American Journal of Evaluation
34(1) 91-98

DOI: 10.1177/1098214012461559
aje.sagepub.com
SAGE

Balancing Evaluation Theory and Practice in the Real World

Eleanor Chelimsky

Abstract

In this paper, the author argues that evaluation theory and practice interact insufficiently today, even though early evaluation theorists expected them to be closely intertwined. She views this limited connection as the result of differing interests on the part of theorists and practitioners, differing frequencies of dissemination, and differing targets of opportunity, with theorists concentrating often on methodological processes internal to evaluation, and practitioners more concerned with the application of these processes to an external environment. The author provides examples taken from current evaluation practice in which theory is largely ignored, as well as instances of theoretical writing which fail to consider problems faced by practitioners in their efforts to accommodate current theory. These examples are chosen among many which commonly occur across the planning, implementation and reporting phases of an evaluation in progress. The likely implications of this still unfulfilled relationship for the development of the evaluation field are discussed and some potential ways forward are described.

Keywords

evaluation theory, practice, interaction, integration

It is wise to revisit the issue of balancing evaluation theory and practice, not only because we often have problems in reconciling the two but also because it may be time now to reconsider the relationship and try to improve it. Why do we need to do that? Because, as our work continues to expand—in terms of purpose, perspective, type of evaluation setting, and subject matter—many evaluators want to develop a body of "best practices" to better inform our work, going forward (see the AEA Thought Leader forum, March 25–April 1, 2012, led by Gail Barrington). But because evaluation *theory* is the foundation of evaluation practice, "best practices" cannot really be developed without a careful, systematic look at how well we have been integrating theory and practice, at the places where we have encountered problems in doing that, at the solutions brought by practitioners to those problems, and at the generalizability of those solutions for evaluation as a whole.

In short, "best practices" cannot be developed separately because evaluation theory and practice are interdependent: each one learns from the other and, in that learning process, both are inspired to stretch, to bend a little, and to grow. Further, their mutual dependence endows them both with

Corresponding Author:
Eleanor Chelimsky, 22311 E. Byron Rd, Shaker Hts., OH 44122, USA.
Email: Eleanor.Chelimsky@gmail.com

legitimacy: Theory protects practice from singularity and anecdotalism, and practice protects theory from abstraction. Their relationship is a mediation between principle and context, which lends breadth, depth, and realism to our work.

How then are we doing today in balancing theory and practice? Before trying to answer that question, I would like to look at two subquestions:

—First, the definitional one: what exactly do we mean by theory and practice?
—And second, how did early evaluators *expect* them to interact?

What Do We Mean By Evaluation Theory and Practice?

If we apply Webster's fifth (generic) definition, then theory is that branch of an art or science consisting in the knowledge of its principles and methods. And practice is the application of that knowledge of principles and methods to the real world. Thus, we need to go no further than definition to find our first problem: evaluation practitioners—like applied scientists, engineers, defense lawyers—have the task of applying principles and methods developed by theory to a world that has not always been carefully examined by theory; a world of complex, chaotic, continually moving program or policy contexts, a world of people and places that are often different with respect to tradition, social characteristics, patterns of behavior, and what Lincoln called "the mystic cords of memory" (Gallagher, 2011, p. G5). Yet, this surrounding environment can facilitate or challenge theoretical concepts, and questions of how to adjust or adapt the one or the other inevitably arise. But integrating theory and practice is not a simple matter because the two emerge from very different intellectual places.

For one thing, theorists and practitioners are separated by dissimilar views of what is important. Theorists tend to see practice as a kind of muddling-through process that leans more on performance art than on science, and they look with disfavor on even small deviations from methodological convention. As the historian John McWhorter puts it, theorists can be "pitilessly dismissive of detractors as just not up for serious abstraction" (McWhorter, 2012, p. 16). Practitioners, on the other hand, tend to think of theorists as irretrievably divorced from reality. Toqueville, for example, who saw himself as a man "dedicated to understanding practice," talked of theory as "an elaborate exercise in abstract reasoning" (Wallace-Wells, 2010, p. 54). How abstract is that, in the case of evaluation? One of our basic texts says that evaluation theory "connotes a body of knowledge that organizes, categorizes, describes, predicts, explains and otherwise aids in understanding and controlling a topic. Theories do this in many ways, such as searching for invariant laws, using definitions and axioms to deduce testable propositions, and describing the causal processes that mediate a relationship" (Shadish, Cook, & Leviton, 1991, p. 30). Thus, evaluation theory, by its very definition, processes, and goals, *is* clearly abstract.

