Assessing What Matters

David Geelan

The following article has been adapted from a workshop for MCATA members held at Barnett House in Edmonton on April 22, 2005.

Educators, both those who work in the classroom and those who support them, are increasingly getting involved in assessing and evaluating programs and innovations. This kind of work used to be the province of university faculty and outside consultants. In many ways, the involvement of those who are implementing programs in the assessment of those programs is an exciting development. However, teachers may find it challenging to undertake assessment activities when they feel that they do not have the relevant training and expertise. This article provides an overview of some of the issues in program assessment and evaluation in educational contexts, and points to further reading for those who need support in conducting assessments. It also aims to make you a more informed reader of research and evaluation reports, so that you can critically evaluate the claims made by others.

What Matters?

In conducting an evaluation (I will define the term *evaluation* in the next section), it is important to know what matters. Why is it important to assess this program? What are the goals of the program? To what extent is the program meeting those goals? Are the goals appropriate? Deciding what matters is essentially a value decision: what matters to us is what we value highly. The decision should be made reflectively rather than reactively; that is, rather than letting someone else dictate what matters, we should make a principled choice.

It is also important to ask, Matters to whom? Whose interests are served by the program or innovation, and how well? Whose interests might suffer?

Deciding what is important gives the assessment a centre that it might otherwise lack. Many of the other decisions that must be made in conducting an assessment become simpler when the key values are clear. In a collaborative assessment project, deciding what matters becomes even more important: if what matters remains implicit and is not discussed, the members of the team might be working in different directions and toward different goals, and applying different standards.

Assessment, Evaluation, Research

Perhaps my working definitions of the key terms I will be using—*assessment, evaluation* and *research*—would be useful here. You would think that by now we would have these terms pretty well defined in education, but I have seen them used in different ways in different places. Your own definitions might vary, but I want to make it clear what I mean when I use these terms.

In its simplest terms, *assessment* means measuring something. Assessment does not include a value judgment: it notes that your speed is 80 km/h, but it is not concerned with whether that is good or bad. Assessment consists of finding ways to measure things. It need not always yield a number. A story that richly describes what happens in a classroom, without making judgments about what is described, is a form of assessment.

Evaluation includes the root word *value*, so I define *evaluation* as making a value judgment based on the evidence collected in an assessment. (When someone makes a value judgment independent of the evidence, we call that prejudice!) Value judgments depend on context: on the highway on a sunny day, 80 km/h might be too slow, but on a snowy city street in front of a school, 80 km/h is much too fast.

It is necessary to make value judgments in education. Essentially, we do what we do because we value it; if we do not value it, we stop doing it. For example,

if our assessment of an innovative teaching strategy shows that its use has increased students' grades across the board and has particularly helped students who were failing the course, we are likely to use the results of that *assessment* to make a positive *evaluation* of the strategy. Of course, an evaluation can be more sophisticated than a single measure. We might also notice that students and teachers are more tired and that absenteeism increases when the new strategy is used. That makes the value judgment more difficult and brings us back to the question, What matters (most)?

Of course, in practice it is often difficult to make an assessment without also making an evaluation. Moreover, those who commission an inquiry into a program often *want* an evaluation; they want to know whether the program should be continued, expanded or scrapped. For that reason I will here tend to talk about *evaluation* rather than *assessment*.

I define *research* as seeking to understand something in a new way. That might include discovering genuinely new ideas or theories, but it might also include activities such as re-evaluating old theories or looking at old practices through new theories. Research usually includes assessment, but it usually should not be evaluative in itself; that is, research should aim to present a strong assessment of *what is* rather than focus on *what should be* and whether *what is* measures up. This rule (which might really be merely a preference of mine!) is not always followed, but it is usually better if researchers can assess a situation richly and leave evaluation to the reader.

I have done many kinds of research in many contexts, and I have even written a sort of textbook on qualitative research (Geelan 2003). I have also done a number of program evaluations. The rest of this discussion mingles the two fields, because many of the tools, methods and approaches used in research are also used in evaluation.

