Evaluation’s Generations:
Veneration, Vituperation, or Simply Divestiture

Robert E. Stake, University of Illinois.

I am honored to be with you to open the 32nd annual meeting of the Canadian Evaluation Society. It is a joy to behold, gathering here in Edmonton, members from Halifax to Victoria, from a variety of professions, speaking multiple languages, and ranging the multiple generations. Here is Graphic 1 to remind us of the diversity of our work. The conference theme draws our attention to the generations of evaluation questions and methods.

Graphic 1. Some Workplaces of Evaluation

In my title, I have questioned our attitudes toward generations past. How do we visualize our seniors? How do we visualize our juniors? I have categorized the visualizations as veneration, that is, respect; vituperation, that is, disrespect; and divestiture; that is, disregard. All three are prominent in the conduct of evaluation studies.

We take pride in the durability of ideas from OISE and McGill, from agencies and organizations across the land. We take embarrassment from the misuse of standardized achievement tests and crude indicators. And divestiture: The role of formal evaluation has changed across sixty years, so much that the need for reliability and validity of representations is now repeatedly disregarded. Details later.

---

1 Keynote address at the annual meeting of the Canadian Evaluation Association, Edmonton, May 2, 2011.

2 In the website for the Saskatoon Public School, it is claimed that the Canadian Achievement Test is used to “determine broad educational needs as a basis for setting priorities and allocating resources,” a claim without evidence nor justification in professional practice.
Stake: Evaluation's Generations. CES Edmonton 2011

While I was being introduced today, my modest side whispered, "... you stood on the shoulders of giants."

Many of my sprouts came from seeds planted by Harold Gullickson, Ledyard Tucker, Tom Hastings, Lee Cronbach, Lawrence Stenhouse, Maxine Green and Michael Scriven. Each commands my respect still, but I abandoned some of the wisdom of that generation. Is it the same with you?

Generations. How long is a generation? Take a branch of my family tree. Here is Elija Foster of Pickering, just north of Toronto, born in 1798.

<table>
<thead>
<tr>
<th>Elija Foster, born 1798</th>
</tr>
</thead>
<tbody>
<tr>
<td>son, Elbridge Foster, 1828</td>
</tr>
<tr>
<td>dtr, Margaret Foster, 1858</td>
</tr>
<tr>
<td>dtr, Nelle Coffin, 1891</td>
</tr>
<tr>
<td>son, Bob Stake, 1927</td>
</tr>
<tr>
<td>son, Jeff Stake, 1953</td>
</tr>
<tr>
<td>son, Christopher Stake, 1982</td>
</tr>
<tr>
<td>dtr, Sloane Stake, 2007</td>
</tr>
</tbody>
</table>

From great great granpa Elija to great granddaughter Sloane, 209 years, seven generations. Thirty years per generation.

Evaluation had school grades and accreditation long, long ago, but let’s say evaluation as a profession started in the 1950s. If we were to take 30 years as a generation, I would have only two generations to talk about. That seems too gross. Of course, for many other families, human and otherwise, generations are shorter. Fruit flies and micro-chips, for example.

Perhaps a better definition of generation would feature changes in activity. That generation might end with a significant event or change in practice. In their Handbooks for Qualitative Research (2000, 2006), Norman Denzin and Yvonna Lincoln identified nine intellectual periods since the 1950s. Some good ideas, but I found the slicing too gross.

In my American evaluation world, there have been two significant events, the launch of Sputnik and the election of Ronald Reagan. Sputnik brought a wave of social and education innovation plus a demand and funding for their evaluation. President Reagan brought a realignment of government priorities to benefit corporate economy, while undercutting schools and welfare with demands for evidence of accountability, and driving many Mom and Pop stores of evaluation out of business.

