

The Flashlight Approach: Evaluating Educational Uses of Technology

Stephen C. Ehrmann, Ph.D.¹, Revised March 3, 2011

Let's start by suggesting what *not* to do. It's usually not a good idea to:

- Focus on the technology itself
- Wait until a program is complete, or stable, before doing an evaluation
- Hope for high marks (and, if need be and possible, design the evaluation so that it will produce excellent results).
- Assume that all users want and need exactly the same things from the program or technology.
- And/or collect as much data as you possibly can
- Use reputable instruments that someone else has already validated
- As your 'control group,' use 'no treatment' – in other words evaluate the difference between doing something and doing nothing;

Instead of those ideas, consider these working assumptions instead:

1. **Above all, do no harm.** Personally, when you get a survey or are asked to respond to an interview, what's your first impulse? Hit the delete key? Hang up the phone? Is that because past surveys seemed of little benefit to you personally or to people you care about?

The most important guideline for doing no harm: if you're not prepared to analyze the data (which can be time-consuming and sometimes expensive than you might realize) and make good use of the findings, don't gather the data in the first place.

A second hint for doing no harm: be very cautious about mindlessly adopting a survey, test or other tool that was originally developed by someone else for some other purpose. No matter how valid and reliable it is, there's a good chance that a good fraction of it, or all of it, will be of no use to you. It's surprising how often investigators choose valid, reliable, irrelevant instruments that produce statistically significant, useless results. And, by failing to benefit respondents while wasting resources, such studies do harm.

If you're reading this document, you're probably in an educational institution. Your study ought to make its respondents feel more positive, not negative, about the research process. Two suggestions: (a) design a study that its findings will potentially benefit respondents or informants, and (b) later on, get back to the people who provided the data, and tell them what, if anything, has been done as a result of the study.²

2. **Design the evaluation in order to discover how to improve results.** It's easier to 'do no harm,' if your evaluation is designed to discover how to improve results ("formative"). Evaluations to decide whether something has succeeded or failed ("summative") are sometimes

¹ *In general, and for the unauthorized use of some of his ideas, many thanks to Tom Angelo! To discuss these eleven principles, and for permission to use this as a handout, please contact Stephen C. Ehrmann (ehrmannsteve@gmail.com) or the TLT Group (info@tltgroup.org)*

² For more on this topic, see Section IV.C of the *Flashlight Evaluation Handbook* at <http://bit.ly/FL-TOC>.

threatening to respondents – your findings might be costly for them. If you need to do a summative evaluation, and if you can't be confident that your findings will benefit your informants, you'd better figure out some way to augment your inquiry and/or provide other incentives for their thoughtful, honest participation.

3. Be prepared to “compare apples and oranges.” Naïve approaches to evaluation often look for quantitative gains on traditional outcomes. But innovations usually involve a qualitative change in outcomes. For example, when students began using software to do statistics, it enabled them to learn new styles of research, not simply to score better on old statistics exams. One way to compare apples and oranges is to enlist a panel of judges who represent stakeholders (e.g., employers, teachers of more advanced courses) and show them two bundles of data: the ‘tests’ and results from your experimental and the ‘tests’ and results from your control group. Which do they prefer? You could also add data about costs of the two methods. This allows you to design assessments that are appropriate for each group, rather than force-fitting the old assessment to the new technique, or vice versa.

4. Different kinds of outcomes: **If this is a large scale study you may be interested in study not only *effectiveness* and *costs* but also *robustness* (to what extent does the idea, technology or material work in different settings), *sustainability* (can its use continue after initial funding and enthusiasm fades), *scalability* (to what extent can its use be expanded while (at minimum) not diminishing its effectiveness or outgrowing its supply lines, and *generativity* (does your target stimulate its users to try things they might not otherwise have tried)?** We'll return to these varied outcomes in suggestion #9 below. By the way, this set of suggestions can be used to evaluate ideas, materials, technologies, techniques and programs. For simplicity's sake, from here on we'll refer to the object of the evaluation as the ‘target.’

5. Study *what users did repeatedly with the technology (their “activities”)*. Technology (whether it's a computer, a software package, a camera, a blackboard, a book, or a classroom usually has no direct, predictable impact on learning; instead it's what people do that influences learning. The importance of the technology is to make certain activities easier, less expensive, less risky, more flexible, etc.

It may be counter-intuitive but if, for example, you're considering studying a thing (e.g., a piece of software), instead design a study about the pattern of activity for which that thing might be used. Suppose, for example, that an institution is using lecture capture. The study might focus on the activity of reviewing lectures. So there are at least two families of questions:

- A. What sorts of reviewing of lectures are students doing, and are those activities having a discernible influence on what they learn?
- B. How good a fit is the current technical and social infrastructure of lecture capture for those activities of reviewing lectures?

Why focus on “repeated” activities? Because evaluation studies the past in order to learn how to improve the future. The more frequent the activity, the sooner your findings can be applied to improve it. And the more widespread the activity, the more influential those findings could be. It's also important to do an evaluation that produces results soon enough to be useful for improvement. Some studies do harm simply by taking so long to produce findings that events have moved on without them.

6. Study *why* the activity unfolded that way. Knowing that people acted in a certain way helps you understand what they learned (principle 3). Discovering **why** they acted that way can help you figure out whether and how to change that activity.

For example, suppose you are studying the value of the online discussion capabilities of a learning management system. The return on investment in the LMS will be influenced in part by whether, and how, students take part in online discussion. So it makes sense to study the factors affecting whether and how each student engage in discussion online. For example, they may not contribute because they haven't found other students' contributions helpful, or because they're convinced that the only way to learn is via textbooks and lectures. Neither of those factors relates directly to the LMS, but both can reduce the value of the LMS. If an evaluation uncovers the fact that such factors reduce LMS use, it may then be possible (by altering courses and training) to make the LMS considerably more valuable without investing another penny in software.

