Evaluating Outcomes

William S. Maki, Texas Tech University, Lubbock, TX

The April 2004 CCLI conference focused on building excellence in undergraduate STEM education. The question motivating this chapter (and the workshop on which it is based) was posed by a member of the NSF staff during electronic discussions before the CCLI conference.

> *"... You decided to develop new curriculum materials, or to adapt and implement them in your institution ... What evidence enables (or will enable) you to infer that it is working—and why?"*

The educational context in which the motivating question is supposed to be asked and answered is often seen to be fraught with insurmountable problems.

- What measures should be taken? (Grades? Ratings? Quality of problem solutions?)
- What comparisons should be made?
- What is a suitable control group?
- What if a control group is not possible? (There is only one class!)
- Students are so different. How can you make sense of the data?

In addition to the above generic questions (that I often hear), another question and a comment surfaced in the audience reaction to the CCLI workshop.

- How does one motivate colleagues that assessment is important and can be done to provide useful information?
- Most scientists have no clue as to what is an acceptable design for educational research. (Scientists are not alone in this puzzlement.)

In this chapter, I hope to persuade the reader of three things. First, all these questions can be sensibly answered. Second, with some advance planning, an adequate research design can be devised. And, third, an acceptable evaluation can be conducted.

Typical Problems

Over the years, I've served on both NSF DUE and local grant panels at two public universities. I've reviewed proposals from both STEM and non-STEM disciplines. Believe me when I say that acceptable evaluation plans are rarely proposed. What characterizes an "acceptable" evaluation plan? I'll begin by considering four scenarios. The surface details of each scenario have been modified so as to protect the identity of the proposers and to veil the specifics of their ideas. But the basic structures and proposed measures are real enough for purposes of this illustration. I'll first present a common context for all the scenarios and then go through them one at a time, presenting first the scenario and then my analysis of it and my suggestions for alternative evaluation plans. The reader should approach each scenario as a real problem and attempt to anticipate points made in my critique. So as you proceed, cover up the text following each scenario and make some notes about what is wrong and what you might do to improve upon the evaluation.

Context Common to All Scenarios

Assume the university has a Teaching and Learning Center (TLC). The provost has given the TLC $100,000 this year to support instructional innovations that will improve student learning. The TLC advisory committee agrees with the center director that a grant competition is the best way to allocate the funds. The TLC has issued a request for proposals to the

university faculty. The criteria listed in the Request for Proposals include the "extent to which the project will have a positive impact on student learning," and the format of the proposal includes a required section on the "Evaluation Plan." Ten awards are anticipated.

Scenario 1: A catalog of web resources for American History

American History is a large enrollment course taught as part of the required general education core. However, retention has been poor. The instructors believe that the drop rate is so high because of the outdated materials—slides and transparencies. They propose to scour the Internet to discover linkable graphic resources and to assemble them into an online catalog accessible by History faculty and students in the American History course. The success of this new resource will be evaluated by recording the number of faculty using the catalog each semester.

Analysis 1

Remember that the objective is some demonstrable impact on student learning. Merely documenting faculty use of the new materials does not in any way show an effect on the students. Better dependent variables would be examination scores or a specially made questionnaire about American History. There are some other ways to improve on the design. If archival records are available in the form of examination scores from previous offerings, a *time-series* design could be implemented. The past years' courses provide a baseline against which to evaluate the new materials. You would be looking for a discontinuity in the series that coincided with the introduction of the new materials. I'm told such records often are difficult to acquire. You might then compare scores in the sections using the new catalog against scores in the sections using the traditional materials. Better yet, adding some kind of *pretest* would allow the determination of the beginning equivalence (or lack of it) among those sections exposed and not exposed to the catalog. Comparing the previous retention rates and examination scores across instructors who use and don't use the new catalog would be helpful in establishing the equivalence of instruction across different faculty. All these changes and extensions allow disconfirmation of competing explanations for any improvements seen in the sections using the catalog (such as select students or select faculty).

