives - University Research Center - National Utility Institute staff - NYSERDA project manager
Turnkey Pump and Compressed Air Program	6	4	4	- Out-of-state Investor-Owned Utility Program Director - Compressed Air Challenge board member - Energy engineer, developer of prescriptive programs - NYSERDA project manager

Reviewers were provided with an accomplishment packet that was designed to address the six criteria and a seventh criterion called Overall Value. Some of the criteria were further divided into subcomponents as shown below:

1. Knowledge Creation
   - 1a. Quantity: The number of technical papers, articles, citations, patents (both filed and granted), licenses, and prototypes passing requirement tests.
   - 1b. Significance: The contributions and relevance to the field of the knowledge created.

2. Knowledge Dissemination
   - 2a. Availability of Knowledge Products: Knowledge that has been codified in some form.
   - 2b. Target Audience: Impact on target audience.

3. Commercialization Progress
   - 3a. Capital Attraction: The extent to which the project has attracted capital for advancing commercialization objectives, including resources provided by project partners.
   - 3b. Technical Achievement: The extent to which the project has accomplished important technical achievements, including the development of prototypes.

3c. Market Advancement: The extent to which the project reduced key market barriers, demonstrated new products at customer sites, led to the development of sustainable business models, produced employment changes within the companies leading the projects, increased the number of business recognition awards, and produced sales of the new product.

4. Energy Benefits

5. Economic Benefits

6. Environmental and Health Benefits

7. Overall Value
   7a. The extent to which the value of the project is greater than NYSERDA's costs.
   7b. The extent to which the value of the project is greater than both NYSERDA's costs and those of other funding partners.

The reviewers were asked to rate the significance of the outcomes based on the information presented in the packet. Not all outcomes were measured for all projects. The measured outcomes are shown in Table 2.

Table 2. Outcomes Measured by Project

Outcome	21st Century HVAC	Aggregating DG	EMEP Air Particulate Monitoring	Truckstop Electrification	Compressed Air Program
Knowledge Creation	X	X	X	X	X
Knowledge Dissemination	X	X	X	X	X
Commercialization Progress		X	X	X	X
Energy		X		X	X
Economic		X	X	X	X
Environmental		X	X	X	X
Overall Value	X	X	X	X	X

Phase 1 Results

Shown in Figure 1 are the ratings for the seven assessment areas. The rating scale was from 0 to 4 stars, with 0 representing Not At All Significant and 4 representing Very Significant. Some projects had a score for all 7 criteria whereas others did not because a particular outcome was not applicable. For example, only knowledge creation, knowledge dissemination, and Overall Value were applicable to the 21st Century HVAC project which was a research for policy project.

