Development of Evaluation Centers and Training of Evaluators[1]

Robert E. Stake

Abstract: *Evaluation practice, as a formal activity, has been and is still characterized by three key elements: (i) the supremacy of a way of reasoning based on criteria and standards, (ii) the promotion of the use of tests and other quantitative techniques and (iii) the idea of quality as an intrinsic attribute of programs. Training in evaluation has followed the same line and it has mainly focused on models and methodological approaches. As Stake points out, many problems arise with such a perspective when identifying the effectiveness of an educational system. Another (often devalued) kind of analytical thinking is also needed: that of episodic thinking, which emphasizes (i) that program should be properly contextualized and (ii) that program quality depends greatly on who experiences it. Both kinds of thinking are required. They must be combined as if they were two sides of the same coin. Therefore, "binocular vision" is the best option (commitment) for evaluation.*

Keywords: *Standards based evaluation, responsive evaluation, evaluation centers, training in evaluation, binocular vision.*

1. EVALUATION CENTER

My first topic is the development of Evaluation Centers. Later, the training of evaluators, and finally to talk about conceptualizing school quality. I thought at first this would be just a technical topic, but upon reflection, I realized that thinking about the development of centers causes us to reflect on the whole territory of educational evaluation. That territory includes:

- The search for quality, with wide differences in the definition of quality.
- The assessment of outcomes, particularly student performance outcomes.
- The political nature of evaluation.
- The promotional climate of institutions, with efforts to create images of quality.
- The global nature of education, including working with the internet.

Before moving into that vast territory, I should say an additional word about who I am. I come from the qualitative side of educational evaluation. Once I was associate director of the Illinois State Testing Program, but, disappointed with test-based program evaluation, I looked elsewhere. I came to believe that evaluators need to look closely at the experience of students, teachers and others in order to understand the activity, complexity, and quality of education. So I give preference in my data gathering and in my teaching to dialogues, vignettes, and situational contexts. Here is an example:

[1] The present work is a transcription of a lecture delivered by the author in the International Evaluation Symposium held at the Universidad Nacional Autónoma de México (México D.F.), on October 25, 2007. It has been reviewed by anonymous referees who were informed that it was a specialist lecture.

Mother of twins: You have taught Pepe to use a camera, but not Pepita.
Teacher: But, Mrs. Martin, Pepita. needed to do her math.
Mother: It's not fair.
Teacher: I love them both. I wouldn't be unfair.
Mother: Pepita hates math. Can 't you make school as good for Pepita as you do for Pepe? Teacher: They are very different children.
Mother: You ought to do right by both.

Hearing this small conversation, we pause to think about both educational philosophy and teaching practice. What constitutes equal opportunity in education can become better understood in a collection of dialogues like this one? Is it not the responsibility of educational evaluators to examine such difficult issues as equality of opportunity?

I see evaluation as the business of searching for quality (which means lack of quality as well, of course). By the end of this paper I will be talking about school quality. But wherever it is, I like to examine a number of dimensions of quality, sometimes to represent quality with measurements, but also to speak of episodes as well as measurements that reveal quality.

Dimensions and episodes, qualitative and quantitative, those are mixed methods, yes, but I see qualitative methods as particularly good to help us understand educational quality. Again and again this bias of mine will show up in what I say today.

1.1. The Typical Center

Let us consider evaluation centers on campus. Here are five common expectations:

- The typical center will be staffed with one or a few professors with help from graduate student researchers,
- It is a craftsman shop more than a computation center
- The director will need to be someone who really wants to do a lot of hard work.
- There is funding is seldom sufficient to do evaluation at level of quality of other campus research.
- The center will fit the uniqueness of the local situation, the facilities, the market, the sponsors, the nearby structures of education.

Some centers will be quite different, of course.

Let us think of a conventional way of conceptualizing the work of an evaluation center. We know of a campus ethic that sees the work of the faculty in three dimensions: teaching, research, and service. Service includes assistance to their own department, to the profession, and to the community. Let's look at what an evaluation center does:

- A center participates in campus teaching about both the skills of evaluation and the profession of evaluation, and sometimes teaching the public about evaluation.
- A center does research, sometimes under contract, research both on theory and practice of educational evaluation.
- A center provides evaluation services, some of it unpaid, in the forms of consultation, project assistance, program development and remediation.