Evaluation practice, on the other hand, is never abstract. In its application of theory, it focuses on programs and policies that are planned by people, implemented by people, use services delivered by people, are undergone by people, and are torpedoed by people, all of whom bring their individual biases and experience to the work at hand. Further, the evaluations we produce, that are based on theory and performed in the real world, are also planned, implemented, analyzed, and reported by people. And those evaluations are open, from beginning to end, to political pressures by policy makers, planners, administrators, special interest groups, subject-area practitioners, participants, and all those who may be affected by the results—or feared results—of the evaluation.

Thus, theory and practice are different, but they have at least three important commonalities: theorists and practitioners both focus on achieving a combined function that produces the strongest possible evaluative information; both want to see better use of that information in government or elsewhere; and

both recognize their need to collaborate in creating an evaluation process that improves the already strong process operative today by making it more comprehensive, integrated, systematic, and credible. Still, the effort to bring the two components closer together involves reconciling, on the one hand, a framework of logical propositions, confined largely to *internal* principles of inference, deduction, and methodology, with, on the other hand, a set of *external* imperatives that are forced on evaluation by its environment. Hence, practice, by treating the evaluation *and* its external environment as a single entity, introduces human immeasurables and contextual uncertainties that feed back into the theoretical principles and methods being applied. As a result, balancing the two means that both will be changed.

In sum, theory and practice are different because one is abstract and the other is not; and also because theory has dealt mostly with one segment of the evaluation process, and practice is involved in a much bigger framework. And this larger framework, which includes both the application of principles and methods and the consideration of multiple contextual factors involved in that application, gives rise in turn to the development by practitioners of strategies and tactics aimed at reducing the problems that emerge from that application. These new strategies and tactics again feed back into theory, sometimes *confirming* evaluative precepts, sometimes *conflicting* with them.

For example, with respect to confirmation, practitioners have found that an atmosphere of heavy political contestation reinforces the theoretical requirement for methodological strength and credibility in the evaluative analysis and that arguments in the form of intimidation reinforce the theoretical requirement for independence. On the other hand, the need to include all relevant voices in an evaluation risks coming into tension with theoretical prohibitions of advocacy, and theoretical precepts are themselves often challenged by unexpected barriers or outcomes encountered during the evaluation's implementation. Further, in the effort to promote use, theorists have tended to focus only on the quality of the evaluation: Cronbach, among others, felt that the main reason why evaluation results are challenged, ignored, or discredited is that "no adequate critical process precedes their release" (Cronbach, 1980, p. 2). Practitioners, however, have learned to predict and try to neutralize a whole host of factors in the external environment that may militate against use. Therefore, in thinking about how to balance such apparently inherent differences, perhaps we should go back a little bit in history and ask:

How Did Early Evaluators Expect Theory and Practice to Interact?

My sense is that the size and complexity of the interaction problem were not well understood by early theorists. Just as findings were expected to be used in government—easily, and almost as a matter of course (Cook, 1997), it was also expected that evaluation principles and methods would be applicable, without too many difficulties, to real-world programs and policies (Campbell, 1969). In both cases—use and theoretical applicability—there seems to have been little ambient knowledge or experience of the complex environmental circumstances that could interfere with both processes.

Early theorists did recognize that their knowledge of the external milieu for evaluation was limited, but they thought that as evaluation practitioners gained experience in applying methods and principles, they would feed their new learning back to inform theory on a continuing basis. For example, Kurt Lewin told us in 1936 that evaluation theory should be "empirical, not speculative," and "closely related to the data and experience brought by practice" (Lewin, 1936, p. 4). But, not only has it been difficult to combine theory and practice, as I have just discussed, it is also the case that information from practitioners has emerged only irregularly, one evaluation at a time; much practitioner experience has remained unpublished, and no other forum has developed which could have fostered routine, ongoing feedback and mutual reflection on the general nature of problems encountered by both theorists and practitioners.

In short, I think it is fair to say that the assumptions of early theorists were more than a little optimistic about the likely use of findings, about the ease of applying principles and methods in the

real world, and in the assumption that a natural, informal process would evolve by which theory could integrate the experience of practice.

How Then Do We See Theory and Practice Interacting Today?

