2. Evaluation of Other High School STEM Enrichment Programs

This section of the report summarizes ways that STEM programs have approached evaluation, and some of the pitfalls that prevent the field from capturing truly reliable and cross-cutting data. Where possible, we have included literature on the successes of these types of programs. It is important to state upfront that published results that indicate participant progress in the sciences or overall program impact are difficult to find and, when found, are difficult to compare with results of other programs. Many individual STEM programs are either not required to publish their evaluation data or cannot publish evaluations because they have not received the appropriate human subjects permissions. However, some well-documented evaluations of pre-college STEM programs do exist, and some resources cited in section 7 of the report use anecdotal data to identify "what works."

2.1 Evaluation Approaches

Most programs conduct evaluations because they are required by a funder, because the program directors are interested in learning how their efforts are being received by the participants, and/or because the directors wish to assess long-term impacts. The National Science Foundation (NSF) usually requires that programs devote 10-15% of their overall budget to evaluation activities. Increasingly, even small funding organizations are requiring evaluations. This research is often conducted by third parties, mainly to enhance objectivity, but also because program directors often have competing demands for their time and lack expertise in program evaluation.

STEM programs engage in two main types of evaluations—*summative* and *formative* evaluations. Summative evaluations assess the impact of specific, measurable program goals, some of which may have been jointly set with the funder. Results are generally quantitative, but often include qualitative data that have been collected over the duration of the program. High-quality summative evaluations are generally expensive and have limitations, for example, the absence of comparison groups that would allow attribution of results to a specific intervention. However, if the studies use standard evaluation methods and a funder is willing to pay for dissemination, the results may be published. Compared with summative evaluation, formative evaluation of STEM programs is conducted with higher frequency. The insights gathered during a formative evaluation are meant to inform future iterations of the program rather than inform the field. In order to assist STEM programs in understanding both the value and process of program evaluation, NSF released an evaluation handbook for program directors in 1993.[4]

The Howard Hughes Medical Institute (HHMI) also funds several programs nationwide, including some of the ones included in this review. In July 2005, the first study on the evaluation of these programs was archived.[5] HHMI wanted to evaluate evaluation; literally, it wanted to determine if there were any consistencies across a set of pre-college STEM programs with highly diverse purposes, target populations, and operations. HHMI also wished to learn how evaluations were conducted. Participants in 35 pre-college STEM programs for both teachers and students

---

[4] Frechtling, J.A. (1993) *User-Friendly Handbook for Project Evaluation: Science Mathematics, Engineering and Technology Education.* VA: The National Science Foundation.

[5] http://www.nahsep.org/study.html

and a control group of programs, which were selected as finalists in the proposal review process but ultimately were not awarded a grant, completed surveys and responded to questions about their evaluations. This particular study of the evaluation of STEM programs is the most recent and most comprehensive study of its kind. Below are some notable highlights, although the entire study is worthy of examination:

> Almost all of the sites (n=31) were utilizing surveys/questionnaires. Several sites (n=27) were conducting observations (one project director indicated that the site had conducted informal observations following the implementation of a neuroscience curriculum in participating teachers' classrooms). Twenty-one sites conducted interviews; some of these were informal, representing a way to revise project components to be more effective in the classroom. Nineteen sites were using performance measures/participant portfolios, and 12 conducted focus groups.
>
> [At one site] project staff are still grappling with what key elements of their program most influence an increase in student achievement at the 15 schools with which they are working.[6]

General indicators of program success were shared by the NSF subset of STEM programs known as the Research Experiences for Teachers (RET) programs, where teachers are invited to work alongside scientists in the lab or in the field.[7] According to participants, success seems to be centered on two main indicators—collaboration with university scientists and/or fellow educators and increased confidence in, and awareness of, the scientific enterprise.

2.2 STEM Evaluation Results

Published results that demonstrate participant progress in the sciences or overall program impact are difficult to find and, when found, are difficult to compare with results of other programs, due to limitations such as the following:

- Almost no baseline data are gathered by programs across the literature. Baseline demographic, attitudinal and other pre-participant data, including data for the entire applicant pool could answer questions such as: Who applies? Who was chosen? Why? What changes in attitudes and intentions occur over time?

- The impact of STEM programs on the attitudes, aptitudes, and behaviors of participants may not manifest during the program itself, yet that is when most evaluation activities are conducted in order to gain the highest response rate for surveys, focus groups, and interviews.[8]

---

[6] Ibid.

[7] Emily Driscoll, RET at Northeastern University (2004) http://www.ret.neu.edu/NSTA-Dallas/Carousel.pdf

[8] Frechtling, J.A. (1993) *User-Friendly Handbook for Project Evaluation: Science Mathematics, Engineering and Technology Education.* VA: The National Science Foundation.

