

JMDE

Journal of MultiDisciplinary Evaluation

Number 3, October 2005

ISSN 1556-8180

Editors

E. Jane Davidson & Michael Scriven

Associate Editors

Chris L. S. Coryn & Daniela C. Schröter

Assistant Editors

Thomaz Chianca

Nadini Persaud

Lori Wingate

Ryo Sasaki

Brandon W. Youker

Webmaster

Joe Fee

Mission

—The news and thinking
of the profession and discipline of evaluation
in the world, for the world—

A peer-reviewed journal published in association with

The Interdisciplinary Doctoral Program in Evaluation
The Evaluation Center, Western Michigan University

Editorial Board

Katrina Bledsoe	Shawn Kana'iaupuni
Nicole Bowman	Ana Carolina Letichevsky
Robert Brinkerhoff	Mel Mark
Tina Christie	Masafumi Nagao
J. Bradley Cousins	Michael Quinn Patton
Lois-Ellen Datta	Patricia Rogers
Stewart Donaldson	Nick Smith
Gene Glass	Robert Stake
Richard Hake	James Stronge
John Hattie	Dan Stufflebeam
Rodney Hopson	Helen Timperley
Iraj Imam	Bob Williams

Table of Contents

PART I

In This Issue i

Michael Scriven

Editorial..... iii

Michael Scriven

Articles

Using Test Standard-Setting Methods in Educational Program Evaluation:
Addressing the Issue of How Good is Good Enough1

Paul R. Brandon

The Value of Evaluation Standards: A Comparative Assessment30

Robert Picciotto

The 2004 Claremont Debate: Lipsey vs. Scriven60

Stewart I. Donaldson and Christina A. Christie

Evaluation Capacity Building and Humanitarian Organization78

Ridde Valéry and Sahibullah Shakir

Ideas to Consider

Ethnography and Evaluation: Their Relationship and Three Anthropological
Models of Evaluation 113

Brandon W. Youker

Book Reviews

Revisiting Realistic Evaluation..... 143

Chris L. S. Coryn

In This Issue

Michael Scriven

This is a particularly interesting issue, which is just as well since it's also our longest to date—over 220 pages, and I doubt you can find a way to shorten it without a hundred readers feeling seriously deprived!

Remember that you can arrange to be notified when a new issue comes out by registering at our website (<http://evaluation.wmich.edu/jmde/subscribe.html>); the next issue will be out in a month or so, with some heavy coverage of the 'causal wars'. And we are now officially registered with an ISSN number—can't be done without two issues on record—so that we're in the scientific journal databases, which gives us more status in scholarly circles. In popular circles, we have over 11,000 hits on the two issues that came out before this one, which suggests (but does not prove) that more people look at our pages (perhaps briefly) than all other evaluation journals put together. Keep that in mind as you're thinking about where to publish!

As usual, we continue our coverage of the international evaluation world, with no less than two reports on evaluation in China, a very interesting one on evaluation in Japan, a new correspondent writing about the scene in Germany, and one on New Zealand (where my co-editor runs a consulting business), plus an update on Canada. Our coverage of journals and events of note includes a report on the First International Congress on Qualitative Inquiry, which almost burst the seams at the

University of Illinois last (northern) spring; and a complete list of all international associations from the International Organization for Cooperation in Evaluation, about which we expect to have an article in the next issue.

The major articles are by major authors: the architect of evaluation at the World Bank, Robert Picciotto, writes on “The Value of Evaluation Standards”; Paul Brandon, the standards guru, addresses the great problem of high-stakes testing—how do you set the lines between the grades—and there’s a study of evaluation capacity-building in Afghanistan by two who did it there. That paper illustrates our policy of ‘naturalistic editing’—editing that leaves the flavor of the writing intact, at some cost to the grammar of Standard English—and the description of conditions in Afghanistan will bring tears to many eyes.

A serious paper on ethnography for evaluation by Brandon Youker looks at three anthropological models of evaluation, and Chris Coryn, one of our associate editors who did more than anyone to pull this issue together, reviews *Realistic Evaluation*. The latest issues of the major journals are also reported on by our best reporters.

Next issue we switch over to the Canadian software for online free journals, a very nice package paid for by the Canadian government, to whom our thanks. It will improve our operations considerably. And don’t forget: this is an evaluation journal, run by evaluators, so *we like to hear criticism*. Tell us how to improve!

Editorial

The Evaluation of Disasters

Michael Scriven

In the last few years, we have seen some mighty catastrophes on the face of the earth, some wrought by human hands directly and others from great natural disasters. Of the latter, the losses from the great tsunami of the Indian Ocean make the others look minor, but to many communities they were a whole world lost. These included huge earthquakes, floods, and wildfires worldwide, and in the U. S. most recently, the hurricanes Katrina and Rita. Where humans were the direct causes, the acts of warmongers and terrorists alike, not too easily distinguished in their impact on the innocent, have altered not just cities but countries forever, and for the worse—usually in the name of improvement. And. Lurking in the wings, are worse possibilities still, widely thought by experts to be inevitable: for example, new epidemics, perhaps as bird flu crosses the species boundary en masse, and mimics or surpasses previous flu epidemics that have killed millions before, perhaps tens or hundreds of millions next time around (because the fast transportation of people, foodstuffs, and other goods make us all neighbors). We are all well aware that global warming, meteor impacts, and black market hydrogen bombs pose great risks of even greater disaster. We must ask, what has

evaluation contributed to aiding humankind cope with these events, and what could it contribute that it has not so far provided?

It's clear that these events pose new challenges for most evaluators, since the usual work of the program evaluator covers only parts of great disasters. We know how to evaluate the relief programs, the health services, the educational makeshift arrangements. But evaluation of the conditions that led to, or exacerbated the impact of these events; evaluation of the developments from them that are aimed to reduce the impact of their inevitable successors: these are a different kind of beast. These call for multidisciplinary effort of considerable novelty, and this journal will try to serve its mission of keeping its readers abreast of efforts to develop good methods and tools for doing this kind of evaluation. Meanwhile, there are a few interesting developments that may inspire us to develop improved models for this new task. Perhaps the time has come to develop what might be called the Failure Case Method?

To take one example of developments that are a possibly relevant to disaster evaluation, there are many of us who feel that one of the most interesting emerging trends in evaluation in recent years has been the emphasis on a systems approach, and surely that is one emphasis that disaster evaluation requires, when we start looking evaluatively at the precursor conditions in preparedness studies. Relatedly, one must view epidemiology, a fast-developing science in its own right, as a model worth considering for its focus on finding and fixing causes of trouble, past and future. The same is true of ecobiology, another of the recent additions to the scientific Pantheon. Television has made us increasingly aware of a third player that values the systems approach—forensic pathology, portrayed on the tube as a science far more sophisticated than its actual embodiment in real labs, where DNA matching is still taking a matter of weeks not hours. And engineering has

contributed a similar discipline in the form of applied research work of the investigation of the accident investigations of the National Transportation Advisory Board. In all of these cases, as with natural disasters and terrorist strikes, one great methodological lesson stands out: they are all primary cause-hunting sciences and none of them has ever felt unable to go to work even though they've never seen a randomly controlled experiment. So, to pick up a theme that recurs briefly in this issue, there are some important issues in evaluation methodology where we may be able to learn something from a study of the existing disaster-hunting and disaster-prevention disciplines. Our nearest approach to date, and a worthy one it is, though low-profile so far, is evaluation of peace-maintenance efforts, with a small appearance at AEA last year.

But perhaps the most important element in disaster evaluation that is familiar to most evaluators is the 'blame game,' the search for responsibility. It's an integral part of aircraft and rail crash investigations, and it poses no insuperable barrier to reliable conclusions there, or in its courts. We must take it in our stride, though of course it helps to arm oneself with the basic tools of ethical and legal analysis. For the bottom line in all of this is simple enough: a good proportion of the disastrous events themselves, and a larger proportion of their terrible consequences, are avoidable by human action. If we take on disaster evaluation and don't step up to do the ethical analysis, and do it rigorously, the job won't be completely done. Evaluators need to grow into this new aspect of a new task as they have so often grown before. It may be the greatest challenge we'll ever face.

Articles

Using Test Standard-Setting Methods in Educational Program Evaluation: Addressing the Issue of How Good is Good Enough

Paul R. Brandon

School districts in the United States and elsewhere commonly use standard setting to assign value to student test and assessment scores. That is, they set standards to show “how good is good enough.” This paper presents a summary of the empirical findings on the most widely-studied test standard-setting method and describes what the conclusions of the summary suggest about the use of test standard-setting in educational program evaluations.

The purpose of setting test or assessment standards is to establish judgmentally the *cutscores* that show the dividing points between levels of student performance such as pass and fail, basic and proficient, proficient and advanced, and so forth. Cutscores are established with methods such as the modified Angoff method, the contrasting-groups method, the bookmark method, and several others (Cizek, 2001). As part of student and school accountability efforts, districts report to students the performance levels at which their scores fall and report to policymakers and to the public the percentages of students achieving at the various

performance levels. The U. S. No Child Left Behind Act has enshrined the use of cutscores, in that schools are required to identify and report student proficiency levels and to increase the levels of students who score below proficiency.

Cutscores are set either by making judgments about test items or about examinees' performance on tests or assessments. Methods for making judgments about test items are known as *test-centered methods*, and methods for making judgments about examinee performance are known as *examinee-centered methods* (Jaeger, 1989). The test-centered method that for years was the most frequently used and that remains the most widely studied method is the modified Angoff method (Angoff, 1971), and probably the most frequently studied examinee-centered method is the contrasting-groups method. In preparation for studying how and when to use test standard-setting methods in educational program evaluations, I conducted exhaustive reviews of the literature on these two methods (Brandon, 2002, 2004).

Before districts or states set cutscores, they first must develop *performance standards*. A performance standard is a statement defining and describing the knowledge or skills that students must show at a particular performance level. Performance standards are developed before cutscores are set; cutscores are the operationalized versions of performance standards. Sometimes policy makers specify performance standards and sometimes the panels of judges that set cutscores develop them.

Under what conditions and for what purposes might it be appropriate to conduct standard setting in program evaluations? This topic has been discussed sketchily by some (e.g., Cook, Leviton, & Shadish, 1985; Rossi & Freeman, 1993; Shadish, Cook, & Leviton, 1991; Worthen, Sanders, & Fitzpatrick, 1997) and somewhat

more thoroughly by a few others (e.g., Fink, Kosecoff, & Brook, 1986; Henry, McTaggart, & McMillan, 1992; Patton, 1997; Wholey, 1979). The inattention given to the topic is unfortunate, because the appropriateness of using standard-setting methods in program evaluation has not been thoroughly discussed, and the types of evaluation instances in which using cutscores would be helpful and appropriate have not been well-established.

This article examines the use of test standard setting in educational program evaluations. It begins with a recounting of the primary findings of my review of the literature on the modified Angoff method (Brandon, 2004). I focus on this method because it has been examined empirically more than any other method. However, despite the relative abundance of research on the method, the empirical literature does not provide strong support for the validity of modified Angoff cutscores. Therefore, in this article, I am cautious about applying the method in program evaluation. I argue that it is appropriate under certain testing conditions in formative evaluation studies or when conducting preliminary summative studies of program outcomes. Studies of these types require a lesser degree of validity than summative evaluations used by policymakers to make go/no-go program decisions. Based on the results of the literature review, I discuss flaws in the methods of modified Angoff studies. I then discuss

1. the types of decisions that might be made when interpreting evaluation results in light of cutscores and the strengths of the conclusions made based on test standard setting in evaluations,
2. the program evaluation scenarios in which it is appropriate to use cutscores for interpreting evaluation results, with a focus on the stage of evaluation and the types of evaluation designs, and

3. four criteria that evaluators should address when using cutscores to help interpret evaluation results.

This article is limited by my decision to base conclusions primarily on empirical findings about the modified Angoff research. Some evaluators might wish to know what standard-setting methods other than the modified Angoff method can be used in program evaluations. Psychometricians and researchers are continually developing new standard-setting methods (Cizek, 2001); many such as the bookmark method are proving promising, and evaluators might wish to learn from the research on them. However, the intent of this article is base conclusions on empirical research, and little sound research has been conducted methods other than the modified Angoff. For example, considerable attention has been paid to the contrasting-groups method, which for years probably was used more than any other examinee-centered approach, but little research has been conducted on it (Brandon, 2002). I base my conclusions solely on the research on the modified Angoff method because I have adopted a conservative approach to applying the standard-setting literature to program evaluation. I limit myself to the best research available; the body of modified-Angoff research may be less comprehensive than desirable, but it is broader and goes deeper than the research on other methods.

The article also is limited because it does not suggest how to apply standard setting methods for purposes other than test standard setting in program evaluation. Other than brief comments in the final paragraph of the article, I do not speculate about using the method for other purposes. Very little program evaluation research has been conducted on using standard-setting methods for purposes other than testing. (I have experimented in two evaluations with applying standard-setting methods to judging how well the evaluated programs were implemented, but the success of the efforts was mixed.) There was no research on test standard-setting methods when

they were first put into wide use; I do not intend to repeat that scenario by making recommendations about using standard setting in program evaluation for purposes other than tests without an empirical basis for my suggestions. The place for extensive speculation about other uses of standard setting in program evaluation is elsewhere.

The Methodological Soundness of the Modified Angoff Method

To learn about the soundness of test standard-setting, it is useful to discuss the modified Angoff method, not only because it is an exemplar of one of the two primary types of test standard setting, but also because more empirical research has been conducted on it than any other standard-setting method. As this section shows, the evidence for the effectiveness and validity of the method is less convincing than desirable, the literature is narrow, and many of the studies of the standard-setting method are unsound or incomplete.

The modified Angoff method includes three primary steps. The method is called *modified* because some aspects of it were developed after Angoff (1971) first proposed it. The first step is to select and train judges. The second step is to define and describe the performance level that examinees must meet—that is, to establish the performance standard. Judges can conduct this step, but often policymakers or others provide judges with the performance standard. The third step is to make *item estimates*—that is, to establish estimates of the probabilities that examinees will correctly answer the items on the test or assessment at the level of the performance standard. Usually judges conduct two or three rounds of item estimation. Between rounds, the judges review empirical information such as the difficulty level of each item and have discussions about their item estimates; then, if they wish, they revise their estimates in the next round. After the three steps are

conducted the cutscore is calculated by summing the item estimates for each judge and averaging the sums across judges.

Researchers and practitioners have studied the modified Angoff method more than any other, but some of the findings on the steps are inconclusive:

Selecting and training judges. Some of the research on selecting and training judges provides conclusive findings, but other research does not. Studies suggest that the appropriate number of judges for modified Angoff studies is 10–20. The conclusions of the small number of empirical studies on this topic (Brandon, 2004) generally were within this range.

Selecting judges for their subject-matter expertise can enhance item estimation, but not all judges need have high levels of expertise. Research on this topic is inconclusive because of some of the studies that I identified had methodological flaws and because other studies examined incomplete versions of modified Angoff standard setting.

Very little research has been conducted on training judges, and no results bear summarizing here.

Defining and describing the performance standard. The findings of a small body of studies support the conclusion that definitions and descriptions of performance standards should be made using a set of prescribed steps and that performance standards should be fully explicated. Research on the topic is inconclusive because about half of the studies on it were simulations of standard-setting that did not include or fully implement all the modified Angoff steps (Brandon, 2004).

Defining and describing performance standards is a difficult step to carry out fully and validly. Developing statements of performance standards for high school

graduation tests requires judges to have a full understanding of the knowledge and skills that teenagers must have upon entering the workforce or post-secondary education, and developing performance standards for earlier school grades requires judges to estimate the level of students' knowledge and skills necessary for success in the following grades. In both these standard-setting instances, judges must know what they are setting proficiency scores *for*. That is, they must understand the purpose of the standard setting and the context that students will be in when the students use the knowledge and skills that are addressed in the examination. "To say that adequacy must be defined for some purpose has important implications for validating passing scores as well as validating performance standards. This condition is much more stringent than requiring the passing score to be consistent with the description of performance standards" (Camilli, Cizek, & Lugg, 2001, p. 459). Understanding what scores are set for is not a trivial endeavor; indeed, some would say it is impossible: "Performance standards simply cannot help us decide whether Johnny or PS 19 or Colorado has enough reading skill, because there is no sensible answer to the question, 'Enough reading skill for what?' beyond the trivial level of 'Enough reading skill to answer test question 36 correctly'" (Burton, 1978, p. 270).

There are no well-established developmental theories to guide methods for estimating what students' necessary levels of performance should be upon graduation. What students need to know and be able to do depends upon the educational or vocational paths they will follow upon graduation. The proficiency level necessary for someone to go directly into the workforce is different from level necessary for someone to enter a community college, which in turn varies from the level necessary someone entering a competitive four-year post-secondary educational institution. The minimum levels of knowledge and skills necessary to

succeed in these settings, as well as the highest levels of proficiency that can be expected, vary among these settings. Similar issues apply to setting cutscores for elementary and middle school tests and assessments. Kane (2001, pp. 58, 82–83) said,

There are generally no accepted performance standards for life after high school and no empirical base of information relating performance in history or science in eighth or twelfth grade to success in life (however that might be defined)... Standards seem most arbitrary when the contingencies they are designed to address are very vague and open-ended. The standards set on a high school graduation test are likely to be judgmental, because the level of skill that a graduate will need for work or life will depend on where they work and how they choose to live, and therefore there is no clear focal activity or contingency that can serve as a guide in standard setting. Standard-setting judges must know what students must be proficient for.

A comparison with standard setting in the military is informative. In military settings, training standards are established and applied in personnel decision making. Military training standards address clear external criteria such as the knowledge and skills necessary to operate equipment or perform specialized tasks. This is also more or less the case in standard setting for licensure or certification—a topic addressed in much of the standard-setting literature. It is not the case in K–12 education, where “it is highly unlikely that a teacher will have had experience in the career that his or her students eventually choose to enter. . . . Schools are relatively isolated from the world of work and the consequences of the quality of education they provide, whereas military training centers and operating units are tightly integrated” (Hanser, 1998, p. 82). If traditional K–12 standard-setting methods were used in the military, “the trainers who set the training standards could be quite divorced from field experience” (Hanser, p. 92)—a clearly

unacceptable state of affairs. “Standards that are relatively context free are difficult to set and accept” (Hanser, p. 93).

Making item estimates. More research has been conducted on making item estimates than on any other modified-Angoff step. Some of the findings of this research support the conclusion that cutscores are valid, but other findings make us question the strength of that conclusion.

The findings of research on the extent to which item estimates are correlated with item difficulty levels—a relatively common thread of research in the empirical standard-setting literature—suggest that the estimates moderately mirror item difficulty. This finding is an indication of the validity of the estimates.

Other studies have examined the effects of activities between standard-setting rounds, when judges review empirical information about items and discuss this information and their item estimates. The results of these studies suggest that judges’ between-round activities affect the magnitude of cutscores. However, these results are tentative because about a third of the studies on the topic have not confirmed these findings (Brandon, 2004) .

Other results suggest that judges’ between-round activities decrease item estimates’ variability and increase their reliability from round to round (desirable results). However, the results about decreasing variability are inconclusive because of large standard deviations, and the results about increasing reliability are inconclusive because of the number of studies is small and the methods for calculating reliability varied among studies. Hurtz and Auerbach (2003) found that judges’ discussions among themselves reduced the variability of cutscores but that reviewing empirical information did not.

Researchers also have examined the absolute value of the differences between item estimates and empirical *p*-values. Their studies address *item accuracy*. The rationale behind the studies is that there should be small differences between item estimates and the empirical *p*-values of examinees whose scores are deemed to be close to the cutscore. Although some evidence has been found that judges are able to make estimates accurately, the results of several studies suggest that item estimation might be less valid than desirable because judges tend to underestimate the difficulty of hard items and overestimate the difficulty of easy items. Of all the findings about item estimates, these are the most troubling for the validity of modified Angoff cutscores. Indeed, Shepard (1995, p. 151) concluded that findings such as these showed that “judges were unable to maintain a consistent view of the performance they expected” and thus made judgments that were “internally inconsistent and contradictory.”

Conclusions About the Modified Angoff Method and Its Literature

The findings about item accuracy and the findings about the “proficiency for what” issue lead us to be concerned about using cutscores for a wide variety of program evaluation purposes. These are not the only reasons to be cautious about using the method in program evaluations, however. There also are three flaws in the literature that throw doubt on using the method for a broad array of evaluation scenarios.

The first flaw has to do with the breadth of the literature: It is broader than the research on other standard-setting methods, but it is still narrower than desirable. Insufficient empirical research has been conducted on some steps of the modified Angoff method, particularly on selecting judges, the need for judge subject-matter expertise, judge training, and defining and describing the performance standard.

More research has been conducted on the modified Angoff method than any other standard-setting method, but the findings of the extant research provide only the first few layers of an empirical foundation for making decisions about how to set cutscores. These layers alone cannot serve as the sole basis for deciding about how to go about setting modified Angoff cutscores; clinical guidance by experienced practitioners is also necessary.

(Brandon, 2004, p. 80)

The second flaw has to do with the reporting of studies. Many empirical modified Angoff studies have not reported full descriptions of the standard-setting methods that were used:

The dearth of complete descriptions obfuscates the interpretation of the body of modified Angoff standard-setting literature. If the studies were described more carefully and thoroughly, patterns of interactions among the variations in methods might be discernible. As the research stands now, these patterns cannot be seen.

(Brandon, 2004, pp. 79–80)

The third flaw is methodological. Many of the findings reported in the empirical standard-setting research are from simulations in which only some of the standard-setting steps have been conducted. Research on the method that omits some of the modified Angoff steps is flawed because it does not examine all the key aspects of standard-setting; such research is akin to studying performance assessments in which students are not given instructions for conducting the assessments. Because of the omission of key steps, the findings of some studies are less generalizable than desirable to the fully implemented modified Angoff method.

The primary effect of these three flaws is that we do not have a full understanding of all of the steps of the modified Angoff method. There are not enough empirical

studies to adequately examine all facets of the method, too many of the empirical studies that have been published do not explain how they conducted the steps or else do not conduct some of the steps, and too many studies are analog studies. These flaws, combined with the findings about difficulties in knowing “proficiency for what” and the findings about the difficulty in making estimates for the hardest and for the easiest items, lead me to conclude that it is questionable whether modified-Angoff cutscores are uniformly valid for making summative, high-stakes decisions in program evaluations. Placing great weight on modified Angoff cutscores in high-stakes decisions, as occurs in K–12 education, might be more than their methodological foundation can bear, in part because some of the findings about the method are troubling and in part because the methods and reporting of many modified Angoff studies are flawed.

Evaluation Scenarios Appropriate for Developing and Using Cutscores

Program evaluators might correctly hesitate to use modified Angoff cutscores for high-stakes, summative purposes, but the findings on the validity of cutscores are not so troubling as to refrain from using them in all program evaluations. Evaluators can use them to help interpret student scores for formative-evaluation purposes or to help interpret scores for *suggesting* summative program-evaluation decisions. Cutscores do not have to be interpreted as definitive demarcations of success; “gray areas” about the cutscores can be calculated using the standard error of the mean, resulting in cutbands instead of *cutscores*. This calculation would show a band around the cutscore that would provide an accommodation to the inexactitude of standard setting. Using standard errors in this way, evaluators would have three score bands—one for students who we could reasonably state are

below the desired level of performance, one for those who are more or less at the desired level of performance, and one for those who are clearly above the desired level of performance. Using this analysis, evaluators could report with a reasonable level of assurance the percentages of student scores above and below proficiency. Such descriptive reports could help evaluators understand how well programs are helping students achieve program goals without placing undue emphasis on the cutscore itself. The reports could provide program personnel with general guidance about their programs. Formative evaluation findings and findings that are only *suggestive* of summative conclusions are not used to make go/no-go decisions about programs. When cutscores are used in ways such as these, their precision and validity are less critical than when they are used for making conclusive summative decisions about students or schools.