7. If you need a comparison group, compare this activity or technology with the most credible competitor. For example, if you're studying use of a fancy graphics program to create art, ask yourself what technology (or teaching/learning activity) would be used if your targeted technology (or activity) weren't in use. If your target is expensive or difficult, choose as your 'control' a plausible alternative that's cheaper and/or easier.

8. Remember that education does *not* work like a well-oiled machine. Even when the same faculty member teaches two sections of the same course, and has taught them for years, what students do and what they learn will differ from one section to another. Every instructor knows that. Technology is called "empowering" when it increases options for faculty and students. As a result of the availability of empowering technology, variation in learning among students will probably increase. So your study design, and your interpretation of its findings, both need to treat human variation and choice as fundamental, not as 'noise' that will cloud your findings unless those choices are eliminated.

9. Recognize that different students (and faculty) have different goals, needs, and (therefore) outcomes ("Unique uses"). The beneficiaries of any program or innovation differ, but the goal in most instances is to help them all. For example, students using the same innovation may have qualitatively different skills and preparation, career goals, backgrounds, abilities/disabilities. An evaluation design that begins with the assumption that the students (and faculty) may use an innovation, technology, or program differently, and with different consequences, is called a "unique uses" design.

In contrast, an evaluation design that begins with the assumption that certain outcomes should be studied for all students (and faculty) is called a "uniform impact" design.

A typical program needs to be evaluated both ways. If the program is mainly for producing identical benefits for everyone, the evaluation should be mainly uniform impact. If the program or technology is mainly empowering and responsive, the evaluation should be mainly unique uses. Most programs, however, are a mix of the two. So the evaluation should use both methodologies.³

³ For more on unique uses and uniform impacts and how to evaluate each of them, see the *Flashlight Evaluation Handbook*, Chapter III, Section A.

We can now put together those different kinds of outcomes (suggestion #4) with uniform impact and unique uses:

	Uniform Impact	Unique Uses
Effectiveness	Deductively assessed, starting with developer's goals	Inductively assessed (starting with evidence from individual users)
Costs	Assume that time and money are used the same way, or in some ideal way, in every adoption of the target.	Assume that different adaptations in different settings will use resources in different ways. In fact, adaptation 1 in setting A may use fewer resources than adaptation 2 in setting B. Yet adaptation #1 may be judged "too expensive" while adaptation #2 may be judged as "a real money-saver" because of such differences.
Robustness	Can the idea and its goals work 'as is' in varied settings?	Can the target (including its goals) be adapted effectively in a variety of settings?
Sustainability	Evidence that the target can be supported over the long term in its setting (without altering the goals or strategy)	Evidence that the target can be supported over the long term in its setting (while perhaps altering the goals and strategies, but maintaining effectiveness)
Scalability	Evidence that, if demand for the target increases, supply can increase as well (without altering the goals or strategy)	Evidence that, if demand for the target increases, supply can increase as well (while perhaps altering the goals and strategies, but maintaining effectiveness)
Generativity	Evidence that users are stimulated to try a kind of approach that they might not otherwise have tried.	Evidence that innovators and early adopters (at least) are stimulated to create surprising, effective variations on the target

10. Confront the dark side. Too many evaluations organize their inquiries only around the hoped-for gains from technology use. But loss and damage are inevitable, too. *Every* shift in the use of technology results in losers as well as winners. And each participant usually experiences losses as well as gains. These losses can't be reduced or compensated unless they are recognized. An evaluation can help improve results by minimizing loss, just as it can by discovering how to increase gains. And evaluations that explicitly recognize loss are more credible and compelling than those that focus purely on gain.

11. Design a study that can influence action, *no matter what you discover.* Too many evaluators and researchers have a finding in mind when they design the study. If the study finds something else, all they recommend is ‘fund more research.’

Similarly, if your evaluation is designed to inform and influence someone else (e.g., a budget-maker), even in part, work with them to design the study. It should be clear **in advance** that the study’s findings are capable of convincing everyone concerned to change their minds, even the findings happen to contradict their preconceptions.

This usually isn’t done. So the following scenario (yet another example of ‘doing harm’) often plays out. An educator wants evidence to convince a decisionmaker to make more money available because the educator thinks a technology is useful and the decisionmaker isn’t yet convinced. The educator designs the evaluation without involving the decisionmaker. Amazingly, the findings indicate that the technology deserves more money. The decisionmaker disregards the evaluation because the decisionmaker believes that the evaluator has designed the evaluation specifically to produce that finding. And the time of the educator (and the respondents) is wasted. Moral: the educator and decisionmaker should agree in advance on a design that is capable of convincing the educator that the innovation has failed (if that’s what the evidence shows) or convincing the decisionmaker to increase the budget (if the evidence comes out that way.)

The most ambitious form of this approach is to teach users how to do their own studies, or to collaborate with them. Then the data will truly be theirs. That’s one rationale behind the spreading practice of the scholarship of teaching and learning: if the instructor has actually asked the questions, and gathered the data, it’s more likely that the data will be directly useful, and credible, than if the instructor looks up a study done by someone else, somewhere else.

12. Start evaluating now! For example, if you're interested in using peer instruction and student response systems such as clickers to foster conceptual learning, start evaluating conceptual learning (including faculty development) now, whether or not you're using clickers yet, whether or not you're experienced in their use yet, whether or not you're considering replacing one kind of clicker with another. Such a study can have many benefits including improving courses and faculty development, helping you choose products, and providing benchmark data that can help show later on whether your next changes in practice had the desired benefits.