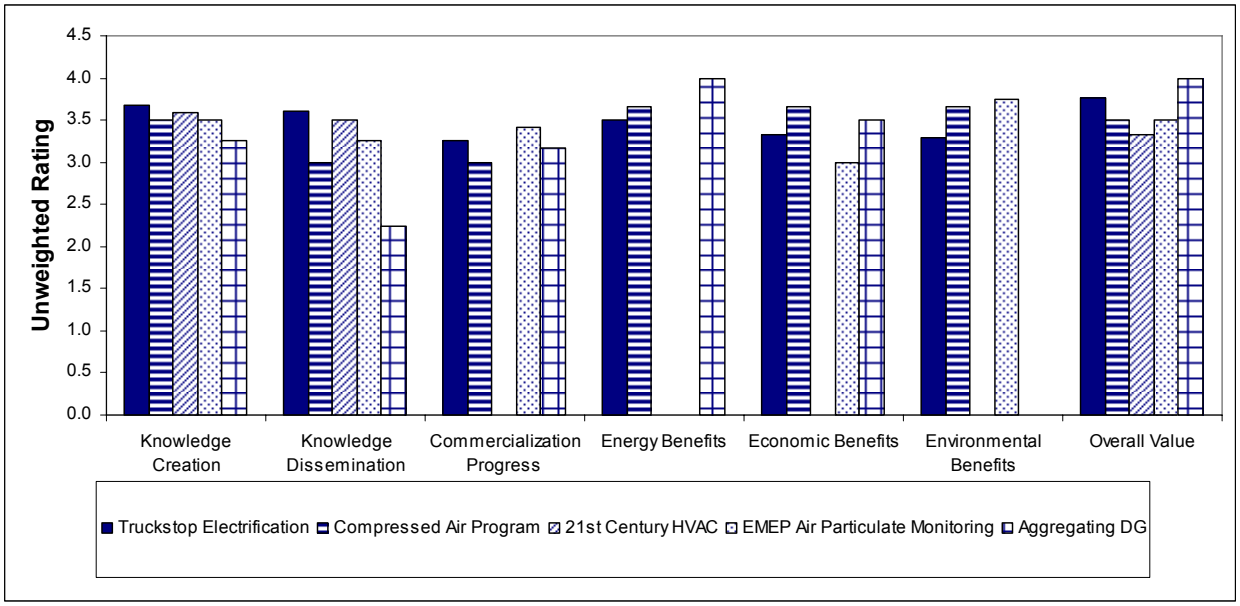
As expected, the Overall Value for all projects were high due to the project selection criteria previously stated. Aggregating DG had the highest Overall Value score followed by Truckstop Electrification. Both these projects encompassed both product development as well as demonstration activities. Aggregating DG had the highest score for energy benefits, presumably due to its potential to reduce peak electric demand. Compressed Air had the highest economic benefit, presumably due to the

potential energy savings at industrial facilities and Air Particulate Monitor project had the highest environmental benefit.

Shown in Figure 2 are the ratings by criteria subcomponents for Knowledge Creation, Knowledge Dissemination, Commercialization Progress, and Overall Value. The correlation between the two knowledge creation subcomponents was 0.13. The correlation between the two knowledge dissemination subcomponents was 0.51, and the average of the correlation coefficients among the three commercialization progress subcomponents was 0.29. The low degree of correlation among the subcomponents indicated that the subcomponents did not represent a cohesive measure and that the scores should not be combined. Also, quantity of knowledge does not appear to be a useful indicator of project success.

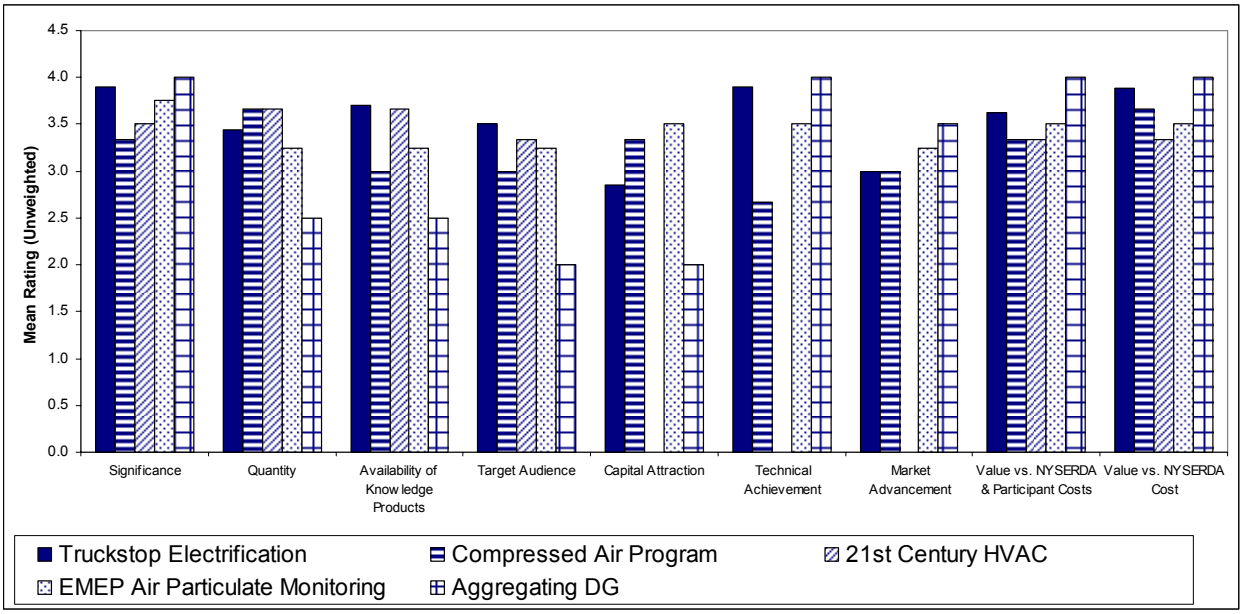
A strong correlation (r=.79) was found between the subcomponents 1b (knowledge significance) and 3b (technical achievement), suggesting that knowledge creation and technical achievement represent a single construct. Knowledge significance was also highly correlated with 7a (value compared to total cost) and 7b (value compared to NYSERDA cost) (r =.68, r = .75, respectively). These results suggest that the reviewers perceived knowledge significance and technical achievement to be the most important criteria for evaluating R&D projects.

