2010; 32: 296–301 MEDICAL TEACHER

TWELVE TIPS

Twelve tips for evaluating educational programs

DAVID A. COOK
Mayo Clinic College of Medicine, USA

Abstract

At one time or another, nearly all educators will need to evaluate an educational program to determine its merit or worth. These tips will help readers collect information to inform a meaningful evaluation, whether for local use or broad dissemination (i.e., research). The two most important questions in any evaluation are, 'Whose opinion matters?' and 'What would really be meaningful to them?' Other key steps include getting input from others, focusing on desired outcomes before selecting instruments, considering the validity or trustworthiness of the data, and pilot testing the evaluation process.

Introduction

For the past 3 months, Judy and John have been planning a workshop to teach postgraduate physician trainees how to examine the thyroid gland. They have spent countless hours studying different examination techniques, searching the literature for similar courses, gleaning teaching tips from colleagues, lining up patients with thyroid abnormalities, arranging for rooms and refreshments, and inviting trainees to attend. At about 3:00 p.m. the day before the workshop, John calls Judy and asks, 'How are we going to know if this workshop is any good? How will we evaluate this program?' There is a moment of silence on the line, and then Judy replies, 'That's a good question. I don't know. I don't even know where to start.'

Sound familiar? I hope not – but it is something I have seen too often. At one time or another, nearly all educators will need to evaluate an educational program. Why? Because we want to know the *value* of the activity into which we have invested time, energy, and other resources. As one classic text states, 'Many different *uses* may be made of those value judgments . . . but the central *purpose* of the evaluative act is the same: to determine the merit or worth of some thing.' (Worthen et al. 1997, p 8) Yet, as attention focuses on effective program development and implementation, the program evaluation may get neglected. Alternatively or additionally, an outstanding teacher might not possess the skills to conduct an effective evaluation.

My purpose in this article is to help you plan effective evaluations of educational programs. These tips may not rescue last-minute emergencies (although they might!), but if applied early and consistently, they will help you collect the information you need to 'determine the merit or worth' of your program.

Tip 1. First ask, 'Whose opinion matters?'

The most important step in planning your evaluation is to identify for whom the information is intended (Figure 1). Who will read the final report? An evaluation intended for the medical school dean will look very different than one intended for publication in a peer reviewed journal. An evaluation intended to help yourself improve a course for the next go-round will require very different information than that in a final report to a funding agency demonstrating your program's success. Of course, you might entertain multiple audiences as you plan the evaluation, but this should be a conscious decision. Stakeholders – people with an interest in the program and its evaluation – might include administrators, students, teachers, secretaries, funding agencies, and the educational community at large. Dissemination to the community at large constitutes a critical element of scholarship (Glassick 2000; Beckman & Cook 2007).

Tip 2. Next ask, 'What would really be meaningful to them?'

The second most important step is to determine what would really be meaningful to your audience. This, together with the answer to the first question, influences everything that follows. Different types and quantities of information will be more or less valuable in different situations. For example, although learner knowledge is often considered more important than satisfaction, in evaluating a new program to orient students to medical school, satisfaction may be more meaningful than performance on the end-of-year comprehensive exam.

Correspondence: D. A. Cook, Division of General Internal Medicine, Mayo Clinic College of Medicine, Baldwin 4-A, 200 First Street SW, Rochester, MN 55905, USA. Tel: 1 507 5380614; fax: 1 507 2845370; email: cook.david33@mayo.edu

ISSN 0142–159X print/ISSN 1466–187X online/10/040296–6 © 2010 Informa Healthcare Ltd.
DOI: 10.3109/01421590903480121

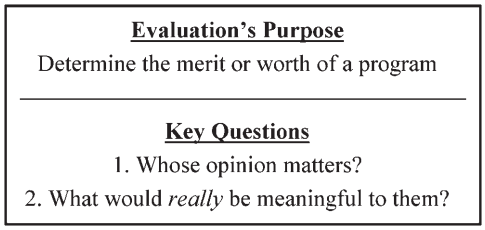

Figure 1. Key considerations in planning an evaluation.

An important factor in this decision is how the evaluation will be used. Evaluations are generally used to inform policy and guide decisions, such as

- Determining effectiveness,
- Identifying areas for improvement,
- Optimizing resource allocation, or
- Empowering individuals (teachers, students, administrators, policy makers, etc) in their respective roles.

For example, a dean might use the evaluation of an elective clinical attachment to decide whether to adjust the ratio of clinical and lecture time, offer it at a different time of year, assign a new course director, or cancel the elective altogether. Each of these uses would require different information. Alternatively, students might want to use an evaluation on the same course to decide whether it will meet their needs, when to take it, what books to buy, and what they will need to do to get a good grade. Different information would be required to inform these decisions.