In sum, a center has more work than it can do! Each center is different, choosing among its opportunities. The center works to make itself distinct.

1.2. All Departments Evaluate

As has been eloquently explained by Michael Scriven (1996), evaluation is a fundamental human process. Most evaluation is informal, using intuition and common sense, but much of it is formal. There is no fixed line between formal and informal. Formal evaluation is more deliberate, more reliant on record keeping, more disciplined.

Across the campus and in every agency, there is a panorama of evaluation scenes. Much evaluation work is not called evaluation, so few people look to the evaluation center for expertise they could use. Whatever he or she calls it, the evaluator searches for quality, somehow acknowledging local and universal definitions. Consider some of the workplaces:

- In colleges and agencies of education, they evaluate student and teacher success.
- In colleges of medicine and hospitals, they evaluate health and treatment. In
- In colleges of law, students learn to evaluate happenings in life as legal and illegal.
- In military schools and camps, professionals at all levels evaluate strategy & tactics.
- In intelligence agencies, they evaluate security risks and threats.
- In regulatory agencies, they evaluate risks, protections and adherence to regulations
- In business schools and companies, they evaluate work routines, management, sales, and advertising.
- In journalism schools, they teach investigative journalism, highly evaluative work, (but the media conglomerates now seldom hire them).

There are many workplaces, each infused with evaluation, and more: engineering, the ministry, music and art, workplaces of all kinds.

Most presume the word "evaluation" means something either technical or managerial. Somehow, most people who work on campus have concluded that general evaluation training is not of value. But any of these are fields to work with, if the people of the evaluation center choose to extend in those directions.

Lots of places need help with evaluation. An evaluation center will compete with other evaluation shops to get contracts. Sometimes the competition is rough. Some competitors will submit low bids and high promises, even promises to do assessments more precisely than has ever been done before. I have lost competitive bids to bidders who later came to me to find out how to do the work.

Forty years ago in the USA, most externally-funded educational programs were evaluated by campus-based centers. The staffs at these centers were able and polite people. But when funding for evaluation got much bigger, and the consequences more serious the competition grew and the bidders became less polite.

Let us identify places offering evaluation services in many countries today. In other words, here are six sites where evaluators do formal evaluation work:

- Big evaluation corporations, such as American Institutes for Resarch, Abt Associates, Rand Corporation.
- Internal program evaluation offices in large government agencies and corporations, set up just to work on the inside, such as the US General Accounting Office, and evaluation offices in the World Bank.
- Small evaluation businesses, often set up by someone leaving a larger center.
- Small evaluation centers. One of the small centers grew big, Dan Stufflebeam's at Western Michigan University, but eventually they had to bid for some contracts they didn't want because they needed to pay the permanent staff.
- And there are large numbers of individual persons who have little or no training in evaluation but are visible and available to do evaluation in their fields .
- And another group of in-some-ways distinguished people are willing to serve on blue ribbon panels or crisis groups or special boards of inquiry, people who have never heard of the word "evaluand," most of whom are honored to be asked to help study a problem and evaluate the situation.

So, to make a long story short, everyone is an evaluator, and a great many are evaluators for hire. An evaluation center should have ties with many of them, and offer training and assistance to lots of others.

On technically-advanced campuses in economically rich countries today, the way I see it, there are opportunities to set up campus evaluation centers. But I hear few requests for them and little expression of need.

It is widely assumed that these campus centers will have values similar to the disciplinary areas, that is, the social sciences and the physical sciences, faculties digging for truth. It is not assumed that they will be highly sympathetic to ordinary people facing obstacles in ordinary living.

The high majority of program administrators asking an evaluation center about doing an evaluation study are seeking an endorsement of their program more than they are seeking a greater understanding of it. They seek promotion more than insight.

And most of us evaluators see their need for support and offer our assistance, especially if paid. And we find all the good things we can. But we also want to identify the bad things we perceive. The question arises, will the evaluation center have the courage to point out trouble? Will we properly assess shortfalls of quality, I think that most campus evaluation centers are less secure than individual evaluators, and show reluctance to find fault.