- Participants' self-reported data often exaggerate the positive effects of a given program because they have invested time and energy to participate.[9] Although this does not negate the value of probing participants about their perceptions of certain tactics, strategies or presentations, results need to be interpreted with caution.

- When data are collected over a considerable length of time, consideration may not be given to collecting the data with standardized instruments and interviews that will allow results to be pooled over time and allow correlations to be computed.

Nonetheless, some evaluation literature has been published despite these limitations. The following results from several programs are particularly worthy of mention. The HHMI study that queried its own program pool about evaluation practices also collected data on student motivation to study sciences. (See Table 1 for findings and a comparison with the national averages). Even though the motivating factors for students to pursue science in later educational levels are numerous, and sometimes even undetectable, HHMI did make an effort to include "the data only if they included a control group of similar participants or if results on the same students were collected before and after the intervention to demonstrate the change."[10]

Table 1—Percentage of students who became science majors post-program in selected STEM programs

Grantee	**%**
University of Cincinnati College of Medicine	83
University of Nevada School of Medicine	63
Robert C. Byrd Health Sciences Center of West Virginia University	59
University of Mississippi Medical Center	59
Cleveland Clinic	53
National average	32
National average for underrepresented minorities (most program participants)	5

It is interesting that many of these programs accept participants through a competitive recruitment process, tapping those who already have demonstrated aptitude or interest in the sciences. It is unclear how far along in their science studies the students were when each survey was taken.

[9] Cameron, J. & Pierce, D. P. (1994). "Reinforcement, Reward, and Intrinsic Motivation: A Meta-analysis." *Review of Educational Research, 64* (3), 363-423.

[10] http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=520842

Evaluation results from a 1993 study of the Dartmouth Thayer School of Engineering program showed that 75% of the 1993 teacher participants conducted an engineering experience or parts of that experience in their own classrooms and schools, using the pedagogy presented in the summer program. Further, 75% of participants made presentations to their colleagues about their program and classroom experiences within 16 months of the summer session, reaching an additional 974 teachers. The evaluation also found that 75% of student participants implemented the Dartmouth/Thayer problem-solving methods upon return to their high schools.[11] These findings show that because teachers were able to replicate this specific pedagogical practice, the goals and intent of this program were achieved.

In 1998, a multsite study of Scientific Work Experience Programs for Teachers (SWEPT) was funded with a four-year $1.6 million grant by NSF. The goals of this study were to measure the effect of SWEPT on students in the classroom, in a way that helps identify key variables, regardless of geographic location or particular facilities/personnel. The College of Physicians and Surgeons at Columbia University, coordinated this eight-site effort, summarized briefly below:

> Data has been collected on the more than 30,000 students who have been in the classes of participating teachers since 1993 and on approximately 600,000 students in the science classes of nonparticipating teachers in the same schools and science departments.
>
> The researchers found a three-fold increase in the number of students of participating teachers who undertake a competitive science project. The number of students participating in after-school science programs has grown from about 10 percent to about 13 percent in the classes of participating teachers, while the average in classes of non-participating teachers remained about the same at 3.5 percent. They also found a significant increase in the number of students of participating teachers who passed the science Regents exams. The researchers plan to submit their findings for publication.[12]

2.3 Resources

Program directors from the aforementioned HHMI study were asked to recommend print and online resources they found useful for planning and conducting their evaluation efforts, and these are footnoted here.[13] Another set of references, compiled by program directors and funders of SWEPT programs is also provided as a footnote.[14] A bibliography relevant to goal-setting, evaluation and effective professional development for science educators may also be useful.[15]

---

[11] http://fie.engrng.pitt.edu/fie95/4b1/4b14/4b14.htm

[12] http://www.cumc.columbia.edu/news/in-vivo/Vol2_Iss13_aug18_03/science-outreach.html

[13] http://www.nahsep.org/study_results#sites and http://www.nahsep.org/study_results#assets

[14] http://www-ed.fnal.gov/trc/program_docs/biblio_trp.html

[15] http://demeter.hampshire.edu/~manual/back.html#Gibson,%20Helen%20L.%201998.

3. Evaluation Methodology

3.1 Approach and limitations

The purpose of this evaluation was to assess the extent to which the ARC-ORNL Summer Math/Science/Technology Institutes met the program's stated goals. ARC also asked that AED staff compare the evaluation findings with similar programs. In addition, ARC asked AED staff to make recommendations for establishing an ongoing evaluation capability for the ARC-ORNL Summer Institute.

This evaluation employed a mixed-methods approach, including surveys that collected quantitative data through questions with fixed-choice responses. Qualitative data were collected through open-ended survey questions and through interviews with former student and teacher participants. In this way quantitative data are illuminated by more in-depth perspectives offered by participants.

