Presented at the Sept. 7, 2007
Hawaii-Pacific Evaluation Association Meeting
Draft for Comments

Panel on Ensuring Credible and Useful Information

Squishy and Marvin, Same Old/Really Different, 21 Dead Babies, and Other Adventures in Evaluationland

Lois-ellin Datta*
Datta Analysis

Some days, evaluators want to pour soda on the PC, flame every eejit over There making rules for over Here, or join a more tranquil profession such as electric eel-watching. Some days, only the mortgage, the habit of eating, and hope keeps us going.

But some days, our work makes an observable good difference, a sudden insight sweeps us off our feet, and the adventure of being an evaluator is grand. Every evaluation is a glorious opportunity to learn more about how to ensure credible and useful information: tweaking what we've done before, trying out new ideas.

In the next ten minutes, I'll share a few stories about tweaking and trying from the small part of Evaluationland in which I live. First, though, please take a few minutes to jot down something from your part of Evaluationland, be it what drove you to eel-watching or an epiphany. Then please pass them to our Facilitator so we can share your stories later.

In thinking about credible and useful information, I've been focusing (a) on credible methods appropriate for jazz-riff project ideas being tried out in a 76 other types of music world, and (b) on useful information for the people most likely to pay attention: the evaluands.

Squishy and Marvin: The first report is about Squishy and Marvin, something that worked out well in the field of useful information, and about Complex Adaptive Systems, that didn't, in the appropriate methods arena. The project is a $3,000,000 National Science Foundation effort to help M.A. and Ph.D. students intending to become research scientists to communicate better with non-scientists. The Fellows are paired with classroom teachers. They jointly develop a six-week curriculum unit in science and teach it.

How did the Fellows do? And how to communicate what I saw? Some protocols have nifty frameworks, require tedious observations, and yield eye-glazing data. So I adapted a science education framework for classroom observations that had been well-tested in Arizona as a guide to writing stories, like the story of Squishy and Marvin, told here in abbreviation. Almost all of us listen to stories. It's culturally appropriate in this Pacific Rim area, and comparing Aesop to Plato, Aesop gets the points across faster and more indelibly, if the points are fairly straight-forward.

We're in a 5th grade classroom, Friday afternoon, after lunch. The young scholars came *bouncing in, much to my surprise, eager as beavers to receive their very own, very dead squid on a paper plate. The students' task was to observe external and internal features, using structure to infer the critter's function and its' adaptations to the squidly environment; the teachers' task, to use inquiry-based learning. About 15 minutes into the event, a girl asked, "Do dead squid squirt ink?" A not-so-good instructor would have answered the question. This pair asked the class, "What do you think, and why?" Forest of hands. Then they asked, "So how would you find out?" Another forest of hands. "Do it." As the students were leaving, a girl asked a boy she'd helped to find the feather cartilage behind the yucky tentacles, "What did you name your squid?" "Squishy," he said. "What did you name yours?" "Marvin," she answered firmly.*

The stories were sent in draft to the Fellows and Teachers for their comments, analyzed, and integrated with other data into a more traditional report for NSF. This was fun for me, easily understood and found provocative by the Fellows, Teachers, and project directors, and I'm going to use the approach again this year as one part of the multi-method evaluation.

The project also seemed like a grand opportunity to try using complex adaptive systems to help sort out what was influencing which in schools experiencing everything from a 6.7 earthquake to the fifth principal in four years to a cornucopia of other science initiatives. The first step, I thought, was to find a colleague who would develop an inventory of what else was happening in the schools and the community, using the Complex Adaptive Systems (CAS) framework. The second step, to integrate this landscape with the in-school observations and focus group interviews to situate the project within the complex system.

Great idea, close to utter failure: CAS is not an intuitively understandable framework, and without some understanding---more than I could convey or coach or share with my historian colleague through readings and discussion---data collection is dry leaves. The inventory was an eye-opener, though: over 50 other science-related initiatives were making music in both random and non-random ways, underscoring the fragility of the central "all other things being equal" assumption of randomized designs to control biases. I will try again this year, but will do it solo or hope for a local colleague with a passion to try CAS. My conclusion: adapting CAS to practical evaluation may be a work in progress, even with the great first book by Williams and Iman (2006), and I'm gonna keep at it.

Same Old/ Really Different. Another project is a meta-evaluation of a National Science Foundation award to an internationally distinguished evaluator for the purpose of field-testing an innovative approach to evaluation. Applause here, for NSF, an agency that consistently supports appropriate evaluations as their gold standard and encourages innovation. In a pretty straight-forward way per Chapter 26 of Stufflebeam and Shinkfield (2006), the meta-evaluation involved descriptions of the adequacy of the field test sites for field test purposes, document analyses, interviews with all concerned, and listening to the annual advisory board meeting. The abbreviated *Educational Evaluation Standards* and a specially prepared checklist based on the innovative approach guided inquiry.

We got lucky in one unexpected way. NSF wanted to know the innovative approach from an evaluee's perspective, particularly the demands on and the benefits to the evaluee. Again, to help evaluation usefulness through effective communication, we asked the evaluands about their *previous* experiences with being evaluated and how, if at all, the current experience differed. Wrote up the interviews, checked them with the interviewees for completeness and accuracy, got a good night's sleep, and looked forward to an advisory board meeting. Good meeting, relevant to meta-analytic interests, but not particularly edge-of-the-seat until an academically genteel version of "And so's your mother" erupted about the distinctiveness of the "innovative approach to evaluation." About four hours later, the advisory board was not convinced and the innovative evaluators were not heartened.

