Editorial

The "Baggaging" of Theory-Based Evaluation[1]

E. Jane Davidson

Davidson Consulting Limited

Program theory and its use in evaluation seems to be an argument that just won't go away—depending on which part of the world one is in. What is it about theory-based evaluation (or the different understandings of it) that polarizes some but brings others together? A little digging shows that there are some serious misconceptions among both the program theory evangelists and the "theoro-skeptics," while some of the best innovations are coming from those who understand program theory's potential *and limitations* and are just getting on with using it to move our discipline forward.

[1] Some of the ideas in this editorial are drawn in an earlier piece that appeared in the 2002 edition of *Mechanisms*, the newsletter of the American Evaluation Association's Program Theory and Theory-Driven Evaluation TIG.

The evaluation community in the United States has traditionally been divided into two camps—those who believe that theory-based evaluation is the wave of the future and that virtually all evaluations should be conducted in this way (e.g., Chen, 1994, 1996; Donaldson, 2003) and those who believe it to be a usually unnecessary addition of bells and whistles that fails to enhance the quality or value of evaluations (e.g., Scriven, 1994, 1997) or who think that it is simply a flawed approach to evaluation altogether (Stufflebeam, 2001).[2]

In other parts of the world, such as Australasia and the UK (and perhaps others can comment on how things are in their part of the world), there has been quite a different discussion around program theory. In the conversations I hear, the focus has been not on *whether* program theory should be used in evaluation, but on *how* the multitude of different options can be used or not in particular cases, innovated on, and improved to maximize program theory's value added (e.g., Pawson & Tilley, 1997; Rogers, 2000).

As I look across the polarized spectrum, from the most ardent advocates of theory-based evaluation (TBE) to its most critical critics, I can't help but notice one thing—baggage. By this, I mean that those who have promoted TBE over the years have failed to keep the fundamental concept clear, but have instead loaded it up with all sorts of extras that reflect the way they happen to use it. These add-ons run the gamut, from wedding it to a particular analysis technique (such as structural

[2] There are, of course, a couple of other camps—those who show little interest in the debate, and a small number who are pushing hard to identify TBE's weaknesses and find innovative ways to address them.

equation modeling) to selling it as something far narrower than it is (e.g., an inherently participatory approach that seeks to uncover stakeholder theory).

The inevitable result of the “baggaging” of theory-based evaluation has been that the critics have targeted not the fundamental idea behind TBE (which is actually really useful), but the particular “TBE + baggage” version they most strongly object to. Let’s take a few examples of criticisms to illustrate the point.

I couldn’t help chuckling at yet another sideswipe at TBE in a recent chapter from Stufflebeam (incidentally, on a topic completely unrelated to TBE). Stufflebeam (2004) argued that “the now fashionable advocacy of theory-based evaluation” makes little sense because it “assumes that the complex of variables and interactions involved in running a project in the complicated, sometimes chaotic conditions of the real world can be worked out and used a priori to determine the pertinent evaluation questions and variables” (p. 253).

I am quite certain that most users of theory-based evaluation would NOT claim that they had worked out in advance “the complex of variables and interactions involved in running a project.” In fact, most would wonder where on earth Stufflebeam got the idea that anyone thought this was possible. While I can’t speak for him, I can identify some contributions to the literature that might well lead someone to believe this was being claimed. Take, for example, Chen’s (1990) definition of program theory as “a specification of what must be done to achieve the desired goals, what other important impacts may also be anticipated, and how these goals and impacts would be generated” (p. 43). One could certainly be forgiven for thinking that Chen’s version of program theory has almost crystal ball-like properties that allow it to predict any and all potential side effects as well as the mechanisms by which they (and any main effects) will occur.

It is quite true that, in the process of developing program theory, the critical thinking exercise involved quite often leads the evaluators (and others working with them) to identify a number of potential side effects that might not have otherwise occurred to them. So, it is easy to see the source of Chen's claim. But it is just as easy to see that the statement can be construed as meaning that evaluators using program theory effectively would always be able to identify *any* important side effect, nor the mechanism by which it occurs. Let's call this the Crystal Ball Baggage.

The above quote from Chen also identifies a second piece of unnecessary baggage that seems to accompany theory-based evaluation—that it is, by definition, a goal-based evaluation approach. It is true that most exercises in developing program theory start with the question "What is the program trying to achieve here?" Particularly if stakeholders are involved in theory development, the content of the program theory inevitably skews itself toward intended consequences. As Chen implies in his definition, program theory can be used to systematically identify potential side effects. But the important point is that *it is actually possible to do goal-free theory-based evaluation* (Davidson, 2000, 2004). The fact that *most* program theories begin with intended consequences (or goals, if any have been formally stated) does not mean that theory-based evaluation, by definition, takes that approach. So, we also need to uncouple TBE from its Goal-Based Baggage.

