4

Program Evaluation

Resources for federal earth science education and training programs are generally limited, so it is important for agencies to invest in programs that work. Program evaluation provides a means for determining whether a program is succeeding and why. However, only a few of the education and training programs considered in this report have been formally evaluated or are structured in a way that facilitates evaluation, making it difficult to address Task 3 (identify successful programs) or Task 4 (determine what made these programs successful) as formulated. This chapter describes effective methods for evaluating programs, the limitations of evaluation approaches used in the federal earth science education and training programs considered in this report, and evaluation of these programs in the context of the Chapter 3 framework of education and training opportunities. Evaluation at each stage of the framework is illustrated with examples of effective practices, drawn from the literature, workshop discussions (Box 4.1), and other sources.

USING LOGIC MODELS FOR EVALUATION

Program evaluations generally focus on understanding program goals, establishing criteria for success, and gathering data to compare program performance to the criteria for success (NRC, 2009). Both formative evaluation (done while the program is under way with the goal of improvement, usually for internal audiences) and summative evaluation (done at the end of a program to determine its worth, often to external audiences; see Scriven, 1991) are needed to help providers develop effective programs and to determine the extent to which those programs met stated goals. Logic models are commonly used in program evaluation to understand how the program is supposed to work (e.g., McLaughlin and Jordan, 1999). They define who the program is trying to reach and what it is trying to achieve, and describe how to translate program resources into near-term results and long-term impacts. Logic models are often represented graphically as shown, for example, in Figure 4.1.

The logic model consolidates information on the inputs, activities, outputs, and outcomes of the program. Inputs are the resources used, such as people, time, or exhibit space. Activities are what the program does, such as attract visitors, air on television, provide summer experiences, or teach

BOX 4.1 Workshop Discussions on Criteria for Evaluating Program Success

Key points raised by individuals at the workshop included the following:

- Success can be defined in many ways (e.g., short term vs. long term, individuals vs. cohorts vs. the organization vs. the profession).
- Criteria for success depend on the goals of the program.
- Measuring the impact of informal programs as well as the connectivity among programs and between programs and a career path is difficult.
- Suitable performance measures include both quantitative data (e.g., number of participants) and qualitative data (e.g., depth of experience) and trends.

Example criteria included the following:

- Nature of the opportunity (e.g., career relevant, culturally relevant, hands-on, real world)
- Number of participants
- Diversity of participants, partners, or the resulting workforce
- Appropriate time and effort to achieve stated goals
- Use of best practices
- Increase in participants' earth science knowledge, skills, or identity with the field
- Intervention of program at critical junctures
- Connectivity of opportunities to keep participants moving along earth science pathways
- Ability to obtain other support or partners (professional societies, private companies, universities)
- Preparation of participants for employment
- Sustainability or longevity of the program
- Ability to scale from local to regional or national interests and issues

skills. Outputs are the immediate, tangible results of the program, such as the number of visitors who viewed the exhibit or the new skills learned by students. Outcomes are the longer term changes that the program aims to achieve. Earth science education programs generally aspire to three types of outcomes: awareness, engagement, or professional preparation.

To determine whether a program has achieved its objectives, each outcome variable must be measured either for a group of individuals before and after they participate in the program or for participants and an appropriate group of nonparticipants. Many measures of baseline awareness, engagement, and professional preparation can be made, but some form of survey or pretesting is likely to be needed to assess an individual's change. For example, one might test geological knowledge before a student took an upper-level earth science course, and then measure the student's geological knowledge after that course was completed.

To determine why a program worked or did not work, rather than just whether it did, the evaluation covers the activities themselves. For example, did visitors who spent more time at a geological exhibit show greater awareness on leaving it than those who spent less time? Did it make a difference whether they participated in hands-on elements in the exhibit? Examining the organizational context of a program may also provide important insights on why some programs work and others do not. For example, are programs that work with educational standards movements in schools more effective than those that blaze their own pathways? Best practices can be developed from program activities that have been demonstrated to produce the desired outcomes.