However, because of the limitations in the research and because of concerns about invalidity, I conclude that the modified Angoff method should be used primarily when other approaches are unavailable for interpreting student scores. That is, cutscores should be developed and used only with some kinds of evaluation designs and only in some evaluation stages. Evaluators should consider using test cutscores to help interpret test or assessment program outcome scores when no comparison or control groups are available. This scenario occurs when educational programs are implemented at all program sites, when administrators and faculty at non-program sites are unwilling to let evaluators use their sites for comparison or control groups, or, in the evaluations of small programs, when evaluation funding is too limited to have comparison or control groups. Cutscores developed when no comparison or control groups are available could help evaluators decide the extent to which children are performing at or near the desired level of performance. Cutscores might particularly be useful during the first year of an evaluation, when

no year-to-year effect sizes can be calculated. Effect sizes showing annual growth are valuable for year-to-year comparisons, because they can be compared with published effect sizes about similar programs studies (Lipsey, 1990; Lynch, 1987), and because they probably are more defensible than cutscores. The two analyses together might also be useful, of course; cutscores used over several years of an evaluation can interpret how high or low program students are performing, irrespective of the size of year-to-year effect sizes.

As long as they are interpreted with caution, cutscores might also be helpful even when comparison groups are used. They can help interpret mean scores when the differences between program and comparison groups are not statistically significant. Comparing average scores to a cutscore could help evaluators know the general levels of performance of both the program and comparison groups. Furthermore, using cutscores could help evaluators tie the interpretation of evaluation results directly to program goals. If a program's goal is, say, to have students achieve proficiency in reading knowledge or skills, evaluators could use cutscores to show the extent to which the proficiency goal had been achieved. The same kind of analysis could be conducted for other levels of student performance. Such reports are rhetorically more powerful than simply reporting whether the program group out-achieved a comparison group or surpassed a specified percentile of a norm group, because comparisons of average scores with cutscores tie evaluation results directly to descriptions of desired levels of student performance.

Criteria for Using Standard Setting in Program Evaluations

There are at least four criteria that should be addressed if evaluators use the modified Angoff method in program evaluations:

1. Standards should be set for reliable and valid tests.
2. The program for which standards are to be set should be well defined with concrete objectives that clearly show what is expected of program recipients upon completion.
3. The standard-setting judges should understand the program objectives well, know the socioeconomic and educational context of the program, and understand the context in which program recipients will study or work after completing the program.
4. The standard setting should be feasible. The standard-setting method should not require more time and resources than the program can afford.

The necessity of the first condition should go without saying; cutscores cannot be used validly to make decisions about program success unless the test for which they are set adequately measures subject matter and produces sufficiently precise scores to make decisions about programs. The other three conditions, however, need some elaboration.

Well-defined programs. When using standard setting in program evaluations, the programs should have clear sets of concrete objectives. Clear objectives are necessary if well-defined and well-described performance standards are to be developed. Although the empirical literature on setting performance standards is not extensive, a small body of studies strongly suggests that performance standards must be thoroughly described and well understood by judges if cutscores are to be valid. Indeed, it is commonsensical that performance standards must be thoroughly explicated, because judges need to understand what students must be proficient *for*.

The “proficiency for what” issue need not be as deleterious in program evaluation

standard setting as it is in K–12 accountability standard setting. K–12 public education provides a wide smorgasbord of educational services to all children. In contrast, many educational programs provide narrow, well-defined services to clearly-demarcated populations. Educational programs typically address a single subject such as reading or science or a narrow topic such as safety, drugs abuse, and so forth. Programs are designed for a single grade level or perhaps two or three grades. They often serve subgroups of students with well-described demographic characteristics. If programs are well-designed, it is likely that their objectives will be clear and the goals more clearly defined the goals typically addressed in K–12 standard setting (i.e., advancing students to the next grade or graduating them from high school). Furthermore, judges in program evaluation standard setting can consider the social and demographic context of the schools that a program serves. Programs often serve smaller populations than entire districts. Judges can define performance standards and set cutscores while keeping in mind the population that the program serves, the wealth and the physical condition of the schools that are served, the typical longevity of teachers serving in the district, and other district demographics that evaluators can gather for judges to consider.

Judges who know the program and its context. Standard-setting judges are more likely to have reasonable expectations about student outcomes in a program if they are intimate with the program’s history, aspirations, administration, line personnel, operations, and so forth. The better they know a program, the more reasonable their expectations about program outcomes will be, and the more likely it will be that they will know the answers to a number of questions, Quoting Smith (1981, p. 266), these questions are

- Has what the program is trying to do ever been done before by anyone? (If not, do not expect too much.)

- Has it ever been done the way the program is trying to do it? (Reasonable expectations are lower for innovations.)
- Is the logic which explains why this program will achieve its desired ends compelling? (The stronger the logic, the more warranted high expectations are.)
- Does the scope of this effort, in terms of time and resources, match the level of effect expected? (Real change usually requires a lot of time and effort.)
- Do contextual factors suggest that this effort might be more or less successful than previous efforts? (Higher expectations are warranted if this program is free of previous contextual constraints.)

It certainly would not be impossible to provide standard-setting judges selected from outside the program with the answers to these questions, but the standard-setting training required to address the questions fully would be onerously lengthy and expensive.

Judges are more likely to develop reasonable expectations if they are familiar with the socioeconomic and educational contexts of a program. Programs in economically disadvantaged communities or in schools lacking good equipment and facilities are less likely to show acceptable levels of performance than are programs in less-disadvantaged communities. Judges should know these contexts because of their effects on student outcomes in the program. Judges can take socioeconomic status and school conditions into account when developing performance standards and setting cutscores. Keeping in mind the mix of schools of varying socioeconomic status and of facilities with varying degrees of maintenance will help ensure that judges' standards are well-informed and reasonable.

The need for familiarity with programs and their social and demographic contexts means that standard-setting judges should be program personnel such as developers or teachers. Others might be insufficiently familiar with the program. For example, parents might not understand program expectations. Also, outside educators such as university personnel might be insufficiently familiar with the conditions of the schools in the program. Program evaluators who are not subject to political pressures can select judges on the basis of how well they know the program and understand the school context, including both the schools themselves and the community in which they reside. It is unlikely that evaluators will find qualified personnel of this sort outside of the program setting.

Having to hire program personnel might mean selecting judges who would be inclined to set lenient program performance standards and low cutscores. Judges might establish erroneously easy performance standards and cutscores because they are loyal to the program, do not wish to see it fail, or believe that they might be under pressure to be easy on the program. This is a source of bias that evaluators should consider when developing program standards. Judges should be trained to establish performance standards that reflect the intent of the program and to set cutscores at levels that match the performance standards.

A colleague and I had teachers serve as standard-setting judges for a state-developed writing assessment that we administered during an elementary-school writing program evaluation (Brandon & Higa, 1998). After pilot-testing the standard setting in another school, all seven fourth-grade teachers in the program school set standards for their students. The teachers addressed the question, “If you instructed your students last year as well as possible, what was the best they could have done?” They answered this question for each of five dimensions of writing—meaning, voice, design, clarity, and conventions (grammar, punctuation, and so

forth).

The seven teachers were deemed the only appropriate group to develop standards because other groups had insufficient knowledge about students' achievement and educational background, writing skills, and the context within which they were taught. The school principal did not participate because he might not have known the capabilities of the cohort of assessed students sufficiently well to have set fair standards, and parents did not participate because they knew too little about content-area knowledge or skills or about program context to arrive at fair judgments.

We were concerned that the seven teachers' estimates of how well students could perform might be lenient because they would not want the effects of their instruction to look poor. To address this concern, we examined the differences between the mean estimates for each of the five writing dimensions and the actual performance of students for which the standards were set (Brandon & Higa, 1998). If the cutscores that the teachers set had been far below student averages, it would have suggested that inappropriate methods were used or that teachers had a self-serving bias. The differences between the cutscores and the performance of the program students showed, however, that the cutscores were somewhat above students' performance, suggesting that teachers did not show a self-serving bias. Furthermore, the cutscores were not so high as to suggest inappropriate expectations. These results helped rule out claims of invalid standards.

Feasibility. Program evaluations must be feasible (Joint Committee on Standards for Educational Evaluation, 1994). Sufficient time and resources are necessary for program evaluation standard setting because good standard setting can be a labor-intensive, lengthy activity. Evaluation theoreticians and methodologists often

overlook feasibility issues, but these must be addressed if practitioners are to use the methods.

In standard setting, both the development of the description of the performance standard and the setting of cutscores require sufficient time and resources. Developing performance standards for a moderately long single-subject test can take half a day (Mills, Melican, & Ahluwalia, 1991; Livingston & Zieky, 1989). Furthermore, setting cutscores is clearly not a brief task, as should be apparent from the description presented earlier of the steps of the modified Angoff method. In modified Angoff standard setting, judges review items, make initial estimates, review empirical information about the items, hold discussions about their initial estimates, revise their estimates, and perhaps repeat the review/discussion/estimation activities for another iteration. These activities can easily last for a full day; in some instances, such as standard setting for the National Assessment of Educational progress, they take two days or more.

When setting standards for the elementary-school writing program (Brandon & Higa, 1998), we eliminated the step of having teachers prepare written descriptions of performance standards; instead, we asked them to estimate the best performance that they reasonably thought children could achieve. We eliminated the step because the rating-scale rubrics described the target level of performance for each rating-scale point. Teachers knew the rubrics well because they had used them to score student papers; they were asked to use the rubrics to substitute for performance standards. When trained in the standard-setting procedures, they simply had to review some of the materials that they had used when doing the assessments. This efficiency contributed to the feasibility of the standard setting. The standard setting method was implemented in a reasonable period of time (less than half a day). The teachers' comments, made during and immediately following

the standard setting, suggested that they understood and fully used the standard-setting methods. Some teachers commented that they were unsure about the percentages to estimate for the scale points, but none resisted participation. None of the comments suggested that teachers found it difficult to apply knowledge of the assessment to the standard-setting task.

Summary and Conclusions

Standard setting, which is widely used by school districts and states to hold students and schools accountable for their educational performance, has not been widely used by program evaluators as a means for helping decide whether a program has performed sufficiently well. Furthermore, the topic has been covered minimally in the program evaluation literature. This is unfortunate, because evaluators could use cutscores to help interpret program outcomes during the first year of an evaluation in which there are no comparison groups. They might even be useful when comparison groups are used, for they help show how high program and comparison groups are performing, irrespective of which group is performing the best.

Standard-setting consists of establishing performance standards, which are statements describing the knowledge and skills that students must attain if they are to perform at a specified performance level (basic, proficient, advanced, and so forth), and it consists of setting cutscores. The modified Angoff method is the most widely studied standard-setting method. As used in the test and assessment standard setting that schools, districts, and states conduct for accountability purposes, the modified Angoff method has three steps. Very little research has been conducted on the first step, which is to select and train the panels of judges who establish performance standards and set cutscores. Other than showing that

10–20 is an adequate range of the number of standard-setting judges, the empirical research literature is of little assistance in identifying the best mix of procedures for this step.

More research has been conducted on the second step, which is to define and describe the performance standard (i.e., the statements describing the level of knowledge and skills that students should attain). The findings are inconclusive but commonsensically suggest that the better that performance standards are defined and explicated, the more valid cutscores are likely to be. Performance standards for educational accountability purposes are murky by nature, however, because it is impossible to know what comprises an adequate level of performance. If a performance standard is defined for graduation, should it be set for students who are going to trade schools, community colleges, state colleges, or private elite universities? What should the performance standard be for students who do not participate in any post-secondary education? If a performance standard for a particular school subject is defined for an elementary- or middle-school grade, what is the developmental or pedagogical basis for deciding what constitutes adequate performance? These questions have not been adequately addressed in the literature, and because of the epistemological complexity of the topic, are unlikely ever to be.

More research has been conducted on the third step of the modified Angoff method than on the other two steps. In this step, judges set estimates of the percentages of students who should pass each item at the level of the performance standard. During this step, judges are given empirical item p -values so that they know the difficulty levels of the items they are judging. The empirical research suggests that judges' discussions make a difference, but the research is not conclusive. Probably the most conclusive research about the third step has to do with the accuracy of

item estimates, which is established by examining the absolute value of the differences between judges' item estimates and item p -values. This research suggests that judges tend to underestimate the difficulty of hard items and overestimate the difficulty of easy items. That is, the range of judges' item estimates is less than the range of empirical p -values.

The research on the three steps of the modified Angoff method has not been conclusive in part because (a) the literature is more narrow than desirable, (b) some of the literature is not reported fully, and (c) the methods of the research have been of low quality. Because of problems with the methods and findings of the empirical research on standard setting, as exemplified by the research on the modified Angoff method—the most-studied of all test and assessment standard-setting methods—it might be concluded that program evaluators should avoid using the method to help make judgments about program success. However, the methods are not so unsound as to preclude their use for formative program evaluation purposes or for making *suggestive* (rather than conclusive) summative evaluation decisions. If cutscores are interpreted with caution and are considered to be suggestive of the success (or lack thereof) of a program, they can help evaluators make conclusions in evaluations that lack comparison groups.

Even though the empirical test and assessment standard-setting literature does not provide convincing evidence about the strength of standard-setting methods, it nevertheless is sufficiently thorough to help us know the conditions that should be present if evaluators use the method in program evaluations. There are at least four of these conditions. The first is that standards should be set only for valid and reliable tests. Evaluators are best advised to set standards for commercially published tests or assessments or for other carefully crafted instruments. Second, cutscores should be set only if program objectives are clearly stated. Otherwise,

performance standards will be difficult to develop. Third, judges should be familiar with the program and the context within which it is taught. The task of setting performance standards for a program is conceptually less complex than the task of setting standards for a school district, because programs (at least those that well-developed and well-run) have clear sets of methods and objectives that standard-setting judges can keep in mind when setting cutscores. This assumes that the judges know the program well and eliminates the possibility of having people outside the program serve as judges. Of course, the charge might be made that program faculty, developers, or administrators who serve as standard-setting judges might set lenient standards. However, in a trial application of standard setting in a program evaluation, it was shown that this need not be the case (Brandon & Higa, 1998). The fourth condition is that the standard setting should be feasible. Evaluators should not assume that they can set standards without proper preparation and full understanding of the mechanics and theory of the procedures. In our trial application of standard setting in a program evaluation (Brandon & Higa, 1998), we showed that it was feasible in a small school-level evaluation.

This article shows that standard setting methods have value in evaluations. They can help evaluators make decisions about program success in the first year of an evaluation that has no comparison groups. In this scenario, other means for deciding about program success are unavailable; therefore, standard setting helps address an empty slot in evaluators' methodological toolbox. The fact that there are weaknesses in the argument for using methods such as the modified Angoff method to make high-stakes decisions need not deter evaluators from using the method during programs' early years, when summative decisions are infrequent. Standard-setting methods also can help evaluators make decisions about program success in later years of evaluations that do have comparison groups. In this

scenario, cutscores can help determine the extent to which both the program group and the comparison group have achieved at sufficiently high levels. In both these scenarios, cutscores should not be interpreted rigidly; they should be used to arrive at *suggestions* about program success. This use of cutscores helps make up for the procedural weaknesses of the method. As long as (a) cutscores are set for valid and reliable tests, (b) program objectives are clear, (c) program personnel serve as standard-setting judges, and (d) there are sufficient resources to conduct the standard setting well, standard setting can contribute to evaluators' decisions.

As stated at the beginning of this article, standard-setting is a means of answering the question, How good is good enough? The conclusions about standard setting given in this article can serve as suggestions about other methods for addressing the question in evaluation studies. First, the stage of the evaluation should be considered. In the case of developing cutscores in program evaluations, the argument for using standard setting to help make evaluation decisions is the strongest in the first year of an evaluation. Other methods for deciding the quality of a program are appropriate in other phases. By way of contrast, experimental and quasi-experimental methods are appropriate when programs are mature. Second, the method for answering the question depends on the use of evaluation findings. Standard-setting methods used for deciding about program success need not be free of flaws when the decisions are formative or when the findings are used to make suggestions, as opposed to conclusive statements, about program success. Experimental and quasi-experimental approaches to evaluation are appropriate for providing conclusive findings about the quality and effectiveness of a program. Third, the context of the program should be taken into account (Smith, 1999). Evaluators using standard setting methods need to find judges who understand the context of the program, or else cutscores will not be well-informed. The

importance of knowledge about context applies to all discussions about how good is good enough. Fourth, the method for answering the question must be feasible. It will not do to require, for example, that all studies use experimental or quasi-experimental designs when the setting or the resources of the evaluation do not allow them. The current push by federal educational research funding agencies to require these designs ignores the feasibility issue—particularly since these same officials do not back up their call for experimental and quasi-experimental designs with funding for expensive evaluations. These four aspects of evaluation should be considered when developing a minimal set of guidelines that evaluators should take into account when establishing the level of performance that a program should show if it is to be considered good enough.

References

- Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.
- Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: A review. *Measurement and Evaluation in Counseling and Development*, 35, 167–181.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59–88.
- Brandon, P. R., and Higa, T. F. (1998, April). *Setting standards to use when judging program performance in stakeholder-assisted evaluations of small educational programs*. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Burton, N. W. (1978). Societal standards. *Journal of Educational Measurement*, *15*, 263–271.

Camilli, G., Cizek, G. J., & Lugg, C. A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives. In G. C. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 445–475). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. C. (2001). (Ed.). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Cook, T. D.; Leviton, L. C., & Shadish Jr., W. R. (1985). Program evaluation. In G. Lindzey and E. Aronson, *Handbook of social psychology* (3rd ed.). New York: Random House.

Fink, A. Kosecoff, J., & Brook, R. H. (1986). Setting standards of performance for program evaluations: The case of the teaching hospital general medicine group practice program. *Evaluation and Program Planning*, *9*, 143–151.

Hanser, L. M. (1998). Lessons for the National Assessment of Educational Progress from military standard setting. *Applied Measurement in Education*, *11*, 81–95.

Henry, G. T., McTaggart, M. J., & McMillan, J. H. (1992). Establishing benchmarks for outcome indicators: A statistical approach to developing performance standards. *Evaluation Review*, *16*, 131–150.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, *63*, 584–601.

- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: American Council on Education/Macmillan.
- Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Newbury Park, CA: Sage.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. C. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum.
- Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.
- Livingston, S. A. & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, 2, 121–141.
- Lynch, K. B. (1987). The size of education effects: An analysis of programs reviewed by the Joint Dissemination Review panel. *Educational Evaluation and Policy Analysis*, 9, 55–61.
- Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice*, 10(2):7–10.
- Patton, M. Q. (1997) *Utilization-focused evaluation: The new century text*. 3rd ed. Newbury Park, CA: Sage.
- Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Newbury Park, CA: Sage.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991) *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels. In *Joint conference on standard setting for large-scale assessments. Vol.2. Proceedings* (pp. 143–160). Washington, DC: U.S. Government Printing Office.

Smith, N. L. (1981). Constructing reasonable expectations in evaluation. *Evaluation News*, 2, 265–267.

Smith, N. L. (1999). A framework for characterizing the practice of evaluation, with application to empowerment evaluation. *Canadian Journal of Program Evaluation, Special Issue*, 39–68.

Wholey, J. S. (1979). *Evaluation: Promise and performance*. Washington, DC: Urban Institute.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guideline* (2nd ed.). New York: Longman.

The Value of Evaluation Standards: A Comparative Assessment

Robert Picciotto

Following an exposition of the ethical dimension, professional role and democratic rationale of standards in the evaluation community, this paper proposes an assessment framework for rating evaluation standards, illustrates its use on a sample of published norms¹ and offers lessons for the participatory elaboration of global evaluation standards.

The Meaning of Standards

Dictionaries do not draw sharp distinctions between principles, guidelines and standards. According to the Oxford English Dictionary, a *principle* is a proposition serving as the foundation of belief or action; a guideline is a general rule or piece of advice; and a standard means a thing serving as recognized example or *principle* to which others conform or should conform or by which the accuracy or quality of others is judged.

Thus, the words tend to be used interchangeably although the notion of principles is commonly perceived as aspirational; guidelines are frequently intended as

¹ The sample reviewed in this paper includes Australia/New Zealand, Canada, France, Germany, Switzerland, the United States and the United Kingdom.

recommendations that do not take precedence over the judgment of experienced practitioners² while standards is the preferred term for mandatory norms, accompanied by enforcement or certification mechanisms.

Since this paper evaluates the intrinsic value of the norms rather than their application it makes no distinction between principles, guidelines or standards. In any event, since no enforcement or certification mechanism exists within the fledgling evaluation profession, all published evaluation principles, guidelines or standards are predicated on voluntary rather than mandatory compliance³ so that the difference between the terms is largely stylistic.

The Ethics of Standards

In industry, standards are used to impose uniformity in design characteristics or processes. They are needed to meet the demands of mass production and/or international commerce for goods and services. As a social practice on the other hand, standard making is designed to shape human behavior and interaction⁴. They

² For more precise definitions see: American Psychological Association, Board of Educational Affairs, *Developing and Evaluating Standards and Guidelines Related to Education and Training in Psychology, Context, Procedures, Criteria and Format*, Approved by the APA Council on February 20, 2004.

³ Principles and guidelines can be made mandatory by including them in contractual agreements between commissioners and evaluators.

⁴ Using the taxonomy of Marie-Louise Bemelmans-Videc, Ray C. Rist and Evert Vedung, *Carrots, Sticks & Sermons: Policy Instruments and their Evaluation*, Transaction Publishers, New Brunswick. 1998, guidelines are carrots, standards are sticks and principles are sermons.

help to achieve explicit or implicit policy goals. Intendedly or not, they promote the interests of particular groups and can restrain competition and creativity.

Hence, standard setting is legitimate only if provides for lack of coercion, equal treatment and the informed consent of participants in an open process. By clarifying expectations and setting rules of conduct, professional standards promote accountability, facilitate comparability and enhance the reliability and quality of services provided. They imply shared values, dedication to professional excellence and voluntary compliance with ethical guidelines. In democracies, standards are set in the public sphere and usually involve the civil society.

According to Jurgen Habermas, rational discourse among principled individuals is the only way to generate sound standards for knowledge creation: *“Representations and descriptions are never independent of standards. And the choice of these standards is based on attitudes that require critical consideration by means of arguments, because they cannot be either logically deduced or empirically demonstrated.”*⁵ This means that standards are context dependent and dependent on the outcome of deliberative processes that are shaped by specific cultural environments.

The Professional Dimension

Whatever their label, all existing evaluation norms have been socially constructed through rational deliberation and context dependent processes. No consensus has

⁵ Jurgen Habermas, *Knowledge and Human Interests*, Polity Press, 1968

yet been reached within the global evaluation profession as to the desirability of complying with internationally accepted norms. Thus, this paper is only meant as a contribution to an on-going debate about the future of the evaluation profession.