I think it is clear that some movement, up and back, between theory and practice, some debate, and some reconsideration of earlier ideas has occurred, willy-nilly, over time. Theorists are no longer telling us, for example, that timing in an evaluation does not matter; nor do they talk so much about "benign policy-shaping communities," nor do they counsel "amicable negotiations" with fire-eating special interest groups. As the result of work by Valerie Caracelli, Jennifer Greene, and others, we have seen mixed methods become accepted in some quarters, but not by many highly skilled evaluators for whom only randomized controlled trials will do. We have also seen important efforts by Ray Pawson, George Julnes, Melvin Mark, and Gary Henry, and especially Michael Quinn Patton, with his utilization-focused evaluation, to bring some sense of real-world context to the concepts and processes of evaluation. Indeed, theorists are looking more closely at practitioner observations from the real world that expand on earlier theoretical concepts about evaluation purpose and about the kinds of policy questions evaluators are likely to be asked. However, all of these ideas remain in the air, there is little ongoing debate about them, no dialogues have ensued about modifications in the methodological toolkit as a result of these new understandings, and so practitioners are left hanging.

For example, it is fairly common now to find theoretical acceptance of perspectives like accountability, management improvement, knowledge gain, and evaluative development to guide the elaboration of an evaluation design, rather than the single purpose of establishing merit and worth, but there is still controversy about what these perspectives mean for methodology. Many theorists, for example, continue to support the randomized controlled design in circumstances where mixed methods might be more specifically appropriate to the particular perspective and the particular question under study.

Practitioners, for their part, understand better that they are dependent on theoretical principles and methods to do strong, credible evaluations, but some have been traumatized by past problems with the experimental design and have failed to use it when it was highly appropriate to do so. Also, practitioners now recognize that they cannot count on theory to help them with contextual issues involving people, subject-matter complexities, history, or politics. As a result, they have developed their own procedures, usually unvetted (and unprotected) by theory, for dealing with these issues. Yet, as I mentioned before, these procedures often have ramifications of their own for the application of principles and methods, and as such, are good indicators of where the interactions between theory and practice have been weak. These procedures can be found at almost any phase in the evaluation process. Here, I consider three of them, for illustration.

The Design Phase

This is the place, at the very beginning of an evaluation, where evaluators have traditionally looked at the question posed to them, thought about methods for answering it, and decided whether to do a study. But today, practitioners have shown that, for an evaluation to be viable, the design must also examine many contextual factors, set up a plan for dealing with potential problems of credibility and use, and lay a foundation for predicting and tracking the key external factors likely to affect the evaluation from beginning to end. Again, these practitioner-developed procedures should interact with theory by forcing reconsideration of: (1) the *kind* of evaluation that may be feasible, based on what has been learned about the program context and especially its history; (2) the *types* of evaluation questions that will be possible to answer; and (3) the methods—individual or combined—that are appropriate (Chelimsky, 2010).

The Preimplementation or Implementation Phase

This is the place where evaluators outline the steps they will need to take in a particular evaluation to promote use of the findings, once they have been established. This work is based on two assumptions: the theoretical belief that it is high quality in the evaluation that chiefly determines use; but also, the practitioner experience that there are powerful external factors that can impede or distort use, regardless of evaluation quality. So in each evaluation, practitioners must not only assure the methodological excellence sought by theory but also estimate in advance—based on the subject matter, its history, and its politics, for example—the likelihood of use problems related to external pressures. They need to identify the likely origins of those pressures; develop a deterrent or neutralizing strategy to preempt or disarm them; and finally, determine whether the evaluation is still worth doing, given external oppositions that are just too numerous or too powerful to fight. Two concepts require integration with theory, here: first, the practitioner recognition of external pressures on the use of findings, which needs to be added on to the theoretical tradition of evaluation excellence *uber alles*; and second, the practitioner strategy of defending an evaluation in the face of those pressures, which needs to be included in evaluation planning, along with the methodological choices that may be involved.

What has happened here is that, over time, practitioners have been improving the likelihood of a better fit between the evaluation and its environment, and in the process, have done three things: (1) They have specified new types of evaluation questions likely to be posed by sponsors in different political settings (their questions are sometimes about merit and worth, as theory would have it, but they can also be descriptive, normative, prospective, or knowledge-seeking, as well as focused on cause and effect); (2) they have clarified gaps and difficulties experienced with traditional methods (e.g., the problems related to external validity in the experimental design when the question posed is about scale-up, or the transfer of a successful program to other places); and (3) they have invented new methods to deal with some of these questions, all of which would benefit from theoretical examination and debate.