From Sputnik to Nixon-Reagan, say from 1957 to 1975: eighteen years of plenty. Then, a second generation: eighteen years of public services famine, from Reagan to 1993, a second generation. And the third generation lasting from 1993 to the present. What characterizes evaluation in this third generation? Globalism and the hockey stick effect (Monford, 2011), using educational policy for political ends. Evaluation has always been political, but the exponential rise in evaluation politics could be said to start with the Bologna Accords, the efforts of the European Union to standardize education to facilitate the growth of European industry. Okay, arbitrary. It’s a reach. But it gives us three tidy generations to talk about, as displayed in Table 1. [on handout in English and French]

Table 1. Three Generations of Evaluation Theory and Practice

<table>
<thead>
<tr>
<th>Generat’n</th>
<th>President</th>
<th>Ethic</th>
<th>Assets</th>
<th>Obstacles</th>
<th>Premier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1957-1975</td>
<td>Eisenhower Kennedy</td>
<td>... that public programs should be technically and</td>
<td>Demand for validity of measurement and</td>
<td>Willingness to accept designs suitable for social</td>
<td>Diefenbaker</td>
</tr>
<tr>
<td>Response to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(I will not be surprised if you find it too artificial to invent “generations” like these. We classify for convenience, inventing a hypothetical structure, and then we get to thinking we have discovered something, and try to persuade others to use our invention. This method of categorization and staging is to be found in many evaluation reports, sometimes with good logic, but I often find inventing labels a superficial way of displaying insights about our evaluands. But it might work this time.)

**Trends across Three generations.**

In Figure 1, I characterize the Post-Sputnik Generation as an explosion of formal evaluation, especially educational program evaluation. It was dominated by psychometric thinking, emphasis on customized testing of outcomes, with reverence to the Campbell and Stanley chapter (1963) and the *Joint Standards for Educational and Psychological Testing* (1969). They tried too hard to measure goal fulfillment and build theory at the same time.

I characterize the Accountability Generation as a reaction to the expense and, departures from tradition of the Sixties. Change efforts were battled by clamoring for quantitative evidence that money was well invested. Evaluators grossly over-claimed what they could measure. A major reaction was the emergence of qualitative evaluation methods (House, 2006) and mixed methods (Greene and Caracelli, 1997) which provided experiential accounts of programs, products and personnel.

The present generation is in tune with the economic control of government and industry, increasing the use of simple indicators to indicate the quality of complex operations. The World Bank and International Monetary Fund have dominated the major growth area, developmental evaluation (Rist, 1997). Cultural perspectives are awakened (Denzin and Lincoln, 2006). School operations are greatly influenced by standardized testing which pushes internal validity with little regard for external validity. Evaluators have not been able to provide checks and balances on institutional and media uses of evaluation data.

In spite of these macro-changes, evaluation practice has, in my view, remained rather stable. There are some subtle changes, changes across these generations. I identify four changes:

a. an increasing exchange of evaluation information
b. small changes in tools
c. a diminished embracing of the evaluand
d. a diminishing appetite for understanding

<table>
<thead>
<tr>
<th>Sputnik</th>
<th>Johnson Nixon Ford</th>
<th>indepen-dently evaluated</th>
<th>other represen-tations</th>
<th>science theory building</th>
<th>Pearson Trudeau</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975-1993 Accountability Carter Reagan Bush</td>
<td>... that public programs should be held to preordained, formal standards</td>
<td>Qualitative interpreta-tions of merit and shortcoming</td>
<td>Belief that single indicators can represent complex evaluands</td>
<td>Clark Trudeau Turner Mulroney</td>
<td></td>
</tr>
<tr>
<td>1993-2011 Global Economy Clinton Bush II Obama</td>
<td>... that macro-efficiency and productivity rate above social welfare</td>
<td>Multi-ethnic alternative perceptions and knowledge</td>
<td>Weakness of professional standards to oppose misuse of evaluation</td>
<td>Campbell Chrétien Martin Harper</td>
<td></td>
</tr>
</tbody>
</table>


**Exchange of information.** Some of us are slow to move to texting and twittering, but the communication tide raised all our boats. It used to be that workplace contracts were negotiated on the issue of increased productivity, but productivity has increased so greatly in these three generations that productivity is seldom argued. Greater productivity does, of course, not mean higher quality production, but the opportunities to increase quality through communication are great. William Schaffer, international business development manager for Sun Microsystems, has said: "If there's one skill that's required for success in this industry, it's communication skill."