Part of the business of educational centers is to teach courses, to train evaluators, and to help the academic disciplines, the professions, and the public to understand that formal evaluation is a discipline. Evaluation has theories, methods, practice, and ethics. Informal evaluation is a trait. Formal evaluation is a discipline.

2. THE TRAINING OF EVALUATORS

Just as evaluation centers are different, adapted to their situations, the training of evaluators has been, and should continue to be, situational, adapted to the particularities of the students, the instructors, and the institution. In my view, it would not_serve our profession well to further standardize the practice of evaluation. We can imagine another world where the practice of evaluation is so diverse, so uncommunicative, so customized, that it confuses clients and the public. That isn't the world I know. So training also should be diversified.

Mentoring and project training should serve to diversify the education of evaluators, and so should classroom instruction. Opportunity should be provided in the classroom to allow the trainee to adapt concepts and methods to his or her special field of work.

Many courses in evaluation emphasize different models or methodological approaches. I think thinking about these models distracts new trainees and professional development participants from conceptualizing the big questions of evaluation. As I see it, the big questions include:

- Does evaluation always make comparisons to standards?
- Should evaluation be limited to the questions desired by the client?
- What is quality?
- Can evaluation be too closely linked to program improvement?
- What are the political influences on evaluation?
- Are professional standards for evaluators always helpful?

The training of evaluators should include such questions. But when I wrote an evaluation book (*Standards-Based and Responsive Evaluation)* based on questions like these, it didn't sell well. I don't know why.

The book also claimed that being explicit about criteria is not a necessary part of good evaluation, saying sometimes we need to let behavioral descriptions, descriptions of activities, provide the definitions more than words. Part of disciplining is, I believe, relying less on nomenclature and explication.

2.1. Disciplining informal evaluation

One of my strategies for training is to emphasize the continuing disciplining of informal evaluation more than the replacement of it. If informal evaluation could speak, it might argue that it is better than formal evaluation--because with it humans survived for thousands of years. Until recently, humans survived without formal evaluation. Luck was probably involved but it is safe to say that survival meant making a lot of the right choices. And good choosing is a matter of evaluation. Since there wasn't much formal evaluation, as we know it today, before 1950, and even the courts, hospitals, priestly councils and sciences go back only a few centuries, we might well conclude that it was informal evaluation that got us this far, not formal evaluation.

For formal evaluation to be better than informal, we need to be satisfied that it helps us to better recognize goodness, and of course without causing us to change our minds as to what goodness is. Or without incurring other costly side effects.

But even if we persuaded ourselves that usual1y it would be better to evaluate our programs informal1y rather than formally, much of today's world expects formal evaluation to be more trustworthy, more likely to provide evidence of quality or dysfunction that can be relied on. It is usual1y a good idea to think carefully as to whether formal or informal evaluation is a better emphasis. But when the contract says program evaluation is needed, it means formal evaluation. Still there is room within formal evaluation for informal data, judgments and interpretation.

In my mind, I am continuing to talk about creating an evaluation center and the training of evaluators.

2.2. Roles and Styles of Evaluation

Evaluation is the pursuit of knowledge about value. Much of that knowledge comes out of personal experience, often from our own awareness. I know that Allopurinol is good medicine for me because without it I have kidney stones, and with it I do not. What the doctor tells me is important and I can read *The Wellness Encyclopedia*, but I pay a lot of attention to my own experience. I can't always depend on experience but I will use it constantly. And upon the experience of people I know, and the indirect experience and reasoning and research of people I do not know. I put all of the knowledge together, and with a mixture of crudeness and precision, I appreciate the value. It helps also to drink lots of water.

Evaluation is always a determination of merit and shortcoming. Sometimes evaluation does a lot more, but its essential function is the determination of merit. That is the first purpose. That is the definition. That is the *sine qua non*.

> I overheard a woman in the waiting room telling an acquaintance she was an evaluator. "What do you do?" "I help people lose weight." "Why do you call yourself an evaluator?" "People are more inclined to pay attention to what I say if I am an evaluator rather than a dietician."