Figure 1. Mean Outcomes Scores by Project



The rating scale ranged from 0 to 4.

Figure 2. Subcomponent Scores by Project



The rating scale ranged from 0 to 4.

Phase 1 Follow-up Survey

Approximately six weeks after the peer review assessments, a follow-up survey was conducted to collect data on the review process, resulting in 18 completed surveys out of 27 reviewers (67% response rate). Three projects had 50% response rates and one project had an 80% response rate. According to the survey, the reviewers felt that the information in the Accomplishments Packet was adequate for their review purposes, that the assessment instructions were clear, and that the assessment criteria were clearly defined.

Peer reviewers also reported that they spent an average of 2.6 hours on the review, with a range of 0.5 to 5.5 hours. Time spent included reviewing instructions, reviewing the information packet, completion of the Assessment Form, and completion of the peer review questionnaire.

When asked whether their organization would allow them to receive a stipend for such work, 56% indicated that their organization would not allow them to accept a stipend. Reviewers were also asked the extent to which an offer of a stipend would affect their willingness to participate in future evaluations. The mean, on a five-point scale, where 0 = "Would Not Affect My Willingness" and a 4= "Would Significantly Affect My Willingness," was 1.1 indicating that the offer would have little effect on the willingness of most reviewers. Even those reviewers, whose organizations would allow them to accept a stipend, reported a low mean of 1.6. They were then asked how large the stipend should be. The mean across all reviewers was $128 with a range of $0 to $400. The mean across those who could receive a stipend was somewhat higher at $158 with a range of $50 to $350.

Finally, reviewers were asked for suggestions to improve the effectiveness of the peer review process. Several reviewers suggested a desire for more interaction among the reviewers. The Delphi approach, focus groups, and conference calls were mentioned.

Interrater Reliability

The degree of interrater reliability was measured by the average measure intraclass correlation (AMICC) (Fleiss, 1981). Table 3 shows the AMICC for each project. Fleiss (1981) suggests that values greater than 0.75 may be taken to represent excellent agreement. Values between 0.40 and 0.75 may be

taken to represent fair to good agreement. Values less than 0.40 represent poor agreement. The reliabilities were in the acceptable range for two of the six projects. The low reliabilities for the Aggregating DG and Turnkey Pump projects were not surprising due to the small sample of reviewers. However, the low reliability found for the 21st Century HVAC project was not anticipated and may have been due to a lack of understanding about the goals of the consortium and NYSERDA's relationship to these goals. Next, for each project, the scores of the rater whose elimination resulted in the highest AMICC were discarded, resulting in all projects, except 21st Century HVAC, having reliabilities in the acceptable range.