In most societies, principles, guidelines and standards are what distinguish a profession from a mere occupation. For some occupations, formal barriers to entry (e.g. academic degrees; certifications or licenses) help to protect the integrity of the profession. For others, informal criteria (e.g. a period of apprenticeship or a record of competitive achievement) suffice. But invariably the franchise enjoyed by a professional group is grounded on the presumption that its members are committed to live up to rules of conduct that protect the public interest⁶.

Such rules underlie the social contract that allows professionals (and the organizations that employ them) to enjoy public trust, practice their craft without undue interference and charge for services rendered. On the supply side, standards enhance the professional stature of those who operate in conformity with them and promote good practices. On the demand side, they facilitate comparisons among providers of services, thus helping customers secure value for money.

Even if the case for evaluation standards is accepted in principle, there are differences of views on their desirable range and scope. Evaluators are still debating whether it is appropriate to set uniform standards to guide or control how evaluation professionals, commissioners, participants and users should behave

⁶ According to the Wikipedia Encyclopedia, to conduct oneself as a professional is to act in accordance with specific rules, written or unwritten, pertaining to the standards of a profession. Evaluation being a young profession, it has yet to develop internationally agreed standards.

(ethical norms), what concepts and practices evaluators should use (methods), the benchmarks their products should meet (quality), the outcomes they should achieve (utilization) or the instruments needed to ensure that agreed standards are met and results achieved in the public interest (verification).

Standards as a Democratic Imperative

According to David Marquand⁷, democracy is characterized by a public domain where “citizens collectively define what the public interest is through struggle, argument, debate and negotiation.” Central to this process is an ethic of public service that “puts public duty and the public interest before market rewards and private interests”. In turn, the effective functioning of a pluralistic society requires professionals that operate autonomously, according to ethical standards. In such a system, professionals are motivated not by “patronage, clientism, connection” but by “professional pride, professional duty, professional authority and ...professional career paths”.

The mandate of evaluation is to assess the merit and worth of public policies and programs on behalf of citizens and with their participation and the credibility of evaluation hinges on proper assurances of quality, objectivity and independence.

Public trust is the critical ingredient. Absent certification and accreditation, evaluators must take it upon themselves to “internalize a set of norms precluding them from abusing their monopoly position and exploiting their clients, and enjoining them to promote the public good”.

⁷ David Marquand, *Decline of the Public*, Polity, 2004

Evaluation standards in and by themselves do not generate good policy outcomes. Indeed, they may have unintended consequences depending on how they are generated and used. If they are centrally imposed and coercively implemented, they may have a chilling effect on creativity and innovation. They may also be viewed as redundant where the value of the evaluation services provided can be reliably gauged in terms of the impact on the quality of decisions reached (ascertained as an integral part of the evaluation process).

On the other hand, just as one does not judge auditors by the profitability of the companies they serve, it is inappropriate to judge evaluators by the effectiveness of the programs and policies being evaluated. While a byproduct of independent evaluation is to assist policy makers and program managers do a better job (the advisory dimension of the profession), the primary responsibility of the evaluator in a democratic society is to enhance accountability, tell truth to power, illuminate policy options, promote public involvement and contribute to the transparency of decisions taken in the public interest.

Equally, asking the clients of evaluations to rule on their usefulness involves moral hazard. It may lead to “feel-good” evaluations tailored to what program managers want to hear rather than forthright assessments that protect the public interest. Evaluation needs to be responsive to the public interest and to operate without fear or favor. As for the accounting profession, the legitimacy of evaluations carried out by (or on behalf of) program managers requires independent validation against agreed standards.

In other words, for evaluators just as for accountants, the client is not always right. Evaluation needs to be independent of vested interests, including those of

sponsors⁸. Irrespective of the funding source, evaluators are ultimately accountable to the public.

While they should give close attention to issues raised by stakeholders, their credibility and integrity hinges on their objectivity and impartiality. This is why evaluation consultants as well as public officials that carry out self-evaluations need independent oversight that attests to the professionalism of their behavior and the validity of their findings.

Do as I Say – Not as I Do?

Ultimately, the case for evaluation standards rests on the golden rule: evaluators should practice what they preach. It would be perverse for evaluators to reject the discipline that they impose on others.

In order to determine the merit, worth and value of an activity, evaluators routinely identify relevant criteria of merit, i.e. they use standards to assess the results of programs and the performance of public officials. Thus, Carol Weiss⁹ refers to standards in her definition of evaluation as “the systematic assessment of the operations and/or the outcomes of a program or policy compared to a set of explicit or implicit standards.”

For Evert Vedung¹⁰, “the value component of evaluation presupposes at least one criterion of merit against which public interventions are judged”. In turn, Michael

⁸ This is why the notion of collective responsibility in evaluation is inappropriate.

⁹ Carol H. Weiss, *Evaluation, Second Edition*, Prentice Hall, Saddle River, 1998

¹⁰ Evert Vedung, *Public Policy and Program Evaluation*, Transaction Publishers, 1999

Scriven¹¹ observes that: “evaluation has two arms, only one of which is engaged in data gathering. The other arm collects, clarifies and verifies relevant values and standards”. With implacable logic, he adds: “anything can be evaluated, including evaluation”.

The design and adoption of commonly agreed evaluation standards would help to resolve the dilemma the profession currently faces in managing an exploding demand for evaluation services within an operating environment characterized by widespread unease about the uneven quality of evaluation products and services and the limited utilization of evaluation results. In setting standards, the emerging discipline of evaluation would emulate the experience of its forebears in the social sciences and the accounting profession.

But in order to make progress along the road of common evaluation standards, a paradox must be explored: if the case for evaluation standards is so strong, why has progress in formulating and endorsing evaluation standards been so slow, halting and partial? What are the limits of standards and what risks must be managed while designing and using them?

The Limits of Evaluation Standards

Evaluation determines the merit, worth and value of things¹². It consists in collecting relevant evidence, identifying suitable evaluative standards and using methods of analysis that are valid and fair. Assuming a stable and predictable

¹¹ Michael Scriven, *Evaluation Thesaurus: Fourth Edition*. Sage Publications. Newbury Park, London and New Delhi. 1991

¹² Michael Scriven, opus cit.

operating environment and provided the causal links that make up a results chain are known (and all evaluation actors are willing and able to comply with the reciprocal obligations that the agreed rules of the game imply) it should be enough to control the quality of outputs or to verify the ultimate impacts of an intervention to create the incentives needed to achieve the desired results.

In other words, the notion of standards is often associated with a linear conception of society predicated on rational behavior and predictable consequences of public policy actions. But in the real world, unintended consequences prevail, the operating context is unstable and the behavior of social actors is influenced by vested interests. The causal links between policy actions and impacts are subject to a wide range of external influences. Lack of precise knowledge regarding the connections between inputs, outputs, outcomes and impacts distorts decisions. This means that evaluation standards must take account of the volatility, complexity and contingency of public service tasks.

Like other social rules and protocols, standards are justified only if they generate social benefits in excess of their costs. Inappropriate standards can cause substantial harm by providing unwarranted assurances. Thus, crude performance indicators, simplistic league tables and performance assessments connected to inappropriate goals can destroy trust and erode the public service ethic. In particular, standards focused on only one stage of the results chain and rigidly applied (e.g. budget controls; quality assurance; inspection; auditing or evaluation) can create perverse incentives.

Conversely, applying and verifying standards at all stages of the results chain can lead to excessive rigidity and prohibitive transaction costs, especially where standards are mandatory and controls are tight. The dogmatic use of standards is

evoked by the alternative dictionary definition of standards as “a document accepted by a church as the authoritative statement of its creed”. Concerns about its restrictive implications can also be traced to the original meaning of the term: “a flag or figurehead attached to the upper part of a pole and raised to indicate a rallying point”; the “distinctive ensign of a sovereign, commander, nation”; the “flag of a cavalry regiment as distinct from the colors of an infantry regiment, etc.”

Such martial images have threatening connotations for evaluators committed to freedom of thought, diversity of perspectives and creativity in methods. They evoke the specter of coerced uniformity, mindless orthodoxy, methodological rigidity and bureaucratic interference. Hence, the strong resistance to mandatory norms among “free thinking” professional evaluators who treasure the integrity of their craft and the freedom to select the methods most relevant to the evaluation challenges they face. This also explains the predilection of most professional associations for terms that are less threatening (i.e. charter, guidelines, principles, etc.) than the word standards.

Diverse Doctrines

The potential chilling effect of standards justifies a gradual and participatory approach to their design and adoption. A pluralistic approach, sensitive to cultural differences is fundamental. Special care is needed to avoid favoring one evaluation school over another. Not all evaluators endorse the notion that evaluators have a mandate to judge the performance of public policies and programs¹³. Some

¹³ See Michael Scriven, *Hard Won Lessons in Program Evaluation*, Sage, New Directions Publication No 58.

conceive of evaluation as a tool for understanding social phenomena. Others hold the view that evaluators are not entitled to question the framework of values or objectives pursued by program managers. Still others view evaluation as a tool for facilitating the achievement of consensus among groups.

Given this diversity, it is entirely legitimate for distinct evaluation schools to elicit different criteria of merit. This is why universal agreement for methodological norms has proven elusive. But all evaluation doctrines endorse judicious rules of conduct with respect to the ethical collection and interpretation of evidence, the validity of findings, etc. Thus, no ontological rationale exists for rejecting evaluation standards based on the notion that evaluation doctrines are manifold. On the other hand, due care should be taken to preserve the space that evaluation professionals need to practice their craft in line with their distinctive values and principles.

Beyond a central “core” of evaluation principles endorsed by all, each of the schools that make up the multi-faceted evaluation profession may choose produce its own principles and methods. Indeed, transparency about the methodological standards used in evaluations ought to be encouraged: clarity about the purposes and roles the evaluation methodologies is designed to serve would help users make a reasoned choice among evaluation suppliers, in line with the “truth in labeling” principle.

Lessons from Accounting and Auditing

Standards are “rules of the game”. They define roles as well as desirable outcomes. They set the voluntary restraints that govern the behavioral relations among individuals or groups. These must be meaningful but not so strict as to paralyze action or hinder innovation. They can be used to assess the performance of all

parties to an evaluation. Whether carried out by individuals or evaluation organizations, they guide the design of evaluation processes.

Since evaluation is to the public sector what accounting and auditing is to the private sector, the lessons gained in the process of developing accounting and auditing standards are instructive. In corporate finance, reporting standards combined with verification and enforcement guarantee consistency and comparability of accounts. The vigorous debate about financial reporting standards between the United States and Europe has centered on the design of standards – not on whether they are needed. Judicious accounting and auditing standards help in the effective and smooth functioning of private markets.

Professional associations of accountants and auditors devote considerable resources to standard setting and certification. They endorse the concept of international accounting standards. Such standards are meant to overcome the problems faced by multinational companies that operate in diverse national jurisdictions. The preparation of reports based on different national principles undermines public trust in corporate accounts since different judgments of financial performance for the same multinational company resulted from inconsistencies in national accounting standards. Thus, the pressure for uniformity in accounting rules rose to protect the credibility, comparability and efficiency of business transactions and facilitate cross border investments.

Similarly, with the internationalization of evaluation under the pressures of globalization, national policies and programs can no longer be held to account without a clear set of benchmarks or without reference to their cross-border implications. A global evaluation community is emerging, peer reviews of national policy performance are becoming routine and a growing international consensus

behind new public management principles is fueling a demand for cross border consistency and transparency in public policy and program evaluations.

Thus, the lessons that have emerged in the process of generating universal accounting standards may have relevance for the design of evaluation standards¹⁴:

- **Ownership:** for legitimacy, standards should gain broad acceptance by professional associations and public authorities at all levels and this in turn requires that they be transparent, enforceable and useful.
- **Tradeoffs between uniformity and relevance:** the advantages of credibility, comparability and efficiency that uniformity delivers may conflict with the quality of the rules and their adaptability to different operating contexts. Principled compromises are needed and, where necessary, second best solutions should be adopted.
- **Need for restraint in prescriptive rules:** Diminishing returns result from efforts to make standards ever more precise and rigorous. Standards should be clear, broad and indicative rather than obscure, detailed and mandatory. They should be as simple as possible but not simpler¹⁵.
- **Independence and competence:** the standard-setting body should be representative, independent and isolated from vested interests. It should have access to expert advisors and have the support of high quality staff. It should include users as well as suppliers of services. Members should be selected

¹⁴ John Flower with Gabi Ebbers, *Global Financial Reporting*, Palgrave, New York, 2002.

¹⁵ From this perspective, Alexander Hamilton's wise words about constitutions are relevant: "(they) should consist only of general provisions: the reason is that they must necessarily be permanent and that they cannot calculate for the possible change of things".

for their technical knowledge, experience and skills operating in their personal capacity rather than as advocates of any national, regional or functional interest.

The Genesis of National Evaluation Standards

For evaluation then, just as for accounting and auditing, standards are part of the social context of the profession. But to a far greater extent, the evaluation profession should adapt its methods to the unique features of individual evaluation assignments. This makes uniform standards for evaluation of public policies and programs far harder to develop than for accounting and auditing. On the other hand, most evaluation practitioners agree that good and bad practices can be distinguished. They accept the judgment of their peers about the quality of their work and they appreciate guidance about the conduct of their work.

In 1994, the Joint Committee on Standards for Educational Evaluation in the United States issued program evaluation standards¹⁶. Since then, other professional associations have generated their own guidelines, principles or standards. The formulation and publication of standards by professional associations has been welcomed even as their limits have been acknowledged. Thus, the American Evaluation Association was responding to a felt need when it developed *Guiding*

¹⁶ The Joint Committee was created in 1974. It published standards for evaluation of educational programs, projects and materials in 1981 and personnel evaluation standards in 1988. The Joint Committee was accredited by the American National Standards Institute (ANSI) to work on program evaluation standards in 1989. Student Evaluation Standards were published in 2003. The Joint Committee Program Evaluation Standards were approved by ANSI in 1994.

Principles for Evaluators (1994) that specify basic criteria for the professional and honorable conduct of evaluations¹⁷.

The principles are very general and cannot be relied upon to provide pointed advice in individual cases. But this does not detract from their usefulness when supplemented by case studies, training and guidance by experienced evaluation managers. Within their limits, they have provided the evaluation profession with a serviceable framework for learning and sharing of experience about the ethical conduct of evaluations.

Similarly, the *Joint Committee (JC) on Standards for Educational Evaluation* published influential standards for the conduct of program evaluations. It deals with ethical, contractual and methodological aspects. The standards were generated through debate among leading exponents of different evaluation persuasions. They were framed in consultation with policy makers and users of evaluation in the education profession. A third edition is under preparation. It is expected to improve attention to cultural diversity issues.

The Swiss Evaluation Society, the German Evaluation Association and the African Evaluation Association have published official standards inspired by the US Joint Committee's work and subsequently streamlined, refined or adapted¹⁸. By contrast, the UK Evaluation Society's good practice guidelines issued in 2003 address

¹⁷ Peter H. Rossi, Howard E. Freeman, Mark W. Lipsey, *Evaluation: A Systematic Approach*, Sixth Edition, Sage Publishers. Thousand Oaks, 1999

¹⁸ The African Evaluation Association guidelines (not reviewed in this paper) include modifications in thirteen out of thirty US PES standards.

explicitly the distinctive roles of evaluation commissioners, evaluators and participants and they also provide guidance for institutional self-evaluations.

The concise charter issued by the French Evaluation Society in the same year stresses the commonality of obligations of evaluators and evaluands while ongoing work by the Australasian Evaluation Society is expected to reach well beyond the rights and responsibilities of commissioners, evaluators and other stakeholders codified in its 2002 *Guidelines for the Ethical Conduct of Evaluations* in order to address more explicitly issues of utilization and integration of evaluation within the policy process.

Criteria of Value for Evaluation Standards

Standard setting in evaluation needs to address social learning as well as individual learning. As Oscar Wilde famously observed, “experience is the name everyone gives to one’s mistakes”. But evaluation helps individuals, groups and organizations learn both from their own mistakes and from the mistakes of others. This is far cheaper and less painful than trial and error. At the level of the individual decision maker, accountability for decisions taken provides incentives for learning while learning improves the quality of decisions and broad based participation helps to protect the public interest¹⁹.

Accordingly, the three main functions of evaluation are (i) to reduce errors in decision-making through knowledge, (ii) to make authority responsible through

¹⁹ The notion that one needs to trade off accountability for learning is mistaken. It reflects a common confusion between the distinctive roles of evaluation consultants (who are loath in their capacity to criticize the performance of their employers and the policies they pursue) and independent evaluators (who face no such constraints).

increased accountability, and (iii) to promote public involvement in public affairs. These three roles (accountability, learning and participation) are inextricably linked. Thus defined they help to determine how the profession should be judged. Specifically, evaluation standards should address three distinct dimensions: (a) collective decision-making; (ii) enhanced accountability; (iii) informed participation.

First, evaluation contributes to effective decision-making by nurturing principled solutions to complex public policy problems. Since neither the state nor private agents have the power to unilaterally define their actions, social decision-making involves bargaining. Evaluation improves the social rewards of the game by providing impartial evidence to all parties and facilitating progress towards agreed solutions. This is where standards of *propriety* come in. They ensure that evaluation is conducted with regard for the welfare of all those involved in the evaluation and affected by its results. Resolving conflicts of value in a constructive way and achieving shared objectives among group members facilitates collective action. Through *transparency*, standards ensure that all parties and the public are adequately informed about evaluation findings.

Second, evaluation levels the playing field of the collective action game by helping to increase the responsibility of authority. By providing objective validation (or censure) of the actions taken by those in authority, it encourages the powerful to behave responsibly and fairly. Evaluation standards add value by discouraging the capture of the process by vested interests. By dispassionately examining the record of past interventions and putting social science disciplines to work, evaluation helps as a counterweight to the prejudices and self-serving opinions of decision makers. This is why *independence* and *integrity* standards are needed to help protect the value of evaluations.

Third, evaluation contributes to public involvement in decisions by reducing information asymmetries and providing advice to the public and to decision makers that can be used to enhance the outcomes of policies and programs. Through participation, evaluation amplifies the influence of those who may not have direct access to decision makers—employees, clients, the poor, etc. Access to relevant information is often too costly to be secured by individual actors. Effective collection and interpretation of the evidence requires *competence*. In this context, standards of *quality* and *comprehensiveness* of guidelines (to cover all stages of the evaluation cycle) provide comfort about the validity, accuracy and objectivity of evaluation findings.

Beyond the seven criteria implied by the three main functions of evaluation (propriety, transparency, independence, integrity, competence, quality and comprehensiveness), the net value added by evaluation is dependent on the *efficiency* of the evaluation process and the *utilization* of evaluation findings. These are only under the partial control of evaluators. Evaluation *governance* factors also intervene.

Getting Results

In learning organizations, evaluation helps to keep transactions low. In rigid, poorly managed, unaccountable organizations, evaluation leads to tense interactions, “cover ups” and even intimidation—so that evaluation ends up raising transaction costs with limited benefits in terms of improved organizational effectiveness. In learning organizations, objective evaluations are used to improve the relevance and impact of interventions and, as a result, unlock enormous benefits at modest cost. Not so in poorly managed organizations where the

evaluation process tends to be captured for the personal benefit of those in authority.

This means that, beyond evaluation supply factors (addressing the right issues, conducting evaluations with efficiency, ensuring fairness and professionalism, etc.) the economics of evaluation hinge on demand factors. How commissioners and program managers behave before, during and after the evaluation process is critical to the derivation of social benefits out of the evaluation process. This is why for evaluation standards to improve the functioning of society, it is important for evaluation standards to deal explicitly with the distinctive accountabilities and the reciprocal obligations of evaluators, evaluation commissioners and program managers.

In particular, the roles of the independent evaluator in the construction of results based management systems and other real time monitoring and evaluation processes (that are an integral part of public sector management) would benefit from codification. Not enough attention has been given to this interface in traditional evaluation standards. Nor have the responsibilities of program managers been explicitly considered in the definition of evaluation standards. Conceptually and operationally, this gap has been filled by the “*evaluability*” doctrine²⁰. Looking ahead, it should have a place in evaluation standards geared to the achievement of results.

Ideally, evaluation standards should relate good practices to the institutional conditions under which evaluation actually takes place. This means that evaluation

²⁰ Joseph Wholey, *Handbook of Practical Program Evaluation* (Chapter 2), Jossey-Bass Publishers, San Francisco, 1994.

governance as well as evaluation practice would benefit from well-conceived standards. Without organizational ethics, the moral hazards of evaluation rise. In poorly managed organizations, evaluation is used punitively to name and shame. In such operating environments, managers use evaluation to censure those who are low on the totem pole. This is why evaluation standards should enhance accountability of authority as well as assist authority through learning from experience.

In other words, evaluation standards do not deliver optimum results unless they focus on the reciprocal obligations of all those involved in an evaluation. Institutions as well as individual evaluators should be guided by evaluation standards. Ethical considerations and technical soundness of quality standards matter but they should be embedded within suitable governance frameworks²¹. Furthermore, it is desirable that the standards be “owned” by the individuals, groups and organizations that use them. Only then are they likely to devote the resources and the skills needed to abide by the standards and make use of evaluation findings.

It follows that the very process of formulating and implementing standards should be viewed as a social learning mechanism (that is subject to evaluation).

²¹ According to Donald T. Campbell, “while all nations are engaged in trying out innovative reforms, none of them are yet organized to adequately evaluate the outcome of these innovations”. This observation led him to pioneer the concept of the “experimenting society” in which “policy relevant knowledge is created, critically assessed and communicated in real life or natural (not laboratory) settings, with the aim of discovering through policy experimentation, new forms of public action which signify a gain in the problem-solving capacities of society”. See William N. Dunn, Ed. *The Experimenting Society, Essays in Honor of Donald T. Campbell*, Policy Studies Review Annual, Volume 11, Transaction Publishers, New Brunswick, 1998

Experience from accounting and other professions suggests that the usefulness of standards hinges on their actual use and, in due course, their independent verification. Effective utilization of standards is facilitated by authoritative accumulation of evidence from adjudicated cases, especially those where the application of agreed standards has been contested.

Finally, tacit norms tend to spread from organizations that are recognized as leaders in their field to the rest of the profession. Thus, sharing of good practice and training programs act as transmission belts between standards and their effective utilization. This is one reason why professional associations have a comparative advantage in the formulation and verification of standards.

A Comparative Assessment

A cursory comparison between national evaluation standards brings out the following common features:

- **Brevity:** the standards in the sample are invariably stated in concise and non-technical terms; they contain 3-6 principles and 11-44 guidelines. The listing of standards varies between one and six pages²².
- **Scope:** Most guidelines focus on the ethical conduct of public program and policy evaluations while the UK product also addresses institutional self evaluation standards.
- **Multiple audience:** all standards in the sample are directed to the individuals and organizations that commission, prepare, conduct and use

²² Additional space is often devoted to commentaries about the guidelines.

evaluations as well as to stakeholders affected by the evaluation or who have an interest in the results.

- **Process orientation:** the standards tend to eschew methodological aspects; instead they concentrate on behavioral, contractual and ethical considerations.
- **Asymmetry:** most standards give far greater weight to the responsibilities of evaluators and the rights of other stakeholders than to the obligations of evaluation commissioners and program managers with the notable exception of the UK guidelines.