This evaluation is a first step for ARC and offers an objective assessment by an outside evaluator. It is important to keep in mind that the Summer Institute program is relatively modest in scope, in that it is a one-time, two-week program without connection to the sending schools or the institutions of higher education to which the students may apply. While we have explored outcomes that ARC hopes to achieve, these outcomes are incredibly ambitious given the scope of the intervention. The real strength of this evaluation lies in its attention to the program's impacts as perceived by the participants.

It is nonetheless important to note some of the limitations of this evaluation.

Absence of a comparison group. One of the major limitations is the absence of a comparison group that would have made it possible to say with a high degree of certainty that outcomes in this report can be attributed solely to the Summer Institute. Also, because baseline data were not available, we were unable to measure, over time, changes in participants' knowledge, attitudes, intentions, and behaviors.

Small sample size. Each year the group that attends the Summer Institute is small—no more than 60 participants. In order to conduct quantitative analysis and analysis involving subgroups (e.g. gender or length-of-time teaching), it is necessary to have a large sample. Because we were unable to locate many participants despite intense efforts, the sample size remained small permitting only a few subgroup analyses.

Time needed to measure long-term outcomes. Outcomes such as completing higher education and beginning a career may take many years. Thus it made sense to look at college completion outcomes and employment for students who had attended ORNL Summer Institutes in 1997 and 1998. For those in later cohorts, it was necessary to explore shorter-term, mediating outcomes, such as high school completion and college enrollment.

Uncertain reliability of self-report. The survey findings are entirely based on self-report which may not be completely reliable. Evaluation staff were unable to confirm these responses either through record review or observation.

3.2 Student and Teacher Surveys

Survey questions were adapted, where possible, from questions used in evaluations of other pre-college STEM programs. New survey items were added when needed. The questionnaires were reviewed by ARC staff and pilot-tested on 2005 student and teacher participants. Student survey questions were designed to obtain basic demographic information, as well as information on educational attainment, career and employment choices, and the perceived influences of the Summer Institute on student attitudes and college-going. Teacher surveys collected demographic data and included questions pertaining to teaching experience and how the Summer Institute directly influenced teaching practices. Both surveys had open-ended questions to allow participants to describe in greater depth the influence of the Summer Institute on them and the aspects of their experience (e.g., people, projects, cultural programming) that they considered to be most influential.

Survey data collection commenced October 2005 and concluded in December 2005. Participants were given the option of completing hard copies of the survey and returning it in a pre-addressed, stamped envelope or completing the survey on-line.

3.3 Student and Teacher Interviews

During December 2005 and January 2006, staff conducted semistructured, 15-30 minute telephone interviews to explore selected responses to the survey in greater depth. The student interviews covered recruitment, in terms of how the student heard about the institute; the application process and his/her decision to attend; the overall experience of the Summer Institute; how the institute influenced college decisions; perceptions of college-going attitudes in their school and community, as well as the perceived presence or absence of support structures; and recommendations for improving the effectiveness of the institute. The teacher interviews covered teacher recruitment, how the institute influenced their approaches to teaching and interacting with students, their perspectives on whether the institute had an impact on their career, and recommendations for improvement.

3.4 Study Population

All 254 student and 132 teacher participants who attended the Summer Institute during the eight years spanning 1997-2004 were eligible to complete the surveys. One of the greatest challenges to conducting the evaluation was locating the participants, many of whom, especially students, may have moved away from home. We expected that some young women had married and changed names. Because the survey was conducted during the fall, we also needed to be able to contact students who were away at college. Accordingly, we employed a variety of techniques to ensure the highest possible response rate. These included searching the Web to confirm or revise contact information; e-mailing to addresses provided to ARC by participants (in the 2003-04 cohorts) or found on the Web; phoning the participants or their families; or enlisting the assistance of the sending school. In some cases, we asked participants we located to help find others who had attended with them.

Through the exhaustive use of these techniques, we were able to successfully "find" 63% of the students and 80% of the teachers on the lists ARC provided AED. AED staff sent surveys to all participants but sought to obtain the highest possible response rate from the “found” participants. Two to three weeks after the initial survey mailing, we sent reminder postcards and emails to the participants. In addition, staff made follow-up phone calls to all the teachers and students who

had not yet returned the surveys (and for whom we had correct contact information). We received surveys from 92 students and 71 teachers. Using a denominator of those with confirmed contact information, the response rates were 58% and 67%, respectively. Of the surveys returned, there were 89 usable student surveys and 67 useable teacher surveys. Response rates varied by cohort.[16] (See Tables 2 and 3.)