Table 3. AMICCs, by Project

Project	AMICC	AMICC Deleting One Rater
21st Century HVAC Research Consortium	0.000	0.061
Aggregating Distributed Generators	0.244	0.410
Continuous Ambient Particulate Monitor	0.528	0.628
Truck Stop Electrification	0.663	0.763
Turnkey Pump and Compressed Air Program	0.235	0.510

Conclusions from Phase 1

The Phase 1 study was instrumental in providing confidence about the applicability of the criteria to the portfolio of projects. The pattern of the responses indicated that the reviewers could adequately differentiate among the criteria. An additional finding was that the cost of gathering the information for the accomplishments packet was higher than expected. For example, for the 21st Century HVAC project, a total of 39 projects were listed and described, along with information about the knowledge gained, presentations resulting from the project, and the number of web site hits received by the paper. Applying this methodology to all the individual projects in the portfolio would be cost prohibitive.

Phase 2

In the second Phase, the framework was applied to whole programs, rather than individual projects. Two programs were selected:

- Combined Heat and Power (CHP) Program. This program provides funding to projects that demonstrate CHP at individual sites. Spending totaled $53 million and average funding per project was approximately $500,000.
- The Environmental Monitoring, Evaluation, and Protection (EMEP) Program. This program provides funding for environmental research. Spending totaled $21.8 million and average funding per project was approximately $400,000.

For the CHP program, data were gathered through document reviews and discussions with the NYSERDA project managers. In addition, results of market assessment surveys conducted in 2004 showing change in awareness of CHP were included (NYSERDA, 2005). For the EMEP Program, program staff prepared and provided sections of the accomplishments packet using the template designed for Phase 1.

For CHP, data were collected on all six assessment criteria. For EMEP, commercialization progress, energy, and economic benefits were omitted because these components are not major issues for the program. Furthermore, for EMEP, the impact of the program on policy makers was added to the knowledge dissemination scale. The data were organized into accomplishments packets.
The program manager of the CHP Program was asked to provide names and contact information of individuals representing various stakeholder groups including:

- CHP developers and system installers
- Architects, engineers and designers involved in recommending or specifying CHP systems
- Owners of CHP-eligible facilities with electric and thermal energy loads
- Utilities, governmental units and other stakeholders, including heads of other CHP programs

Five reviewers were selected from this pool.

The pool of potential reviewers for the EMEP Program was limited due to the program's scientific focus. In the end, six reviewers were selected for the panel consisting of one EMEP science advisor, one EMEP program advisor, one former EMEP program advisor, and three individuals unrelated to the EMEP program (one from EPA, one from NARSTO[2], and one from a regional organization).

In Phase 2, program staff made presentations of their programs explaining the contents of the accomplishments packet. In the case of EMEP, the meeting with reviewers was held at NYSERDA offices and lasted approximately five hours. In the case of CHP, the presentation was conducted via teleconference and lasted for approximately 60 minutes. During these meetings, the reviewers asked a number of clarifying questions.

Following the presentations, program staff left the room and the peer reviewers were allowed to discuss the program among themselves with guidance from a member of the evaluation team. This session lasted for approximately 60 minutes. Following the closed-door session, program staff were invited back and the reviewers provided initial feedback and asked follow-up questions.