This said there are substantial differences among the published standards. In order to carry out a comparative assessment among them, based on the considerations elaborated in the above sections of the paper, the ten criteria of merit that were identified above were used by the author as an evaluation template:

- **Propriety:** preservation of the dignity, security and privacy of people; protection of confidential information; prior informed consent of participants.
- **Transparency:** agreed expectations about objectives and methods are disseminated to stakeholders; evaluation reports disclosed to stakeholders and the general public without modification.
- **Independence:** adequate safeguards provided to ensure that vested interests do not influence the evaluation; distinct accountabilities are attributed to evaluation commissioners, program managers and evaluators; full protection is provided against intimidation and interference; adequate funding without strings is provided.

- **Integrity:** disclosure and avoidance of actual or potential conflicts of interest; contestability of evaluation judgments; evaluator’s access to relevant information.
- **Competence:** requirements about the knowledge, skills and experience expected of evaluators.
- **Quality:** guidance about the practices needed to achieve evaluation relevance, credibility, rigor and objectivity; norms for achieving fair and valid evaluation findings and recommendations; practices that generate constructive relationships among participants.
- **Comprehensiveness:** coverage of all phases of the evaluation cycle – from commissioning to dissemination and utilization.
- **Efficiency:** economy in use of skills, funds and administrative resources; limited burden on participants.
- **Utilization:** responsiveness to the public interest and to users’ needs; participation of stakeholders in the evaluation; timely delivery; clarity of presentation.
- **Governance:** clarity of roles between evaluation commissioners, evaluators and participants; appropriate contractual relationships; oversight of self-evaluation by independent evaluation; “evaluability” norms for program and policy design.

Table 1 presents the summary results based on the admittedly subjective judgment of the author. Equal weights were ascribed to each category. Based on a textual analysis of their content, each of the national standards was rated from 1 to 6—where 1 presents a highly unsatisfactory treatment and 6 a highly satisfactory

treatment. The exercise was carried out for illustrative purposes only, i.e. to demonstrate that evaluation standards can be valued. No claim is made for their accuracy. And it goes without saying that the process followed does not comply with sound evaluation process norms. Validation of the criteria by an expert panel combined with independent scoring by qualified evaluators would be required to attest to the reliability of the individual ratings.

Table 1. Ratings of National Evaluation Standards

	Australia	Canada	France	Germany	Switz.	UK	USA	Average
Propriety	6	5	5	4	5	6	5	5.1
Transparency	4	4	4	4	4	6	4	4.3
Independence	3	1	3	1	1	4	1	2.0
Integrity	2	2	2	2	4	4	2	2.6
Competence	5	5	5	5	5	5	5	5.0
Quality	5	3	3	6	6	4	6	4.7
Comprehensiveness	6	2	3	5	5	5	5	4.4
Efficiency	1	4	1	5	5	1	5	3.1
Utilization	2	2	2	5	5	6	5	3.8
Governance	4	3	3	3	3	4	3	3.3
Average	3.8	3.1	3.1	4.0	4.3	4.5	4.1	

Most national standards give considerable emphasis to the imperative of doing no harm and stress the rights of evaluation participants and the protection of confidentiality. Some mandate a right of prior informed consent for evaluation participants preferably in writing. In general, the standards give salience to the necessity of ensuring that evaluators have appropriate knowledge and skills. The critical importance of quality standards is also stressed, except for the charters of Canada and France that treat this aspect very lightly.

The lowest ratings are those related to the independence and integrity of the function. Only the United Kingdom achieves a satisfactory rating for independence given the scrupulous balance it strikes between the reciprocal obligations of evaluation commissioners and evaluators in achieving good evaluation outcomes. The Australasian guidelines and the French charter come next—the former because of its strong focus on fair and precise contractual relationships and the latter because it states unambiguously that arm’s length relationships between evaluators and program managers (“distanciation”) are needed to ensure the credibility of the process.

None of the other national standards address the risks inevitably associated with cases where evaluation commissioners have major executive responsibilities for the programs being evaluated. Integrity ratings are relatively low because conflict of interest problems tend to be treated lightly if at all and contestability procedures are not provided for. Where conflicts of interest are treated it is in terms of requiring their disclosure rather than on automatic disqualification from taking part in the evaluation.

Nevertheless, Switzerland and the UK achieve satisfactory ratings in this category, the former because of its emphasis on neutral reporting, the latter because it makes clear that the commissioners have a responsibility to provide evaluators with access to relevant documentation and data. The lack of reciprocity in the obligations of evaluation commissioners and evaluators is especially striking with respect to information disclosure. Evaluators are invariably instructed to make transparent the evaluative information on which they have based their findings.

On the other hand, the unimpeded access to relevant information (an acid test of independence for evaluators) while encouraged in some cases is not guaranteed by

any of the national standards. Nor under the rubric of transparency does the inalienable right of the public to access uncensored evaluation reports figure explicitly in any of the national standards although the UK guidelines discourage the quotation of evaluation results out of context and suggest that the final reports should “normally” be lodged in the public domain.

The Canadian guidelines do not address the disclosure of evaluation reports. Instead they emphasize the responsibility of evaluators to their clients with respect to confidentiality, privacy and ownership of findings and reports. The US guidelines (along with the German and Swiss versions that they have inspired) take a middle and somewhat ambiguous course by requiring that the “results” of the evaluation be made available to all potentially affected persons as well as to all others who have a legitimate claim to receive them.

Similarly, the French charter opines that public dissemination of evaluation results is desirable but reserves actual disclosure modalities to a negotiated outcome at the time of evaluation commissioning. By contrast, the Australasian guidelines are explicit in requiring the consent of the evaluator for any amendments to the final report but they do not compel the commissioners to disclose evaluation reports to the public. Instead, they enjoin commissioners not to breach the integrity of the reports in their pronouncements.

The Road Ahead

The above analysis brings out the following conclusions that may help trace a road map for future work on evaluation standards. Utilization ranks highest for the UK guidelines because they devote a full section to self-evaluation. Concern with utilization is also high in the US standards as well as the German and Swiss standards that they have inspired. Nevertheless, most of the evaluation standards

are not oriented to results. They give considerable weight to the contracting phase of the evaluation process, a stage when the commissioners have enormous leverage over the evaluator. They stress early identification of risks and promote good communications between evaluators and commissioners without specific provision for contestability, arbitration mechanisms or independent oversight of the executive branch by the legislature.

Most national standards give far more weight to the obligations of evaluators than to policy makers, program managers and evaluation commissioners. They do not address criteria of program “evaluability” or the measures needed to ensure effective utilization of evaluation results. They do not instruct evaluation commissioners to support evaluators in their evaluative work; provide them with unencumbered access to data; protect their independence; and avoid retribution, intimidation and other means of evaluation capture.

None of the standards makes public officials accountable for the effective use of evaluation results in the public interest. This would require the formulation of standards that address explicitly the institutional prerequisites of organizational learning, e.g. based on the accountability principles of the new public management movement. This would involve codification of the distinctive roles of independent evaluation, self-evaluation, inspection and auditing in various administrative environments.

For evaluation standards to be fully relevant, evaluators, evaluation commissioners and program managers would need specific guidance with respect to results based management systems, quality assurance processes, results based scorecards and selection and use of performance indicators in public service delivery. This is where the current frontier of program evaluation activities lies. Similarly, the

regulatory interface between citizens, government, voluntary organizations and the private sector would benefit from judicious guidelines. Other promising areas for standard setting include the design of appropriate linkages between independent evaluation, budget allocation processes and personnel evaluation practices.

National audit offices have often trespassed profitably into the evaluation domain through value for money and comprehensive audits. Conversely, systematic evaluations of the effectiveness of public auditing and inspection processes would have considerable merit and should be encouraged. In most industrial democracies, public officials feel victimized by “inspection overload”, taxpayers feel powerless to influence the quality of public services and performance indicators are widely criticized because they fail to encourage beneficiary involvement and genuine quality assurance. This suggests a need for more systematic evaluations of control functions, including of inspection and auditing...and of evaluation itself.

Towards Global Standards

Finally future work on evaluation standards should take account of the *transnational* features that now characterize the profession. Increasingly, evaluators are called upon to assess public policies and programs that extend beyond national borders. As a result, a global evaluation community is in the making. It is seeking a common language in order to facilitate evaluation assignments across national boundaries. Given this new context, harmonization of evaluation standards across national boundaries would be desirable. Demands for cross border consistency and transparency in evaluation have become more pressing.

But a global approach not grounded in national and regional experience would involve risks of coercion, rigidity and misplaced homogeneity. To achieve

credibility and legitimacy, global standards should be grounded in the initiatives of national evaluation associations. Consistent evaluation standards that would respect universal ideals of peace and justice would promote high quality work in evaluation, foster trust in the profession and contribute to the advent of an international evaluation community. There is now a wealth of experience in the design and implementation of national standards. It would be appropriate for such standards to be used as building blocks for a global initiative.

The global standards should be inclusive, embrace new stakeholders, accommodate all evaluation doctrines and focus on institutions rather than the individual evaluator. A comprehensive approach to standards (capturing its ethics, its governance, its methods and its linkages to policy making and resource allocation processes) would be desirable so that the sterile debate between principles based and rule based standards that has plagued the development of universal accounting standards is not repeated and the results based approach that is the hallmark of the evaluation profession is given a chance.

National evaluation societies should take the lead in the design of global evaluation standards. A gradual, organic progress is more likely to yield greater ownership than hasty standardization. To provide credibility to the formulation of evaluation standards, policy makers and representatives of the private and voluntary sectors should have their say and due process, including broad based public consultations, will have to be observed. Last but not least, in order to ensure legitimacy, special efforts should be made to involve evaluators of the developing world where 85% of the world's peoples live.

About the Author

Robert Picciotto is Visiting Professor at King's College, University of London. He served as Director General, Operations Evaluation in the World Bank Group from 1992 to 2002.

The 2004 Claremont Debate: Lipsey vs. Scriven

Determining Causality in Program Evaluation and Applied Research: Should Experimental Evidence Be the Gold Standard?

Stewart I. Donaldson and Christina A. Christie

Claremont Graduate University

While there is little disagreement about the need for, and value of, program evaluation, there remain major disagreements in the field about best practices (Donaldson & Lipsey, in press). For example, Donaldson and Scriven (2003) invited a diverse group of evaluators to Claremont in 2001 to share their visions for “how we should practice evaluation” in the new millennium. Theorists and practitioners discussed a wide range of views and evaluation approaches, many at odds with one another, on how best to improve evaluation practice (e.g., the experimental paradigm, evaluation as a transdiscipline, results-oriented management, empowerment evaluation, fourth generation evaluation, inclusive evaluation, theory-driven evaluation and the like). In response to some of the heated exchanges, Mark (2003) noted “it seems ironic when evaluators who espouse inclusion, empowerment, and participation would like to exclude, disempower, and see no participation by evaluators who hold different views.” He

further concluded that whatever peace has been achieved in the so-called quantitative-qualitative paradigm wars remains an uneasy peace.

This uneasy peace seemed to revert back to overt conflict in late 2003, when the U.S. Department of Education's Institute of Education Sciences declared a rather wholesale commitment to privileging experimental and some types of quasi-experimental designs over other methods in evaluation funding competitions. At the 2003 Annual Meeting of the American Evaluation Association (AEA), prominent evaluators discussed this new level of support for experimental designs as a move back to the "Dark Ages" of evaluation. Subsequently, the leadership of the AEA (supported by Michael Scriven among many others) developed a policy statement opposing these efforts to privilege randomized control trials in education evaluation funding competitions:

AEA STATEMENT

November 24, 2003

Dear Colleagues,

We encourage AEA members to share their views on Scientifically Based Evaluation Methods with the U.S. Department of Education. Up to now a number of members have shared their views with other members on EvalTalk. This discussion has been helpful in clarifying our thoughts and in presenting potential arguments, but NOW it is time for AEA members to share their views directly with the Department of Education.

A statement has been prepared by a team of distinguished evaluators including: Randall Davies, Ernest House, Cheri Levenson, Linda Mabry (chair), Sandra Mathison and Michael Scriven. This team received valuable assistance from: Lois-ellin Datta, Burt Perrin, Katherine Ryan and Bob Williams. We are grateful

to this team for their rapid response to this proposal. This statement has been approved by the current and future Executive Committees of the Board of the American Evaluation Association, including:

Molly Engle, 2002 President

Richard Krueger, 2003 President

Nick Smith, 2004 President

Sharon Rallis, 2005 President

Nanette Keiser, 2002-2003 Treasurer

Kathleen Bolland, 2004 Treasurer

We encourage AEA members to share their thoughts directly to the U.S. Department of Education and possibly with legislative leaders. If you agree with the AEA statement, you might indicate your support of the AEA statement.

OR

If you wish to offer other arguments or points of views, please submit those as well.

Responses are to be sent to:

Margo K. Anderson, U.S. Department of Education, 400 Maryland Avenue, SW.,
Room 4W333, Washington, DC 20202-5910

Or by internet to: comments@ed.gov and include the term "Evaluation" in the subject line of your electronic message. Comments must be received on or before December 4th.

Sincerely

Richard Krueger, President

American Evaluation Association

* * * * *

American Evaluation Association Response

To U. S. Department of Education

Notice of proposed priority, Federal Register RIN 1890-ZA00, November 4, 2003

"Scientifically Based Evaluation Methods"

The American Evaluation Association applauds the effort to promote high quality in the U.S. Secretary of Education's proposed priority for evaluating educational programs using scientifically based methods. We, too, have worked to encourage competent practice through our Guiding Principles for Evaluators (1994), Standards for Program Evaluation (1994), professional training, and annual conferences. However, we believe the proposed priority manifests fundamental misunderstandings about (1) the types of studies capable of determining causality, (2) the methods capable of achieving scientific rigor, and (3) the types of studies that support policy and program decisions. We would like to help avoid the political, ethical, and financial disaster that could well attend implementation of the proposed priority.

(1) Studies capable of determining causality. Randomized control group trials (RCTs) are not the only studies capable of generating understandings of causality. In medicine, causality has been conclusively shown in some instances without RCTs, for example, in linking smoking to lung cancer and infested rats to bubonic plague. The secretary's proposal would elevate experimental over quasi-experimental, observational, single-subject, and other designs which are sometimes more feasible and equally valid.

RCTs are not always best for determining causality and can be misleading. RCTs examine a limited number of isolated factors that are neither limited nor isolated in natural settings. The complex nature of causality and the multitude of actual influences on outcomes render RCTs less capable of discovering causality than

designs sensitive to local culture and conditions and open to unanticipated causal factors.

RCTs should sometimes be ruled out for reasons of ethics. For example, assigning experimental subjects to educationally inferior or medically unproven treatments, or denying control group subjects access to important instructional opportunities or critical medical intervention, is not ethically acceptable even when RCT results might be enlightening. Such studies would not be approved by Institutional Review Boards overseeing the protection of human subjects in accordance with federal statute.

In some cases, data sources are insufficient for RCTs. Pilot, experimental, and exploratory education, health, and social programs are often small enough in scale to preclude use of RCTs as an evaluation methodology, however important it may be to examine causality prior to wider implementation.

(2) Methods capable of demonstrating scientific rigor. For at least a decade, evaluators publicly debated whether newer inquiry methods were sufficiently rigorous. This issue was settled long ago. Actual practice and many published examples demonstrate that alternative and mixed methods are rigorous and scientific. To discourage a repertoire of methods would force evaluators backward. We strongly disagree that the methodological "benefits of the proposed priority justify the costs."

(3) Studies capable of supporting appropriate policy and program decisions. We also strongly disagree that "this regulatory action does not unduly interfere with State, local, and tribal governments in the exercise of their governmental functions." As provision and support of programs are governmental functions so, too, is determining program effectiveness. Sound policy decisions benefit from data illustrating not only causality but also conditionality. Fettering evaluators with unnecessary and unreasonable constraints would deny information needed by policy-makers.

While we agree with the intent of ensuring that federally sponsored programs be "evaluated using scientifically based research . . . to determine the effectiveness of a project intervention," we do not agree that "evaluation methods using an experimental design are best for determining project effectiveness." We believe that the constraints in the proposed priority would deny use of other needed, proven, and scientifically credible evaluation methods, resulting in fruitless expenditures on some large contracts while leaving other public programs unevaluated entirely. Statement prepared by: Randall Davies, Ernest House, Cheri Levenson, Linda Mabry (chair), Sandra Mathison and Michael Scriven. This team received valuable assistance from: Lois-ellin Datta, Burt Perrin, Katherine Ryan, and Bob Williams.

Opposition to the AEA Statement

An influential group of senior members of the American Evaluation Association opposed the AEA Statement, and did not feel they were appropriately consulted as active, long-term members of AEA. In response to President Krueger's call for members to share their individual views on this matter, a new statement now referred to as the "NOT AEA STATEMENT" (as seen on Evaltalk) was submitted to the U. S. Department of Education:

NOT THE AEA STATEMENT

Posted on Evaltalk on: 12-3-2003

AEA members:

The statement below has been sent to the Department of Education in response to its proposal that "scientifically based evaluation methods" for assessing the effectiveness of educational interventions be defined as randomized experiments when they are feasible and as quasi-experimental or single-subject designs when they are not.

This statement is intended to support the Department's definition and associated preference for the use of such designs for outcome evaluation when they are applicable. It is also intended to provide a counterpoint to the statement submitted by the AEA leadership as the Association's position on this matter. The generalized opposition to use of experimental and quasi-experimental methods evinced in the AEA statement is unjustified, speciously argued, and represents neither the methodological norms in the evaluation field nor the views of the large segment of the AEA membership with significant experience conducting experimental and quasi-experimental evaluations of program effects.

We encourage all AEA members to communicate their views on this matter to the Department of Education and invite you to endorse the statement below in that communication if it is more representative of your views than the official AEA statement. [Comments can be sent to the Dept of Ed through Dec. 4 at comments@ed.gov with "Evaluation" in the subject line of the message].

This statement is in response to the Secretary's request for comment on the proposed priority on Scientifically Based Evaluation Methods. We offer the following observations in support of this priority.

The proposed priority identifies random assignment experimental designs as the methodological standard for what constitutes scientifically based evaluation methods for determining whether an intervention produces meaningful effects on students, teachers, parents, and others. The priority also recognizes that there are cases when random assignment is not feasible and, in such cases, identifies quasi-experimental designs and single-subject designs as alternatives that may be justified by the circumstances of particular evaluations.

This interpretation of what constitutes scientifically based evaluation strategies for assessing program effects is consistent with the presentations in the major textbooks in evaluation and with widely recognized methodological standards in

the social and medical sciences. Randomized controlled trials have been essential to understanding what works, what does not work, and what is harmful among interventions in many other areas of public policy including health and medicine, mental health, criminal justice, employment, and welfare. Furthermore, attempts to draw conclusions about intervention effects based on nonrandomized trials have often led to misleading results in these fields and there is no reason to expect this to be untrue in the social and education fields. This is demonstrated, for example, by the results of randomized trials of facilitated communication for autistic children and prison visits for juvenile offenders, which reversed the conclusions of nonexperimental studies of these interventions.

Randomized trials in the social sector are more frequent and feasible than many critics acknowledge and their number is increasing. The Campbell Collaboration of Social, Psychological, Educational, and Criminological Trials Register includes nearly 13,000 such trials, and the development of this register is still in its youth.

At the same time, we recognize that randomized trials are not feasible or ethical at times. In such circumstances, quasi-experimental or other designs may be appropriate alternatives, as the proposed priority allows. However, it has been possible to configure practical and ethical experimental designs in such complex and sensitive areas of study as pregnancy prevention programs, police handling of domestic violence, and prevention of substance abuse. It is similarly possible to design randomized trials or strong quasi-experiments to be ethical and feasible for many educational programs. In such cases, we believe the Secretary's proposed priority gives proper guidance for attaining high methodological standards and we believe the nation's children deserve to have educational programs of demonstrated effectiveness as determined by the most scientifically credible methods available.

The individuals who have signed below in support of this statement are current or former members of the American Evaluation Association (AEA). Included among us are individuals who have been closely associated with that organization since

its inception and who have served as AEA presidents, Board members, and journal editors. We wish to make clear that the statement submitted by AEA in response to this proposed priority does not represent our views and we regret that a statement representing the organization was proffered without prior review and comment by its members. We believe that the proposed priority will dramatically increase the amount of valid information for guiding the improvement of education throughout the nation. We appreciate the opportunity to comment on a matter of this importance and support the Department's initiative.

Signed by:

Leonard Bickman

Professor of Psychology, Psychiatry, and Public Policy at Vanderbilt University, Associate Dean, and Director of The Center for Mental Health Policy at the Vanderbilt Institute for Public Policy Studies; Coeditor of the Sage Publications *Applied Social Research Methods Series* and the *Handbook of Applied Research Methods* and the editor of the Journal, *Mental Health Services Research*; recipient of the American Psychological Association's Public Interest Award for Distinguished Contribution to Research in Public Policy and the American Evaluation Association Outstanding Evaluation award; past president of the American Evaluation Association.

Robert F. Boruch

Professor in the Graduate School of Education, Fels Institute for Government, and the Statistics Department of the Wharton School of Business at the University of Pennsylvania; Fellow of the American Statistical Association and the American Academy of Arts and Sciences; recipient of the American Evaluation Association Myrdal Award for Evaluation Practice and the Policy Studies Organization's Donald T. Campbell Award; founder of the Evaluation Research Society, a parent to the current American Evaluation Association.

Thomas D. Cook

Joan and Serepta Harrison Chair in Ethics and Justice and Professor of Sociology, Psychology, Education and Social Policy at Northwestern University; Coauthor of *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, and *Foundations of Program Evaluation: Theories of Practice*; Fellow of the American Academy of Arts and Sciences and the American Academy of Political and Social Science; recipient of the American Evaluation Association Myrdal Award for Evaluation Science, the Donald Campbell Award for Innovative Methodology from the Policy Sciences Organization, and the Distinguished Scientist Award of Division 5 of the American Psychological Association.

David S. Cordray

Professor of Public Policy and Psychology at Vanderbilt University; Coauthor, *Evaluation methods for social intervention*, *Annual Review of Psychology*; past President and Board Member of the American Evaluation Association.

Gary Henry

Professor of Public Administration and Urban Studies, Political Science and Educational Policy Studies at the Andrew Young School of Policy Studies, Georgia State University; Coauthor of *Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs*; former Editor-in-chief of *New Directions for Evaluation*; recipient of the American Evaluation Association Outstanding Evaluation award and the American Society for Public Administration and Center for Accountability and Performance Joseph Wholey Distinguished Scholarship Award; Board Member of the American Evaluation Association.

Mark W. Lipsey

Director of the Center for Evaluation Research and Methodology and Senior Research Associate at the Vanderbilt Institute for Public Policy Studies; Coauthor of *Evaluation: A Systematic Approach*; former Editor in Chief of *New Directions*

for Program Evaluation; recipient of the American Evaluation Association Lazarsfeld Award for Evaluation Theory.

Peter H. Rossi

Stuart A. Rice Professor of Sociology and Professor Emeritus at the University of Massachusetts at Amherst; Fellow of the American Academy of Arts and Sciences and the American Association for the Advancement of Science; Coauthor of *Evaluation: A Systematic Approach*, *Thinking About Program Evaluation*, and *Program Evaluation in Education, When? How? To What Ends?*; recipient of the American Sociological Association Commonwealth Award and the American Evaluation Association Myrdal Science Award.

