“But what good came of it at last?”[1]–

How to Assess the Value of Undergraduate Research

Ani Adhikari and Deborah Nolan
Department of Statistics, University of California, Berkeley

First Revision: June 14, 2002
Second Revision: July 2, 2002

Every year a variety of special programs introduce undergraduates to research in the mathematical sciences. There is a general belief that such programs are good – many undergraduates apply, and being admitted is something of an academic distinction.

But with millions of dollars and tremendous effort being poured into these programs, a general belief in their worth is not enough. Granting agencies, and the rest of us, need a little more substance. Assessment plans have become a required element of most program proposals. But oh, that assessment section – scattered, vague, hurriedly cobbled together on the day before the proposal is due. We took a look online at several evaluation plans for the undergraduate research components in NSF VIGRE proposals. What we saw hardly inspired confidence. One plan promised only “the usual student evaluations, with some supplementary questions.” Another was reluctant to promise anything: “The perceived desirability for the student of these choices of employment or further education will be given where possible.” And a third simply assumed what the results would be: “Qualitatively, there will be enormous number of benefits to the department, industry, and profession itself.”

All these proposals were funded. But wishy-washy evaluation plans lead to wishy-washy assessments, making it hard to decide whether the programs are worthwhile.

As a pair of statisticians who ran a summer seminar-type program for a number of years, we had to do better. In this article we provide strategies for assessing the value of the undergraduate research experience, and include the evaluation of our own program as a case study.

The structure of our evaluation scheme can be adapted to a variety of programs. The most direct application is to short-term intervention programs which involve a selection process for the participants. Some examples of these are the Research Experience for Undergraduates (REU) where students work on research problems, either singly or in small groups, under the supervision of faculty; summer math programs where students take seminars or “reading” courses and write expository papers; and intensive short courses. Our example evaluation plan can also be used, with modifications, for “in house” programs where all participants are from the host institution. These include research apprenticeships and seminars, and other innovative programs such as those funded by several VIGRE grants.

Planning the evaluation. The first rule of evaluations is: plan ahead. Not only is this good for the evaluation, it can also benefit the program. For one thing, it forces the program’s goals to be absolutely clear from the outset. Only when a program has a clear objective is it possible to design an evaluation to show whether and to what extent that objective was met. For example, “We want to encourage our students to go to graduate school,” is not the same as, “We want to help our students to become research mathematicians.” Evaluating the latter involves more long-term tracking than the former. Planning ahead can be good for the program in other ways – if organizers feel that an important question cannot be answered clearly under the proposed design, then perhaps the design needs a change.

Most evaluations try to answer some form of the question “Did the program work?” A sensible answer can only be obtained if the right questions are asked. We recommend listing the evaluation’s goals early. Common examples are: to secure continued funding, to improve the program for the future, or to see if the program achieved a particular objective. All these are different, and having them clearly in mind helps to define the scope of the assessment and to sharpen its focus.

[1] Little Peterkin’s question to his grandfather, in *After Blenheim* by Robert Southey. Grandad’s answer: “Why that I cannot tell,” said he, “But ’twas a famous victory.”

Every evaluation requires some fundamental decisions about its structure. We discuss the most common decisions in turn.

When to ask. When and how often surveys are conducted depends heavily on the nature of the program. However, the basic timeline laid out here is common to most programs.

The effect of a program is best determined by comparing its participants to a "control group" who were not in the program. This is not always easy, as we will discuss later. But at the very least, an evaluation should try to determine whether it caused any change in its own participants. Baseline information gives a picture of the participant before the program – background, prior knowledge of areas which the program is trying to address, expectations of program, and so on. Some of this may be available in application materials, though applications are not entirely reliable. Students tend to fill them up with only what they feel will get them into the program. Anonymous questionnaires at the start of the program help to reduce this bias. These should carry an identification code so that each person's answers can be matched with answers given at the end of the program. We found the baseline information about our program participants very revealing. It dispelled many preconceived notions about our students, as we will point out in our case study.

Almost every program asks participants for feedback near the end of the program. This should not be put off until after the program is over, because once people have left it is easy for them to "forget" to return their evaluation forms. All participants should be surveyed: students, faculty, and graduate student assistants if any. Measures such as test scores and grades can be included if appropriate, but the main emphasis as always is on the questions. These should go beyond eliciting cries of, "I loved it!" and cover specific issues. Examples are: whether or not the program met the participant's expectations; whether in the view of the participant the program met its goals; the assessment of program details; and suggestions for changes in the future. Some questions should be designed for comparison with baseline information. Finally, an open-ended "Other comments?" section is highly recommended. The answers can reveal unanticipated flaws or benefits. We got everything from complaints about the quality of the pillows in the dorm to revelations about adversities our program had helped to overcome.