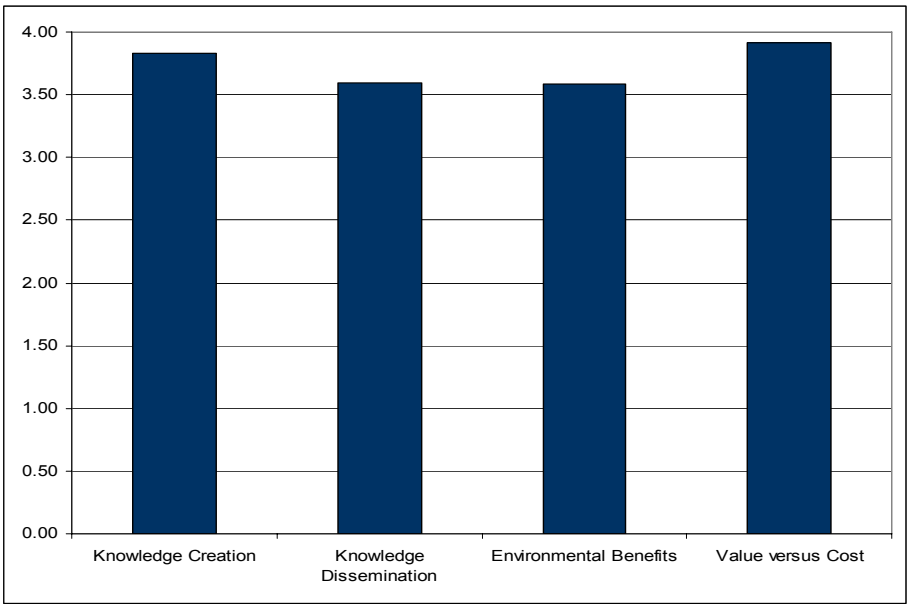
After the presentations, the reviewers were asked to individually score the program in private using the rating form and to provide as many comments as possible so that the staff could benefit from their observations.

EMEP Program Results

The average scores for the EMEP Program are shown in Figure 3. As with Phase 1, the scores were very high. The subcomponent scores are presented in Figure 4. The subcomponent 2b (Target Audience), while still quite high, is the lowest relative to the other subcomponent scores.

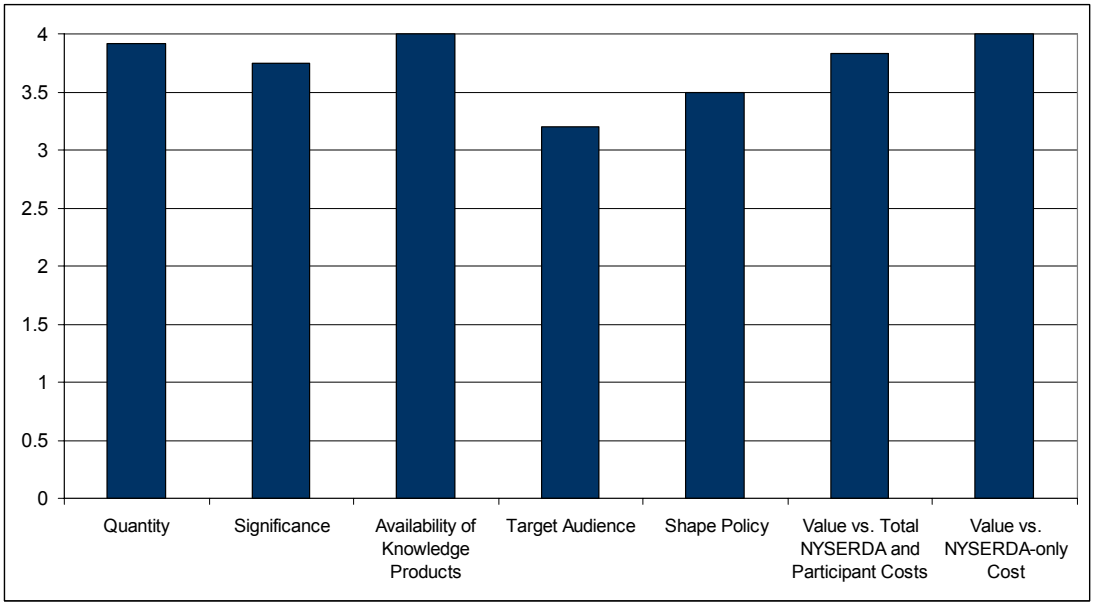
[2] North American Research Strategy for Tropospheric Ozone.

Figure 3. Scores for EMEP



The rating scale ranged from 0 to 4.

Figure 4. Subcomponent Scores for EMEP



The rating scale ranged from 0 to 4.

An analysis of the peer reviewer comments showed that they agreed that the program has produced a large number of papers published in quality journals; has done an exemplary job of making results available to multiple audiences ranging from the scientific community to policy makers; and the reviewers applauded program staff's efforts to translate and synthesize key findings. The reviewers pointed to the adoption of mercury legislation, adoption of the acid rain program, and the quantity of EMEP research cited in U.S. EPA PM Criteria Documents as indicators of the program's impact on policy.