Purposes of Assessment and Evaluation

Assessment in education might be done for any of a number of reasons, including the following:

- *Measuring achievement.* Every teacher conducts simple assessments when grading tests and assignments and writing report cards. The data from these assessments can also be used in research and program evaluations.
- *Comparisons.* Though we might have ethical misgivings about the ways some comparisons (for example, league tables of school achievement) are used, some comparisons between students, between schools, between provinces and so on can help us improve learning.
- *Evidence in research.* Some kinds of assessment are done purely for the purposes of research or program evaluation.
- *Diagnosis, and support for teaching and learning.* Teachers also conduct many formal and informal assessments of student understanding, both formatively and summatively, to improve teaching and learning.

Similarly, evaluation might be done for any of a number of reasons, including the following:

- *Decision making about programs, strategies and technologies.* Should we put energy, money and other resources into a new program, or redirect them elsewhere? Do we need the latest and greatest technology, or will what we already have serve our students' real learning needs? These questions are better answered using carefully designed evaluations rather than ideology, bias or gut instincts.
- *Ranking.* Given that not everyone who wants to go to university can have a place, how do we decide who gets to go? And given that not all schools are the same, what do we do about funding? (Hint: Taking it away from schools that are already struggling is the wrong answer!)
- *Decisions about how to apply scarce resources.* Given that we do not have unlimited resources in education, where can existing resources best be applied? Where will they be the most effective and efficient in supporting better teaching and learning for all?

Research is generally done because the researchers imagine that it will contribute new understandings, but—let's face it!—it is also done because academics must publish or perish.

You can probably think of more purposes to add to these lists.

Types of Research and Evaluation

There is a huge range of types of research and evaluation. We often think first of what Shulman (1986) has described as "process–product research"; that is, we try something new in the classroom and find out what happens. We might do this with all the trappings of experimental models like those in the sciences: a single independent variable (the thing we change) and dependent variable (the thing we measure), controlling as many of the other variables as possible,

including using a randomly chosen experimental group and control group. Or, realizing that humans are not as simple as atoms and that they do not quite fit into a true experimental framework, we might use a quasi-experimental design in which students' earlier behaviour acts as the control for their later behaviour. Process–product research and evaluation is often (though not always) quantitative; that is, what is measured is expressed in numbers and subjected to various kinds of statistical analysis.

At the opposite extreme in many ways but important in education is action research, in which the aim is to change what is happening in a particular context by understanding it better rather than to merely capture a snapshot of the situation. The evidence used in action research can include numbers, but it may also include teachers' observations of their students, reflective journal entries, interviews, qualitative and open-ended surveys, and a variety of other information. In between these two extremes—experimental, process–product research and action research—remains a range of research commitments, approaches and methods.

A program evaluation can use a wide variety of evidence, from the quantitative to the qualitative (defined below), in seeking to make judgments about the program's value.

Paradigms and Research Methods

A paradigm is a set of related beliefs, theories and assumptions (Kuhn 1970). The two main paradigms in educational research are often called *quantitative* and *qualitative,* although maybe *positivist* and *post-positivist* would be better labels. The basic characteristics of each paradigm are listed below.

The Quantitative Paradigm

- Is modelled on the methods used in the physical sciences
- Measures quantities of things; yields numbers as data
- Uses simple models of cause and effect
- Uses the scientific method—dependent, independent and controlled variables
- Uses validity and reliability as the standards for judging quality

Quantitative methods are usually fairly linear: formulate the question(s), create an instrument (survey or test), gather the data, analyze the data and write a report.

The Qualitative Paradigm

- Is modelled on methods from the social sciences and humanities
- Measures qualities of things; yields sophisticated descriptions
- Recognizes the complexity of educational contexts
- Uses trustworthiness and authenticity as standards

Qualitative methods tend to be iterative: formulate the question(s), gather some data, analyze the data, revise the questions, gather more data, analyze again, gather more data, analyze, report, revisit . . . .