Interview samples were drawn from participants who responded positively to a question on the survey asking if they would agree to be interviewed. From these, we selected individuals who together would represent the diversity of participants on characteristics such as gender, year they attended the institute, and race/ethnicity. The student sample included students who had attended or were attending two- and four-year colleges, as well as those who were employed. The teacher sample included some who were early in their careers and others with many years of experience. (See Tables A2 and A3 in the appendix for a more detailed description of the interview samples.)

3.5 Response Bias

In order to determine whether there was response bias in our survey findings, we compared respondents with the "found" participants who did not complete the survey and with the entire group of participants on key variables, including gender, year attended, and the economic status of the county in which the sending school was located according to a classification system used by ARC.[17] A variety of statistical methods were used to determine whether there was any bias.[18] For students, there were no significant differences among groups by gender, but the more recent their entry into the program, the greater likelihood of their being found. However, cohort had no significant correlation with likelihood of responding to the survey. There were no significant differences among groups for teachers or students with regard to location of the sending school in a distressed county. With regard to teachers, there were no significant differences between the full group and those found or between found participants and respondents. (See Table 2 for data on students and Table 3 for data on teachers.)

---

[16] Our response rate based on the total number of attendees was 36% for students and 54% for teachers. As comparison, a response rate of 48% was achieved in a follow-up study reported in 1996 of 1985 participants of the Student Research Program, a science and engineering summer program for undergraduate students sponsored by the U.S. Department of Education. As reported in a U.S. Department of Energy Working Paper, "Impacts on Participants of DOE Research Participation Programs," prepared by Argonne National Laboratory and Oak Ridge Institute for Science Education (1996).

For another study, the STRIVE Teacher Research Associates Program, 1986-1991, a response rate of 83% of teachers was achieved one year following their participation in an eight-week program. By comparison, our response rate for 2004 teacher participants was 75%. Oak Ridge Institute for Science and Education (1992). *Assessment Summary, STRIVE Teacher Research Associates Program, 1986-1991*.

[17] ARC designates counties as economically distressed on the basis of low per-capita income and high rates of poverty and unemployment. The number of distressed counties changes from year to year, depending on conditions. For this evaluation, we used the county's ARC designation the year the participant attended the program. Between 1997 and 2004 the number of counties designated as distressed ranged from 90 to 120. The average number of counties in the region designated as distressed was 26%.

[18] Statistical methods included one-way ANOVA, bivariate Pearson correlations, two-tailed t-tests, and binary logistic regression (the latter because some of the variables were dichotomous).

Table 2—Comparison of respondents with all attendees and with attendees who were located but who did not respond to the survey–students

	All Attendees (n=254)		Located Attendees who were Nonrespondents (n=67)		Respondents (n=92)*	
	N	%	N	%	N	%
Gender						
Male	123	48%	30	45%	46	50%
Female	131	52%	37	55%	46	50%
Year						
1997	32	13%	5	7%	10	11%
1998	37	15%	10	15%	14	15%
1999	43	17%	7	10%	11	12%
2000	34	13%	10	15%	12	13%
2001	18	7%	6	9%	5	6%
2002	36	14%	10	15%	14	15%
2003	22	9%	7	10%	10	11%
2004	32	13%	12	18%	16	18%
Sending School in Distressed County**						
Yes	94	37%	26	39%	28	31%
No	160	63%	41	61%	63	69%

* Of the 92 respondents, 89 were used in analysis. Two students said they did not attend, and one returned home after attending for only a day or so.

** We could not identify sending school for one survey respondent who did not give a name.

Note: Some percentages in this table do not equal 100% due to rounding.

Table 3—Comparison of respondents with all attendees and with attendees who were located but who did not respond to the survey–teachers

	All Attendees (n=132)		Located Attendees who were Nonrespondents (n=35)		Respondents (n=71) *	
	#	%	#	%	#	%
Gender						
Male	48	36%	12	34%	28	40%
Female	84	64%	23	66%	43	60%
Year						
1997	17	13%	3	9%	10	14%
1998	22	17%	7	20%	9	13%
1999	10	8%	3	9%	5	7%
2000	18	14%	8	23%	3	4%
2001	15	11%	5	14%	8	11%
2002	16	12%	6	17%	9	13%
2003	14	11%	1	3%	12	17%
2004	20	15%	2	6%	15	21%
Sending School in Distressed County						
Yes	59	45%	15	43%	33	47%
No	73	55%	20	57%	38	53%

*The survey analysis in this evaluation uses data from 67 respondents. Four of the 71 respondents were listed in the database of participants more than once because they attended more than one institute between 1997 and 2004.

Note: Some percentages in this table do not equal 100% due to rounding.