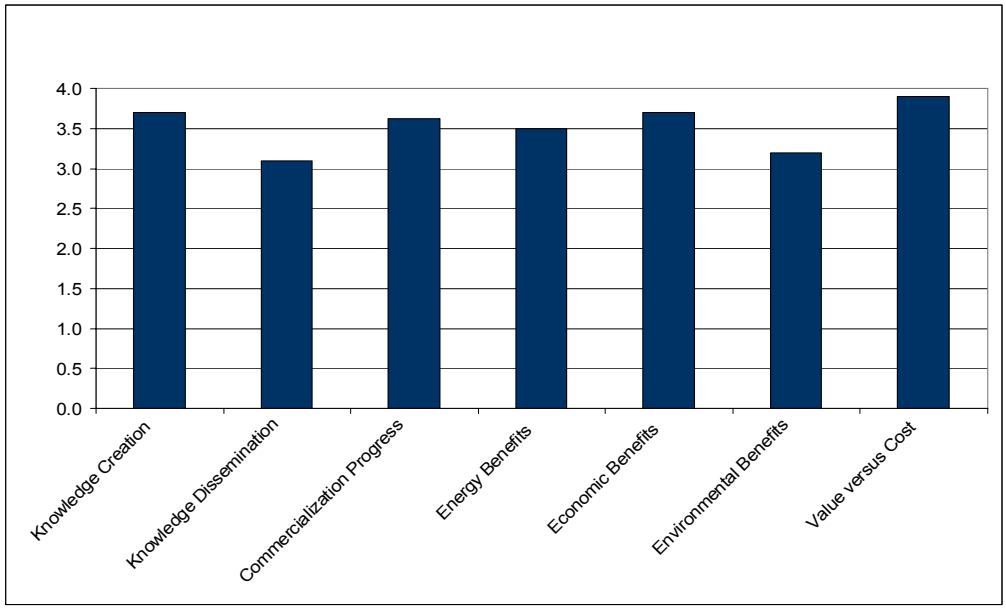
Suggestions for improvements for the program included focusing more strongly on the policy implications of the results, making greater efforts to engage the industrial sector, conducting surveys to determine whether the target audiences have been reached, and considering supporting efforts to quantify (in dollar terms) the potential environmental impacts of their work.

CHP Program Results

The average criteria scores for the CHP program are shown in Figure 5. All exceed 3.0. The subcomponent scores are shown in Figure 6. As with the EMEP results, the score for 2b (target audience) was the lowest. According to the comments, the reviewers' perceptions of the program were very positive. They agreed that the program provided high value relative to the small program budget. With respect to knowledge creation, the reviewers pointed to the diversity of the projects as an important feature. With respect to knowledge dissemination, the usefulness of the web sites had mixed reviews. Recommendations were made with respect to more refined targeting of key groups including policy makers, architects, manufacturers, and industry associations.