Dieticians are evaluators too, but if she primarily helps people with weight loss, she is misleading people by passing as an evaluator. Can people call themselves whatever they want? Pretty much, they can. But we will communicate better if we speak of an evaluator who, within some sphere of activity, seeks out and reports the merit and shortcoming of an evaluand. The value determined can be used for lots of different purposes, such as improving a process, awarding a prize, assigning remedial teaching, recognizing contracts fulfilled. These are some of the many roles in which evaluation is used (Scriven, 1967).

2.3. Criterial and Episodic Thinking

The dictionary doesn't require it but in evaluation circles we conventionally think of a *criterion* as a descriptor describing the evaluand, such as its effectiveness, durability, cost, or aesthetic value. And we think of a standard as the amount of that criterion that classifies the evaluand as at some other level of quality. More jargon, but useful.

Forty-eight years ago I was in graduate school in a Psychology Department. One day, sitting at my desk, I realized that there is a Social Science of Education because educational psychologists and sociologists were able to restructure educational phenomena as variables. They had invented the constructs of education, building blocks for disciplined thinking about education, and they called those constructs *variables*. Sometimes they called these same descriptive constructs: attributes, properties, traits, characteristics, facets, and dimensions. Much of the time they called them *criteria.* By reducing the complex phenomena of the classroom, boardroom, history, and community aspiration, to variables, one could get a handle on things. Translating happenings into variables is called *criterial thinking*.

A variable is an attribute that varies. It can vary in various ways, but social scientists decided to emphasize that it varies in quantity. Amounts are seen to vary up and down a scale. So once we identified the construct, that is, the scale, the important thing was to measure the quantity. We can use these quantities to describe, to distribute, to compare, even to make like we are finding causes, and to interpret the identified causes as bases of control, of improvement, of reform. It looked to me forty-eight years ago like harnessing the atom. With criteria, with *criterial thinking*, we could measure, and with measurement scales we could move mountains (Stake, 2004).

With criterial thinking and sampling as our entree, the study of education could be precise, generalization-producing and useful. Any doubts I had were blown away. I enlisted in the science of testing. I devoted myself to becoming a "measurements man," and I am one still. I am a measurement man. My work is program evaluation. I try to measure the quantity and quality of Education, or of Training, or of Social Service, the merit and shortcoming, the elusive criteria of teaching and learning.

My psychology tells me that the alternative to criterial thinking is episodic thinking. Educational phenomena come to be known through episodes, happenings, prototypes, activities, events. The episodes have a time and context base. They are populated with people having personalities, histories, aspiration, frailties. We sometimes talk about personality and frailty, contexts and episodes, in terms of variables. Almost anything can be converted into variables. But the variable-based conversation often simplifies, under-represents. We gain a handle and lose a situation.

To plan the training of evaluation, and to set the ethic of a center, we should recognize that criterial thinking and episodic thinking exist, side by side, in our culture and in our brains. With some kind of binocular resolution, we sometimes can resolve the disparity, unconsciously, into a unity not attainable from either criterial or episodic thinking alone: seeing depth from putting criterial thinking and episodic thinking together. An example of mixing this thinking is in evaluating school quality.

Let's spend a few more minutes on the concept of *quality.* The word is commonly used in two ways. The word qualities refers to the characteristics of something such as the qualities of a musical comedy or the qualities Isabella's teaching. Speaking of these qualities is more a matter of description than judgment. To describe the qualities of music, we use such variables as lyrical, sonorous, earthy, conventional. To describe the qualities of teaching, we use terms such as creative, conventional, child-centered, and lacking focus. Here's a quote from John Dewey: "self interest and sympathy, opposites in quality." These uses of the word quality refer to the nature or ingredients of something, not about its goodness.

The evaluator's meaning of quality *is* about goodness. The quality of a music performance is its degree of excellence. The quality of a teaching episode is its merit and shortcoming. This is the sense in which we usually use the term in evaluation work. We are looking for program quality, meaning its merit and worth. But that leaves much unsaid. Just what the quality or merit or excellence is often hard to specify and agree upon. The standards we set in words or measurements are often simpler, less complicated, than what we experience personally. When we say that the quality of a student's writing is mediocre, we are sensitive to many aspects: its coherence, topicality, grammar, creativeness, penmanship, timeliness, word play, conformity to assignment, even to some qualities that we do not think about in advance--but not all of these every time, and not weighted the same from time to time. Criterial thinking is important but so is interpretation.