The Reporting Phase

Here, procedures have been developed to address an audience of policy makers, program managers, and the public, without, of course, excluding the traditional audience of evaluators. This involves producing a report that is technically accurate, but is also written clearly, simply, and without jargon. The goal here, of course, is appropriate use, and it leans on the idea that technical excellence cannot be persuasive unless it is, at very least, understood. This translates into a change in reporting style, considerable expansion of briefings and presentations to all who are involved in the evaluation, especially those in a position to affect use, and the relegation of methodological discussion to special chapters or appendices. Indeed, long experience has shown that too much technical analysis tends to alienate rather than convince sponsors; policy makers' eyes have been seen to glaze over at the mere mention of words like "study design" or "measurement"; and so practitioners now focus more on what they have learned, rather than how they learned it.

Sometimes practitioners even use individual cases, or a couple of striking numbers, or anecdotes to explain the significance of their findings, and this can seem strange to theorists, given that doing evaluations in the first place is at least partly to move *away* from the anecdote, the war story, as a credible answer to a policy question, and instead to try to understand the size of a problem, its range, frequency, direction, average characteristics, and so forth. But when the results are in, when the time comes to report on them to policy makers, and when credibility and persuasiveness become the name of the game, it is very helpful to illustrate the general findings via specific cases and analogies that graphically explain and emphasize the larger points, in a language familiar to the relevant audiences. And this is a far cry from the highly quantified,

jargon-filled evaluation reports of the past. So practitioners have rediscovered the anecdote, but an anecdote that does not stand alone as an answer to a policy question, and instead, represents the broader evaluative evidence. In integrating these changes with theory, we may want to incorporate two approaches that run in parallel: a scientific and philosophical approach to achieve the production of strong evaluative information and a craftsman-like, audience-related approach to achieve the best possible use of that information.

In sum, although practitioners are innovating and progress is being made in developing strategies for making evaluation work in the real world, still, in my judgment, this is taking place in only a perfunctory and occasional relationship to theory. This increases the danger that, as practitioners go their own way, they may lessen further their familiarity with theory, already found to be at a low level some years ago (Shadish & Epstein, 1987). We have a long way to go to answer Lewin's call for an evaluation theory that is "closely related to the data and experience brought by practice." At risk of oversimplifying, I would argue that the problem appears to be in three parts: (1) theory has been largely devoted to internal evaluation considerations and has less often considered evidence about the environment in which evaluation theory is expected to be applied; (2) practice brings experience in applying theory in the real world, but it typically brings it one evaluation at a time and communicates it irregularly and sporadically (my own experience has been unusual in that I have been able to speak to issues of practice from a database of nearly 300 evaluations); and (3) no mechanism has been developed for practitioners and theorists to reflect together on the size, scope, and commonality of problems experienced in practice, along with the theoretical modifications, efforts at resolution, and follow-up they may imply.

What Can Be Done to Enhance the Process of Integration?

I have four suggestions for improving the balance between theory and practice. All of them have to do with a belief in the importance of sharing information (i.e., increasing the amount of awareness that theorists and practitioners have about each others' advances so as to achieve new "knowledge that can be empirically generalizable at the same time that it is relevant to specific real-life contexts;" Fischer, 1991, p. 11).

A Specialized, Ongoing Forum

I propose that we organize some type of forum, perhaps at AEA's annual meeting, perhaps here at the Eastern Evaluation Research Society (EERS), that would have as its mission the presentation of often-encountered practitioner experience that appears to challenge theory in some explicit or implicit way. I think our aim should be to surface the various unresolved issues, commonly raised by theorists and practitioners, for discussion by experts in theory and practice, along with a panel of diverse, experienced evaluators, for rethinking, comment, criticism, and debate. We could use such a forum to deal, for example, with tensions such as those that occur: between the idea of the "best" methodology and the idea of the "appropriate" methodology; between the resistance of program subject-area practitioners (e.g., physicians, nurses, teachers, police) to measurement by evaluators, and the accommodations this forces on methodology; between getting a thorough understanding of participant experience in a program, and translating that understanding into "lessons learned" that inform theory, as Hall (2012) advises; between the purpose of "social betterment" and the need for nonadvocacy in an evaluation; and between the theoretical presumption of good faith in the planning of public programs, versus the practitioner experience of political suppression of evidence that threatens use and wastes evaluative resources. Second, I propose, for each evaluation where evaluators have found conflict with theory that those evaluators themselves prepare:

A Brief, Informal Report of the Implications They See for Evaluation Theory

To develop a strong database for the AEA or EERS forum proposed, we would need to collect and analyze problematic practitioner encounters with theory. The experiences of evaluators in a subject area have always been extremely precious sources of information for practitioners beginning a new evaluation of the same program, or of a new program in the same area. I suggest that evaluators write down the problems they have experienced and would like to see examined, and present them as candidates for discussion by the expert panel. A few paragraphs would suffice, and doing this would not only help later evaluators but also (1) enable practitioners to develop, over time, a more systematic database of ongoing problems that theorists should consider; and (2) add to the process of learning how to generalize from the continuing experience of theory and practice, and how to translate that experience into ''best practices.''