The basic assumptions of quantitative research and qualitative research are different. Nevertheless, combining aspects of both paradigms can often be much more powerful than limiting yourself to one paradigm and set of methods.

Context

To be useful, research and evaluation reports in education must explain in great detail the context in which the research was conducted. That allows readers to make sense of the findings and to think about how their own contexts are similar and different in order to determine how useful the results might be to them. Relevant variables include the students' age, grade and socio-economic status; whether the context is urban or rural; the subject areas taught; the characteristics of the teacher(s); whether the students have any special needs; the history surrounding the program or innovation; and a host of other factors. Of course, at some point you have to stop reporting the factors (unless you want the report to be the size of a phone book), so you must think carefully about which contextual factors are the most important in allowing readers to understand and apply the findings. As a reader of research and evaluation reports, be aware of the presence or absence of these contextual cues.

Research Question

Your research question should be rooted in your values—in your own notion of what matters. That does not mean that you should go into research looking for ammunition to shoot down those who disagree with you; rather, it means that the research should be meaningful and important to you. That is what helps keep you interested and committed when the work gets hard (and occasionally boring).

How the research question is phrased will depend on the kind of project you are doing—research, action research or evaluation. Trying to answer too many questions can lead to an unfocused research process, so try to have only one research question or a small set of related questions. Phrase the question early in the process, but realize that it will likely evolve as the project goes on, particularly if you are working in a qualitative, iterative mode.

A key consideration is what evidence you will need to gather to answer the question. I have been known to say, "If you can't get the data you love, love the data you can get"—but that is not really good advice. If we do research that draws on only the information that is easy to get (the low-hanging fruit), whole aspects of education will be ignored or misrepresented. Find the question you *really* want to answer, and then think seriously about what kinds of evidence you will need in order to credibly answer the question.

The final important issue to think about is the scope of the question. Is it big enough to be worth pursuing but small enough to be manageable? Taking on a question that is too big might mean that you never finish, but taking on a trivial question fails the test of catalytic authenticity (discussed later).

Standards

Twenty-five years ago, the education community largely agreed on the appropriate standards that were to be applied to educational research and evaluation—quantitative (positivist) standards of validity, reliability, objectivity and generalizability, defined in terms of how the research accurately represented reality. That meant that, in writing up evaluation reports, no one had to be explicit about what standards they were using; it was assumed that the quantitative standards applied.

That is no longer the case. Below, I outline the quantitative standards in a little more detail and then outline an important set of qualitative standards for contrast.

Quantitative (Positivist) Standards

- *Validity.* Does the study measure what it claims to measure?
  - *Construct validity.* Is this variable related to other variables in the way the theory requires?
  - *Face validity (or content validity).* Is the measure actually measuring variables in the right domain?
  - *Criterion validity.* Does this measure of the variable correlate with a known correct measure of the variable?
- *Reliability.* If we measure the same thing again, will we get the same results?
- *Objectivity.* Are the results independent of the biases of the researcher(s)? Would everyone get the same results? Are the results researcher-independent?
- *Generalizability.* Can these results be applied everywhere? Can they be generalized to other contexts? Are the results context-independent?

In short, validity is concerned with whether a test measures what we think it measures, and reliability is concerned with whether the test measures it the same way repeatedly. It is possible for an instrument to be reliable without being valid. For example, a reliable rifle will have a tight cluster of bullets on the target, whereas a valid rifle will have the bullets clustered around the centre of the target. It is possible to imagine validity without reliability and vice versa. A good quantitative study or evaluation will aspire to both. To extend the rifle metaphor, objectivity measures the extent to which a different shooter would get the same results, and generalizability measures whether changing to a different rifle range would change the results.