Lee Sechrest

Professor Emeritus of Psychology at the University of Arizona and founder of the Evaluation Group for Analysis of Data; recipient of the American Evaluation Association Myrdal Award for Evaluation Practice and the Distinguished Scientific Contribution Award from the Division of Evaluation, Measurement, and Statistics, of the American Psychological Association; past president of the American Evaluation Association and the Division of Evaluation, Measurement, and Statistics of the American Psychological Association.

The 2004 Claremont Debate

The exchange above about the role of randomized control trials in program evaluation practice in educational settings set the stage for the 2004 Claremont Debate.

The apparent resurgence of issues reminiscent of the well-known quantitative-qualitative paradigm wars in evaluation has the potential to be destructive and to

stunt the healthy development of the discipline and profession. In an effort to seek a deeper understanding of the current dispute, and to possibly discover a middle ground or productive resolution, Claremont Graduate University hosted a debate between representatives from both sides. Below, you will find selected excerpts from the opening remarks by Mark W. Lipsey (who plans to publish a more complete version of his thoughts in the near future), followed by excerpts from the response from Michael Scriven.

Selected Excerpts from Mark Lipsey's Opening Comments

“In this context, it seems to me that there are at least three topics that we might discuss.”

“One has to do with the way randomized trials appear in government agencies and the legislation and so on, some of which is simplistic and inept, as uncharacteristic as that is of government activity.”

“Another thing we might talk about is the little flack in the American Evaluation Association (AEA) that involves the stance that was taken last year opposing an obscure division of the Department of Education to try to bring in some randomized evaluations to some of the projects it was funding. Since this event is being sponsored by an AEA Affiliate, that is a possibility. I'd be happy to explain to you why I think the AEA now has the same relationship to the Field of Evaluation as the Flat Earth Society has to the Field of Geology.”

“The third thing we might talk about is the methodological issue and what is actually at stake in these methodological critiques. That is actually what I want to talk about, but if anyone, maybe the audience, or Michael wants to talk about the others, then I'd be happy to do that.”

“We really are poorly served by this gold standard terminology. I think that when you use randomized experiments, which I am basically going to defend in this context, they are much like what Winston Churchill once said about democracy. He said, ‘It’s the worst form of government except for all the others that have been tried from time to time.’ I do not think this is the gold standard. I think that for impact assessment randomized experiments are the worst methodology except for some of the others that have been tried from time to time. That is pretty much my theme here.”

“Experimental and quasi-experimental designs have been around a long time and have well known properties. What’s really new is this broadside against them from certain research communities.”

“This issue has evoked mostly a yawn in areas where intervention research and program evaluation is done broadly. So, in mental health, public health, drug prevention, medicine, chronic delinquency evaluations, and a whole range of areas this is not a particularly exciting topic where randomized field trials are well respected, well known, widely used, and understood to be something of the state of the art for doing impact assessments. The reactions I’ve seen have come predominantly from the education research culture and to a certain extent from one wing of economists that work in this field that have an interesting take on it. I will get to that later on.”

“Let me turn now to the non-experimental approaches. This is an area that has fascinated me. Back when flap was going on, methodological pluralism was all over the Evaltalk. I kept asking respondents and finally gave up on what these other methods were that were supposed to be equally valid, and the most interesting list came out: epidemiological methods, observational correlation

modeling, realist methods, case studies, qualitative, ethnographic, Glasser and Strauss' grounded theory, and from Michael Scriven the modus operandi technique, forensic analysis, direct observation, all put forth in establishing the effects of programs."

"I have in recent years, every time I see somebody putting forward the argument that qualitative methods could be used to assess program effects, I've been writing them for some examples. Show me a case where this was done convincingly."

"Why is the education research culture so riled up about randomized experiments? Here are a couple of possibilities. In all the politics this year, the Bush Administration, the Department of Education, the No Child Left Behind Act, there's a lot not to like there, okay? They have been pushing for randomized designs, so we may as well not like those too. The biggest factor I think is ideological. The education research culture bought into constructivism and post modernist epistemologies and so on really big time and there is a lot of ideological opposition. Tom Cook calls it science phobia to quantitative methods and experimentation and so on. Third, I think that there is a considerable amount of ignorance, not stupidity, not stupidity, but ignorance."

Selected Excerpts from Michael Scriven's Response

"Well, apart from the character assassination at the end, which I can tell you in the education community there may be people in it about which those things can be said, but the greatest attacks on constructivism are from people within the education community. So, there are plenty of others like us who absolutely reject all of that crap and so, it is certainly not true. Some of my friends are also on the side of the angels over there, like Tom Cook, for the new move. So, no, I don't think that is really a very plausible account of the story."

“I think that if you want to look at reasons why people objected, the three big ones are these. One, the objections were not at all against randomized control trials (RCT), they were against the decision to take all \$500 million dollars of their research money and pull it out of anything except randomized control trials. Now, it is quite clear the previous speaker is not identifying himself with this extreme wing, but who is the leader of the extreme wing? It is the guy who is the head of the Institute of Educational Science that has the \$500 million, and what does he say? He says there is no scientific way of establishing causation except by randomized and allocated control group trials, etc. etc. There is no such thing as scientific research in the area of human behavior except by means of RCTs, and that is complete bullshit! It happens to be coming from the guy who has all of the money. So, the sad thing is that this is man killing off alternatives”

“Read Tom Cook on problems in practice of running RCTs. So, this is a very tricky procedure. While it has theoretical advantages, the theoretical advantages in validity aspects of it are undeniable. That is not the issue. The issue is not whether or not there is an alternative that has the same theoretical bulletproof-ness. The question is whether there is an alternative that can get you results beyond reasonable doubt, and that is another story all together. Very often, you can get results beyond reasonable doubt in other ways.”

“First, the concessions. We have not used RCTs when we should have many, many times. There have been many occasions when we could have pulled off RCTs, when we could have staffed them with competent people, and this is still the case in the present, and that was the best design around. The arguments around are sloppy arguments including a number of arguments that Professor Lipsey ran into at the Evaltalk discussion. There was a lot of whistling in the dark going on there and ideological crap going on. You have to get down to the logic of the cases and

you can't just pull this off by waving things like constructivism, observational, or etc. So, this is a situation where there is no doubt at all. This is a very powerful tool, and sometimes much the best tool, but it has as the same value as the torque wrench in a good mechanic's toolbox. For certain tasks, you can't beat it. After all, this is a quantitative instrument. The torque wrench reads out in inches and meters and so on, so this is very important if you are interested in matching the specs that you are supposed to be matching...a very good instrument. Nothing can match it, but it has a very narrow range of uses. Now, that doesn't matter if the alternative approaches aren't very good, but of course there is a lot of them and some of them are very good indeed."

"Well, there's a lot more I'd like to say, but perhaps I can just leave it by saying I think I agree strongly with him. A lot of the attacks have been empty and they have lacked specific examples that will work. A lot of the attacks are based on ideological positions, which are logically unsound. All of this is true, but nevertheless, given the difficulties facing RCTs, one has to be very cautious going to any sort of wholesale commitment to them. I hope in the future we can develop a better kind of existence than what we have at the moment."

Conclusion

Somewhat surprisingly, Lipsey and Scriven agreed that randomized control trials (RCTs) are the best method currently available for assessing program impact (causal effects of a program), and that determining program impact is a main requirement of contemporary program evaluation. However, Scriven argued that there are very few situations where RCTs can be successfully implemented in educational program evaluation, and that there are now good alternative designs for determining program effects. Lipsey disagreed and remained very skeptical of

Scriven's claim that sound alternative methods exist for determining program effects, and challenged Scriven to provide specific examples. Streaming video of the entire Claremont Debate can be viewed at: <http://www.cgu.edu/pages/465.asp>.

About the Authors

Stewart I. Donaldson is Dean and Professor of Psychology, School of Behavioral and Organizational Sciences, at Claremont Graduate University. He has published widely in evaluation, developed one of the largest university-based evaluation degree, certificate, and professional development programs, and has conducted evaluations for more than 100 organizations during the past decade.

Christina A. Christie is an Assistant Professor, Director of the Masters of Arts Program in Psychology and Evaluation, and Associate Director of the Institute of Organizational and Program Evaluation Research in the School of Behavioral and Organizational Sciences at Claremont Graduate University. Her research interests focus on investigating the relationship between evaluation theory and practice and issues related to the development of descriptive theories of evaluation. She has conducted a variety of educational evaluations, and evaluations of social programs targeting high-risk and underrepresented populations.

References

Donaldson, S. I., & Lipsey, M. W. (in press). Roles for theory in contemporary evaluation practice: Developing practical knowledge. In I. Shaw, J. Greene, & M. Mark (Eds.), *Handbook of evaluation*. London: Sage.

Donaldson, S. I., & Scriven, M. (2003). Diverse visions for evaluation in the new millennium: Should we integrate or embrace diversity? In S. I. Donaldson &

M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium* (pp. 3-16). Mahwah, NJ: Erlbaum.

Guiding Principles for Evaluators (1994). *New Directions for Program Evaluation* (No.66). San Francisco: Jossey-Bass.

Joint Committee on Standards for Education Evaluation (1994). *The program evaluation standards: How to assess evaluations of educational programs*. Thousand Oaks, CA: Sage.

Mark, M. M. (2003). *Toward a integrative view of the theory and practice of program and policy evaluation*. In S. I. Donaldson & M. Scriven (Eds.) *Evaluating social programs and problems: Visions for the new millennium* (pp. 183-204). Mahwah, NJ: Erlbaum.

Evaluation Capacity Building and Humanitarian Organization

Ridde Valéry and Sahibullah Shakir

Abstract

This paper documents a process of evaluation capacity building in a humanitarian organization in Afghanistan between 2001 and 2003. The authors carried out an annual evaluation and they undertook evaluation capacity building activities. The analysis of the empirical data shows that in the context of humanitarian organizations, the capacity building process would be improved if it would i) employ a mix of participative and utilization-focused approach, ii) organize participative workshops and on-the-job training, with the continuity of collaborators ensured, iii) use a myriad of dissemination/advocacy activities for a varied public.

Résumé

Cet article vise à expliciter un processus de renforcement des capacités en évaluation de programme d'une organisation humanitaire en Afghanistan entre 2001 et 2003. Nous avons effectué une évaluation chaque année et certaines activités visaient le renforcement des capacités. L'analyse des données empiriques montre que dans le contexte des organisations humanitaires, le renforcement des capacités gagnerait à i) employer une approche participative et centrée sur

l'utilisation des résultats, ii) organiser des ateliers de formation participatifs, former les acteurs sur le terrain et s'appuyer sur des collaborateurs récurrents, iii) user d'une myriade de formes de valorisation des résultats et de plaider en faveur de l'évaluation pour un public varié.

Introduction

The capacity building of humanitarian organizations relates to the multiple functions and activities carried out by these organizations. Literature is rich with articles and chapters depicting poor capacity building practices in these types of organizations, “capacity development has been largely unsuccessful” said ALNAP in the 2003 review of humanitarian action¹. These are often written in a very negative way. In this paper we wish, to provide a more constructive perspective, as Morgan et al.² have done regarding training and education, on the way in which capacity building activities of humanitarian organizations is carried out, while remaining critical and rigorous at the same time. For this purpose, we will present the case of a Non Governmental Organization (NGO) implementing community health programs in Afghanistan. Milstein et al³ said that “an important distinction might have to be made between the conditions that confer evaluation capacity building to an organization and the strategies used to bring about those conditions and sustain them over time. The former is a theoretical question, the latter an empirical and practical one”. This paper deals with the latter case, it does not pretend to provide a theoretical basis, but it exclusively aims to present empirical data concerning the process of capacity building in a particular double context: a country in a transition and a humanitarian NGO. It has been particularly interesting to study a case in this country, as for the past three years, this part of the world has lived upheaval, passing from a situation of war to a situation where democratic elections were organized in a post-conflict country. That being said, we cannot in

these few pages review all the capacity building activities of this NGO, particularly those for medical or administrative activities.

This is why this paper will only focus on the evaluation capacity building (ECB) activities of this organization in Afghanistan. ECB is the intentional work to continuously create and sustain overall organizational processes that lead to quality evaluation and its routine use⁴. In this paper we will handle this topic for three essential reasons. First, experts in this field are asking for more empirical case studies to document the range of practices in order to improve their knowledge³⁻⁶, as ECB is “an emergent field of practice”⁷. For example, the topic of ECB was only brought up in the agenda of the Annual American Evaluation Association National Conference in 2000⁸. Second, it should well be recognized that papers on this subject in a context of humanitarian aid are relatively rare. Donors and NGOs are supporting ECB activities for at least three decades⁹. But most of these activities occur in developing countries and not in conflict or post-conflict settings. Third, we believe that what makes this of particular importance is that the evaluator, the author of this paper²³, has carried out three evaluations in continuation in the same country for the same NGO during three years in 2001, 2002 and 2003. This is a rare situation, and we believe contributes to the abundance of knowledge. We thus consider it is important to share this experience,

²³ The first author knows this NGO well since 1996. He has served as its Head of Mission in Afghanistan from 1996 to 1998, then in Mali and Niger in 1999. He has also conducted evaluation work for this NGO in other countries, like Niger (98), East-Timor (99) and Iraq (2003). But we will focus the case study on Afghanistan. In other words, the CBE activities were not only implemented in Afghanistan. In addition it is thought that other evaluation practices undertaken by this consultant and others contributed to the building of the evaluation capacity of this NGO. These endeavours are clearly beyond the scope of this paper.

and our reflections, with the humanitarian community. One of the limitations of this paper is that it focuses more on process than outcome of ECB, even if some indicators of changes that occurred as a result of those activities will be shared. It seems to be the case most of the time in this kind of papers¹⁰. As this Afghan process is new and recent, outcome based evidence is scarce and its description is considered to be the first stage to climb the evidence-based iceberg.

Context

A thorough description of the context is important as ECB practices are highly “context-dependent”⁷ according to the most cited definition. After more than 20 years of conflict and important economic decline¹¹, chances for development in Afghanistan are impaired by the worsening health condition of the population. Indeed, health indicators, especially maternal and infant mortality rates, are among the worst in the world and some of them are increasing: UNICEF shows a rise from 600 maternal deaths in 1981 to 1,700 deaths in mid-1990. A recent women’s mortality survey, conducted in four provinces of Afghanistan, confirms this scenario: the maternal mortality ratio is 1,600 per 100,000 live births. Even more, the maternal mortality rate reported in Badakhshan province is the highest ever reported globally in the world with 6,500 per 100,000 live births¹². The infant mortality rate is thought to be 165 per 1,000 successful births and the under five mortality rate about 257 per 1,000 live births. The low socio-economic status of women renders them and their children particularly vulnerable¹³. Most of the burden of illnesses stems from infectious diseases, particularly among children, where diarrhea, acute respiratory infections and vaccine-preventable diseases are likely to account for 60% of the children’s deaths¹⁴.

According to a recent report done for the Afghanistan Research and Evaluation Unit (AREU)¹⁵, the health system is adversely affected by major problems: a grossly deficient, and even absent, infrastructure; a top-heavy health system; poorly distributed resources; health care delivered on a 'project' basis by many distinct, relatively uncoordinated service providers; absence of a practical, useful, coordinated information system for management decision-making. In addition, the pre-war in human resource capacity has been eroded and there is scarcity of personnel with managerial and technical skills throughout the country. There is also a lack of training and a lack of public health expertise, for all health staff and doctors are generally not able to deal with the most urgent problems at a community level. Indeed, medical facilities and personnel are very few in number and are primarily found in Kabul; approximately 90% of all trained physicians practice in urban centers, with almost 80% in Kabul itself. In rural areas, NGOs are in charge of the large majority of the health facilities. They have to implement, mostly through a contractual approach¹⁶, the new Basic Package of Health Service defined by the Afghan Government in its new National Health Policy¹⁷. However, access to health services remains appalling for rural populations because of limited public transport, cultural constraints that limit the access to health care for women, high illiteracy levels with lack of knowledge about health care, few hardtop and rural roads and the absence of telecommunications. Moreover, twenty-three years of war and recent droughts have eroded household assets and many families live in abject poverty¹⁸.

In 2004, Afghanistan is not yet safe and secure; tensions still run high in most parts of the country. Moreover, there are signs of nascent problems, notably harassment of the International Community by Government authorities and the potential return to violence in some areas. Current insecurity and political instability will obviously

constrain the pace and geographic scope for extending health services. Intense ethnic rivalries and local conflicts have undermined trust in public and government institutions and will remain a challenge in the years to come.

A French medical NGO founded in 1979, Aide Médicale Internationale (AMI), is acting and working in Afghanistan since the early eighties, undertaking different kinds of activities that focus on the rehabilitation of health care structures and on medical training for health care workers. Initially, all missions were secret ones taking place during the Soviet occupation of the Afghan territory. From 1985 to 1993 AMI ran a training program (Medical Training for Afghans) in Peshawar (Pakistan), and provided the 115 graduated students with medical kits to start their activities inside Afghanistan¹⁹. This was a huge project in term of medical capacity building for Afghanistan. Unfortunately, AMI do not have much information regarding the current position and profession of those hundred medical trainees. In 1993 AMI started two dispensaries in Kunar Province, and a reference Hospital in Logar Province that was linked to a training centre. In 1995 the NGO started two dispensaries in Laghman Province and took over the provincial hospital of Laghman. From 1995 to 1998, AMI ran ten Mother and Child Health (MCH) clinics in Kabul. Then, in 1997 AMI rehabilitated the Central Reference Laboratory in Kabul and still supports it through supply, training and supervision activities. In April 1998 a medical team went to the Upper Panjshir Valley and opened three Dispensaries.

Since 1996, AMI run a multi-disciplinary health program funded by the European Union and implemented in partnership with the British NGO “Sandy Gall Appeal for Afghanistan”, with AMI acting as a primary agency in the partnership. AMI supported different health facilities in three provinces (Kunar, Logar, and Laghman) in the Eastern Region of Afghanistan. From 2001 to 2003, the name of

the program was: “Support to the Health Care system in three provinces, Salamati, a distance-learning magazine for Afghan health workers and The Rehabilitation and Prevention Program for Disabled Afghans in the Eastern Region of Afghanistan”. The general objectives of that program were to improve the quality of services and to improve access to health care for the most vulnerable groups in the target areas of the project, especially women. To reach these objectives, AMI was providing financial, technical and logistical support to implement the following activities in three provincial hospitals and six clinics as well as in the surrounding communities: i) to train the medical and administrative staff; ii) to supply the facilities with necessary medication and equipment to treat the patients; iii) to maintain the buildings in proper conditions and add new constructions where necessary; iv) to train community health workers and organize information meetings in the communities; v) to edit, publish and distribute a quarterly distance-learning magazine.

Evaluation Capacity Building Framework and Practices

Medical and Administrative Capacity Building Activities

As we can see, most of the past and current programs run and supported by AMI have a capacity building component, mostly on the medical and administrative side, like many other organisations in international health development². The training of 115 graduate students during the Mujjahidine times is an earlier one, but in the past years some Afghan employees had the opportunity to reinforce their capacities thanks to three strategies: on-the-job training, formal workshops and courses at the headquarters and formal training abroad. The Afghan responsible for the biology programme spent two months in different hospitals in France in 2000 and he started in the end of 2004 a six-month training program at a French

university. The Afghan financial director worked in dyad during three years with some expatriates and followed distance courses in accounting and finances. He was in Paris for a few weeks in 2004 to pass the national (French) accounting exam. The «Salamati» magazine is one of the famous medical capacity building activities done by AMI in Afghanistan. “Salamati” means «health» in Persian. This journal was created in 1994 as a medium to foster continuous education amongst midlevel Health Care Workers in Afghanistan. The Journal is published quarterly with 6,000 copies. It is distributed all over the country, through the outlets of different medical NGO’s and United Nations agencies.

ECB Framework

During the last three years AMI commissioned one program evaluation a year in Afghanistan and, even if it was not explicitly stated, there were important ECB components established in this exercise. This is what we are going to describe in the following pages. We wish to demonstrate that ECB practices and evaluation practices are two faces of the same coin.

In one of the most recent publications on ECB, experts from the Center for Disease Control (CDC) said “One problem is that the evaluation profession as a whole still lacks a well-developed theory and associated indicators for understanding evaluation capacity at an organizational level, particularly its inherent change over time and “ongoingness.”. This is why, first, this paper does not pretend to provide extensive data on ECB outcomes, and second, we will use a broad framework to make the way in which the ECB activities were held in Afghanistan understandable. Using an adaptation of mainstreaming evaluation and key elements of building evaluation capacity according to Duignan²⁰, we will present in this paper some activities that we implemented during the past three years, in term

of i) evaluation model, ii) evaluation skills, and iii) advocacy/dissemination. Even if for some authors²⁴ mainstreaming and ECB are different⁷, the divergence between these two evaluation streams does not appear so big in term of their main components. Although ECB literature is limited⁷, these three elements which were chosen from a mainstreaming author to depict the Afghanistan process are usually described as part of the ECB practice. According to Bozzo²¹, two of the challenges for ECB in the voluntary/nonprofit sector are evaluation skills and finding the appropriate approaches. The recent conceptual framework and the accompanied extensive review proposed by Cousins et al.¹⁰ regarding the integration of evaluative inquiry into the organizational culture present three key variables of interest in the evaluation dimension: evaluative inquiry, evaluation capacity and evaluation consequences. The first variable corresponds to our evaluation model and approach element, the second to the skills component and the third to the advocacy/dissemination activities. In this paper, the spirit of the use of this last component, according the ECB definition retained^{4,7}, is that we believe that the aim of ECB practices is not only directed to “the ability to conduct an effective evaluation”, as Milstein and Cotton⁸ or Bozzo²¹ said, but also in order to increase the utilization of quality evaluation results by NGOs. This is why we consider that advocacy and dissemination activities could contribute, as a component of ECB, to the utilization of conclusions, lessons learned and recommendations of evaluation.

Appropriate Evaluation Model

Between 2001 and 2003, three evaluations in Afghanistan were conducted by the first author of this paper. The second author is responsible for the programme at

²⁴ Note that if the 2000 Annual American Evaluation Association National Conference was on “Capacity Building”, the 2001 topic was “mainstreaming evaluation”.

the NGO headquarters and he supervises, at distance and a few times per year in the field, the programme in Afghanistan.

We have argued elsewhere²² that in an international situation of humanitarian aid where the context of the evaluation is an essential element, but impossible to manage, it is best to use a participative approach and to minimize the distance between the evaluator and the participants. This evaluation model could significantly increase the probability of appropriation of the evaluation results and the application/adaptation of the recommendations. Thus, NGOs wishing to organize an evaluation in such a context may find it very useful to collaborate with expert-facilitators (as evaluators) who use the participative approach, and who at the time same know well the specific situation and the organization that implements the program. The expertise in evaluation is not its own self sufficient. For all these reasons, we believe that this specific approach is, in this particular context, one of the most appropriate evaluation models to improve and build the evaluation capacity of NGOs. We also argue that this does not only hold true for development projects, as we have known for a long time²³, but also, as is the case in this paper, it holds true for humanitarian projects run by NGOs in complex settings.