We would also need AEA to appoint a panel of evaluators—perhaps like the one appointed to analyze cultural issues in evaluation—to examine the candidate problems, look at their frequency, determine their relationship not only to setting (e.g., a university setting vs. a state or federal agency setting, or an independent private practice, research firm, foundation or other setting) but also to purpose or perspective, and then select candidates for discussion at the AEA forum, based on evaluative importance, commonality, and likely generalizability. Further, to deepen and strengthen this process, we might also want:

A Blog, Listserv or Even Google+Hangout

Such a mechanism would help keep the conversations accessible, current, and immersed in both theory and practice (Fetterman, personal communication, April 18, 2012). It would also serve to remind us continually of Heraclitus' caution that you can never step twice into the same river. Finally, we might think of establishing:

An Annual Debate at AEA on a Specific "Balance" Issue

Another way to move integration a notch forward might be to pick an emerging problem each year for the AEA annual meeting, and organize a three-sided debate about it among a practitioner, an expert on theoretical issues, and someone with strong knowledge of both evaluation theory and current practice. For example, we could orchestrate a discussion around the difficult methodological problems we face in establishing the external validity of a randomized controlled design, and examine the viability of current theoretical and practitioner efforts to remedy them. Or we could look at practitioner experience in developing evaluation capacity in agencies, both in terms of outcomes and in terms of lessons learned. Or we could open a dialogue about what actually constitutes a ''replication,'' for scale-up purposes?

These four suggestions have the goal of focusing attention and discussion on three types of ''balance'' problems between theory and practice in evaluation: the unquestioned need for methodological strength in the evaluation, despite great difficulties for practice in fulfilling the conditions of some theoretical models in the real world; the need to recognize and prepare adequate responses to political and other constraints on the evaluation and the evaluators; and, the need to expand the theoretical focus beyond methodology to at least some aspects of the evaluation environment as a whole. But there is also a process goal. As two of our theorists have written, ''The maturity of a discipline is reflected partly by the clarity with which problems facing the discipline are understood, and partly by the sophistication with which solutions to problems are devised and practiced'' (Shadish & Reichardt, 1987, p. 13). We need to ensure that theory and practice stop existing in parallel, but support each other, and in a manner that is not ad hoc, but ongoing and continuous. If we can achieve this, it will help us in a number of ways: by improving the credibility and usefulness of evaluative information;

by strengthening the evaluation process itself; by creating a foundation for developing "best practices;" and by setting up the conditions for the future strong growth and development of our field.

Author's Note

This study was based on an address given to the Eastern Evaluation Research Society, May 1, 2012.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*, 409-429.

Chelimsky, E. (2010, June 8). *Navigating evaluation complexity in the age of Obama, Plenary Remarks to the Environmental Evaluators Network (EEN) Forum*. Washington, DC.

Cook, T. D. (1997). Lessons learned in evaluation over the past twenty-five years. In E. Chelimsky & W. Shadish (Eds.), *Evaluation for the twenty-first century*. Beverly Hills, CA: Sage.

Cronbach, L. J. (with Ambrom, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S.). (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Fischer, F. (1991). *Participatory expertise: Toward the democratization of policy science*. Piscataway, NJ: Rutgers University Press.

Gallagher, G. W. (2011, December 25). The cause: Why did the north fight? (Citation from Abraham Lincoln's First Inaugural Address in 1861). *The Washington Post*, Forum, p. G5.

Hall, M. (2012). Comment posted to the Thought Leader Forum, April 1, 2012.

Lewin, K. (1936). *Principles of topological psychology*. New York, NY: McGraw-Hill.

McWhorter, J. (2012, April 8). Language: Repeat after me. *New York Times Book Review*, p. 16.

Shadish, W. R., Jr., & Epstein, R. (1987). Patterns of program evaluation practice among members of the Evaluation Research Society and the Evaluation Network. *Evaluation Review*, *11*, 555–590.

Shadish, W. R., Jr., & Reichardt, C. S. (1987). The intellectual foundations of social program evaluation. *Evaluation Studies Review Annual*, *12*, 13–30.

Shadish, W. R., Jr., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation*. Thousand Oaks, CA: Sage.

Wallace-Wells, D. (2010, April 26). Democracy in America: The visitor. *Newsweek (N)*, p. 54.