Measuring short-term outcomes is easier than measuring long-term outcomes, but the latter are more important for determining whether a program is meeting its goals. Follow-up after an

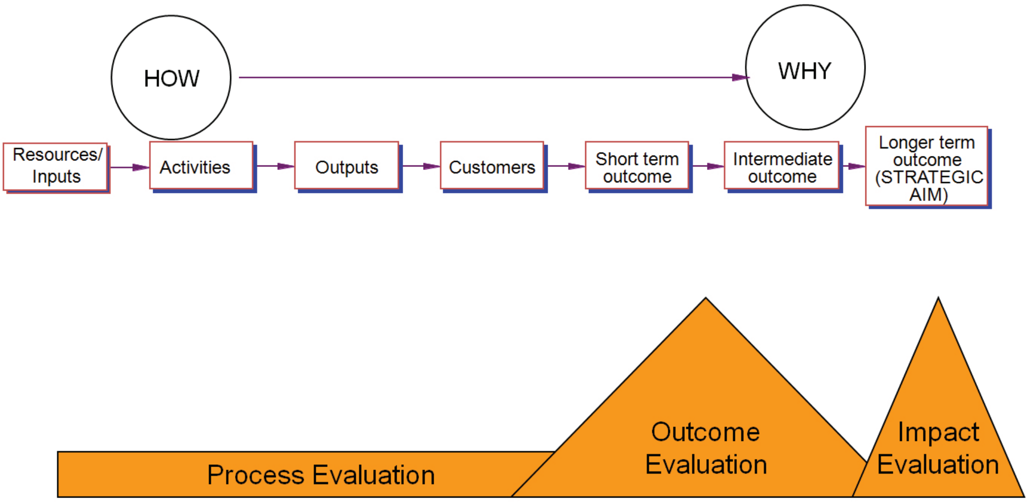

FIGURE 4.1 Example of a logic model illustrating the causal relationships among program elements (boxes) and evaluation stages (orange shapes), which show how the program works and whether and why it succeeds in generating results. SOURCE: Adapted from a 2005 presentation by Federal Evaluators (Evaluation dialogue between OMB and federal evaluation leaders: Digging a bit deeper into evaluation science), www.fedeval.net.

appropriate period of time is therefore important. Tracking individual participants over time is ideal, but even the best surveys lose track of some participants, and participants often lose interest in responding to requests for information. Surveys across similar programs may partially compensate for these problems at the level of individual programs, and they are also more cost-effective.

AGENCY PROGRAM EVALUATION

Two of the committee's tasks concern the evaluation of federal earth science education and training programs. Task 3 was to identify criteria for evaluating success and, using those criteria and the results of previous federal program evaluations, to identify examples of successful programs in federal agencies. Task 4 was to determine what made those programs successful. Important sources of information for these tasks were the workshop discussions (Box 4.1) and the written responses of program managers to the following questions:

1. What are the key goals or outcomes for the program?
2. How is the program evaluated?
3. What are the major successes of the program and what criteria are used to measure success?
4. What things have been essential to the program's success?

The answers to these questions revealed a wide range of criteria for success and evaluation approaches (see Appendix D). As noted above, criteria for success depend on the specific goals of the program. Thus, no single set of criteria can be developed to determine the success of all federal earth science education programs considered in this report. Rather, a comprehensive evaluation approach is needed to demonstrate program success.

Evaluation approaches used by the agencies range from informal assessments by an agency manager or principal investigator to rigorous external review. Few programs considered in this report have been designed to facilitate evaluation (Box 4.2) or collect the data necessary to determine whether the program succeeded or how to improve, sustain, or expand it. Even when data are collected, they are commonly not ideal for evaluation purposes. In addition, the formulation of goals and criteria for success poses problems for evaluation. Some of the stated goals are too broad to measure (e.g., improve understanding, build a community). In some cases, program goals are narrow (e.g., increase the number of participants), but the evaluation criteria are simple enumeration measures, which provide only limited information on program success. Only a few programs try to measure the impact of their program toward long-term, strategic aims (e.g., recruiting and retaining minorities, attitudes toward context-specific activities). Finally, the criteria do not always match the stated goals. For example, measuring participant satisfaction with the program does not indicate whether more students chose an earth science career. The mismatch of goals and measures confounds the ability to define program-level criteria for success.