It is also helpful at this stage to consider whether you need a summative evaluation, a formative evaluation, or both. Summative evaluations typically come at or near the end of a program or course. The intent is to inform a final pronouncement on the course: did it work or not? Formative evaluations, on the other hand, seek to identify areas of strength and weakness so that a course can be improved. Formative evaluation often occurs at the end of the program, occasionally at the beginning, and frequently at various points along the way. Formative feedback is usually an ongoing process, whereas summative feedback typically takes place at a single time point. Of course, the two are not mutually exclusive – you can do both if needed (but see Tip 12).

Tip 3. Do not confuse evaluation with assessment

Educators commonly use the terms evaluation and assessment interchangeably. However, I find it useful to distinguish evaluation, which focuses on programs, from assessment, which focuses on learners (see Wilkes & Bligh 1999). Simply put, you assess learners to determine how well the learner is doing, and you evaluate programs to determine their merit or worth. Since learner assessments often comprise a substantial portion of a comprehensive program evaluation, precise language can help avoid confusion when planning and when presenting results.

Tip 4. Get input from others

The adage 'two heads are better than one' is true in planning and conducting an evaluation. Seek input, not only from other educators, but also from other stakeholders such as students and administrators. Both those who will use the evaluation and those who are providing data (e.g., students and teachers) can offer suggestions on what they might find important. It is also helpful to examine what others have done by, for example, searching the literature. This can provide models to emulate, identify mistakes to avoid, suggest specific questions to answer, and offer specific measures to employ.

Tip 5. Consider various evaluation paradigms and approaches

In determining how to meet your audience's needs and their intended uses, it is helpful to consider a variety of evaluation paradigms and approaches. I will touch briefly on three of the countless approaches that have been described.

Objectives-oriented

First, and probably the best known to educators today, is the objectives-oriented approach. In this method you define the instructional goals or objectives at the start of the activity, and then at the end you evaluate to determine if these goals have been met. The specific outcome(s) studied depends on the objectives (outcomes will be discussed next). The strength of this approach lies in its simplicity – it facilitates a relatively uncomplicated design and straightforward interpretation of results. However, the objectives-oriented approach has several disadvantages. First, it promotes tunnel vision and, by focusing on predetermined objectives, tends to be rather inflexible. It is poorly suited to capturing developments that arise unexpectedly during implementation, and the evaluator can wind up a 'slave' to the objectives. Second, if objectives are not carefully chosen the corresponding outcomes can potentially be trivial ('learners will enjoy the course') or infeasible ('participants will become internationally-renowned experts in this topic'). Third, educators may focus on achieving the outcomes themselves rather than facilitating lasting learning (teaching to the test), and they may inadvertently neglect other important teaching points. Despite these limitations, the objectives-oriented approach has been, and likely will continue to be, a powerful and popular method of summative evaluation.

Process-oriented

Next comes the process-oriented approach. In its most complete execution, this evaluation begins collecting data at the very inception of the idea for the program. It starts by determining if a need exists and if so what is the best way to meet that need. It then tracks the development process, monitors what actually happens during implementation, and typically concludes with a summative objectives-oriented evaluation at the end. The advantage in this approach is its comprehensiveness – providing information on each step in the program from start to finish. It thus provides both formative

and summative information. However, this comprehensiveness comes at a price. It is very resource-intensive and complex, and generates voluminous data that may be difficult to interpret. It also requires tremendous foresight: once the program is underway, it is often too late to go back to the beginning and start collecting data. Thus, although the process-oriented approach presents a powerful technique, it is usually employed in part rather than in full.

Participant-oriented

Finally, we have the participant-oriented approach. This approach seeks to determine how the people involved perceived the program. It typically employs qualitative methods in which data collection and analysis follows an inductive and iterative process, with an ongoing cycle of data collection, data interpretation, recognition of need for specific additional data, and more data collection. Triangulation – the inclusion of data from multiple perspectives – is the key, and the evaluator will often solicit input from participants other than the learners themselves, such as teachers and support staff, as well as non-human sources such as the course syllabus or minutes of planning meetings. This approach captures the complexity of a large program, including the local context, and includes a flexibility that enables it to respond readily to unintended effects. It also tends to be humanistic – focusing on the participants and their needs rather than intangible objectives and processes. However, as with the process-oriented evaluation, this comes at the expense of cost and complexity. Also, both the data and its interpretation are highly subjective and strictly speaking apply only to the local context, which may bother some audiences. Nonetheless, the participants-oriented approach (and qualitative approaches in general) is seeing increased use.