Having said this, we must add that the extent of participation was not the same during the three above mentioned evaluations. Implicitly, we decided to use an evaluation model which employed approaches more and more near the ideal-type of the participative model (practical type and not empowerment type²⁴). The goal was to gradually reinforce competences and knowledge of the NGO stakeholders in terms of evaluation and institutionalization of those activities. Although in the context of international development NGOs have been first to mainly apply this type of pluralist approach^{22,23}, AMI was not truly accustomed to such a process in

Afghanistan. The context of permanent war during more than 20 years, obliged the NGO to work in substitution of the State and without much of participation of the communities in decision making, is one of the explanations to the lack of use of such an approach. It should be noted that the implementation of the participative approach for the first time in 2001 during the first evaluation proceeded in parallel with the will of the NGO to give a wider role to the local populations in the management of health centres. It is as of this time that the first attempts to establish Health Management Communities were tried. Also, the gradual approach with regards to participation is justified by the gradual evolution of the context passing from a situation of war with the presence of Tabebans (2001) to a situation of post-conflict and rebuilding of the State (2003).

Before we show and analyse the depth of the participation, let us summarise in few words the purposes of those three evaluations (see Table 1).

Table 1. The Three Evaluations from 2001 to 2003

Evaluation Component	2001	2002	2003
Context	War, American Invasion	Sporadic conflict, interim government, donors come-back	National health policy, performance-based contract approaches
Evaluation Team	One External Evaluator, two internal data collection supervisor, four internal data collectors	One external evaluator, two internal workshop facilitators, three internal indicators team members	One external evaluator and a team of six internal evaluators
Type	Effectiveness and efficiency	Criterion-focused	Process evaluation
Objectives	Assessment of health care financing mechanisms	Determination of performance indicators for the programs	Analysis of program activities and strategies implemented and development of "lessons learned"
Tools	Household survey, bed census, interviews	Three Regional Workshops with stakeholders, NGO Health Information System, WHO indicators	Evaluation workshop, questionnaires, focus group, interview, documentation, action plan workshop
Data	Mostly quantitative	Mostly qualitative	Mostly qualitative
Duration in the field	One month	Three weeks	Three weeks
Potential Utilization	Change in the user fees schemes	Implementation of a monitoring and evaluation system	Improvement in future programs

In another article where we analyze in depth the 2001 evaluation participative process²² we proposed, following and adapting Patton²⁵, to define participative evaluation according to nine criteria gathered in three categories. We will distinguish three categories of participants whose hierarchy is instituted according to their capacity to intervene in the use of the evaluation results since we are using

an utilization-focused evaluation approach²⁵. Table 2 illustrates the depth of the participation in the three processes and how, gradually, we use the appropriate evaluation model according to the context and the NGO wishes. We will, in the next section, explain in more detail how this progressive practice allowed us to build the evaluation skills of the local staff in order to improve their participation in the process.

The detailed analysis of the elements in Table 2 is beyond the scope of this paper. However, we think that it is useful to give some empirical elements. For that purpose, we are using this table to show how much the degree participation was gradual important from 2001 to 2003. The top of the use of this approach was the evaluation of 2003 which, adapting a method proposed by Aubel²⁶, allowed the utilization of a model close to the ideal-type of the practical participative evaluation model. The details of this last evaluation are presented elsewhere²⁷. We just want to add that to overcome the problem of integration of lessons learned into the program and appropriation of recommendations, it was proposed that the evaluation exercise include a final one-day workshop in which a draft action plan regarding the implementation of recommendations was developed based on the evaluation findings and lessons learned. Then, it was decided to establish an evaluation steering committee in order to organize a participative process to finalize the document of action plan by topic and implement it.

One of the arguments in favour of the utilization of the appropriate evaluation model in order to improve the capacity building activities is that a wrong model will, not only be unable to answer the evaluation question asked by the NGO, but also it would decrease the understanding and the trust of stakeholders regarding the evaluation practice. In others words, as said Bozzo²¹, “the efforts undertaken will be sustainable over the long term”. Table 2 is of special interest with regard to this

point and it demonstrates that the participative approach, in its ideal-type sense, is maybe not the most appropriate model for an effective evaluation. In fact, if the depth in participation gradually increased it was also due to a pragmatic objective: to increase the appropriation of the evaluation model. In other words we can say that if in 2003 AMI wanted an efficiency evaluation, it could be sure that the depth of participation was not as it was for the process evaluation. This observation is not new for evaluation theorists but with this empirical data we confirm it and show that this was certainly one of the elements of the capacity building process.

Table 2. Degree of Participation of Three Categories of Participants According to the Nine Minimal Criteria of a Participative Evaluation

	On the field: head of mission and medical coordinator	Local department responsible, expatriates in the field and directors and staff of clinics/hospitals	Population and patients
	In the headquarters: persons in charge for program and medical		
Content			
The evaluation process involves participants in learning evaluation logic and skills	+/-	+	-
	+	++	+/-
	+	++	-
Participants focus the evaluation process and outcomes they consider important and to which they are committed	++	+/-	+/-
	++	+/-	+/-
	++	++	++
All aspects of the evaluation, including data, are understandable and meaningful to participants	++	+	-
	++	++	+
	++	++	+
Process			
Participants in the process own the evaluation. They make the major focus and design decisions, they draw and apply conclusion	+	++	+/-
	+	+/-	+/-
	++	++	+/-
Participants work together as a group and the evaluation facilitator supports group cohesion and collective inquiry	-	+/-	-
	++	++	+
	-	++	+/-
The evaluator is a facilitator, collaborator, and learning resource; participants are decision makers and evaluators	+	++	+/-
	++	++	+/-
	+	++	+/-
Status differences between the evaluation facilitator and participants are minimized	++	++	-
	++	++	+/-
	++	++	+/-
Finalities			
Internal, self-accountability is highly valued	++	+	-
	+	+	+/-
	+	++	-
The evaluator facilitator recognizes and values participants' perspectives and expertise	++	+/-	+
	++	++	+
	+	++	++

Note. Degree of participation from 2001 (first line) to 2003 (third line) + + = > very intense, + = > intense, +/- = > average; - = > absent.

Developing Evaluation Skills

Since 2001 and throughout the three evaluations, we used every favourable moment to the develop program evaluation skills of the stakeholders engaged in the evaluated projects. Two particular strategies were retained: on-the-job training and workshop training.

On-the-Job Training During the Evaluation Process

Thanks to the fact that the Afghan medical coordinator of the NGO remained in his position during the three years, his presence contributed largely to the NGO capacity building in evaluation. Admittedly, these evaluation exercises were not the only capacity building opportunities, and his work throughout the year with expatriates was as much of an occasion to improve his general knowledge and skills in public health and project management. In the same vein, the three evaluations were particular opportunities for him to learn and use concepts in program evaluation. We use the recommended strategy for adult learners: “learning by doing”². The first evaluation was less participative than the others and more technical, it was also more research oriented. This enabled us to evoke subjects such as the construction of a questionnaire, the constitution of a sample, statistical tests, and concepts like ethics or external validity. This person had also the responsibility for the administration of the questionnaires in villages aided by a team of investigators. This enabled him to become aware of the difficulties on the ground and to assume responsibilities and decisions which could impact on the validity of the evaluation. Since all investigators did not speak English (and we know that ECB is language-dependent²⁸), he had to transmit a certain amount of

knowledge to his colleagues, which certainly contributed to reinforcing it. As an outcome of the ECB process, the medical coordinator was able at the end of 2001 to design and administer a quick survey when a huge number of displaced people reached the Laghman Province during the Taleban departure after the US-Troops attack. He could also contribute largely in the design and the implementation of a drug use survey in 2003 based on the WHO guidelines.

This being said, we should mention that the most significant moment in term of capacity building for him and one other colleague who is no longer with the NGO, was the second evaluation in 2002. The method employed for this evaluation consisted of drawing up a list of indicators through the carrying out of three regional workshops with all project stakeholders. The medical coordinator acted as a translator for the foreign consultant, but the translation of certain concepts required a real understanding of the training contents. How to explain, for example, the difference between output and outcome, or between objectivity and subjectivity. We thus worked together to find useful examples. It was necessary to adapt examples and exercises to the Afghan public, all the more so since the group members had very diverse backgrounds (which we take pride in), with some illiterate members. Having doctors and farmers (or teacher, community health workers) work on the same project is not customary, in Afghanistan or anywhere else! It was therefore necessary to adapt training tools both before and during the workshops in order to take into account the various reactions of the participants to the examples. For instance, it was very useful to illustrate the concepts of the logical model through concrete examples inspired by everyday life, such as the example of seeds (inputs) to obtain apple trees (outputs) then apples (outcomes) used to feed children and reduce malnutrition (impact). To illustrate the concepts of objectivity and subjectivity, we used the example of a judge who had to hear a

case of excessive use of a field by a neighbour who happened to be his brother. Additionally, numerous role-playing sessions, simulation games and practical exercises²⁹ were used to alternate with useful but austere theoretical and conceptual sessions.

This medical coordinator was also part of the third evaluation (2003), but most of his evaluation (and facilitation) skills were developed through collective action, as well as for a large part of, the second evaluation (2002).

Workshops Training During the Evaluation Process

In 2002, three training/action workshops were carried out over three days in Mazar-e-sharif, Gulbahar and Kabul (three regions where AMI is involved) in the presence of 77 people from local communities, the Ministry of Health and AMI (medical and non-medical staff). The aim of those workshops was to make participants aware of the basic concepts of program evaluation and to teach them a logical model to determine what to expect from projects in their local context³⁰. The AMI logical performance model served as a tool for sharing a common vision of projects by identifying the chain of results from input to impact. This method, which aims to create useful and usable indicators of performance through training sessions, appeared somewhat laborious at the time. However, it emphasized the importance of using a participative method. It would have been easier and faster to implement WHO indicators for AMI programs in Afghanistan, but it would have been unnatural and nobody would have actually used this method of performance evaluation. These workshops led to the creation of a list of indicators related to the concerns of local actors. To that list, we added generic indicators usually used on this type of programs and indicators used by AMI. Through the two AMI local experts, a first selection of significant and useful indicators was carried out using

criteria of quality and relevance. This work constitutes an answer to the need of tools to facilitate continuous feedback and periodic production of reporting results.

In terms of the outcome of the ECB process and according to the shortened cascade approach in training², the medical coordinator was able, a few weeks after those three workshops to organize, on his own, the same workshop in another province (Laghman) with 24 participants. He was also in a better position, knowing the logic model approach, to interact with expatriates and contribute to the formulation of new AMI projects and proposals sent to donors. The annual obligatory presentation of NGO program results in the Ministry of Public Health (MoPH) during the National Technical Coordination Committee in front of many stakeholders it was easier to explain the logic of the programmes, performance indicators and the result-based management activities. There were also outcomes for provincial MoPH staff, notably regarding their skills in writing proposals and program planning according to the new health policy (Basic Package of Health Services).

In 2003, the participatory evaluation process started with an evaluation planning workshop held in Kabul. We established an evaluation team composed of six people which was balanced in terms of gender, location and professional status. The purpose of the first workshop was to build consensus around the aim of the evaluation; to refine the scope of work and clarify roles and responsibilities of the evaluation team and facilitator; to review the schedule, logistical arrangements, and agenda; and to train participants in basic data collection and analysis. Assisted by the facilitator, participants identified the evaluation questions they wanted answered. Participants then selected appropriate methods and developed data-gathering instruments and analysis plans needed to answer the questions. Some of the participants already had some knowledge of evaluation and for them this

workshop represented a form of revision. In fact four of them and the medical coordinator were participants in the 2002 workshop in one of the three regions where AMI is involved. During this workshop we assessed whether or not the AMI program was ready for evaluation (evaluability assessment³¹). During the assessment, calls for early evaluation were made, in collaboration with people working on the programs, in order to ascertain whether their objectives are adequately defined and their results verifiable. To do this assessment evaluators used the Logical Framework (LF) Approach³². The evaluation team first reviewed the current LF of the AMI program. For most of the team, it was the first time that they saw the LF with its activities and objectives. After this, it was necessary for the evaluation team to study the LF of the next program financed by the European Union. Indeed, since we had decided to carry out an evaluation of the implementation process of the program, it was necessary to select the relevant fields of activity to be evaluated. In order to use the lessons learnt to improve the program developed in the following months, it was necessary to choose some common activities. Each evaluation group developed a number of evaluation questions for each topic. A maximum of three questions could be answered during the evaluation but each team could start by choosing more than three. Then, the consultant selected the three most important (or feasible) questions and the evaluation team agreed on the choice. Here the role of the consultant, as in other phases of the evaluation process, was both to structure the task for the group and to actively contribute to the development of evaluation questions based on insights from the fieldwork and on their own experience with other programs.

We used different sources of data collected through quantitative as well as qualitative methods. The following methods were used: interview (22), focus group (16), observation (6), document analysis (2), and questionnaire (3). In

addition to the people observed, 205 people (51% of women) had the opportunity to express their views on the implementation of the AMI program in Afghanistan. Once the data was gathered, a participatory approach to analyse and interpret it helped participants to build a common body of knowledge. The consultant allowed the evaluation group to carry out their own analysis but was always present to ensure that the quality of the analysis was of an adequate level. The daily qualitative data analysis process was structured around the interview questions asked of each category of interviewees. A simplified approach to content analysis²⁶ was used by each group.

So, we can say that this whole evaluation process done by an evaluation team from the organization was a perfect approach to develop their evaluation skills in all the evaluation areas, from the evaluability assessment to the data analysis and action plan formulation phase. It is also clear that skills to participate in the whole process were increased, for some, partly due to the capacity building process done over the past two years. Some of them were able in 2004 to use some evaluation techniques (focus group and bed census) during an assessment of the NGO cost-recovery schemes.

Follow-Up of the Baseline Survey in 2004

In addition to those individual and collective training sessions during the last three evaluations, we had another opportunity to develop the evaluation skills of the NGO staff in 2004. During this year, the European Union grant given to AMI covered four clusters of districts spread out among three provinces of Afghanistan. In accordance with the donor, the realization of a baseline survey on the health status of the population in the targeted clusters need to be done at the beginning and at the end of the project by the cluster supervision teams. AMI recruited an

expatriate specifically for this task. She was, not surprisingly, one of the six members of the 2003 participatory evaluation team. This was a good opportunity for her to use some of the knowledge that she had acquired during the previous year. In addition even though she was not part of the 2001 survey using household questionnaires, she was in the hospital, as a physician and not as an evaluator, who serve as an office during the evaluation. For this 2004 baseline survey, a questionnaire was designed and conducted in at least 6 randomly selected villages in each of the districts of the targeted cluster. At the beginning of the project the results of the baseline survey on the health status of the population in the targeted clusters were to be published. These survey results and overall approach need to be readily used to measure the progress at the end of the project, compare the performance of supervisory areas, identify good performers and weak performers and target their resources more effectively.

The expatriate in charge of the survey, asked us to follow the whole process, from a distance in a voluntary and informal capacity. She also solicited our advice and guidance during the evaluation process. As a result many methodological discussions were carried out through e-mail and phone. She decided to adapt the questionnaire that we used in the 2001 evaluation. For some part of the baseline survey, she asked us for some scientific literature (e.g. how to evaluate the quality of health care services) or statistical advice. We also reviewed part of the final report. This 2004 windows was not only an opportunity to develop the staff skills in program evaluation but also to start the building of an infrastructure for data collection, analysis, and presentation that would support program evaluation, in addition to the routine health information system (HIS) which focuses more on input and output than outcome indicators. This infrastructure is now in place and the Afghan collaborators are still in the NGO after the expatriate left. It should be

noted that, even though, the expatriate was involved in the design, coordination and analysis of the survey, she was in the field only in one of the four provinces. Therefore in this three other settings, the process was in the hand of the local staff. The medical coordinator delivered 80% of the training for the surveyors in three provinces and 100% in the other. The ECB of the last three years was surely responsible for this outcome.

Advocacy and Dissemination

The third element which helps us to meet the ECB objective for this NGO consists of myriad activities of advocacy in favour of the program evaluation practice and dissemination of results of various evaluations. As we said earlier, the final aim of those advocacy/dissemination activities are to increase the probability of results utilization per se, following the Patton²⁵ approach.

In terms of advocacy, and in addition to our continual personal interaction in favour of evaluation culture, we produced different papers in order to increase the awareness of the NGO staff regarding different topics in relation to evaluation. These papers, in addition of the evaluation reports, targeted NGO staff directly and more generally the humanitarian community. All these papers carry out a discussion on evaluation in a language that is understood. Some of these papers were published in peer reviewed journals and others in professional reviews or books. The following topics were discussed:

Table 3. Publications in French (F) and English (E) Following the Three Evaluation

Evaluation in	2001	2002	2003
Publication on the results	<ul style="list-style-type: none"> • Book chapter on Canadian humanitarian aid (F) • Poster and proceeding of an international health care financing conference in France (F) 		
Publication on the process or on the general topic	<ul style="list-style-type: none"> • Article in <i>Humanitarian Affairs Review</i> on health financing in a complex emergency context (F, E) • Article in the <i>Canadian Journal of Program Evaluation</i> on usefulness of a participatory evaluation model in an emergency context (F) • Article in <i>The Journal of Afghanistan Studies</i> on the results and on the usefulness of a participatory process to explain changes implemented ,results show 2 years after the evaluation (E) 	<ul style="list-style-type: none"> • Book chapter in the <i>Encyclopedia of Evaluation</i> on participatory determination of performance indicators and utilization-focused evaluation model (E) • Article in the internal newsletter (<i>Tam-Tami</i>) for AMI staff on ethics (F) 	<ul style="list-style-type: none"> • Article in <i>Développement et Santé</i>, on basic concepts in evaluation and the usefulness of a participatory evaluation model (F) • Article in <i>Revue Humanitaire</i> on usefulness of a participatory evaluation model and lesson learned workshop (F) • Article in the AMI newsletter (<i>La Chronique</i>) for donors : advocacy for humanitarian program evaluation (F) • Book chapter in the 25th anniversary book on AMI on the basic concepts in evaluation and the usefulness of a participatory evaluation model (F)

We clearly know that following the different stages of knowledge utilization (from transmission to application), dissemination of results does not mean their

utilization. But, we can also say that these dissemination activities through all these papers published for various members of the public and in different forms could contribute to the installation of an evaluation culture in the organization. Moreover, some articles were specifically written, in their languages, to train the readers and explain to them the logic of evaluation and the importance of practicing it (e.g., ³³). We tried to translate one of these articles in the local language and publish it in the Salamati magazine published by this NGO. But unfortunately, the expatriate in charge on this publication in Afghanistan stated that health workers (the readership) are not prepared to read this kind of material. We are not sure this holds to be true and this story illustrates that ECB “is not “power neutral””⁶ and how an explicit capacity building policy needs to be established in the organization in order to avoid this kind of personal decision which could counter a whole (implicit) process. Fortunately, it seems that the same publication project for medical staff in East-Asia (Saytaman) will translate and use this introductory paper on programme evaluation.

In addition to these publications, during the past four years we conducted various oral presentations to present some evaluation results and to raise the awareness of the NGO staff on the evaluation practice. In Afghanistan, for example, we presented the 2001 evaluation results on health financing for the whole NGO community in Kabul. The presentation was organized in the NGO coordination body office (ACBAR) and around 30 persons represented various NGOs and the Ministry of Public Health. The Afghan medical coordinator took part in it and contributed to the discussions with the participants. Part of the results were used in some preliminary meeting for the development of the National Health Policy, as this was the first survey done regarding this topic in 10 years in Afghanistan. During the same year, the headquarters asked us to train, during one day, all

country projects Head of Mission, about the topic on health care financing. This day was organized in June 2001 in Paris with around 25 people from the field and from the headquarters. In 2002, before starting the criterion-focused evaluation, we spent one day at the NGO headquarters in Paris and organized an oral presentation of the proposal process. This was a window of opportunity to receive feedback and critiques on the proposal and a perfect moment to do some advocacy on evaluation among the staff. In 2003, when the evaluation team presented the results and the recommendations, we started the workshop with a presentation of basic concepts and practices of program evaluation to ensure that the participants had basic notions of evaluation. The same presentation was done in Paris during the monthly board meeting of the NGO where headquarters staff were also present. Most of the people were impressed by the usefulness of the evaluation participatory process and some of them learned some concepts of evaluation.

Last but not least, we took the opportunity of a Canadian bursary program to invite the Afghan medical coordinator, who was present in all evaluations since 2001, to the 2nd International Conference on Local and Regional Health Programmes held in Quebec (Canada) in October 2004. He presented a paper that we co-authored. The topic of this article, then published in the *Journal of Afghanistan Studies*³⁴, was health financing and participatory evaluation. In the paper we tried to demonstrate the relevance of a participative approach in program evaluation and the importance of contextual (local) evidence to make program staff aware of user fees schemes in a complex setting. This conference was an opportunity to share our collaborative experiences on health financing evaluation with colleagues from other countries. In addition, it was an important occasion, even if aid donors are still skinflint, to show that Afghanistan is back in the international public health scientific community, as more than forty countries were present in this conference.

The *Journal of Epidemiology and Community Health* presented this story in its Gallery section in May 2005 (vol. 59). During this meeting, the medical coordinator improved his skills in term of evaluation results dissemination. The presence of this medical doctor in Canada, the first time for him in the “developed” world, was also an empowerment activity and a kind of acknowledgment of his involvement with the NGO for many years, taking into account the turnover problem that NGOs face in the post-conflict settings.

Conclusion

The descriptive elements presented previously clearly show that the implicit step of capacity building was gradual and effective as demonstrated by some of the partial outcomes. Contrary to our definition of ECB which claims that the process need to be intentional, the case shows that a non-intentional process (from the organization point of view) could also have some impact in term of capacity building. The “evaluation capacity building practitioner considers how each study is connected to the development of the organization and to meeting the organization’s goals and mission”⁴. For this reason and to counter the non-intentional process, we (as individual and not as an organization) decided to use all windows of opportunity, or “teachable moments”³, to act in favour of the ECB for the NGO and its staff. One of the recommendations by Gibbs et al³⁵ after their study on 61 NGOs in the USA in terms of ECB was to “take advantage of every available opportunity to use existing evaluation data as a resource for program improvement”. We have tried to implement this recommendation, and more. This strategy was based on three particular components which, in a concomitant way, allowed us to reach this goal, as shown in Figure 1.