External evaluations have demonstrated the success of the Opportunities for Enhancing Diversity in the Geosciences (OEDG) Program, the effectiveness of the selection process for Science to Achieve Results (STAR) fellows, and the progress toward achieving Educational Partnership Program (EPP) goals (Box 4.2). Other federal programs considered in this report cite successes (e.g., participants obtain earth science positions; see Appendix D), but the program information provided by the agencies was insufficient for the committee to make an independent determination. The lack of suitable evaluation data across programs underscores the importance of incorporating evaluation into the program design. By using a logic model in the context of the Chapter 3 framework of education and training opportunities, it would be possible to evaluate success at several levels: (a) whether a program is achieving its particular goals; (b) a program's contribution to increasing awareness, engagement, or professional preparation; and (c) a program's contribution to preparing a skilled and diverse workforce, including which programs work for which target groups and under which circumstances. Program evaluation in the context of the framework is described in the next section.

Because the committee lacked the robust data needed to choose successful examples of federal earth science education and training programs, it could not offer insight on why these programs are successful (Task 4). In assessing their own programs, managers identified several factors for success. The most common were stable funding, cost sharing, the commitment of agency managers or principal investigators, and partnerships. Agency support and community outreach were also important for many programs. Some managers highlighted program design—such as a good fit between participants and providers, flexibility, or institutionalization—as important for success. All of these factors are reasonable and consistent with workshop discussions, which also identified the involvement of families and the motivation for mentors (e.g., recognition of service) as important factors.

PROGRAM EVALUATION IN THE CONTEXT OF THE FRAMEWORK

Key elements of logic models (inputs, activities, outputs, and outcomes) and effective evaluation practices for programs at different stages of the framework are described below. The discussion is illustrated using exemplars that embody at least some elements of logic models in their evaluation.

Awareness

Awareness activities (e.g., formal education, after-school programs or clubs, earth science exhibits in museums) are designed to attract individuals to earth science, often through their own choice to participate. Participants include students and families of students in elementary school through high school, with researchers and scientists providing the content (inputs to the logic

BOX 4.2 Formal Evaluations of Federal Earth Science Education and Training Programs

Most of the federal earth science education and training programs considered in this report use relatively informal evaluation methods (see Appendix D). A few have undergone a more rigorous external review, either as individual programs (e.g., National Science Foundation [NSF] OEDG Program, Environmental Protection Agency [EPA] STAR Graduate Fellowship Program) or as part of a broader education portfolio (e.g., National Oceanic and Atmospheric Administration [NOAA] EPP, NSF Research Experiences for Undergraduates [REU] Program). The methods and results of these formal evaluations are summarized below.

NOAA Educational Partnership Program. A National Research Council review (NRC, 2010) found that a variety of evaluation methods were used for NOAA educational programs, ranging from no formal evaluation (e.g., EPP) to an outcome-based summative evaluation. The NRC evaluated the EPP using information provided by NOAA or gathered in interviews of NOAA staff and site visits. The evaluation found that the EPP had made progress toward achieving its goals. It significantly increased the number of African American Ph.D. graduates in atmospheric and environmental sciences, and many of these graduates took jobs as NOAA scientists. The program also supported more than 150 research collaborations involving NOAA and minority-serving institutions.

NSF Research Experiences for Undergraduates Program. A 2006 evaluation carried out by SRI International examined NSF's REU and other undergraduate research programs (Russell et al., 2006). The effectiveness of the earth science REU program was not specifically examined. The evaluation used surveys of participants and recipients of bachelor's degrees to assess the characteristics of participants, why faculty and students choose to participate, and the impacts of different types of research experiences on students' academic and career decisions. The results showed that undergraduate research experiences increased participants' understanding of the research process and confidence in their ability to conduct research. The experiences also raised their awareness of STEM career options and informed their graduate school and career decisions. Among the report's recommendations was that evaluations could be strengthened by participant feedback on program strengths, weaknesses, and possible improvements.

NSF Opportunities for Enhancing Diversity in the Geosciences Program. The American Institutes for Research assessed the OEDG Program in 2010, based on their annual reviews of the impacts and rigor of evaluation activities of OEDG projects (Huntoon et al., 2010). The report identified successful OEDG projects as well as those that could not demonstrate success because of poor evaluation or data collection, and used these examples to develop best practices. Overall, the review found that the OEDG portfolio has produced an impressive array of successes in meeting OEDG Program goals, which are primarily aimed at exposing or involving underrepresented minorities in earth science. The report also made recommendations for improving data collection and evaluation of OEDG projects (e.g., requiring that proposals identify goals, outcomes, and an evaluation plan; documenting demographics of providers and participants; monitoring impacts).