Probably the most pervasive of all of TBE's accompaniments is something we might call Stakeholder Involvement Baggage. For example, Donaldson (2003) describes program theory as “a common understanding [developed between the evaluator and stakeholders] of how a program is presumed to solve the social problem(s)” (p. 114). To be fair, Donaldson also lists a range of other sources of information that can be used to build a solid program theory, but he (like so many other theory-based evaluators) seems to be quite consistent in the extent to which he promotes stakeholder involvement in program theory development.

Don’t get me wrong here. Stakeholder involvement in the development of program theory is an incredibly powerful method for providing the kinds of insights that can drive needed change, whether that is program improvement, program replacement, or something else. While a good ‘black box’ evaluation can give a conclusion about whether or not a program is effective or worthwhile, we know from the literature (and our experience) with organizational change that there is often very strong resistance to believing a negative conclusion. And in many cases, programs are not as effective as they might be because all or part of them are based on flawed assumptions (or, more often, highly questionable logical leaps, or no logic at all) about why they should work. The “let’s discover this together” approach is far more effective for getting people to see with their own eyes why something doesn’t work and therefore needs to be changed or abandoned. Put another way, participatory theory-based evaluation is one of the most powerful tools we have for ‘organizational unlearning’ (i.e., getting people to stop believing something that is untrue and has been leading them to be less effective in their work).

The drawback of the Stakeholder Involvement Baggage occurs when examples of theory-based evaluation or descriptions of how it is best done are almost exclusively participatory in nature. This leads critics to believe that TBE cannot be used in non-participatory mode, and therefore is of no use in situations where the contract, legislatory requirements, and/or the political climate require a clearly independent, impartial evaluation. Again, this is not true. TBEs can absolutely be done in a non-participatory mode, even if this isn't the preferred approach of most practitioners.

The final piece of unnecessary baggage that so often tags along with TBE is the Hypothesis Testing Baggage. It is well known that one group that gravitates strongly toward theory-based evaluation are the 'quantoids'—social scientists with high levels of expertise in and preference for quantitative methods. Through quantoid glasses, a program theory logic model looks just like something that should be plugged into a structural equation model, thereby yielding a fit index. If the model 'fits' the data ($p < .01$), the program is deemed effective, albeit in the careful "the evidence appears to tentatively support the hypothesis…" kind of way.

TBE's Hypothesis Testing Baggage is actually a lot more problematic than the other kinds of baggage. Although, in the right hands, it *can* be extremely useful, many of the examples I have seen fall into some insidious traps. The most common trap is remembering to evaluate the theory but forgetting to evaluate the actual program. The fact that certain variables are associated with each other might give us some hint that the underlying theory behind the program may not be completely off base, but it fails to tell us whether the *magnitude* of any program impacts were substantial enough to make a difference to anybody's lives, let alone whether the value of the impacts were worth the cost of implementing the program. Put simply, many hypothesis testing approaches to theory-based evaluation are not really evaluations at all; they are research projects.

I am sure there are plenty of other types of baggage that go along with theory-based evaluation; I have simply plucked out a few that I find to be most problematic. But the main point is this. Wherever there is baggage, the critics tend to attack the baggage rather than the approach itself. Or, as Pawson and Tilley (1997) put it:

> We know from long, long experience in trying to develop research practice that methodological positions can become 'badges of honour'. One can easily be misunderstood, debating positions tend to become foreclosed too early, and straw men and red herrings galore litter the path to progress. (p. xiv)

So, what should we do about the baggage, the red herrings, and the straw men?

I would like to argue for a concerted effort to strip down useful evaluation concepts and approaches to their bare bones so that it is easy for all to see what the approach entails and what the potential add-ons are (or are not). Elsewhere I have used a deliberately 'bare bones' definition of theory-based evaluation as "any evaluation that uses program theory or program logic as its guiding framework" (Davidson, 2004, p. 248). Program theory is simply a description of the mechanism by which a program achieves (or is expected to achieve) its effects (p. 38).