Long-term effects of the program can only be detected by following the students' progress for some time after the program is over. Exactly how much time will depend on the goals of the program. Many programs may carry out just one follow-up evaluation, but others may need to carry out several. Follow-up evaluations typically focus on participants' work after the program, but they can also reveal more. We found, for example, that our participants stayed friends and valued the support they got from each other for years after the program.

Whom to ask. Typically the main source of information is the students who participated in the program. However, while many undergraduates can provide detailed and thoughtful comments, they can hardly be expected to come up with unbiased assessments of their own performance. It pays to get a second opinion, for corroboration and balance.

Faculty advisors such as those who wrote letters of recommendation for the student are an excellent resource. The most helpful information comes from those who continue to have some contact with the student after the intervention. They can comment on the effect of the program on the student, and can provide comparisons with others who were not in the program. Programs which anticipate a long-term effect will need input from those who can judge the quality of the student's future work. Graduate advisors and employers are the most obvious people to ask, and we found them very willing to answer.

We tried to be careful about issues of privacy and confidentiality when asking for such evaluations. Before every follow-up survey we asked each student for permission to approach faculty and other advisors. Permission was almost always granted, but we had to be cautious about making conclusions in the cases where it was withheld.

Other sources of information are worth keeping in mind. For example, test scores and grades may provide useful information, especially if there is a way to compare them to those of students who were not in the program. Other relevant information may be found in published tables containing rates of entry into graduate programs, percentages of students of various demographic groups, and so on. Many departments maintain such data; examples of tables are in *Undergraduate Origins of Recent (1991-95) Science and Engineering Doctorate Recipients* (NSF 96-334).

We cannot over-emphasize the importance of staying in touch. Students move, graduate, drop out, spend time in the Peace Corps – there are countless decisions they can make about moving forward. And only the rare student remembers an undergraduate research program during these moves. The task of keeping in touch therefore falls on the program staff. A database of contact information should include current addresses, phone numbers, email addresses, and most importantly, permanent contact information for parents or guardians of students. Parents are much less likely to relocate than their dorm-sheltered children. We found them very well-disposed to help us regain contact with "lost" participants. The contact database should be brought up-to-date from time to time even when no survey is imminent. For example, in the case of programs which anticipate a long-term effect, contact must be made with students when they make decisions about the next stage of education or employment. If a program anticipates surveying mentors or others, information for these people has to be added.

What to ask. Exactly what gets asked during an evaluation depends so heavily on the program that we felt it most useful to provide an example by briefly describing the evaluation of our own program. See the case study below. Changes required for other programs will, in most cases, be clear to organizers. Later we discuss adaptations appropriate for programs that are very different from ours. For excellent general information, consult Davis and Humphreys (1985).

What to do with the answers. Once the responses come in, evaluators are usually faced with large piles of forms or megabytes of email which have to be sifted through and summarized. The summary must state, preferably right at the beginning, the number of people who were surveyed as well as the number who responded. After all, even a statement that starts with, "All the respondents said ..." loses some of its punch if it turns out that only a small percent of the surveyed students bothered to say anything.

As statisticians we are familiar with numerous complex and ingenious methods for summarizing data. We think, however, that in evaluations such as the ones described here, the simplest methods are usually the most revealing. They are also easy to implement.

Standard numerical summaries such as percents and averages are the most common. It is worth noting that percents should be avoided if the total number is small – "4 of the 6 students" is better than "66.67% of the students." And averages can hide variability – if half the students give something a 0 rating while the other half give it 10, then "the average rating was 5" is not a useful statement. Percents are better. Bargraphs are handy for displaying uneven distributions. They are easier to understand than standard deviations or interquartile ranges, and they can be superimposed to show changes, for example from one year to the next.

Restricting the summaries to one-dimensional ones can lose a lot of information. A student's response to two questions (possibly from different questionnaires administered at different times) can be displayed in a table of percents which show how the answers to the two questions vary together. That is, the percents reported should be conditional ones such as: "Among the students who gave a rank of 1 or 2 at baseline, 20% gave a rank of 4 or 5 at the end of the program." These percents can also be displayed in a bargraph, where one bar would represent each baseline rank and the height of the bar would give the proportion of "rank 4 or 5 at the end of the program" among the students who gave that baseline rank. Of course the heights of the bars will not add up to 100

The SMI/SIMS case study. Our own experience in evaluating special programs for undergraduates grew from running the Mills Summer Mathematics Institute (SMI) and later the Berkeley Summer Institute in the Mathematical Sciences (SIMS). These programs ran for six weeks each summer from 1991 to 1997. Each year about 20 undergraduate women participated. The programs were designed to prepare and motivate them to attend graduate school and pursue careers in the mathematical sciences. The students were selected through a nationwide application process. For more information on the programs see Nolan (2000); details of our evaluation appear in Adhikari et al. (1997) and on the Web at `www.stat.berkeley.edu/users/sims/`.