EPA Science to Achieve Results Graduate Fellowship Program. An NRC review committee developed four metrics and gathered information, primarily surveys of former fellows, needed to evaluate them (NRC, 2003). The metrics focused on the selection process and outcomes (completion of a degree, publication of research, and a career in environmental science). The review found that the program's peer review process was effective in selecting high-quality fellows. Nearly all recipients completed their research and received a degree, and most had at least one peer-reviewed publication about their fellowship research. In addition, most were employed in an environmental science field. The report recommended that EPA collect information to quantify these results and better document the success of the program.

model) through intermediaries such as curriculum developers, exhibit designers, video producers, and print editors.

The goals of participants and providers differ for awareness activities, as do the outputs. In general, participants are looking for "fun" through positive interactions with peers and adults in novel contexts, while providers are looking to share research findings and the excitement of discovery or creation (Dierking et al., 2004). If an awareness activity is successful, participant outputs include enthusiasm and excitement for the positive interactions and some satisfaction for knowledge gained. The provider outputs include the number of individuals participating in the awareness activity and the participants' attitudes, intentions, and satisfaction with the activity.

Free-choice learning opportunities (i.e., those that take place outside the classroom) are a productive area for federal agencies to raise student awareness, but outcomes can be difficult to measure. Falk and Dierking (2000), for instance, noted that visitors to a museum exhibit often have difficulty expressing what they learned, unless they are asked to provide their own descriptions of the content of an exhibit. Furthermore, it is difficult, if not impossible, to randomly assign individuals to treatment groups (i.e., those receiving a specific intervention or experience) and control groups (i.e., those not receiving the specific intervention or experience) if they are voluntarily approaching a learning opportunity (Gaus and Mueller, 2011; Tucker et al., 2011). Observational, survey, or interview methods can return data useful for evaluation (Hein, 1998), but these methods are often expensive.

In the absence of adequate resources, the simplest method for evaluation is enumeration: counting participants or characteristics of participants. Enumeration data are useful for determining the scope and character of the participant pool, but they provide little information on how well an awareness program is working (Korn, 2012). To determine outcomes, the intentions of the program developers have to be aligned with the intentions of the participants through planned cycles of learning and practice. In such cycles, steps taken for planning, action, evaluation, and reflection are documented to show how results, drawn from evaluation data of different types, can be matched to the overall effort.

Another best practice is to carry out audience research. Researching the needs, interests, motivations, expectations, and learning styles of the intended audience enables the program design to be calibrated to the mission of an agency relative to the transaction (Seagram et al., 1993), in this case, raising awareness of earth science. Through audience research, agencies can generate evaluation data that match program content to the needs, interests, and capabilities of the intended audience (Kelly, 2004). Recruitment of participants is a critical and a constant activity, and provider organizations that share participant goals and accommodate group learning styles are among the most successful (Dierking et al., 2004).

Example Evaluations of Awareness Programs

Many of the earth science awareness programs discussed at the workshop employ enumeration of participants as the primary evaluation mechanism (e.g., NSF's Geoscience Education Program, USDA's Agriculture and Food Research Initiative programs). A few programs also make an effort to understand what participants have gained. For example, the National Park Service's (NPS's) National Fossil Day includes an online survey that allows participants to share what aspects of the program met their expectations and what they took away from the experience. Such efforts enable a closer alignment of the goals of the provider with the goals of the participants.

A comprehensive evaluation strategy is being employed by the Trail of Time project, an NSF–NPS–university collaboration not discussed at the workshop (Karlstrom et al., 2008). The project is aimed at helping visitors interpret Earth history along the south rim of the Grand Canyon. The project's evaluation plan includes both formative and summative evaluation, adjusting the content and design of exhibits based on participant learning outcomes. Although limited by sample size and potentially intrusive to the participant experience, the robust evaluation design allows providers to match content to participant motivations, capturing fine details of participant responses that would otherwise be lost.