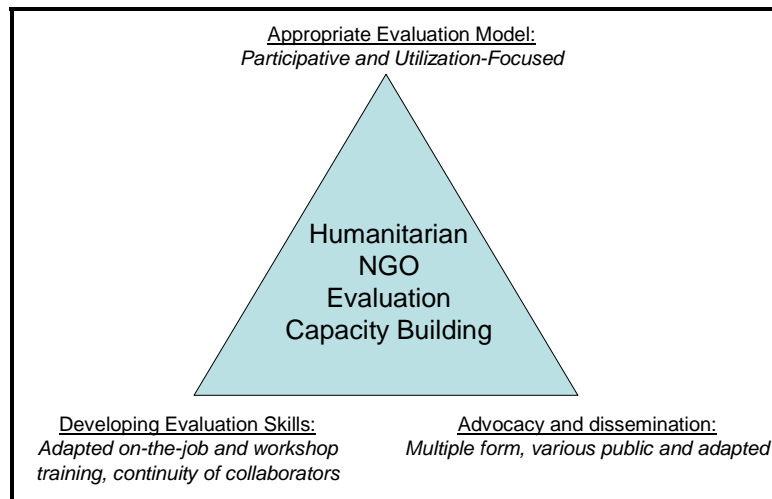


Figure 1. Evaluation Capacity Building Components

The implementation and the quality of the baseline survey planned in 2006 will be a good test for these capacity building activities. For the moment, this paper has highlighted some the ECB outcomes, mostly at the individual level for the Afghanistan staff that we previously mentioned: understanding of evaluation concepts and practices, use of evaluation techniques (logic model, data collection and analysis), ability to facilitate training and disseminate results, etc. But at the organization level, two learning organisation indicators lead us to believe that our approach caused that the actors of this NGO to become more attentive to the importance and the necessity of quality program evaluation. First was a request by the president of the AMI board to produce a chapter devoted to the topic of program evaluation in a book to celebrate its 25 year anniversary intended for general public¹⁹. This testifies the degree of importance granted today to this practice. The second indicator relates to the realization of an evaluation in Thailand another country where this NGO intervenes. The NGO granted a significant amount of money for this evaluation. Then, contrary to the past practice, detailed care was given to the selection procedure of consultants. A detailed term of reference was written and one of the persons in charge (who is based in France but

was, by chance, in Afghanistan during the evaluation lesson-learnt workshop in 2003) asked us for some advice on this matter. Moreover, whereas usually one is satisfied with only the resume of the consultant, it was required that the consultant send some pages of an evaluation plan. Also, the practice that we implicitly employed was intentionally institutionalized, which is a good indicator for continuation and organization learning.

Now, it remains for the NGO to pass from a process of non-intentional ECB program level (Afghanistan) to a process at agency level as a whole. This does not mean that there nothing left to be done at the program level in Afghanistan to improve the current evaluation capacity (“building capacity for evaluation never ends”, Milstein et al³), as there is much that needs to be implemented at the organization level. This Afghanistan case study allows us to draw some lessons in terms of the three ECB components processes. The most significant and useful processes for this purpose can be adapting from some recommendations from the literature^{4,21,35,36}. AMI and other NGOs need to consider:

- Designating organizational (independent) evaluation leader at the headquarters and in the field
- Locating those leaders in the organization hierarchy
- Formulating and adopting an evaluation policy (stated for example the preferred evaluation model, the choice for internal or external evaluation, the way for results dissemination and capacity building, etc)
- Producing internal material
- Developing an evaluation consultants network
- Coordinating evaluation activities around projects countries

- Training expatriate and national staff
- Sustaining leadership

References

1. ALNAP. ALNAP Review of Humanitarian Action in 2003. Field Level Learning. London: Overseas Development Institute, 2004.
2. Morgan CJ, Deutschmann PW. An evolving model for training and education in resource-poor settings: teaching health workers to fish. *Med J Aust* 2003;178(1):21-5.
3. Milstein B, Chapel TJ, Wetterhall SF, Cotton DA. Building capacity for program evaluation at the Centers for Disease Control and Prevention. In: Stockdill SH, Baizerman M, Compton D, eds. *The Art, Craft, and Science of Evaluation Capacity Building. New Directions for Evaluation*, n°93, spring 2002: Wiley Periodicals, Inc., 2002: 27-46.
4. Baizerman M, Compton D, Stockdill SH. Capacity Building. In: Mathison S, ed. *Encyclopedia of Evaluation*. Thousand Oaks: Sage Publication, 2004: 38-39.
5. Sanders RJ. Mainstreaming Evaluation. *New Directions for Evaluation* 2003;99(Fall 2003):3-6.
6. Lusthaus C, Adrien MH, Perstinger M. Capacity development: Definitions, issues, and implications for planning, monitoring, and evaluation. Montreal: *Universalialia Occasional Paper* n°35, 1999: 21.
7. Stockdill SH, Baizerman M, Compton D, eds. *The art, craft and science of evaluation building*: Wiley Periodicals, Inc., 2002.

8. Milstein B, Cotton D. Defining Concepts for the Presidential Strand on Building Evaluation Capacity. American Evaluation Association. Available at www.eval.org/eval2000/public/presstrand.pdf. 2000.
9. Schaumburg-Müller H. Evaluation Capacity Building. Donor Support and Experiences. Report for the DAC Expert Group on Aid Evaluation, OECD. Copenhagen: Danida, 1996: 30.
10. Cousins JB, Goh SC, Clark S, Lee LE. Integrating evaluative inquiry into the organizational culture : a review and synthesis of the knowledge base. *Canadian Journal of Program Evaluation* 2004;19(2):99-141.
11. Marsden P, Samman E. Afghanistan : the economic and social impact of conflict. In: Fitzgerald V, ed. *War and Underdevelopment*. Queen Elisabeth House: University Oxford Press, 2000.
12. UNICEF. Maternal Mortality in Afghanistan : Magnitude, Causes, Risk Factors and Preventability, Summary Findings. Kabul: UNICEF, CDC, MoPH, 2002: 7.
13. van Egmond K, Naeem AJ, Verstraelen H, Bosmans M, Claeys P, Temmerman M. Reproductive health in Afghanistan: results of a knowledge, attitudes and practices survey among Afghan women in Kabul. *Disasters* 2004;28(3):269-82.
14. World Bank. Joint donor mission to Afghanistan on the health, nutrition, and population sector. *Aide-Memoire*, 2002: 19.
15. Waldman R, Hanif H. The public health system in Afghanistan: Current issues. Kabul: Afghanistan Research and Evaluation Unit;, 2002.

16. Ridde, V. (2005). "Performance-based Partnership Agreements for the reconstruction of the health system in Afghanistan." *Development in Practice* 15(1): 4-15.
17. Ministry of Health. A Basic Package of Health Services For Afghanistan. Kabul: Transitional Islamic Government of Afghanistan-Ministry of Health, 2003: 51.
18. Ridde, V. (2002). L'aide humanitaire et la santé de la population afghane sous le régime des Tâlebân. L'action humanitaire du Canada. Histoire, concepts, politiques et pratiques de terrain. Y. Conoir and G. Vera. Québec, Presses de l'Université Laval: 545-566.
19. AMI. Aide Médicale Internationale : 25 ans d'ingérence médicale (provisional title). In press. Paris, 2005.
20. Duignan P. Mainstreaming Evaluation or Building Evaluation Capability ? Three Key Elements. *New Directions for Evaluation* 2003;99(Fall 2003):7-21.
21. Bozzo SL. Evaluation capacity building in the voluntary/nonprofit sector. *Canadian Journal of Program Evaluation* 2002;17(3):75-92.
22. Ridde, V. (2003). "L'expérience d'une démarche pluraliste dans un pays en guerre : l'Afghanistan." *Canadian Journal of Program Evaluation* 18(1): 25-48.
23. Cracknell BE. *Evaluating Development Aid. Issues, Problems and Solutions.* New Delhi.Thousand Oaks.London: Sage Publications, 1996.

24. Cousins JB, Whitmore E. Framing Participatory Evaluation. In: Whitemore E, ed. *Understanding and Practicing Participatory Evaluation*: Jossey-Bass Publishers, 1998: 5-23.
25. Patton MQ. *Utilization-Focused Evaluation*. 3rd ed. Thousand Oaks-London-New Delhi: Sage Publications, 1997.
26. Aubel J. *Participatory Program Evaluation Manual. Involving Program Stakeholders in the Evaluation Process*. Calverton, Maryland: Child Survival Technical Support Project and Catholic Relief Services, 1999: 86.
27. Ridde, V. (2004). "L'évaluation de programme en santé internationale : qu'est-ce que c'est, comment la planifier et utiliser une approche participative ?" *Développement et Santé* 169: 23-29.
28. Toulemonde J, Bjornkilde T. *Building Evaluation Capacity : Experience and Lessons in Member States and Acceding Countries*. Budapest: Fifth European Conference on Evaluation of the Structural Funds. 26-27 june 2003, 2003: 13.
29. Patton MQ. *Creative evaluation*. 2nd ed. Newbury Park, Beverly Hills, London, New Delhi: Sage Publications, 1987.
30. Ridde, V. (2004). Seeds against malnutrition in Afghanistan: an experience in participative performance evaluation training. *Encyclopedia of Evaluation*. S. Mathison. Thousand Oaks, Sage Publication: 433-434.
31. Thurston WE, Potvin L. Evaluability Assessment: A Tool for Incorporating Evaluation in Social Change Programmes. *Evaluation* 2003;9(4):453–469.

32. Sartorius RH. The logical framework approach to project design and management. *Evaluation practice* 1991;12(2):139-147.
33. Ridde , V. (2004). "L'utilité d'un processus participatif lors des évaluations de programmes humanitaires." *Revue Humanitaire* 11: 59-65.
34. Ridde, V., P. Bonhoure, et al. (2004). "User fees and hospital health care utilization in Afghanistan : lessons learned from a participative evaluation." *Journal of Afghanistan Studies* 2: 93-109.
35. Gibbs D, Napp D, Jolly D, Westover B, Uhl G. Increasing evaluation capacity within community-based HIV prevention programs. *Evaluation and Program Planning* 2002;25(3):261-269.
36. Boyle R. Building effective evaluation capacity : some lessons from international practice. Presentation to IDEA Seminar, Rome, october 24th, 2003.

About the Authors

Valéry Ridde is a postdoctoral researcher at Université de Montréal (GRIS/USI). He works and studies in global health, community health and evaluation. His PhD in community health thesis (Université Laval, Canada) is on health policy and equity in West-Africa (Burkina Faso). His research interests focuses on global health, health policy, program evaluation, equity, health care organization and financing. He is an Editorial Board member of the *Canadian Journal of Program Evaluation*. He has already published several papers in books, academic and professional journals. His teaching as a lecturer in the academic world and as a private consultant with communities focuses on global health and program evaluation/planning in community health.

Shakir Sahibullah is a medical doctor living in Afghanistan. He was during the last years the general medical coordinator for Aide Médicale Internationale – France in Afghanistan. He is now a Ministry of Public Health staff, partly as an advisor in health financing topic regarding a pilot project implemented by John Hopkins University. He is a coauthor with Valéry Ridde of various papers and was in 2004 a recipient of a bursary to present a paper and assist at the 2nd International Conference on Local and Regional Health Programmes in November in Quebec (Canada).

A special thanks to all AMI staff in Paris and Afghanistan, Lara Arjan for her help in the English translation and Ian Christoplos for comments about an earlier version of this paper. This case-study report was commissioned by ODI (UK) for the ALNAP Review of Humanitarian Action in 2004 (see <http://www.alnap.org>).

Ideas to Consider

Ethnography and Evaluation: Their Relationship and Three Anthropological Models of Evaluation

Brandon W. Youker

Abstract

This paper examines the relationship between ethnographic research methods and evaluation theory and methodology. It is divided into two main sections: (a) ethnography in evaluation and (b) anthropological models of evaluation. Three levels of the leading anthropological models of evaluation are summarized, which include responsive evaluation, goal-free evaluation, and constructivist evaluation. In conclusion, (a) there is no consensual definition of ethnography; (b) in many circumstances, ethnographic evaluation models may be beneficial; and (c) ethnography can be used in evaluation but requires a high level of analysis to transform ethnographic data into useful information for eliciting an evaluative conclusion.

**The author would like to thank Daniela C. Schröter, Chris L. S. Coryn, and Elizabeth K. Caldwell for editing this paper and for their extremely useful comments and suggestions.*

Introduction

Ethnography²⁵, an applied qualitative social science research method, is often employed in program evaluation. Ethnography, alone and according to its pure anthropological definition, is not a research method capable of being the sole method implemented in an evaluation. Ethnography may prove advantageous to evaluators as an additional method to be employed or considered. However, sound evaluation typically requires multiple data collection methods and a higher level of analysis than ethnography alone can provide. Evaluation synthesizes the narrative and develops an evaluative conclusion. There are various instances when the implementation of an evaluation model that relies heavily on qualitative methods based in the tradition of anthropological research is beneficial. As an evaluator, at minimum, familiarity with these models should be in one's repertoire.

The paper is divided into two main sections: (1) Ethnography and Evaluation and (2) Anthropological Models of Evaluation. The first section presents a summary definition of ethnography, its theories, concepts, and benefits; and the difference between ethnography and anthropology. The author then provides a brief definition of evaluation and discusses the relationship between ethnography and evaluation. There are three anthropological models of evaluation in which the author summarizes, discusses the strengths and limitations, and reflects on their

²⁵ AUTHOR'S NOTE: The author of this paper uses the terms “ethnography,” “ethnographic techniques,” and often “qualitative research methods” interchangeably. Additionally, the term “program” is used generically, to refer to the *evaluand**. Ethnography in the context of this paper is primarily in regards to program and policy evaluations. Ethnography may also be used in product, personnel, and performance evaluations.—* “*Evaluand*: That which is being evaluated (e.g., program, policy, project, product, service, organization)” (Davidson, 2005, p. 240).

relationship with ethnography. The paper concludes with a synopsis of the author's main impressions and key points.

Ethnography and Evaluation

*Ethnography*²⁶ is an applied research method most often associated with anthropology, where it was developed to study cultural interpretation. Ethnography, also called field research, is the process of describing a culture or way of life from a folk peoples' perspective. Anthropologist Clifford Geertz described the ethnographic method as "thick description." It provides detailed notes and descriptions of everything that occurs without attempting to summarize, generalize, or hypothesize. In fact, with traditional ethnography, as a rule of thumb, for every half hour of observation a researcher writes for two hours. The researcher focuses on factual description to allow for multiple interpretations to later infer cultural meaning. To obtain this description of a population's perception, the principle of 'naturalism'²⁷ is assumed. Thus, trust and rapport are essential between the researcher and the population being studied.

Ethnographers, if following the constructivist²⁸ philosophy, believe that pure

²⁶ Alternative definitions: Ethnography is "a descriptive study of an intact cultural or social group or an individual or individuals within the group based primarily on participant observation and open-ended interviews. Ethnography is based on learning from people as opposed to studying people" (Beebe, n.d.). Ethnographic research "involves the study of groups and people as they go about their everyday lives" (Emerson, Fretz, & Shaw, 1995). "Ethnography is the art and science of describing a group or culture" (Fetterman, 1989, p. 11).

²⁷ Naturalism: Leave natural phenomenon alone.

²⁸ Constructivist philosophy maintains that the researcher manufactures knowledge through her interaction in the field and that there is no objective truth to be uncovered (ontological

objectivity is impossible as: (A) ethnography is an interpretive endeavor by fallible human beings; (B) not all field sites are “foreign” for ethnographers in the same way; (C) ethnography is not replicable; and (D) ethnography is not based on a large number of cases. The epistemology of ethnography is typically a model based on a phenomenologically oriented paradigm, which focuses on multiple perspectives and multiple realities of a phenomenon. Phenomenological inquiry seeks to answer the question: “What is the structure and essence of experience of this phenomenon for these people?” (Patton, 1990, p. 69). Constructivists take a heuristic²⁹ approach to answering this phenomenological question. According to Fetterman (1989), most ethnographers subscribe to ideational theory, which suggests that change is the result of mental activity—thoughts or ideas—versus materialists who believe that “material conditions—ecological resources, money modes of production—are the prime movers” (Fetterman, 1989, p. 16). The most popular ideational theory is cognitive theory, which assumes we can infer peoples’ thoughts from hearing what they tell us.

While many theories, concepts, and methods (e.g., in-depth, open-ended interviews, direct observation, written documents, triangulation) resulting in narrative description commonly recur in the literature, consensus on any one set of fundamental principles of ethnography cannot be found (Genzok, 2001; Patton, 1990; Payne, 1994). For example, ethnographic theories, concepts, and data collection techniques are also used in non-ethnographic qualitative research and

relativism) (Maxwell, 1998 in Bickman & Rog, 1998).

²⁹ Heuristics is a form of phenomenological inquiry focusing on the personal experiences and insights of the researcher—it considers researcher’s experience in addition to other observers that experience the phenomenon.

distinctions between ethnography and other qualitative theories, concepts, principles, and methods is not clearly evident. Instead, there are copious combinations of varying concepts considered fundamental to ethnography from researchers and anthropologists alike.

The key in understanding the differences between ethnography and other qualitative social science research methods is to understand the multiple combinations of techniques, concepts, and data collection methodologies encompassed under the term “ethnography.” As with all research methodologies, each philosophical and theoretical decision is located on a spectrum or continuum. Thus, the definition of ethnography and what it entails is idiosyncratic to the ethnographer or researcher depending on her degree of commitment to a hodgepodge of “fundamental” concepts. Past and current literature presents definitions and concepts of ethnography differing by technique, values emphasized, time allotted, data analysis procedures, and commitment to the purist practice of anthropological ethnography. Therefore, ethnographic techniques are qualitative in nature but distinct. Below are a few of the reoccurring concepts specific to ethnography (Fetterman in Bickman and Rog, 1998; Genzuk , 2001; Hall, n.d.):

- ✓ The focus is on culture and cultural interpretation.
- ✓ There is an emphasis on an emic³⁰ perspective.
- ✓ The holistic perspective is often of greater depth than other qualitative research methods.
- ✓ Sampling measures are conducted over a longer period of time.

³⁰ Emic perspective is that of the insider and includes the acceptance of multiple realities.

- ✓ The researcher herself is the primary tool for data collection.

In contrast to ethnographic methodology, *evaluation* methodology commonly comprises the use of both qualitative and quantitative techniques. Commonly defined as the systematic determination of something's merit, worth, or significance (Davidson, 2005; Sanders, 1994; Scriven, 1991). Scriven (1991) claims that evaluation is not only a methodology, but a distinct multi- and transdisciplinary field of study not to be regarded as merely a branch of applied social science. As an independent discipline, evaluation may utilize applied social science research methodology, but it is distinct by its unique purpose and methodology (e.g., ranking, grading, and scoring). The determination of merit, worth, and significance of an evaluand requires evaluators to consider the relevant values and to make judgments based on those values. Autonomous subspheres of evaluation are program, policy, product, personnel, performance, and proposal evaluations as well as metaevaluation (i.e., the evaluation of evaluation) and intradisciplinary evaluation. Often but not always, evaluations are based in social science research methods including both qualitative and quantitative data collection procedures. Subcomponents of program evaluations, for example, may include the assessment of context, resources, processes, immediate outcomes (outputs), intermediate and long-term outcomes and impacts considering costs, comparisons to best and worst practices of other programs (Davidson, 2005; Scriven, 1991). Moreover, evaluation may be formative, summative, or ascriptive (Scriven, 2005).

Multiple factors may guide evaluators and researchers alike toward choosing quantitative or qualitative evaluation methodology. In the following, qualitative ethnographic evaluation models will be introduced.

Ethnographic Evaluation Models

Ethnographic evaluation methodologies have been discussed for over thirty-five years and came about as a response to the more traditional evaluation approaches which were overly committed to the scientific paradigm of inquiry. According to Guba and Lincoln (1989), an extreme dependence on the methods of science demonstrated some negative results. For example, reliant on primarily quantitative measurement, evaluands were stripped of their context as if they were not entwined in a highly specific one, resulting in irrelevant or non-useful findings (cf. Seafield Research & Development Services). Moreover, scientific truth is non-negotiable, thus all alternative explanations must be in error.

Ethnographic evaluation methods, in contrast, utilize stakeholders' claims and concerns. For example, Guba and Lincoln (1989) insist upon ethnographic methods for determining what information is necessary in an evaluation and provide five reasons:

1. Stakeholders are placed at risk by an evaluation.
2. Evaluation exposes stakeholders to exploitation, disempowerment, and disenfranchisement.
3. Stakeholders represent an "untapped market" for the use of evaluations that are responsive to self-defined needs and interests.
4. Stakeholder input expands the scope and meaningfulness of the evaluation, in addition to contributing to the dialectic process that is necessary in conducting a sound evaluation.
5. All individuals and parties can be mutually educated toward more sophisticated personal constructions and they may gain enhanced

appreciation of the constructs of other individuals or parties.

Other strengths of implementing ethnographic methods in evaluation are exemplified in connecting quantitative data to observed actual outcomes; the flexibility of design; the ‘thick description’ of program impactees; the clarification of processes; the study of participation; and the identification of unintended positive and negative side effects. The weaknesses in utilizing ethnographic methods in evaluation consist of such problems as introducing complex threats to validity; increasing the time and cost demands compared to other methods; raising the potential for impactee reactivity to the evaluator; and limiting the ability to compare the data from different measurement instruments.

There are many considerations that will need resolution before deciding if an ethnographic method is an appropriate method for an evaluation. Considerations include the purpose of the evaluation; whether the evaluation is formative or summative; the amount of time allocated for the evaluation; the financial and other resources available; and the level of expertise and competence of the evaluation team. Prior to adopting a specific methodology or model, all the typical issues regarding methodology, conceptual context, validity, ethics, etc. must be discussed.

Relationship Between Ethnography and Evaluation

In evaluation, ethnography should be viewed on a spectrum. One end of the spectrum consists of the pure anthropologically-defined ethnography and on the opposite end are various ethnographic techniques of data collection and methodologies loosely defined, combined, and flexibly implemented. Many researchers and evaluators implement one or a few qualitative data collection methods and then claim their research to be ethnographic. However, most agree that ethnography is defined by the rigor of the data collection procedures.

Fetterman (1982) identified a study that called itself ethnographic although the researchers were on site for only five days. Deneberg (1969) and Fetterman (1982) claim that these researchers are fickle to scholastic fads and refer to them as “Zeitgeister-Shysters.” Zeitgeister-Shysters become involved in research that is a hot topic or trendy and the result is superficial research. Such researchers contribute minimally to the field and often tarnish the reputation and credibility of ethnography. In describing the Zeitgeister-Shysters, Fetterman stated, “rather than conducting ethnographies, they are simply using ethnographic techniques” (Fetterman, 1982, p.2). Wolcott (1980) concluded that “much of what goes on today as educational ethnography is either out and out program evaluation, or, at best, lopsided and undisciplined documentation” (p.39). Fetterman warns that the adoption of random elements of ethnography without emphasis on the whole, results in “the loss of the built-in safeguards of reliability and validity in data collection and analysis” (Fetterman, 1982, p.2). Researchers often use anthropological tools (ethnography) without understanding the values and cosmology underlying the ethnographic techniques. Wolcott (1980) reminds the reader that the purpose of ethnography is cultural interpretation and this requires the researcher to examine the whole trait complex rather than a few single traits. Still many evaluators study single traits and call their evaluation ‘ethnographic’.

The importance of ethnographic data sources in the evaluation of social programs and policies is rarely argued (Agar, 2000; Fetterman, 1982; Fetterman, 1984; Guba & Lincoln, 1989; Hopson, 2002; Patton, 1997; Posavac & Carey, 1997; Scriven, 1991; Swartzman, 1983; Shadish, Cook, and Leviton, 1995; Stake, 1975; Wholey, Hatry, & Newcomer, 2004; Wolcott, 1982; and Worthen, Sanders, & Fitzpatrick, 1997). Hopson (2002), for example, cites a report by Nastasi and Berg (1999) who urge evaluators to “capture views of program participants about their experience of

a program, its acceptability, and whether or not they were influenced to modify behavior or thinking” (p. 45). This has always been a consideration for evaluators, as it pertains to, or affects the program's quality, significance, or merit. Experienced evaluators typically employ several qualitative data collection methods in an evaluation in hopes of understanding some of these cultural issues, albeit less in depth than with pure ethnography.