Tip 6. First select the outcome, then the measurement method, then the instrument, then the modality

When planning an evaluation, resist the tendency to start by selecting a specific instrument or tool (such as 'licensure exam scores' or 'Mini-CEX' or 'SurveyMonkey'). Rather, first identify the *outcome(s)* you feel will be meaningful, then select the *measurement method*, then the *instrument,* and finally, the *modality*. I will discuss each of these in turn.

Outcomes are conceptual and often intangible constructs – things like 'knowledge,' 'communication skill,' 'patient satisfaction,' and 'mortality.' Kirkpatrick (1996) identified four broad classes of program outcomes: Reaction (satisfaction), Learning (knowledge, skills, and attitudes in a test setting), Behaviors (in practice), and Results (effects on patients and society, such as satisfaction, compliance, or health). Concentrating first on the outcome at a conceptual level rather than jumping to a specific instrument helps focus attention on what would be most meaningful rather than what is convenient or familiar. For example, if communication skill will provide the most meaningful information to evaluate my course, then I should figure out how to assess communication skill rather than administer a written multiple-choice test (which would probably be measuring knowledge rather than skill).

There are almost always multiple *methods* to measure a given outcome. For example, to assess knowledge, one could use self-report, a multiple-choice test, or faculty judgment. Each of these approaches has strengths and weaknesses, and each could be implemented in various ways.

We refer to specific implementations of a given method as *instruments*. Specific multiple-choice tests include the United States Medical Licensing Examination Step 1, an end-of-year cumulative exam, and a self-assessment quiz. Using assessment of clinical skills as another example, one method is direct observation of behavior, and one specific instrument using this method is the American Board of Internal Medicine's Mini-CEX (Norcini et al. 2003).

Finally, many instruments can be implemented using different *modalities*. For example, the Mini-CEX can be implemented on paper, on the Internet, or on a personal digital assistance (PDA). A course evaluation survey could be administered on paper, via e-mail, by telephone (not commonly done, but possible), or using an Internet-based tool such as SurveyMonkey (www.surveymonkey.com).

Again, first identify the outcome, then the method, then the specific instrument, and finally the modality.

Tip 7. Consider many different outcomes (and measures and instruments and modalities)

Before selecting the outcome(s) for your evaluation, spend some time thinking about alternatives. I often find it useful to enlist two or three colleagues (see Tip 4) to brainstorm ideas with me, focusing on what will inform the evaluation most meaningfully. I always come away with outcomes I had not originally thought to use, and occasionally I find that an outcome I initially thought would be useful is no longer on the list! Once you have selected the outcome(s), then follow the same procedure with measures, then instruments, then modalities – at each stage considering various alternatives with an open mind before making final decisions. You may find the exercise described in Table 1 helpful in this process.

Tip 8. Select outcomes that align with educational goals

It should go without saying that the evaluation outcomes should align with the educational goals, but unfortunately this is not always the case. For example, I have seen an evaluation in which the educational goal was to enhance learners' communication skills, but the outcome assessed was actually knowledge about effective communication techniques. This knowledge might translate into enhanced communication, but there is no guarantee. It might have been better to assess an outcome more closely aligned with the educational goal.

Today, we see a growing push toward outcomes that reflect impact on patients (Chen et al. 2004). Since the ultimate goal of medical training is to improve the health of patients, this makes sense, and may enhance alignment of curricula with practices

Table 1. Thinking creatively about outcomes, methods, and instruments in evaluating educational interventions.

Outcome level	Outcome	Method	Instrument	Advantages and disadvantages
Satisfaction	1.	1.	1.	1.
			2.	2.
		2.	1.	1.
			2.	2.
	2.	1.	1.	1.
			2.	2.
		2.	1.	1.
			2.	2.
Learning	1.	1.	1.	1.
			2.	2.
		2.	1.	1.
			2.	2.
	2.	1.	1.	1.
			2.	2.
		2.	1.	1.
			2.	2.
Behavior	1.	1.	1.	1.
			2.	2.
		2.	1.	1.
			2.	2.
	2.	1.	1.	1.
			2.	2.
		2.	1.	1.
			2.	2.
Results	1.	1.	1.	1.
			2.	2.
		2.	1.	1.
			2.	2.
	2.	1.	1.	1.
			2.	2.
		2.	1.	1.
			2.	2.
Behavior (example)	*1. Frequency of thyroid exam*	*1. Self-report*	*1. Monthly e-mail*	*1. Limited recall*
			2. Paper form at end of clinic day	*2. Burdensome*
		2. Patient report (was thyroid examined?)	*1. Paper form at end of visit*	*1. Efficiency of visit*
			2. Letter survey	*2. Cost, recall*
	2. Thyroid exam correctly done	*1. Observation by preceptor*	*1. Mini-CEX*	*1. Global*
			2. Thyroid exam-specific checklist	*2. Will have to develop*
		2. Incognito standardized patient	*1. Checklist*	*1. Development*
			2. Global rating	*2. Nonspecific, unstructured*