Engagement

Engagement activities (e.g., earth science projects at science fairs, enrollment in an earth science major) provide opportunities for participants to develop their understanding of the Earth and the nature of earth science. Provider inputs to the logic model include specific content knowledge and skills as well as pedagogic expertise in designing engaging experiences. Outputs include participants' increased motivation to engage in learning activities beyond the formal science curriculum, increased understanding, and a more complete sense of ownership of a specific work product, project, or artifact through the application of new skills. Outputs for providers include the development of scientific habits of mind by participants, helping them to understand through participation in a professional community what it takes to become a scientist. Providers usually seek to enumerate participant characteristics, but they can also provide feedback that would further refine the interests of participants. Such feedback can be a critical incident that draws students into earth science.

The range and complexity of engagement activities present challenges to evaluation because short-term outputs may differ substantially from long-term outcomes. Nevertheless, common methods of assessment can generate useful data. The assessment systems used by state education agencies, for example, provide substantial data on the knowledge gained by students through formal instruction and some data on scientific skills. Positive feelings are commonly used as a proxy for assessing interest and motivation, but better indicators are available, including time on task; stored knowledge and value; responses to novelty, challenge, and complexity; and goal setting and self-regulation (Renninger, 2011). Evaluation models for experiential learning contexts (e.g., Fetterman and Bowman, 2002; Cachelin et al., 2009) can be used to assess knowledge, skills, and feelings. These approaches provide a strong basis for determining how to successfully engage students in earth science.

Example Evaluations of Earth Science Engagement Programs

Some federal engagement programs considered at the workshop specify outcomes focused on local, place-based needs (e.g., NSF's OEDG and Geoscience Teacher Training programs). Two programs use critical incident theory to understand how engagement opportunities influence subsequent academic and career choices. The NPS Geoscience-Teachers-in-Parks Program documents teacher feedback, the persistence of teacher's use of instructional materials, and student familiarity with the material to determine the importance of critical incidents in students' academic careers. Some projects in NSF's OEDG Program use critical incident theory to understand how and when students choose to engage in earth science and then pursue a career. The OEDG Program is a good example of a federal earth science education program that has been able to demonstrate success through a good evaluation strategy (Box 4.2).

An example of a successful engagement program not discussed at the workshop is the International Ocean Drilling Program's School of Rock, which is supported by NSF and uses data from ocean floor drill cores to document changes in the Earth system over time. A pilot evaluation of the ocean-going research experience was based on daily teacher connections journals, which recorded past experiences and knowledge, people, memorable events, instructional ideas, frustrations, and missed connections (St. John et al., 2009). A subsequent summative evaluation was based on interviews of teachers, who reflected on the efficacy of program implementation in their classrooms, and continued communication with participants. A 5-year follow-up (Collins et al., 2011), conducted through online surveys, included enumeration, an analysis of teacher lesson plans, and opportunities for professional development enabled by the experience. This evaluation identified critical elements of the program (e.g., teacher access to data and scientists) and acquired skills (e.g., knowledge transfer) and attitudes (e.g., science as a collaborative enterprise) through the material presented in classroom lessons.

Professional Preparation

Professional preparation opportunities (e.g., formal education, participation in professional society meetings, involvement in research, internships, postdoctoral fellowships) are aimed at a wide range of participants. High school students and undergraduates seek opportunities that provide a taste of the profession and help them acquire the knowledge and skills needed for an earth science career. Undergraduate and graduate students and new Ph.D. recipients seek opportunities that provide the full workplace experience or help them identify a suitable introductory position. These diverse audiences and objectives require a range of evaluation approaches. Approaches used in the two most common professional preparation activities—research experiences and internships—are described below.

Research Experiences

Among the reasons students get involved in undergraduate research are to experience what it is like to do science, to test their interest in an earth science career, or to develop specific job-related skills (e.g., Manduca, 1999). Providers of these experiences, namely researchers, seek to promote research activities, impart context-specific skills and scientific habits, and obtain results from specific learning goals. Inputs to the logic model include participants' interest and enthusiasm for "doing" science as well as providers' research interests and desire to mentor students as they enter the field. Outputs for undergraduate research projects include new knowledge and skills, increased persistence and interest in science careers, graduate school attendance, and higher graduation rates, especially among groups underrepresented in science (Thiry et al., 2011).