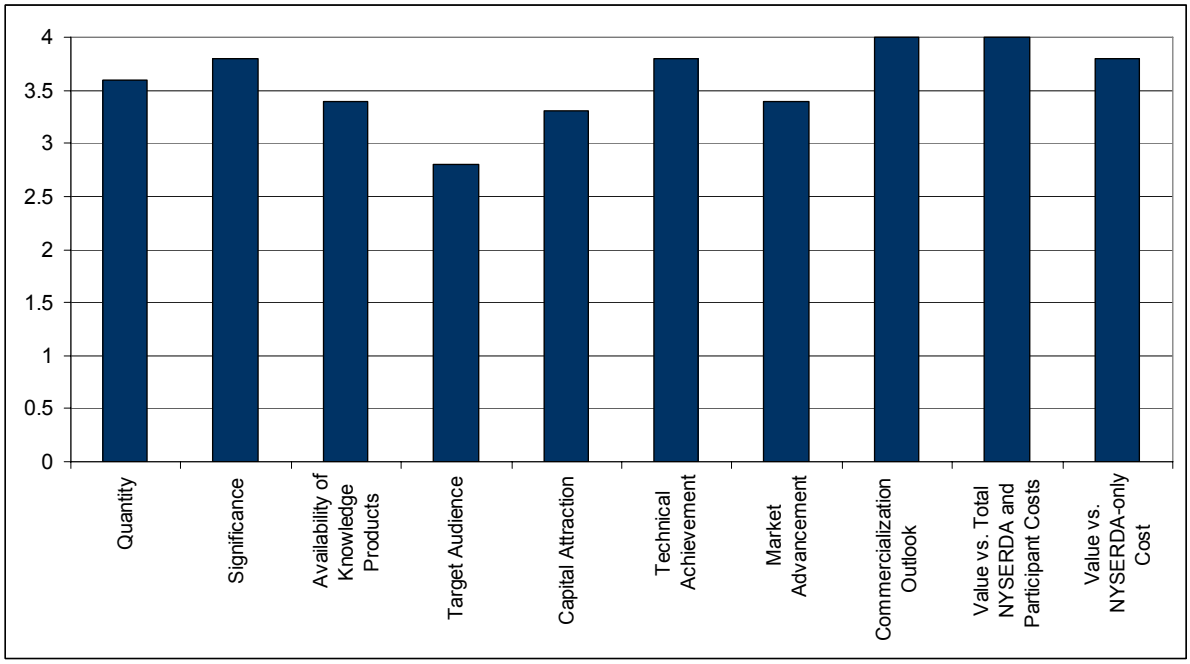
With respect to commercialization progress, key outcomes referenced by the reviewers included resolution of interconnection issues in Con Edison service area and creation of information that will avoid expensive mistakes by CHP developers and owners. With regard to energy and economic benefits, the reviewers felt that the demonstrations will result in replications that will produce significant benefits. Reviewers recommended that staff provide information to the public that will enable them to better quantify energy and economic benefits.

Figure 5. Scores for CHP



The rating scale ranged from 0 to 4.

Figure 6. Subcomponent Scores for CHP



The rating scale ranged from 0 to 4.

Phase 2 Follow-up Survey

Peer reviewers were asked, on a scale of 0 to 4 (with a 4 being the most positive):

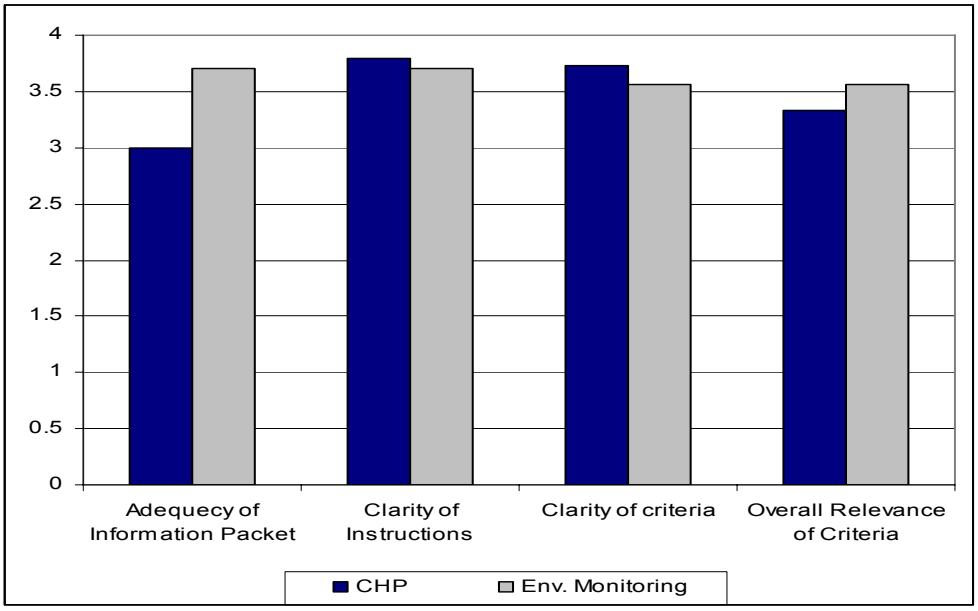
- How adequate, in terms of the quality, breadth, and depth, was the accomplishment packet in allowing them to make informed assessments and provide supporting comments?
- How clear were the peer reviewer instruction and response forms?
- How clearly defined were the six assessment criteria?
- How relevant were the six assessment criteria for documenting the accomplishments and impacts of the program?