Later I will speak of methods of evaluation and comment that one of the most underused techniques evaluation is the use of reviewers, member checking and specialist panels. I do not mean focus groups. We need to clarify the meaning of the data we collect and to challenge the interpretations we make. This should be done with insiders, critical friends, and even competitors. The changes in communication resources across these three generations has made it easy to do better.

**Small changes in tools.** Except for advances in communication, I am of the opinion that the advance in tools across the three generations has been small. We in evaluation continue to draw heavily upon social science methods, seldom on such fields as chemistry, pharmacology, and nanotechnology, where advances have been rapid. I am not suggesting that those technologies have potential use for Canadian evaluators, but that such fields as sociology, economics, and political science, from which we draw heavily, have few new tools, other than those in communication. I do not think the quality of a profession is necessarily indicated by the invention of tools. New tools are as much an expression of vanity as a response to need.

Even if so, qualitative inquiry has made great strides since the post-Sputnik period. The greatest move perhaps has been to design studies less as search for generalizations (as promoted by social science) and to design more as search for understanding of particular cases (as promoted by social work and astronomy). In his book, *Making Social Science Matter* (2005), Bent Flyvbjerg said that we advance toward understanding the world in general by case studies as well as by experimental methods.

**A diminished embracing of the evaluand.** In my book, *Qualitative Research*, I refer to an embraceable study, an evaluand you can get your arms around. You can come to know some evaluands in an experiential way, up close and personal. I have written my ideas of “responsive evaluation” with a leaning toward that kind of program or product. Others might call it “local,” or “situational” or “primitive.” An embraceable study is one that you observe and talk with and come to know its habitat. I published my meta-evaluation of Cities-in-Schools published as *Quieting Reform*, illustrating my preference for such designs. Will Shadish, Tom Cook and Laura Leviton (1991) took that to mean my views of evaluation were “fundamentally conservative.”

Among funded studies, embraceable studies have become rare since the first generation. It could be because they are too conservative but, to my knowledge, conservative spokespersons have not endorsed the personalized evaluation advocated by Saville Kushner (2000) and the democratic evaluation of Ernest House and Kenneth Howe (2006). It could be the embrace is too subjective or too expensive. There is one venue the energy supply is still abundant. It is dissertation research, and there, embraceable studies have become more prominent, partly because they are more acceptable to doctoral committees in this third generation.

**A diminishing appetite for understanding.** All of us have an immense lack of curiosity about some things. I myself do not understand my retirement annuities. And even with Google and its relatives, much of what is knowable we ignore. It is partly the complexity, partly the doubt that explanations are free of distortion. Explanation is a "power construct."

Among the first generation of evaluators, it was pretty common to believe that the truth could be dug out and that our truths would be appreciated by some audiences, near and far. It was believed that a good evaluation study would have diagnostic usefulness, or at least experiential value and that program improvement could follow. That ideology has waned. Political campaigns, the media, and other superficialities have educated many people to doubt that investing in formal evaluation will be primarily useful to other people.
Increasingly we have a crisis of representation. In the small prinng, guarantees are limited. Indicator
variables change meaning in different times and places. People will see what they want and what they don’t want
in photographs and testimonials. Often the public is suspicious. All too often people learn that they should have
been more suspicious.

Evaluators have not done much to oppose this crisis in confidence. They know that internal validity does
not assure that indicators can safely be used for a variety of purposes. For example, they know that the ratings of
colleges and universities are flawed. They know that student test scores do not indicate quality of teaching. But
few are protesting.

The Mega-Discipline of Evaluation

One of the disappointments in this last generation is that few, to my knowledge, have followed up on Michael
Scriven’s grand idea of evaluation as a mega-discipline. Michael (1991) has claimed that because evaluation is similar to
logic, history, language and use of computers and libraries, because it is essential to all scholarship and professional
practice, it should be conscientiously developed throughout the liberal arts and sciences. Law professors, theoretical
physicists, and music ethnologists do teach evaluation intensively, and in substantive ways, but do not examine the theory
and practice of evaluation as well as the students in our evaluation classes.

Informal and Formal Evaluation. To recognize evaluation as a mega-discipline, I think we need to make at
least one change. We make too much distinction between formal and informal evaluation. Education is infused
with evaluation. Social science is infused with evaluation. All professional work is infused with evaluation. It is not a
matter of choice. One cannot do one’s work without sensing the quality of things (Eiseley, 1962).