As an evaluator, I do not find it necessary to be explicit as to the quality I am looking for. Some evaluators try hard to be explicit. But I am wary of using a single or even just a few criteria. I am wanting to teach my students to become experientially acquainted with a collection of aspects of the program. Some evaluators prefer to devote their resources to measuring the best single or the best several criteria they can.

Another difference among some of us evaluators is our conception of the locus of quality. When you think of the quality of a cantaloupe, you may think of goodness as a property of the melon. Or you may think of quality as defined by responses to the eating of it. Does the quality belong to the melon or to the experience? Notice that the latter requires more attention to who is doing the eating. Evaluators differ as to how much attention they give to who is experiencing the evaluand. When you think of the quality of a performance in performance testing, you may think of it as a property of the performer or some kind of interaction between the performer and the examiners. Its part of your training.

You don't have to agree with me but I like to think of quality as originating in human experience. I see little use for the concept of quality without human response as the key reference. There is quality because people experience quality. And over the ages, the positive experiences that has counted most perhaps are comfort, contentment, and happiness. And of course, the negative experiences of discomfort, anger and fear. What we quickly recognize as high quality teaching or cantaloupe is rooted, I think, in present and past experience with them. By now, we may have developed formal or informal standards, conventions, and traditions, to grade melons and music, but the roots of the meaning of quality are in the emotional experiences evoked over time. A watch is of high quality partly because it keeps good time and has finely tooled parts but also because people find it highly satisfying and superior to other watches they have known. What this makes us evaluators sensitive to, as we go about the business of evaluating things, is to recognize that quality depends greatly on who has been experiencing it.

2.4. School quality

Educational evaluators do not agree on what makes a good indicator of school quality. Many in the USA are happy with student performance scores. Many others prefer school accreditation or school inspection. All agree that the schools are complex and that any single indicator will greatly over-simply the concept of school quality. Yet there is globalized pressure for seeing school quality on a single dimension.

Measuring quality is made extra difficult by the fact that different people want different things from their schools, even some contradictory things, such as more student conformity and more student expressiveness. Many school policies presume these differences are mistakes in communication rather than artifacts of democracy. Quality is a manifold of complexity.

Teachers know much about school quality, usually more than administrators, visiting evaluators and members of governing bodies. But teacher judgments and critical episodes are undervalue. And they are often accused of being self-serving, which sometimes they are. But a good evaluation process can use the criterial and episodic thinking of teachers.

But, hold on! We are talking about training. The training of school evaluators should include questioning of the need for knowledge of overall school quality. Modernist thinking and the management ethic presume there is a need for simple comparison of schools. But as a professional evaluator, I have become persuaded that there is little proper use for measurements of institutional quality. We evaluate school quality as a matter of control, competitiveness, punishment and pride. We too seldom question the need for it.

It is important for everyone in a school to recognize quality, to see problems, and to seek improvement, but improvement comes from the study of particular problems, not by tracking school summary scores over time. Improvement comes from self-study and action research, and sometimes broadens into case study or participant evaluation. The evaluation trainee should learn that there should be no single view of school quality. The school can well be served by people who have quite different views of its quality. School quality and student achievement can better be understood through extended observations of classroom activity. Consider this episode.

> The science class began with examination of seed growth. Each student retrieves a couple of half-pint milk cartons, their own, from the windowsill. Today they are to record the vertical growth of one or more shoots. They had planted forget-me-nots, cosmos, and something called perennial mix. Each student 18 recording in a journal started earlier. The journal cons18ts of unruled paper folded once to make a booklet.