Calibrating the goals of undergraduate research with student expectations remains a significant evaluation challenge, although provider outcomes more consistent with participants' interests have been documented in supervisor evaluations of participants (Hunter et al., 2006). Relatively few empirical studies have examined whether students with undergraduate research experiences acquire higher order thinking skills in science (Kardash, 2000). Lopatto (2007) used the *Survey of Undergraduate Research Experience* (Lopatto, 2004) to investigate whether undergraduate research enhanced students' educational experience and attracted or retained them in science, technology, engineering, or mathematics (STEM) career paths. The surveys showed that the undergraduate research experience clarified or solidified students' graduate school plans. Participating students reported greater learning gains and a better overall undergraduate experience than nonparticipants. Students from underrepresented groups also showed greater retention rates than nonparticipating groups. These results were partly corroborated by Seymour et al. (2004), who found that participation in undergraduate research confirms students' prior career choices, increases their capacity to deal with ambiguity, and provides them with opportunities to take greater initiative for their own learning. Through a detailed review of the literature and a rigorous evaluation design, Thiry et al. (2011) found little evidence for the notion that participation in undergraduate research succeeds in recruiting students, attracting them to graduate school, or changing their choice of subjects. Thus, providers of professional preparation experiences may need to adjust their inputs into their logic model.

Research experiences are commonly evaluated by enumerating participation. For example, in a recent study, 85 percent of responding STEM graduates reported participating in some form of research experience (Thiry et al., 2011). However, a more effective approach is to match evaluation strategies to changes in experience format and duration. Research experiences range from projects with a research component that last a few weeks or a semester (Wagner et al., 2010; Gibson and Bruno, 2012) to fully immersive research experiences for undergraduates (e.g., Jarrett and Burnley, 2003; Gonzales-Espada and LaDue, 2006) to research at field stations and marine laboratories that last multiple semesters (Hodder, 2009). Efforts to define excellence in undergraduate research (e.g., research skills) and the logistics and infrastructure necessary to support high-quality work (e.g., Hensel, 2012) may help inform a comprehensive evaluation of undergraduate research experiences. Defining excellence requires both quantitative data (including enumeration of participants and their characteristics) and qualitative data (including surveys and interviews) and a careful matching of data to the goals of the program (Gonzalez-Espada and Zaras, 2006). Russell et al. (2006) concluded that there is no single way to define (and, by extension, to evaluate) the research experience, but that the sustained inculcation of enthusiasm for research provides the greatest impact.

Internships

Undergraduates seek internships to gain specific skills that will make them more competitive in the workplace, access to potential employers, and references to support their applications. For scientific internships, students seek broadly defined employment opportunities (Taylor, 1988) and the development of a scientific identity (Hsu et al., 2009). Providers, on the other hand, seek the successful completion of specific work products, the transfer of context-specific workplace skills, and access to a larger pool of suitable candidates for employment. Outputs include the acquisition of skills desired by the providers or themselves, a sense of ownership of the work product, and clarification of professional goals, even when the desired permanent job is not obtained. Providers gain work products at potentially lower costs, access to what they believe are top candidates for available positions, and satisfaction in providing a service to the profession. Reconciling the goals of the providers and the participants for evaluation purposes is aided by the transactional relationship between participants and providers.

The impact of internships on students and their hosts has been evaluated in a variety of ways, including interns' evaluations of their experiences, which provide useful feedback to the hosts (Morris and Haas, 1984; Girard, 1999), and supervisors' assessments of students' performance using the traditional academic grading structure (Cutting and Hall, 2008). Less available are clear evaluation findings that indicate whether the programs work or are cost-effective or whether interns gain knowledge, skills, and disposition in their chosen field (Schultz, 1981). The literature in science education (Schultz, 1981; Cutting and Hall, 2008; Hsu et al., 2009) and psychology (Shoenfelt et al., 2012) suggests that formative evaluation frameworks could be developed based on interactions between interns and their supervisors using an analysis of verbal transactions, work products, or written documentation. Key elements for summative evaluation include the appropriateness of the internship, provider and participant obligations and responsibilities, participant qualifications and expected competency gains, onsite supervision frameworks, and participant performance evaluation.