But we, the meta-evaluators, think they should be, according to the data from the interviewees---classroom teachers and on-site program directors--- collected to answer that related but somewhat different question. To them, this approach was different---way different, and they liked it, because the approach involved deep, respectful understanding and observing what was happening.

> *Other evaluators come in, spend a few days, and walk out. Sometimes we don't even hear anything from them. We're expected to change what we're doing to accommodate them, and that doesn't even give them a good picture of what usually happens. This group begins by discussing with us our values and theirs, taking time to understand. They are unobtrusive, like when they get students' ideas over a hamburger. They are here a lot. In conversations and briefings, they share what they observe and engage us in discussions that are full of tidbits of questions, ideas, thoughts. This is way different from other evaluations. This information is believable, and this approach is helpful to us.*

The good difference, which may or may not be reproducible on a larger scale, was not the vertical dimension---the framework, the constructs, and the philosophy---but the horizontal---the time line in which the work is carried out and the way in which the values of the philosophy, admittedly those of many theorists, are realized in action. The innovative evaluators are now working on better describing (or understanding) themselves and yes I'll use the question again in the June2008 visit.

21 Dead Babies: Perhaps the evaluation work that has been most arduous this year came from my concern for the starting point in evaluation policy when attribution is desired. For many human service areas, I hold with Scriven that the randomized experiment is appropriate in few settings if the criteria its strongest advocates endorse are actually applied. A starting point better combining the need for reasonably strong attribution may be quasi-experimental designs. That is, a good quasi-experimental design perhaps should be the standard for appropriateness, with no extra points given even if a randomized design turns out to be OK, rather than starting with the randomized design as the standard with penalties and points deducted for quasi-experimental designs.

I was floundering about how to say this more convincingly when I came across the 21 dead babies. The Early Head Start Evaluation (2004a, 2004b) was a congressionally mandated randomized experiment. Programs over-recruited and assigned families prenatally to E or C conditions. Data were collected through the three program years and into the pre-KG years. But what data? The appendices way in the back indicated that about a third of the E families experienced only half or less of the full program and that about half of the C families found similar services. That's a non-trivial cross-over. However. using imputation techniques, *regardless of actual experiences,* the data were analyzed by original random assignment categories---including imputed data at all the time periods for 21 children who died before birth or during the evaluation period. When--also in the appendices---data could be compared for imputed program effects and for effects based only on services actually received, analyses by actual services received showed a more effective program. Yes, I know that the same imputation approach is used when the children move away, when they drop out, when they aren't around at testing time, and so on, but somehow, imputing results to dead babies and analyzing the data according to the original random assignment grabbed me as pseudo-science. The agency was perfectly, explicitly aware that analysis by original assignment would yield what they proudly described as a conservative estimate of program effects, and proudly decided that for policy purposes, a conservative estimate was appropriate. This stance said more to me about Politics than policy, but that's another paper.

With this as a clue, I found other studies which had data on cross-overs and comparative analyses from areas such as alternative sentencing programs, youth mental health programs, health, and early education. Consistently, treatment actually received shows program effectiveness when analyses by original assignment, however massaged by propensity scores and imputation, do not. I have yet to find an instance when the reverse is true. I don't know whether my paper in the *Journal of Multidisciplinary Evaluation* describing the effects of active control groups and what to do about them will immediately change Department of Education evaluation policy, but it sure felt good to have followed that thread further.

Riding the Tiger: I'm not sure yet how this is going to work out, an evaluation of a fairly large leadership development project. It's been tricky to design an evaluation plan because the budget is shoe-string, because the evaluative questions require both knowledgeable observations that can help determine what the project is in its adaptation to Hawaii, because some deep water quantitative analysis of someone else's data is necessary, because the "It" is turning out to be four quite different "its," one of which has decided to involve every possible agency in its jurisdiction rather than staying within the field test complexes, and did I mention that the evaluees are a bit allergic to having the value, merit, and worth of the effort systematically assessed? I'm trying here 101 different ways to reduce allergies and to leverage efficient use of resources, aka as subcontracting with the evaluation equivalent of 500 pound gorillas that are already almost over the finish line in relevant knowledge from other sites or studies. That and expanding face time, providing appreciative inquiry-based formative information, and hoping the tigers don't totally outrace the evaluation. We'll see.

Now to read about what's happened to you this year….and to remind each other that unless we do take up eel-watching, evaluation is still a grand adventure, and one whose social consequences have merit, worth, and value.

References

Administration for Children and Families. (2004a) *Making a difference in the lives of infants, toddlers and their families: The impact of early Head Start, Volume I: Final technical report.* Washington, D.C.: U.S. Department of Health and Human Services.

Administration for Children and Families. (2004b). *Making a difference in the lives of infants, toddlers, and their families: The Impact of Early Head Start, Volume II: Final technical report appendices.* Washington, D.C.: U.S. Department of Health and Human Services.

Datta, L. (2007) "Why active control groups matter and what to do about it," *Journal of Multidisciplinary Evaluation.* 4, 12-24.

Stufflebeam, D. L. and Shinkfield, A. (2007) *Evaluation theory, methods, and application.* San Francisco: Jossey-Bass.

Williams, B. and Iman, I. (Eds.) (2006) *Systems concepts in evaluation: An expert anthology.* Point Reyes, CA: EdgePress

* Lois-ellin Datta
Datta Analysis
78-7054 Kamehameha III, #1304
Kailua-Kona, HI. 96740

Datta@ilhawaii.net