All this seems like a lot of effort. True, but if you are more comfortable knowing whether or not the new materials work than wondering about whether they work, then you will figure out ways to bear the cost. In my experience, even modest grant-supported activities include a teaching or research assistant in the budget. The record-keeping and test administration are tasks that can be assigned to that assistant.

Scenario 2: Response technology to increase active learning

Lecture classes, especially large ones offered to nonmajors, tend to promote passive listening and inattention. Class discussion may be effective but difficult to implement in large classes. Response technology offers an alternative way to engage students in these courses. Each student is equipped with an electronic keypad that sends infrared signals to a receiver wired into the instructor's computer. Student responses are summarized and displayed on an overhead projector. The instructor can then comment on the response trends, correcting misimpressions on the spot or posing other challenges. Student's grades in the semester in which the response technology is introduced will be compared with those grades received by students in the previous semester.

Analysis 2

Individuals who worry a lot about assessment are generally reluctant to accept course grades as a dependent measure; grades are determined by too many factors. Some other measure of student learning is needed. This design only has a two-point comparison—last semester versus this semester. A *selection* problem is a serious confound; students may differ across two semesters. (My colleagues notice different class "personalities" from semester to semester.) It may be possible to dig deeper into the past records to establish a baseline. Or, it might be possible to teach concurrent sections, one with the response technology and one without. (The exact approach here will depend on details of the local context that are not given.)

Scenario 3: A photographic library for ichthyology lectures

The lectures in the Ichthyology course in the Zoology department could be improved by increasing the variety of visual aids beyond the single examples (photos and/or drawings) of

various fishes presented in the textbook. The instructor is a scuba diving/underwater photography enthusiast. He plans to solicit 35-mm slides from other underwater photographers using the underwater photography list. These slides will then be introduced in his lectures to improve the quality and quantity of examples. The amount learned will be assessed by student ratings. The university course evaluation that is administered at the end of each semester in each course contains an item that asks the students to rate the amount learned.

Analysis 3

What students say they learn and how they actually perform are often quite different. Having some evidence of what the students know (e.g., number of fish correctly identified) would be a much better measure. Here, too, different evaluation designs are possible. The instructor could mine course records from previous offerings to establish a time-series baseline. If his teaching load permits multiple offerings, he might opt for a pretest-posttest nonequivalent groups design. At the start of the semester, all students are tested on their ability to identify a random selection of fish from some larger pool of test items (the pretest). At the close of the semester, the remaining items from the pool are presented as the posttest. One section chosen at random receives the new slides; the other section does not. A differential change from pretest to posttest would be evidence consistent with an effect of the new materials; the gain score for the section with the slides should be (statistically significantly) higher than that for the section not exposed to the new slides.

But suppose the instructor's teaching load does *not* permit his teaching of multiple sections. One design that is not desirable here is the *one-group pretest-posttest* design in which the students are assessed at the beginning and end of the course and the new pedagogy is implemented during the course. This design admits *maturation* as a competing hypothesis. Any change from pretest to posttest could be attributed to learning that occurs during the semester with or without the new materials. Instead, an *equivalent time-samples* design might be possible. If the course is organized around some classification scheme (e.g., based on biology or geography), then the course may be segmented into experimental and control units. A pretest covering all units would begin the course. Then each unit could be followed by a different posttest (or a larger posttest could be administered at the end of the semester). Some units would be supplemented by the new slides and some would not. To further strengthen the design, the instructor could reverse the ordering of experimental and control treatments in a subsequent offering of the course.

Scenario 4: Online grading of calculus homework problems

Immediate feedback is important for learning skills such as those important for solving mathematical problems. However, grading of calculus homework problems is at present labor-intensive and time-consuming. Thus, students in our Introductory Calculus courses do not receive immediate feedback on their homework problems. We plan to introduce a new online software system that will automatically grade homework problems. Use of the new system will be at the discretion of calculus instructors, so there will be sections that use the system and sections that do not. All the instructors have agreed to place a common set of problems on their final exams so that problem-solving of students in sections that use the software will be compared with problem-solving of those students in sections that do not use the software.