Notes: When planning an evaluation, I find it useful to think creatively about the outcomes, methods, and instruments that might be useful. To do this, I try to think of at least two outcomes for each outcome level (I like Kirkpatrick's (1996) hierarchy), then think of at least two methods for each outcome, and two instruments for each method. Finally, I list advantages and disadvantages for each approach. It is often helpful to engage the help of colleagues, students, and other stakeholders in this brainstorming activity. Once the list is complete, you are well poised to select outcomes and instruments that will most meaningfully inform the evaluation you conduct.

In the last row, I provide an example of how one might complete this task when planning the evaluation of a course teaching postgraduate physicians how to examine the thyroid. Objectives for this hypothetical course are that trainees will: (a) Enjoy the course, (b) Correctly perform the thyroid exam, (c) Examine the thyroid on each patient presenting for a general medical exam, and (d) Be able to identify diffuse thyroid enlargement and discrete thyroid nodules. Objectives (b) and (c) lend themselves to evaluation using behavior-level outcomes.

that improve patient care. However, I see at least three reasons not to treat patient-related outcomes as the holy grail of educational program evaluation. First is the risk of misdirected emphasis. For example, if a course's educational goal is to improve knowledge, a focus on patient-care outcomes might cause educators to emphasize algorithms that improve the measured outcome rather than facilitating deep understanding of underlying principles. Second, since some patient-related outcomes are very difficult to assess, there is the risk of selecting an outcome measure (e.g., hemoglobin A1c levels) because it is measurable rather than because it is most important. Third, measuring patient-related outcomes is simply not feasible in many instances. Issues such as statistical power (sample size), outcome sensitivity to change, dilution of effect (students' decisions will be diluted as supervising physicians, other team members, health systems, and patient preferences come to bear (Shea 2001)), and insufficient resources all reduce the chance of demonstrating an impact using patient-related outcomes. Thus, rather than targeting an outcome high on Kirkpatrick's hierarchy presuming it is better, you may have more success if you simply select those outcomes that best align with your objectives.

Tip 9. Consider the validity and reliability (or trustworthiness) of instrument scores

Once you select the outcome and the method, you will begin to consider various instruments. It is best to start with

instruments that already exist, not only because it can save you the work of developing an instrument *de novo*, but because published instruments usually have accrued evidence to support the validity of the information. It is important, however, to select an instrument that really meets your needs (see Tip 2) rather than pulling something off the shelf that does not quite align with your evaluation plan (Tip 8). Frequently, the best solution requires a compromise – adapting one or more existing instruments, in combination with your own original contributions, to create a new instrument tailored to your evaluation needs.

Regardless of the origin of the instrument, when it comes time to interpret the data, you will need to know how well you can trust the results. This requires evidence to support the validity of your interpretations (Cook & Beckman 2006). Note that we do not speak of the validity of the instrument itself, but the validity of interpretations (Downing 2003). The same instrument, applied to different uses, may provide for more or less valid interpretations. Moreover, validity is not a yes/no variable – it is a matter of degree. Higher-stakes evaluations will require greater validity evidence.

To support a proposed interpretation (such as communication skills), we can collect validity evidence from five sources: content, relations to other variables, internal structure, response process, and consequences (for details on these evidence sources see Downing 2003; Downing & Haladyna 2004; Cook & Beckman 2006). Validity evidence is often available (published, or unpublished from authors) for existing instruments. However, you should still collect fresh evidence to support the validity of inferences in your educational context. For new or modified instruments you will need to collect original evidence. Again, other sources describe this in greater detail.

Although these principles have generally been applied to quantitative measures, evidence should also be collected in qualitative studies to support the *trustworthiness* and *meaningfulness* of the data and interpretations (Lincoln 1995; Côté & Turgeon 2005).