The results are presented in Figure 7. The ratings for both programs equal or are above 3.0.

Conclusions from Phase 2

The results of the Phase 2 work reveal that the framework can be applied at the program level. The projects in the EMEP Program are less homogeneous than the CHP projects and required more effort to summarize. The EMEP program staff spent many hours distilling the learning that resulted from the various projects. This activity could not be assigned to evaluators. For the product development program, the lack of a model for determining economic impacts hindered development of an accomplishments packet.

Figure 7. Follow-up Survey Results for Phase 2



The rating scale ranged from 0 to 4.

Next Steps

Based on the favorable impressions of the selected programs received from the reviewers, no changes to program designs and deliveries are anticipated. The procedures used in this analysis will be reviewed in the second half of 2007 along with other procedures for R&D evaluation, particularly, methods designed to assess economic impact of product development projects. Possible hypotheses to be addressed include whether NYSERDA's product development programs have influenced companies to remain in New York or locate in New York as a result of NYSERDA funding. Another hypothesis asks whether companies assisted by NYSERDA grew more rapidly than similar companies in other states that do not provide R&D funding. The mechanisms for the differing rates of growth may be examined, perhaps through case studies.

References

Boardman, A. E., D. H. Greenberg, A. R. Vining, and D. L. Weimer. 1996. *Cost-Benefit Analysis: Concepts and Practice.* Upper Saddle River, NJ: Prentice Hall.

Bozeman, B. & J. Melkers. 1993. *Evaluating R&D Impacts: Methods and Practice*. Norwell, MA.; Kluwer Academic Publisher.

Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions, 2nd Edition*. New York: John Wiley & Sons.

Lee, R., G. Jordan, P. Leiby, B. Owens, and J. Wolf. 2003. "Estimating the Benefits of Government-Sponsored Energy R&D." *Research Evaluation* 12 (3):189-195.

Link, A. N. 1996. *Evaluating Public Sector Research and Development*. Westport CT: Praeger Publishers.

Mahajan, V. and R. A. Peterson. 1985. *Models for Innovation Diffusion.* Newbury Park, CA: Sage Publications.

Mahajan, V., E. Muller, and R. K. Srivastava. 1990. "Determination of Adopter Categories by Using Innovation Diffusion Models." *Journal of Marketing Research* 27: 37-50.

McGraw, K. O. and S. P. Wong. 1966. "Forming inferences about some intraclass correlation coefficients." *Psychological Methods* 1 (1): 30-46.

NYSERDA. 2005. "DG-CHP Market Characterization and Market Assessment and Causality Study." Prepared by Skumatz Economic Research Associates, Inc., Summit Blue Consulting, LLC, and Quantec, LLC, Project Number 7721, May 2005.

Ridge, R. 2003. "Evaluation of the 2002 California Emerging Technologies Program." Submitted to Southern California Edison.

Ruegg, R. and I. Feller. 2003. *A Toolkit for Evaluating Public R&D Investment: Models, Methods, and Findings from ATP's First Decade*. Prepared for the Economic Assessment Office, Advanced Technology Program, National Institute of Standards and Technology, Gaithersburg, MD.

Shipp, S., A. Kirtley, and S. McKay. 2004. "Evaluating Individual Projects and the Portfolio of Completed Projects in the Advanced Technology Program." Unpublished manuscript.