Improvement in quality in the arts and sciences requires sensitivity and refinement—refinement in the
recognition of quality and worth (Scriven, 1996). With good meta-evaluation, i.e., with good management of
evaluation, the eye becomes sharper and the will becomes stronger.

Formal evaluation is the conscious disciplining of judgment. We train ourselves better to see value. A
conscious step taken to manage, to refine, to validate judgment makes it more formal. Formalization is disciplined
thinking. But often formalization is a way of making thinking impersonal and artificial. Soon we have styles,
protocols, models, and mechanisms to refine our formalizations, and we argue about which are the better.

Formal evaluation and informal evaluation are part of the same act. People who become professional
evaluators, or experts in any way, such as administrators or arbitrators, formalize their ways of recognizing quality.
They rely still on the informal, as
Mixing the two is a skill to be sharpened and sensitized to complication, resistant to simplification. As professionals, we should avoid disdaining informal evaluation. We should welcome the mix and teach and facilitate the use of informal evaluation.

Under-representing complex evaluands. For my assignment today, it was suggested that I speculate on what would happen in the next generation to come. I shrink from the challenge, expecting future changes to be

well as the formal. The boundary between formal and informal is indistinct. Mixing the two is a skill to be sharpened and sensitized to complication, resistant to simplification. As professionals, we should avoid disdaining informal evaluation. We should welcome the mix and teach and facilitate the use of informal evaluation.
I want to share with you a table on the evaluation of teaching from a paper by Gloria Contreras, Isabel Arbesú and myself (in press). Our main purpose in preparing the paper was to counter the impulse of many agencies and people, particularly psychologists and economists, to use a single construct or set of competencies, to indicate quality. I single out psychologists and economists because we have been spectacular in representing complex human activities, syndromes, and phenomena with indicator variables (Shavelson, McDonnell and Oakes, 1991). Teaching is a complex enterprise, sensitive to its situation and cast of characters, and is, I think, hurtfully oversimplified by our indicators of quality.3

Figure 1 is a graphic representation of both the formal and informal evaluation of teaching in higher education. It is more a representation of responsibility than a guide to assessment. All these responsibilities are exercised in informal evaluation of teaching, but in research and contracts, we have settled for less. For example, take U. S. News and World Report rankings of universities. Quality of teaching is taken as represented in opinions of reputation (Gladwell, 2011). Similar oversimplifications occur in many of our evaluation studies.

Reflecting on our profession. It seems to me that the “variance within” greatly exceeds the “variance across” the generations. People evaluate differently, informally and formally, from situation to situation, for different purposes. We invent a few new approaches but the practices of today are not much different from the practices of Cold War times.

We lack a system of assessment and remediation perhaps similar to what medicine lacked six generations ago. Some say (Morra Imas and Rist, 2009) we are moving toward greater links between evaluation and development, but I do not see it. There are two products that keep us afloat, the satisfaction of greater understanding of quality and the sales value of the evaluation system. Right now, the public is not getting much from either. We have today a third generation, a third generations that Thomas Kuhn (1962) might have called, “normal science,” a maintenance of techniques and services. The post- Sputnik generation had much in the way of “extraordinary science,” with invention and innovation. But things changed. I was talking with Nick Smith a couple of weeks ago and he said, “In the first years we drew ideas from a hundred different lines of work (Smith, 1981). In the last years we have been consolidating the profession.” He admired the integrity of the Canadian Evaluation Society and was proud of his work with evaluators, junior and senior, from Toronto and Queens Universities. I also feel pretty good about our juniors. Some veneration, not much vituperation, but clearly some divestiture.

Bibliography


3 After reviewing the literature on social indicators, Richard Jaeger (1978) concluded that indicators are "anything but clear and consistent. ... I would not require that reports of status or change in status be in quantitative form, for narrative is often a better aid to comprehension and understanding of phenomena than is a numeric report" (pp. 285-287).


Joint Standards for Educational and Psychological Testing.


Robert Stake, Gloria Contreras and Isabel Arbesú, in press. Evaluando la calidad de la universidad, particularmente de su docencia. *Perfiles*.