I also like the way Rogers, Petrosino, Huebner, and Hacsi (2000) clearly explain what is in and what is out of their definition:

> We consider [Program Theory Evaluation] to have two essential components, one conceptual and one empirical. PTE consists of an explicit theory or model of how the program causes the intended or

> observed outcomes and an evaluation that is at least partly guided by this model. This definition, though deliberately broad, does exclude some versions of evaluation that have the word *theory* attached to them. It does not cover all six types of theory-driven evaluation defined by Chen (1990) but only the type he refers to as *intervening mechanism evaluation*. It does not include evaluations that explicate the theory behind a program but do not use the theory to guide the evaluation. Nor does it include evaluations in which the program theory is a list of activities, like a "to do" list, rather than a model showing a series of intermediate outcomes, or mechanisms, by which the program activities are understood to lead to desired ends. (pp. 5-6).

Note that the above definitions include no assumption that outcomes are derived from program goals,[3] or that the theory would be able to predict side effects, or about where the theory comes from (e.g., stakeholders, the literature), or how the theory is tested. These elements are deliberately left open in recognition of the diversity of ways in which practitioners actually use evaluation. And this is the way it should be.

[3] The term "desired ends" in Rogers et al.'s (2000) definition *could* be construed as referring to goals; however, the beginning of the definition clearly and deliberately states that outcomes in program theory may be intended *or* observed. To my mind, the reference to desired ends simply reflects the broader observation that most program theories include outcomes that are implied by the very nature of the program or of its participants or their documented needs (e.g., reduction in violent offending as an outcome for a cognitive behavioral program delivered to violent offenders).

I know that readers who subscribe to EVALTALK and some of the various evaluation listservs around the world probably roll their eyes every time yet another discussion comes up about the meaning of key terms such as evaluation, research, theory-based evaluation, and so forth. But unless we can get these fundamentals right, we can end up being distracted by the baggage and—more importantly—not making use of a potentially valuable approach simply because someone has erroneously defined “the way it is” (i.e., its essential and distinguishing nature) as “the way I happen to do it.”

As evaluators, we need to keep a critical eye on this habit of using our own job descriptions as [supposedly universal] definitions, as well as definitions that omit a crucial element (e.g., explaining what a key term such as ‘evaluation’ means). After all, how many times have we seen something like this: “Evaluation is defined as the process of working collaboratively with stakeholders in social programs to identify their questions and then apply social science research methodologies to answer those questions, for the purpose of program improvement”? Interesting description of a particular individual’s consulting practice, but a definition it is not, let alone a baggage-free one!

References

Chen, H. T. (1990). *Theory-driven evaluations.* Newbury Park, CA: Sage.

Chen, H. T. (1994). A panel of theory-driven evaluation and evaluation theories. *Evaluation Practice, 15,* 73-74.

Chen, H. T. (1996). A comprehensive typology for program evaluation. *Evaluation Practice, 17*, 121-130.

Davidson, E. J. (2000). Ascertaining causality in theory-based evaluation. In P. J. Rogers, T. A. Hacsi, A. Petrosino, & T. A. Huebner (Eds.), Program theory in evaluation: Challenges and opportunities [Special issue]. *New Directions for Evaluation, 87,* 17-26.

Davidson, E. J. (2004) *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Donaldson, S. I. (2003). Theory-driven evaluation in the new millennium. In *Evaluating social programs and problems: Visions for the new millennium.* Claremont Symposium Series, Claremont, CA.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London: Sage.

Rogers, P. J. (2000). Causal models in program theory evaluation. In P. J. Rogers, T. A. Hacsi, A. Petrosino, & T. A. Huebner (Eds.), Program theory in evaluation: Challenges and opportunities [Special issue]. *New Directions for Evaluation, 87,* 47-55.

Rogers, P. J., Petrosino, A., Huebner, T. A., & Hacsi, T. A. (2000). Program theory evaluation: Practice, promise, and problems. In P. J. Rogers, T. A. Hacsi, A. Petrosino, & T. A. Huebner (Eds.), Program theory in evaluation: Challenges and opportunities [Special issue]. *New Directions for Evaluation, 87,* 5-13.

Scriven, M. (1994). The fine line between evaluation and explanation. *Evaluation Practice, 15,* 75-77.

Scriven, M. (1997). Minimalist theory: The least theory that practice requires. *American Journal of Evaluation 19*(1), 575-604.

Stufflebeam, D. L. (2001). Evaluation Models. *New Directions for Evaluation, 89* (complete issue).

Stufflebeam, D. L. (2004). The 21st century CIPP model. In M. C. Alkin (Ed.), *Evaluation Roots* (pp. 245-266).