Analysis 4

The use of a common set of problems for assessment is a good move; this posttest should assess the problem-solving skills germane to the course objectives. The principal problem here is the determination of group equivalence. The Mathematics department may be able to acquire student grade-point averages, entrance scores, and/or scores on math placement examinations. The instructors might be able to invent some practical problems appropriate for a pretest in the course.

Varieties of Evaluation Designs

The analyses of the foregoing scenarios introduced some terms and induced some research designs. Here, I'll review the lessons learned from the scenarios and consider those designs in a more formal way (modeled after the classic texts by Campbell and Stanley, 1963, and Cook and Campbell, 1979).

In my experience, probably the most common approach to evaluation is to propose some innovative approach to

instruction to be implemented within a single academic term and evaluated in some way at the end of that term. This is the case-study approach (design 1 in Table 1). The design suffers from many threats to internal validity (confounds). The students may end up where they do at the end of the class because they change during the academic term in spite of the treatment ("maturation"). That particular group of students may be particularly receptive to the treatment or be special in some other way by virtue of their enrollment ("selection"). (For other threats to internal validity, consult either of the two texts on quasi-experimentation in the Bibliography.)

Another common approach is to add a control group to the case study (design 2). Most often, the proposed control group is a different section of the same class (sometimes varying with respect to time of day, academic term, and/or instructor). In the absence of random assignment, and without a pretest, there is no way to ensure equivalence of the groups at the start of the term. Hence, the selection threat is still a problem; preexisting differences may contribute to the difference in scores at the end of the term.

Simply adding a pretest to the case study in which the students are tested at the beginning of the term doesn't help much either. Design 3 controls selection in that each student's score at the end of the term is compared with that student's score at the beginning of the term. If these change scores are consistent across students, self-selection cannot be responsible for the changes. However, the two-point comparison cannot rule out maturation; the students could still show those changes from pretest to posttest in spite of the treatment. Thus, just adding a single pretest is not an ideal solution. Adding several observations before treatment is much more useful in that a baseline trend is established against which to evaluate treatment effects. Variations on the time-series design, as in analyses 1 and 3, control for maturation and other threats to internal validity.

The strongest research design is a true experimental design in which students are randomly assigned to different treatment groups. Random assignment is intended to equate the groups before the introduction of the intervention for one of the groups. Such designs are rare. Students are not random numbers; they select courses for all kinds of reasons (such as class schedules and outside employment), with the result that, most often, we are faced with comparisons between nonequivalent groups. However, a reasonably strong design can be created by adding both a control group and a pretest (design 4). This pretest-posttest nonequivalent groups design controls for both maturation and selection. The use of a pretest establishes a baseline for each student against which to evaluate that student's posttest score (thus ruling out selection as a threat). If only maturation were operating to produce pre-post changes, then the same average change should be observed in both groups. Thus, the treatment becomes the more plausible cause of differential changes from pretest to posttest in the two groups. This kind of design was used in recent work on evaluation of web-

Table 1. Schematics for Various Research Designs

Design	Group	Pretest	Treatment	Posttest	Treatment	Posttest
1			X	O		
2	Experimental		X	O		
	Control			O		
3		O1	X	O2		
4	Experimental	O1	X	O2		
	Control	O1		O2		
5	Experimental	O1	X	O2		O3
	Control	O1		O2	X	O3

O symbolizes observations, and X symbolizes treatments.

based instruction in a general psychology course (Maki et al., 2000; Maki and Maki, 2002).

Suppose that Professor Jones has an idea about how to improve problem-solving performance in his Quantitative Methods course. He intends to implement his idea this semester. He proposes to test his students at the beginning of the semester and again at the end of the semester. Jones knows that he still needs a control group. Professor Smith is scheduled to teach another section of Quantitative Methods this semester, so Jones asks Smith to help out with the evaluation. Smith agrees to let her students receive the same pretest and posttest. This is an example of design 4, which controls many threats to internal validity. But a problem remains with this design. Jones, the believer in his innovation, may be the energizing factor responsible for larger change scores in his students. Or, students may become aware of the new method and opt for Jones' section during registration; those self-selected students may be particularly receptive to the new method of instruction.