The questions we addressed in the evaluation were formulated by considering both the concerns of reviewers for the NSF and the goals and possible side-benefits of the program. The main issues were:

- What was the impact of the program on the student's decision to apply to graduate school?
- What was the program's success rate for students entering and completing advanced degrees in the mathematical sciences?
- How does the success rate compare to other rates of attendance and completion of graduate school?
- Did the program improve a student's self confidence?
- Did the program increase a student's knowledge about and preparation for graduate school?
- Do students use the network of peers, graduate students and faculty formed at the program?
- How does the program compare to an REU?
- What impact did the program have on the faculty and graduate students?

To answer these questions, we devised an evaluation plan that included:

- a baseline survey of students on the first day of the program;
- an end-of-program evaluation;
- a two-year-out survey of students as they made decisions about and entered graduate school;
- a survey of the faculty who wrote letters of recommendation for the students;
- a four-year-out survey of students when they would be well into graduate school;
- a survey of their graduate advisors, where applicable.

In addition, we surveyed faculty and graduate students who worked with the students in the program, and we collected information on the numbers of undergraduates going on to graduate school at the participants' home institutions.

Brevity is a virtue in questionnaires, so student questionnaires were approximately two pages long, and the faculty surveys were under one page. In addition to open-ended questions, there were several questions that asked the respondent to rate some aspect of the program on a numerical scale. Some questions in the baseline and end-of-program surveys were the same in order to make comparisons (with phrases like "Do you expect ..." in the baseline survey being replaced by "Did you find ..." in the later questionnaire). The two-year and four-year surveys had similar questions for comparison purposes as well. In addition, some questions on the faculty survey mirrored those on the student survey to allow comparison of the faculty member's perception with the student's.

We phrased these questions carefully to minimize bias in the response. Two examples appear below; in each, the student was asked to provide a rating on a scale of 1 (little or none) to 5 (a great deal). Notice the words "if any" in the first and "from your perspective" in the second which neutralize the questions.

To what extent, if any, did the program affect your:					
Self confidence	1	2	3	4	5
Motivation to do graduate work	1	2	3	4	5
Knowledge about what graduate school is like	1	2	3	4	5

From your perspective, how important was it that the program involved only women as:					
Students	1	2	3	4	5
Graduate students	1	2	3	4	5
Faculty students	1	2	3	4	5

An important benefit of the baseline survey is to get an accurate reading of who the participants are, and to check whether prior assumptions about the participants are grounded in reality. Our baseline evaluation produced a revealing example of false assumptions. Entry to the program was based on an application process, and a natural query was revealed in an NSF review of the program's proposal: "For such an expensive and selective program should the success rates in turning out top notch female graduate students be higher?" To answer this question it was necessary to have some idea of the proportion of participants who firmly intended to go to graduate school even before they applied to the program. The anonymous baseline survey produced results that laid to rest the assumption that most of the students were pre-disposed towards graduate school; indeed, only 7 of the 20 respondents were definitely planning to attend graduate school and expected financial support in the form of a fellowship or assistantship. With this concrete information at hand it was easier to measure the success of the program.

We conducted all our two-year and four-year surveys by email. Eighty percent (34 out of 43) of the participants returned completed questionnaires, a very high response rate for a survey involving a long-term component. Of course the majority said that the program had a great impact on their motivation to do graduate work, but the detailed answers were more interesting. Over two-thirds reported a great increase in self-confidence after the program, and almost as many strongly agreed with the statement, "My work in the program showed me I enjoyed doing challenging math." Two students strongly disagreed, and said that the program showed them that graduate school was not their goal; even though this is a negative lesson, it is better learned in the summer than in the first year of a Ph.D. program.

Surprisingly, more than half the students strongly agreed with the statement that the program "showed them how to learn advanced math." Given that the students were selected specifically for their ability to do mathematics, this proportion is very high, and shows how work in the program differed from standard undergraduate fare.

But the main message from the students was that they were deeply impressed by the women on the faculty, and by the group of talented women who had all gathered in the program to do mathematics. Over 80% of the respondents have stayed in touch with other students in the program, and over 70% have stayed in touch with program faculty. The students' opinion is best summarized by one of their own: "Until attending the SMI, I had only one female math professor. Ever. I think now I have a great advantage in having discovered some positive female role models in mathematics ... I found the program ... to be extremely helpful to seeing myself as a mathematician."