Example Evaluations of Earth Science Professional Preparation Programs

For the federal professional preparation programs discussed at the workshop, the most commonly employed evaluation strategy is the enumeration of participants. In addition to collecting enumeration data, the U.S. Geological Survey (USGS) Hydrologic Technician Internship Program, Youth Internship Program, and EdMap collect participant reports of satisfaction, provider evaluations of participants, and participant work products. The EdMap data show a relatively strong correlation of participant and provider responses on performance evaluations, onsite supervision frameworks, and obligations and responsibilities. However, the relationship between the expected competency gains and the appropriateness of internship opportunities is less clear. Adding an examination of work products and longitudinal tracking of participants as they move into the workforce would improve the evaluation with little added effort. Programs that collect these data include NOAA's Educational Partnership Program, which analyzes participant work products and tracks the transition of participants to the workforce, and NSF's Earth Sciences Postdoctoral Fellowships program, which collects some information on the workforce transition. Overall, providers that collect all of the information described (enumeration, self-reports, supervisor evaluations, work product analysis, and tracking) in a systematic, rigorous manner have a greater chance of aligning their goals and outputs with those of the participants.

SYSTEM-LEVEL EVALUATION

Evaluations at the various stages of the framework provide important information on how well an education and training program is achieving a goal of awareness, engagement, or professional development. Evaluations encompassing all activities in the framework could be used to find imbalances in effort and connections and gaps between activities at different stages of the framework. It could also provide a measure of the extent to which the portfolio of education and training programs offered by various organizations is changing earth science pathways.

In a system-level evaluation, the size and effectiveness of individual programs is viewed in the context of information about (a) levels of activity at various points along the path and (b) the status of the system objective. Broad indicators of program activities at various stages of the framework can be obtained by aggregating information from individual program evaluations. For example, the sum of earth science exhibits or classes and the number of people exposed to them can provide a measure of national awareness of earth science. Such measures can be supplemented with in-depth evaluations aimed at providing insight on the dynamics of the system at the various stages. Targeted program evaluations that measure activities and outcomes would increase understanding of how to create effective programs, and qualitative studies would show how individuals find the opportunities and what they learn from them.

As noted in Chapter 3, individuals travel different pathways to an earth science career, sometimes skipping stages or moving back and forth across stages of the system. A system-level evaluation would take account of the networks that help individuals find a path through the system. The presence, size, and interconnectedness of organizations in the various networks (e.g., university consortia, cultural and ethnic affinity organizations) can all be measured. Network analysis of the connections can be based on unobtrusive indicators such as Web links and common themes in public statements. Communication and dissemination efforts are particularly easy to measure, and they intersect with the awareness indicators described above.

SUMMARY AND CONCLUSIONS

Program evaluations provide a means for determining whether a program is succeeding and why. External evaluations have demonstrated successes in the OEDG, EPP, and STAR programs. The other federal programs considered in this report have not been evaluated and most were not designed to facilitate evaluation: some program goals are too broad to develop criteria for success; the goals and criteria do not always match; and the criteria and data collection emphasize what is easy to measure, not what the program is trying to achieve. These programs may be successful, but the data were too sparse and uneven in quality to make that determination. The difficulty of identifying successful programs (Task 3) and determining what made them successful (Task 4) underscores the importance of incorporating evaluation into program design.

Rigorous evaluation approaches commonly use a logic model to define who the program is trying to reach, what it is trying to achieve, what resources it requires (inputs), and how to translate program resources into near-term results (outputs) and long-term outcomes. Each program needs its own evaluation design and criteria for success. Enumeration, pre- and post-testing, observations of participants or providers, work product analysis, and determination of long-term plans and satisfaction with experiences are all useful tools for evaluation.

The framework of opportunities described in Chapter 3 can be used to conceptualize evaluation of individual programs and suites of programs with a collective goal of building earth science pathways to careers. Each stage of the framework (awareness, engagement, professional development) has its own input, activity, output, and outcome measures. Careful attention to input and activity measures would ensure that the goals of participants and providers are aligned. Measures across several fiscal years are commonly needed to assess long-term outcomes. Although more time-consuming and costly, long-term measures can demonstrate program impact as well as its sustainability.

A system-level evaluation, encompassing all activities within the framework or at a stage of the framework (e.g., engagement), could be used to identify imbalances in effort and gaps, enabling agencies to determine where future education and training efforts may be useful. Broad indicators of program activities could be developed by aggregating relevant information from individual program evaluations, and supplemented with targeted program evaluations aimed at understanding how to create effective programs. Network analysis of the programs in the system could reveal which connections among participating organizations help move individuals through the system, and qualitative studies would help show how individuals find education and training opportunities and what they learn from them.