Tip 10. Pilot test the evaluation process

Once you have identified or created your instrument(s), take time to pilot test the instrument and the data collection process prior to full-scale implementation. This allows you to detect and correct poorly worded questions, suboptimal formatting, and problems in administration and collection. In reviewing the results of the pilot test, consider each of the sources of validity evidence in turn. While the importance and relevance of a given piece of evidence will vary depending on the situation (i.e., you do not necessarily need to collect validity evidence from each source), this exercise helps identify problems that might otherwise go unnoticed. It also anticipates the final evaluation report, in which the pilot test data should be reported as an important step in the evaluation process.

Tip 11. Obtain a sufficiently large and representative sample

The saying goes, 'If you've seen one case of chest pain, you've seen one case of chest pain.' Being able to identify myocardial infarction in high-risk men does not guarantee that I would recognize ischemia in a moderate-risk woman, or pulmonary embolism or gastroesophageal reflux. The concept – known in many fields as content specificity (Norman 2008) – that it often takes more than one question (frequently many more) to appropriately assess a learner or evaluate a course holds true across most outcomes, methods, and instruments. Even using qualitative methods, a single question will rarely suffice. Hence, it is important to adequately sample the content domain, whether it is learner perceptions of the course, knowledge of chest pain, communication skill, or patient satisfaction. Obtain enough information to see a clear picture of what you are trying to evaluate. The instrument should be as long as needed – but no longer.

Similarly, when using evaluation surveys or assessments of learning, be sure you have collected data from enough individuals. Sampling methods and sample size will depend on the evaluation design. Quantitative outcomes typically involve statistical tests of inference, and sample size can be calculated using procedures available in standard texts. Qualitative studies might intentionally select participants to provide contrasting perspectives (purposive sampling), and continue obtaining information until no new themes emerge (saturation).

Tip 12. Plan ahead and be realistic (you can not have it all)

You have realized by now that a lot of work goes into planning an evaluation. Yet it is better to invest energy up front, and save yourself the headache and frustration that comes when a poorly planned evaluation fails to provide the information required to satisfy the audience and facilitate necessary decisions.

One of the most difficult tasks in the planning process is determining where to draw the line. As the plan evolves, the list of desired information tends to grow longer and longer. It could easily reach the point that data collection instruments (e.g., questionnaires) exceed a reasonable length, or that demands on time and other support surpass available resources. You will have to make choices that retain the highest quality and most important (most meaningful) data. Referring back to Tips 1 and 2 will facilitate this decision process.

Acknowledgments

I thank Karen Mauck, MD, MSc and Richard A. Berger, MD, PhD for their constructive comments on this manuscript.

Declaration of interest: The author reports no conflicts of interest. The author alone is responsible for the content and writing of the article.

Notes on contributor

DAVID A. COOK, MD, MHPE, is an Associate Professor of Medicine and Director of the Office of Education Research, Mayo Clinic College of Medicine, and consultant in the Division of General Internal Medicine, Mayo Clinic, Rochester, Minnesota.

References

Beckman TJ, Cook DA. 2007. Developing scholarly projects in education: A primer for medical teachers. Med Teach 29:210–218.

Chen FM, Bauchner H, Burstin H. 2004. A call for outcomes research in medical education. Acad Med 79:955–960.

Cook DA, Beckman TJ. 2006. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med 119:166.e167–116.

Côté L, Turgeon J. 2005. Appraising qualitative research articles in medicine and medical education. Med Teach 27:71–75.

Downing SM. 2003. Validity: On the meaningful interpretation of assessment data. Med Educ 37:830–837.

Downing SM, Haladyna TM. 2004. Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ 38:327–333.

Glassick CE. 2000. Boyer's expanded definitions of scholarship, the standards for assessing scholarship, and the elusiveness of the scholarship of teaching. Acad Med 75:877–880.

Kirkpatrick D. 1996. Revisiting Kirkpatrick's four-level model. Train Dev 50(1):54–59.

Lincoln YS. 1995. Emerging criteria for quality in qualitative and interpretive research. Qual Inq 1(1):275–289.

Norcini JJ, Blank LL, Duffy FD, Fortna GS. 2003. The Mini-CEX: A method for assessing clinical skills. Ann Intern Med 138:476–481.

Norman GR. 2008. The glass is a little full – of something: Revisiting the issue of content specificity of problem solving. Med Educ 42:549–551.

Shea JA. 2001. Mind the gap: Some reasons why medical education research is different from health services research. Med Educ 35:319–320.

Wilkes M, Bligh J. 1999. Evaluating educational interventions. BMJ 318:1269–1272.

Worthen BR, Sanders JR, Fitzpatrick JL. 1997. Program evaluation: Alternative approaches and practical guidelines. 2nd ed. New York: Addison Wesley Longman.