Just about 75% of the faculty mentors responded, and provided valuable input on the effect of the program: over 80% of the faculty respondents in the short-term survey said the program was very good for the student, and indeed, about 50% said that the program was a shot in the arm for their department as a whole. When asked whether participating in the program distinguished this student from other women math majors in their department, two themes recurred in their answers. They noticed a tremendous increase in the self-confidence and in the mathematical maturity of the student upon her return. This shows the benefit of surveying faculty as well as students: while students may be able to tell us they feel more confident, a professor is surely a better judge of mathematical maturity. The faculty were also able to provide us with information on the proportion of their students who go on to graduate school.

Adaptations: Small programs and "in house" programs. The details of an assessment depend partly on the size of the program and the composition of its participant group. Specific issues arise in programs which have a small number of participants, and in those which only involve students from a single institution.

Many REUs have fewer than 10 participants each summer. In programs as small as these, yearly percentages are not meaningful. Instead, data from more than one year can be aggregated. Of course, responses should be pooled according to the same relative participation time, e.g. one year out of the program or first year in graduate school.

In addition, small programs typically involve closer interaction between students and faculty than programs with many participants. When a student works on a research project one-on-one or in a small group with a professor, the student-faculty interaction is very different from the classroom or seminar format, and the nature of the interaction is an important element of the program. The effectiveness of these programs may be best assessed through in-depth personal interviews rather than questionnaires. These interviews may, for example, be conducted in person at the end of the program, or over the phone one to two years after participation. They should not be conducted by the faculty or directors of the program as it could bias student response. Someone outside the program, preferably someone who is trained in interviewing people, should conduct the interviews.

Programs where the students are from the host institution, such as many of those supported by VIGRE, have a big advantage in evaluation. Because the students are local, it is easier to maintain a database of past participants and to observe participants after the program. For example, a student's choice of future course work and course grades can be excellent sources of information. In addition, a control group of students who are similar to program participants can be found in recent graduates or peers at the home institution.

Difficulties. Even with enormous effort put into the design of an evaluation and its questionnaires, it is not always possible to get clean and unambiguous results. We found two major problems.

Nonresponse. People do not like to fill out questionnaires. Indeed, even the most enthusiastic participant may need some prodding before he or she will answer questions on an evaluation form and send the form back. To reduce nonresponse bias, we recommend the following.

- Keep the contact database up to date.
- Keep the questionnaires short, and have an email or Web-based version. This means that the questions have to be very well chosen, and some work may have to go into creating a Web form. But the increase in response rate that results from forms that are easy to fill out and return is well worth the effort.
- Follow up nonrespondents. This means sending reminders, and reminders of reminders, by every means available – phone, email, and letters. This task is time-consuming and can be disheartening, but it is part of the difficulty of carrying out any survey.

In spite of all efforts, most surveys still suffer from some nonresponse, and the response rate needs to be reported in the analysis of the evaluation.

Lack of comparison. Ideally, an evaluation would compare the performance of program participants to what their performance would have been had they not taken part in the program. This is of course impossible. It is also often difficult, if not impossible, to compare the performance of program participants to that of a control group, that is, students who were not in the program. The main problem lies in ensuring that participants and controls are similar in every respect other than the program.

For example, if entry into the program is the result of a competitive application process, it is not easy even to identify a control group. Students who did not apply to the program are clearly different; so are those who applied but were rejected. In any case, neither of these groups will be highly motivated to respond to questionnaires sent by a program they did not attend. In the absence of a reliable set of student controls, we strongly recommend getting information from external sources such as mentors (usually faculty), who know the participants as well as other students with similar backgrounds. This is a good way of obtaining the comparisons which are crucial for a meaningful evaluation. Available data, such as tables on the number of graduates from a student's institution who go on to graduate school or academic careers, may also be used as comparison figures for success rates.

Summary. Evaluations can provide concrete evidence of whether or not a program has achieved its stated goals. Such evidence can be used to improve the program, to support requests for continued funding of the program, and to give granting agencies something to which new programs can be compared.

Our experience shows that serious evaluation of undergraduate research programs is hard work, but well worth the effort. We found that our evaluations influenced others to adopt successful aspects of our program and to start similar programs. A thoughtful and comprehensive evaluation, such as we have described here, helps not only to validate an innovative program, but also to chart a course for future improvements of the program and, ultimately, for the way the U.S. trains undergraduates in mathematics. Updated information about our assessment materials is available on the Web at `www.stat.berkeley.edu/~sims/`.

References

Adhikari, A. and Givant, S., and Nolan, D. (1997) The Mills Summer Mathematics Institute, in *Women in Math: Scaling the Heights*, MAA Notes, no. 46, pp. 97-104.

Davis, B.G. and Humphreys, S. (1985) *Evaluating Intervention Programs: Applications from Womens' Programs in Mathematics and Science* New York, NY: Teachers College, Columbia University.

Nolan, D. (2000) Evaluating summer math programs. In *Proceedings of the Conference on Summer Undergraduate Mathematics Research Programs*, American Mathematical Society, pp. 323-329.