> A student asks, *Do the plants grow faster if you talk to them?*[7] "That's a good question, one scientists have studied. Perhaps they have considered music as a stimulus to growth more than talking. (pause) Kevin, you're putting too much water on it."
>
> *How do they make plants?* "That's something we want to ta1k about today. First let's check the growth. Raise your hand if you have one sprout at least 1&1/2 inches or more tall. ... how about two inches or more? … three inches? (several hands, finally just one.) Syliva, yours aren't that tall." *Yes they are.*
>
> "All right. How many have seen changes since last Monday? (pause) What changes?" *"Two more sprouts." "One caught up. It was smal1er then." "Mine was growing crooked and it straightened up." "Maybe it was too dry, now getting more water."*

The class continues. The teacher gets them to talk about comparing growth. They talk about measuring. She eases them into thinking about an experiment. She designs one. They talk about comparing measurements and averages. She decides to teach them about finding the median, and she teaches it wrong. When hearing the mistake, her mentor is very upset. An evaluator has to raise the question, "Is her teaching about an experiment nullified by teaching the median wrong?"

2.5. Scholastic Aptitude

Evaluation training should be adapted to the individual trainee. But in education, most trainees need to know the invalidity of assessing schools with standardized student achievement tests. Most school assessment in the USA uses scholastic aptitude as an indicator of school quality. That is right, scholastic aptitude of its students.

Scholastic aptitude can be defined as readiness to learn in school, i.e., to profit from academic experience. Aptitude is partly a function of native intelligence, from the immeasurable capabilities drawn from genetic codes--when given ample opportunity to develop. Much of life shapes the development of aptitude, especially early nurturing. Nurturing is visibly the domain of mothers and fathers, but siblings, peers, extended family, then teachers, counselors and coaches--plus social groups and co-workers and all the rest of each culture. All contribute to scholastic aptitude. Our trainees know this.

Since 1900 we have had intelligence tests, created to measure native ability as it developed in our culture. Partly because declaring some groups more intelligent than others is hurtful, the name of the test was changed from intelligence test to scholastic aptitude test. The items were changed slightly if at all, just the label changed. Intelligence is something real. Aptitude is real. Achievement is real. But aptitude and achievement are quite different things.

Political pressure and assessment contracts pushed evaluators to emphasize the measurement of school quality. These pressures create increased public questioning of what the schools are accomplishing. Schools are expensive and they don't keep all their promises and they don't meet all public expectations. They have been pressed to be more "accountable." To measure accountability, our fellow evaluators in the USA took the scholastic aptitude test items (that were once called intelligence items) and started cal1ing them student achievement items. What was once measuring aptitude was now measuring achievement.

In the USA we tested students by the millions. They were mandated by each state. It was ignored that "achievement tests" have very little diagnostic value, very little external validity. Mean scores are given attention in some educational policy discussions. Sometimes sanctions are imposed on those with low scores. It is a fantasy world. Sam Wineburg explained it well in his paper "Crazy for history." (2004). So did Neil Postman in his book, *The End of Education* (1995). The tests do not identify educational effectiveness. And yet U.S. federal education policy and policy in most of the 50 states are largely based on their scores.

These measurements are allusions to achievement, not measurements of achievement. Even so, maybe they help motivate the children. We don't know how much good they are doing our children, or how much harm, but we evaluators should not speak as if we are measuring children's achievement or school quality that way. We are not.

3. FINAL WORD

The training of evaluators is not just about finding performance that meets quality standards but about experiencing situations of merit and shortcoming. Rather than concentrating on good evaluation examples, courses in program evaluation, I think, should pay lots of attention to evaluator's mistakes, obstacles, and coping with problems.

Good evaluation centers also need to locate problems, mistakes, and coping behavior. They need to develop ways of talking about problems, and poor quality, with audiences who have little appetite for that talk. Making our communications contextual, problem-sensitive, and experiential is a challenge to all of us who would create evaluation centers and train evaluators.

BIBLIOGRAPHY

Margen, S. (1992), *The Wellness Encyclopedia of Food and Nutrition*. Berkeley: University of California *Wellness Letter*.

Postman, N. (1995), *The End of Education.* New York: Vintage.

Scriven, M. (1967), "The methodology of evaluation". In Stake, R.E., editor, *Perspectives of curriculum evaluation*. Chicago: Rand McNally.

Scriven, M. (1996), "The skeleton in the discipinary closet". Millercom lecture, University of Illinois, April 23.

Stake, R. E. (2004), *Standards-Based and Responsive Evaluation.* Thousand Oaks California: Sage.

Wineburg, S. (2004), "Crazy for history", *Journal of American History*, March: 1401-1414.