This is the problem my colleagues and I faced when evaluating training research methods skills using the web (Maki et al., 2004). To control for these potential threats, we took advantage of the fact that most courses consist of modules or units. We extended design 4 by treating the control group after a first posttest and then administering a second posttest to both groups. The extended design 4 is diagrammed as design 5 in Table 1. Two research methods skills (course units) were targeted in each section. The first module in one section (web group) was exposed to the web-based method, and the second module in the other section (control group) was exposed to the web-based method after the first posttest. Both sections received a second posttest. The design was repeated in a second semester and the groups (web versus control) were swapped across instructors. Figure 1 shows the results. The two groups were not significantly different in their performance on an initial set of problems (the pretest). The web group was significantly better than the control group on the posttest (showing the effect of the web-based training). The control group, after its exposure to the web-based training, converged on the web group in the second posttest. Instructor influences and student selection factors seem unlikely to be the cause of this pattern of results in which the significant gains were associated with the web-based treatment regardless of semester and regardless of instructor.

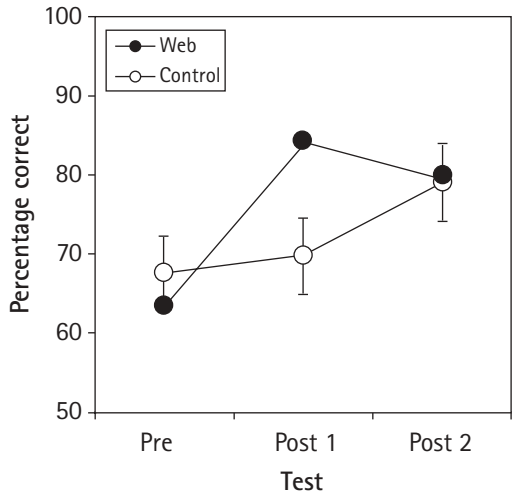

Figure 1. Average percentage correct on solving problems in a psychological research methods course. Web-based instruction was introduced in the web group between pre- and post1-tests. That same treatment was given to the control group between the two posttests. The 95% confidence limits are shown for the control group. Data are averaged across two replications. See Maki et al., 2004, for details.

Concluding Observations

Faculty, regardless of discipline, always are inventing or adapting or adopting instructional methods. The desired outcome is that these new methods improve student learning. But observing high performance (or high satisfaction) at the end of a course or module is not in itself evidence of such an outcome. A good evaluation design is needed to answer the question, "What evidence enables you to infer that it is working?" True experimental designs are difficult and may be impossible to implement in most educational settings, so quasi-experimental designs were the focus of this chapter. These designs can be thought of as adaptations to nonrandom assignment, as cobbled together to rule out competing explanations of observed student performance. The designs are not conceptually difficult, but they do require advance planning. The result is greater confidence in the conclusion that your instructional innovations make a difference in your students' learning.

BIBLIOGRAPHY

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston, MA: Houghton Mifflin.

Maki, R. H., W. S. Maki, M. Patterson, and P. D. Whittaker. 2000. Evaluation of a web-based introductory psychology course: I. Learning and satisfaction in on-line versus lecture courses. *Behavior Research Methods, Instruments, and Computers* 32: 230–239.

Maki, W. S., F. T. Durso, and S. Alexander-Emery. 2004. Using the world-wide web to build research skills from multiple examples. Poster presented at the Conference of the CCLI Program, Crystal City, VA, April 16–18.

Maki, W. S., and R. H. Maki. 2002. Multimedia comprehension skill predicts differential outcomes of web-based and lecture courses. *J Exp Psychol Appl* 8: 85–98.