Qualitative (Postpositivist) Standards

There are many sets of quality standards for qualitative research, but Guba and Lincoln's (1989) parallel criteria, or trustworthiness criteria, have been very influential. These criteria attend to the same issues addressed by validity and reliability in the quantitative paradigm, using the assumptions of the qualitative paradigm.

Trustworthiness Criteria

- *Credibility (parallels validity).* To what extent can the research credibly claim to be measuring what it has set out to measure?
- *Transferability (parallels generalizability).* To what extent are the research results useful in contexts other than those in which they were obtained?
- *Dependability (parallels reliability).* To what extent will the results be similar if the research is done again?
- *Confirmability (parallels objectivity).* To what extent will the results be similar if the study is done by another researcher or team?

The concerns of qualitative inquiry, however, are broader than these technical standards for the quality of the research. Qualitative research also applies moral standards in relation to protecting the research participants (the term *participants* is preferred over

subjects) and to the vexed question of how we as researchers can be so audacious as to claim that our work is a fair representation of the views and needs of others. Guba and Lincoln (1989) add the following authenticity criteria to remind researchers to pay attention to these ethical and political dimensions of their work.

Authenticity Criteria

- *Fairness.* Are the representations of others fair? Will the participants recognize themselves in the accounts of them?
- *Educative authenticity.* Does everyone involved (the researchers and the participants) learn something?
- *Ontological authenticity.* Does the research enhance understanding of its social context (the constructed realities in which the research is occurring)?
- *Catalytic authenticity.* Does the research make something happen? Does something change because the research was done?
- *Tactical authenticity.* Are the methods used in the research consistent with the values implicit in the work and in the educational context of the work?

A variety of standards corresponding to the two paradigms are available now. Thus, it is crucial to carefully choose the standards you will apply, and to clearly state in the evaluation report which standards you chose and perhaps why.

Writing and Publishing

Academics are driven by publication, but classroom and district educators are far less so (unless they are doing graduate studies). So why write a report of your results? Where is the payoff?

The payoff may not be great in financial or career terms (although a publication listed on a CV is increasingly coming to mean something for teachers). However, you will be reporting your results as a service to the profession. Writing a high-quality report that outlines the context of the study in detail and in clear, accessible language allows you to share your ideas, experiences and results with colleagues in similar and different contexts. We all value learning, and we are obliged to share what we learn with others who are trying to enact similar values in their teaching.

For teachers, publishing in peer-reviewed academic journals is much less important than it is for academics. It is more valuable for teachers to publish their work in publications that make it accessible to other teachers—such as *delta-K*. Web publishing is another way to share your knowledge and experiences as widely as possible.

Conclusion

An evaluation is about assessing what matters. Do not do an evaluation with only the evidence that is easy to find; rather, use the evidence that allows you to do a high-quality, well-supported evaluation of a program or innovation that is important to you, using clearly defined standards. Then, share your findings as broadly as you can.

References

Geelan, D R. 2003. *Weaving Narrative Nets to Capture Classrooms: Multimethod Qualitative Approaches for Educational Research.* Boston: Kluwer.

Guba, E G, and Y S Lincoln. 1989. *Fourth Generation Evaluation.* Thousand Oaks, Calif: Sage.

Kuhn, T S. 1970. *The Structure of Scientific Revolutions.* 2nd ed. Chicago: University of Chicago Press.

Shulman, L S. 1986. "Paradigms and Research Programs in the Study of Teaching: A Contemporary Perspective." In *Handbook of Research on Teaching,* 3rd ed, ed M C Wittrock, 3–36. New York: Macmillan.

David Geelan is Sue's husband and Cassie and Alex's dad. He reads two or three novels every week and spends more time playing computer games than he should. He has taught junior and senior high school chemistry, physics, math and general science in several Australian states, and has worked as an educational researcher and teacher educator in Papua New Guinea (1993/94), Australia (1995–2001) and Canada (2001–06). He is currently an associate professor of science education at the University of Alberta, where he conducts research in high school physics classrooms using video analysis and helps prepare the next generation of Alberta teachers.