Focusing on context is crucial in all evaluations and the utilization of qualitative methods is fundamental to any good program or educational evaluation; however, the title 'ethnographic evaluator' may be problematic or misleading. Many readers may assume that the term “ethnographic evaluator” implies the use of ethnography in conducting an evaluation. This is false and arguably not possible. Ethnography is a social science research method that emphasizes cultural interpretation. The product of ethnography is a non-judgmental description of context and then a cultural interpretation of the program.

Evaluation is the systematic process of determining the merit, worth, significance or importance of the evaluand. To evaluate something, the relevant values are determined and used to place judgments regarding the overall quality of the program. Ethnography and other qualitative research methods are instrumental in collecting data for determining the most important values to use as criterion for success. Ethnography may uncover unanticipated costs, processes, and outcomes; however, other qualitative methods may reveal similar side-effects but take much less time. There is a point of saturation when a researcher gets the sense that it is unlikely that further study will uncover significant new information that will be important to include in the evaluation. Extended time in the field may not be necessary or feasible for many evaluations.

To summarize, ethnography is a research method and evaluation uses multiple research methods to collect information for determining the merit or worth of a program. As Fetterman (1984) points out, the distinction between ethnography and evaluation is regarding the level of analysis and objective. Evaluators take the ethnographic data to a higher level of analysis by extracting data which is relevant to some standard; comparing it with data from other methods and sources; and judging the program accordingly. Therefore, I conclude that “ethnographic evaluation” is a misnomer or false label for what some evaluators do. Moreover, evaluations claim to use ethnographic methods while in reality, they simply employ varying degrees of qualitative methods. Anthropologically, pure ethnography may serve useful when analyzed further by an evaluator to examine actual processes and outcomes. Anthropological evaluation techniques may be best when conducted independently of the more quantitative research methods, similar to Scriven's (1991) goal-free evaluation. Therefore, in an evaluation which uses multiple research methods, ethnography serves as a way of triangulating these methods. Furthermore, ethnographic data is useful in triangulating data sources adhering to the *principle of critical multiplism* (c.f. see Shadish, 1994). An examination of three evaluation models which are based in anthropology will further illustrate the relationship between ethnography and evaluation.

Anthropological Models of Evaluation

Payne (1994) categorizes 4 evaluation models, the fourth of which contains anthropological approaches (see Figure 1).

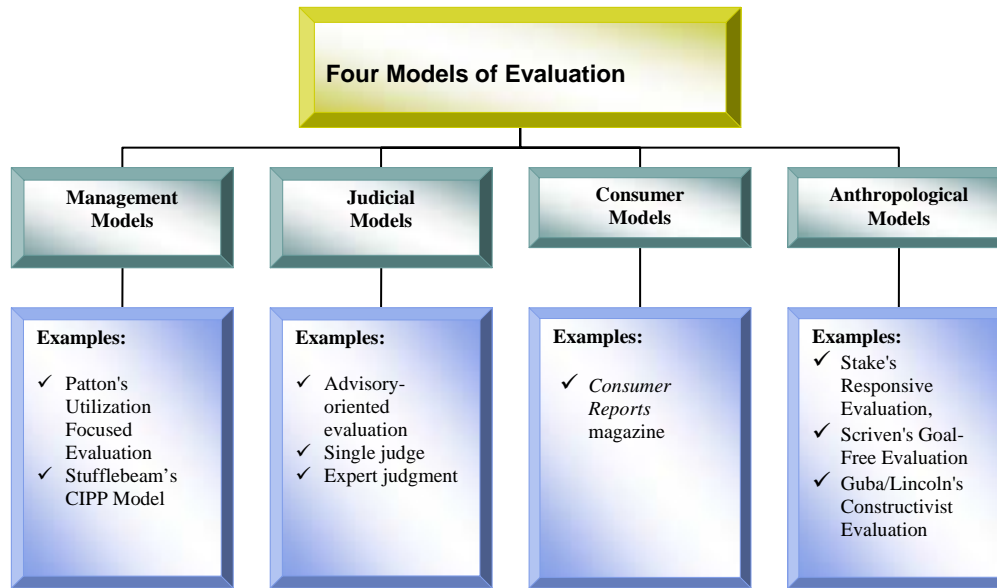


Figure 1. Models of Evaluation (adapted from Payne, 1994)

The anthropological models of evaluation—responsive evaluation, goal-free evaluation, and constructivist evaluation—have many similarities. They tend to be qualitative, exploratory, highly descriptive, and take an inductive approach to understanding the program under evaluation. Each model was created in the post-positivist value-pluralist perspective, focusing on the question: *whose values and methods should shape or have shaped the evaluation?*

The anthropological models protect against any of the evaluator's personal opinions from being used to determine the values and methods emphasized in the evaluation. However, Scriven separates goal-free evaluation from the other two anthropological models by contending that the stated goals of the client should also not be known or utilized by the evaluator. The three models re-examine the ontology³¹ of evaluative interpretations. In both responsive evaluation and constructivist evaluation, the selection of relevant values and the determination of

³¹ Ontology: The nature of the real.

the merit of outcome measures are decided by the program impactees and stakeholders. Evaluators are partners with the stakeholders in the creation of data and they orchestrate the consensus building process. By contrast, in goal-free evaluation, program success is decided by examining change relative to the identified needs through a comprehensive needs assessment. Lastly, all three models rely on an evaluator with significant commitment to and experience with ethnographic and qualitative methods.

The remainder of the paper will discuss each anthropological evaluation model and illustrate its relationship to ethnography and the qualitative research paradigm of evaluation.

Responsive Evaluation

Stake (1975) called his approach to evaluation responsive evaluation to stress flexibility and responsiveness to the concerns and issues of the program stakeholders. Responsive evaluation is less reliant on formal communication such as the statement of goals, objective tests, standards of program personnel, and research-type reports. Rather, it focuses on gathering the observations and reactions of the program stakeholders, which as Stake claims, is the way people naturally evaluate things. Stake believes this and other qualitative methods are not frequently employed in evaluation due to “subjectivity.” Responsive evaluation is poorly suited for evaluating formal contracts, and there lies potential to uncover negative side effects or raise embarrassing questions.

Stake suggests examining a program by organizing the evaluation into four components: environment, workspace, output, and support (see Figure 2).

<p>Environment</p> <ul style="list-style-type: none">• Quantity (investigate for quantity including the counting of frequencies, occurrences, products, performances, participants, resources, etc.).• Diversity (diversity in artistic products, performances and participants).• Excellence (refers to technique or quality of execution/performance; has a varying threshold of acceptability).• Originality (separate from quantity and diversity; referring more to creativity and inventiveness; the ability to make someone “catch their breath”; best measured by degree on a variable range).• Vitality (changeability of physical environment measured over time; encourages regular review of the physical conditions and aesthetics of environment). <p>Workspace</p> <ul style="list-style-type: none">• Space and content - suitability and accessibility• Quantity and quality of equipment and supplies <p>Output</p> <ul style="list-style-type: none">• Measure outputs with careful consideration of the threshold of acceptability• Incorporate experts in the field <p>Support</p> <ul style="list-style-type: none">• Within the program and from the community, the school or organization as a whole• Investigates how outputs are regarded and rewarded

Figure 2. Four Components of Evaluation: Environment, Workspace, Output, and Support (adapted from Stake, 1975)

Exemplifying educational evaluation, Stake states,

[A]n educational evaluation is a *responsive evaluation* if it orients more directly to audience requirements for information; and if the different value-perspectives present are referred to in reporting the success and failure of the program.

(Stake, 1975, p. 14)

It is not critical to be explicit about purpose, scope, or causation in determining worth, according to Stake. In conducting responsive evaluation, the evaluator observes the program to gather narrative and descriptive information from program stakeholders; and negotiates values in which to judge the program. An evaluator should not presume that only the measurable outcomes provide evidence of the program's worth. Outcome evaluations tend to negate the idiosyncratic and unique

ways people benefit from involvement with the program and among each other; furthermore, they are not sensitive to changes in program purpose. Stake cites Scriven (1967) and suggests that it may be preferable to evaluate the “intrinsic merit of the experience rather than the more elusive payoff” (p. 27). Stake feels that less emphasis on preconceived notions of success will allow for increased stakeholder flexibility in determining the purposes of the evaluation and criteria by which to measure success. In a responsive evaluation, the evaluator has the ability to respond to emerging issues, rather than sticking to a strict evaluation plan or structure. This ultimately leads to an increase in the evaluation's utility to the program stakeholders. Recurring events in responsive evaluation (Stake 1975):

1. Talking with clients, program staff, and audiences.
2. Identifying program scope.
3. Providing an overview of program activities.
4. Discovering purposes, concerns.
5. Conceptualizing issues, problems.
6. Identifying data needs regarding issues.
7. Selecting observers, judges, and instruments if any.
8. Observing designated antecedents, transactions, and outcomes.
9. Providing a theme; preparing portrayals, case studies.
10. Winnowing, match issues to audiences.
11. Formatting for audience use.
12. Assembling formal reports, if any.

Data is collected through direct personal experience or the second best option,

vicarious experience. Observations are not only conducted by the evaluator, but the evaluator enlists program stakeholders according to the issues being studied and the audience being served. Having multiple observations and observers increases data reliability; observations continue to be subjective but through replication random error is reduced. The bias of direct or vicarious experience decreases as repeated observation and diverse points of view are attained. The evaluator produces portrayals typically featuring descriptions of persons, such as a five-minute script, a log, scrapbook, multi-media or audience role-plays. The small number of case studies is often criticized for sampling error, but Stake attests that the error may be minimal and that it is a small price to pay for potentially substantial improvements in communication. Moreover, Stake assumes that case studies of several students are more interesting and representative of a program than a few measurements on all program participants. Therefore, the reader benefits by a more comprehensive understanding of the program.

The evaluation encounters two pluralisms of values: (1) in context, or in Stake's terms the "antecedent condition in which the program is found" (p. 23) and (2) the personal outcomes or outcomes of the program. The evaluation team should not impose its values on the "actors," "spectators," and/or "critics" of the program during the consensus building process. Stake identifies two measures of the value of evaluation: its increment of added experience and its enhancement of responsive alternatives.

Strengths of responsive evaluation include it being flexible, adaptable, and good in providing cultural explanation and recognition of diversity. It may be particularly useful in evaluating programs where the stakeholders generally agree on the intrinsic value rather than the instrumental value of the program. For example, many people will discuss the importance of music and art "because they're good

things to do” (ibid, p. 16). Furthermore, in formative evaluations, responsive evaluation is useful in monitoring the program and to identify positive and/or negative side effects. It is helpful in summative evaluations by giving the reader an understanding of the program's activities, its strengths and weaknesses, and by providing a vicarious experience in the evaluation.

Limitations of responsive evaluation include the difficulty in making comparisons to standards; it serves the immediate audience and may not fulfill distant or future needs. In today's world, funding constraints on arts education programs, for example, has led to an increased demand for quantifiable outcomes and results which are not emphasized in responsive evaluation. Moreover, responsive evaluations may be less objective, reliable, and generalizable as compared to traditional evaluations, or as Stake calls them preordinate evaluations. Responsive evaluation is not useful when it is important to measure goal attainment, whether promises were kept, or in cases where predetermined hypotheses are to be examined.

Ethnography, in the more traditional sense, has compatibility problems with responsive evaluation but there may be potential for combining them. A primary distinction is that with responsive evaluation, the evaluator solicits the observation of stakeholders, thus making the stakeholders part of the evaluation team and adding them as additional data collection instruments. Nevertheless, with some concessions on both sides, the two may be combined.

Goal-Free Evaluation

The evaluator, in a goal-free evaluation (Scriven, 1991), intentionally enters the field without being aware of the specific stated goals and objectives of the program. The evaluator learns about the program and its outcomes inductively.

This means that all program materials are screened either by a non goal-free evaluator on the evaluation team, an administrative assistant, or by the client to ensure that none of the stated goals or objectives are described to the goal-free evaluator. The purpose of this is:

...finding out what the program is actually *doing* without being cued as to what it is *trying* to do. If the program is achieving its stated goals and objectives, then these achievements should show up; if not, it is argued, they are irrelevant.

(Scriven, 1991, p. 180)

Goal-free evaluations can assist in determining whether the objectives are worthwhile; instead of “are the objectives being achieved?” It is similar to the double-blind pharmaceutical study; just like the drug evaluator, the goal-free evaluator does not have to know the direction of the intended effect or the intended extent of the outcomes (Scriven, 1973). The evaluator intends to find the program’s actual outcomes and then works backward to determine if the effects were caused by the program. The goal-free evaluator is like the crime scene investigator who tries to eliminate rival explanations which may have led to the outcome under investigation. Information regarding the stated goals of the program is withheld from the evaluator. However the evaluator is able to review some program documents, budgets, schedules, recorded observations, profiles of participants and staff, etc. as long as there is no implication of any stated goal.

A comprehensive, fair, and accurate needs assessment is essential in conducting a goal-free evaluation. Merit is determined by comparing the actual program outcomes to the relevant needs of those impacted, instead of to the program goals or consumer wants or desires. The program is evaluated according to the level of fulfillment of the consumers needs. Scriven believes by keeping the goals vague, a

less pure goal-free evaluation still makes finding outcomes difficult and encourages the evaluator to connect program effects to recipients' needs instead of the stated goals of the program. Altschuld and Witkin (2000) state that the needs at the primary level (i.e., recipients of the program) are the most critical concern, and from there the needs assessment can consider the needs of the service deliverers and the program delivery system. They argue that the primary needs are the "raison d'être" or the "rationale for the existence" of the service deliverers and delivery systems (Altschuld & Witkin, 2000, p. 10).

There are also relative degrees to which an evaluation may be goal-free. Goal-free evaluations may be combined, in full or in part with other evaluation methods (e.g., "qualitative versus quantitative, survey versus experiment, multiple perspectives versus one right answer, etc.", Scriven, 1991, p. 182). Additionally, an evaluation may begin goal-free and then become goal-based; the reverse is not possible. It is also suggested that goal-free evaluation can be used as a supplement to a traditional outcomes evaluation conducted by a separate evaluator. The evaluator implementing the goal-free evaluation collects exploratory data to supplement and provide context to another evaluator's goal-oriented data. Goal-free evaluators observe the program in an attempt to understand the culture, meanwhile considering needs, processes, and outcomes. Below, the author provides a simplified illustration of a goal-free evaluation using a physical education and training program.

The evaluator of a physical education and training program enters into the evaluation without any prior knowledge of the program's goals. She would likely be capable of directly observing changes in health-related knowledge, strength, and endurance, which are the program's stated goals. However, the goal-free evaluator might also discover changes in endurance, flexibility, physique, changes in

behavior, social status, networking with other students, finding new supportive workout partners, sharing of dietary and nutrition tips, increased self-esteem, etc. all of which were not original goals of the program and would be considered positive, unintended side-effects. They would likely have been missed if the evaluation solely examined the stated or preordained goals.

Arguments for the utilization of goal-free evaluation include (Scriven, 1991):

- It may identify unintended positive and negative side-effects and other context specific information.
- As a supplement to a traditional evaluation, it serves as a form of triangulating both data collection methods and data sources.
- It circumvents the traditional outcome evaluation and the difficulty of identifying true current goals and true original goals, and then defining and weighing them.
- It is less intrusive to the program and potentially less costly to the client.
- It is adaptable to changes in needs or goals.
- By reducing interaction with program staff, it is less susceptible to social, perceptual, and cognitive biases.
- It is reversible; an evaluation may begin goal-free and later become goal-based using the goal-free data for preliminary investigative purposes.
- It is less subject to bias introduced by intentionally or unintentionally trying to satisfy the client because it is not explicit in what the client is attempting to do; it offers fewer opportunities for evaluator bias or corruption because the evaluator is unable to clearly determine ways of cheating.

- For the evaluator, it requires increased effort, identifies incompetence, and enhances the balance of power among the evaluator, the evaluatee and client.

Scriven and other users of goal-free evaluations have provided minimal direction regarding operational methodology in conducting the model. The only known attempt to develop an operational methodology for goal-free evaluation was by Evers (1980) in a doctoral dissertation. Evers outlined a goal-free evaluation methodology consisting of six components each of which comprising several sub-categories. The six main components were: (1) Conceptualization of Evaluation; (2) Socio-Political Factors; (3) Contractual/Legal Arrangements; (4) The Technical Design; (5) Management Plan; and (6) Moral/Ethical/Utility Questions. The success of a goal-free evaluation is dependent upon the quality of the needs assessment. If there is not an accurate comprehension of the program participants' needs then the entire evaluation may be at jeopardy.

A goal-free evaluation could feasibly be ethnographic. However, goal-free evaluation focuses on using observation to connect needs to actual program activities, rather than for thick description. Furthermore, traditional ethnography focuses on culture which is always goal-free in nature.

Constructivist Evaluation

Guba and Lincoln's (1989)³² fourth-generation or constructivist evaluation

³² The new meaning of constructivist methodology: Truth is determined by consensus building among informed constructors, not of correspondence with an objective reality. Facts are meaningless without a value framework; therefore, no proposition can be objectively assessed. Causes and effects do not exist; accountability is relative and implicates all interacting parties equally (Guba & Lincoln, 1989).

approach outlines five generations of evaluation: (1) measurement (e.g., IQ testing); (2) description (e.g. formative evaluation of programs); (3) judgment of merit and worth; (4) *constructivist* (negotiated co-creations of social reality); and (5) meta-evaluation (the evaluation of an evaluation). The constructivist evaluation was created in response to the perceived failure or critical flaws of the first three generations of evaluation. The fourth-generation evaluator may use any of the earlier evaluation techniques as appropriate. Carney (1991, p. 35) reports that the underlying method in fourth-generation evaluation is known by other names:

British scholars call it 'human inquiry' (inquiry conducted in human ways for humane ends); American scholars call it 'action research' (research which aims to produce action on or through its findings, and third world or developmental evaluators call it 'developmental evaluation' (evaluation which develops the understanding, and resources to respond, of those evaluated). A common generic term for it is 'collaborative inquiry' (which simply describes what goes on when you use the method).

In constructivist evaluation, evaluation is:

- a. A process that combines data collection and data valuing (interpretation) into one inseparable.
- b. A local process.
- c. A sociopolitical process.
- d. A teaching and learning process.
- e. A continuous, recursive and divergent process.
- f. An emergent process.
- g. A process for sharing accountability.

h. A hermeneutic dialectic relationship.

In collaborative inquiry the people being evaluated participate as informed collaborators rather than research subjects. The purpose of a constructive evaluation is to attain a deeper comprehension of all the issues encountered by all the stakeholders and consumers; while the goals comprise mutual education, improved awareness, and increased motivation to utilize the evaluation results. Most constructivist evaluators are relativist and implement qualitative methods; however, the evaluation is conducted in a disciplined manner and it produces an audit trail to ensure transparency and credibility of its findings. The realities discovered by the constructivist inquiry are the constructions of the reality proposed by the evaluatees themselves. They develop into co-constructions and subsequently reconstructions, as both evaluators and evaluatees mold them. The constructivist evaluation assumes that evaluators are unable to maintain distance from the evaluatees. Therefore, it accepts a hermeneutic dialect. Guba and Lincoln continue by rejecting the positivist assumptions, which they claim are embedded in most evaluation methodology. They use “validity” as an example of a term that evaluators are socialized into accepting as the positivist definition. Furthermore, they feel that the relationship between the evaluator and the program managers is often characterized by disenfranchisement and disempowerment. The constructivist evaluation, in the same vein as Scriven's goal-free evaluation, aims to restore the balance of power.

The process of Constructivist Evaluation can be illustrated in nine steps:

1. Identify all relevant stakeholders.
2. Elicit from each stakeholder group their construction and concerns regarding the issue at hand.

3. Provide a context and methodology which allows for multiple constructions, claims, concerns, and issues that can be comprehended, critiqued, and factored in the evaluation as appropriate.
 - Conduct this methodology within each stakeholder group.
 - Cross fertilize each group with the constructions, claims, concerns, and issues identified by other stakeholder groups, or issues drawn from the literature or other sites. All view points are taken into account as long as they are open to critique and criticism.
4. Generate consensus.
5. Prepare an agenda for negotiation on items where there is little or no consensus.
6. Collect and provide the information requested in the agenda for negotiation.
7. Establish and facilitate a forum of stakeholder representatives where negotiation can occur.
8. Develop a report, or several reports, that communicate any consensus on constructions and resolutions. Additionally, the report should communicate the pertinent issues raised by other stakeholder groups.
9. Recycle the evaluation to continue working on unresolved constructions.

The main limitations of Guba and Lincoln's model is that it minimally acknowledges the fundamental role of evaluation in determining the merit, worth, significance, value, quality, or importance of the program, which are core elements within the definition of evaluation (see for examples Scriven, 1991; Davidson, 2005; and Sanders, 1994). Guba and Lincoln claim to offer a formative evaluation

model placing little emphasis on making an evaluative conclusion and more on program improvement through consensus-building; however, to suggest improvement, they must determine deficits in the evaluand, thus they do actually evaluate. A second weakness with the constructivist model is in assuming that stakeholders will always offer the reliable, valid, and honest information. There may be many factors contributing to a stakeholder's knowledge, ability, and candor that must be weighed relative to the observed program impacts in providing a valid evaluative conclusion.

Conclusion

Ethnography is an applied social science research method, while evaluation incorporates various research methods, one of which may be ethnography. The purpose of ethnography is thick description and cultural interpretation; evaluation's aim is to systematically judge a program's merit and develop an evaluative conclusion. The qualitative evaluation approach has demonstrated benefits for evaluators, and three of these approaches are epitomized in the anthropological models of evaluation. Responsive evaluation, goal-free evaluation, and constructivist evaluation have conceptual and methodological similarities. An evaluator should be able to recognize when one of these ethnographic or anthropological models may be feasible and appropriate in evaluating a program. The evaluator should then present the model and its strengths and limitations to the program stakeholders to be considered when selecting the most appropriate evaluation methodology. Sound evaluation typically requires the employment of both quantitative and qualitative research methods. Ethnography and the anthropological models of evaluation may be best suited as a supplement to the quantitative components of an evaluation and serve as a way of triangulating data

collection methods and data sources. A competent evaluator should be informed of these ethnographic techniques and the anthropological models of evaluation.

References

Altschuld, J. W. and Witkin, B. R. (2000). *From needs assessments to action: Transforming needs into solution strategies*. Thousand Oaks, CA: Sage Publications, Inc.

Beebe, J. (n.d.). *Rapid assessment process*. Gonzaga University website. Retrieved January 27, 2005 from http://208.164.121.55/reference/SOME/Outlines/rapid_assessment_process.html

Beebe, J. (1995). Basic concepts and techniques of rapid appraisal. *Human Organization*, 54(1), 42-51.

Carney, T. (1991). Fourth generation evaluation. *Canadian Journal of Communication*, 16(2).

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Denzin, N. K. & Lincoln, Y. S. (Eds). (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage Publications Inc.

Deneberg, V. (1969). Prolizityies A. Zeitgeister. *Psychology Today*, 311(50).

Evers, J. W. (1980). *A field study of goal-based and goal-free evaluation techniques*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Fetterman, D. M. (1982). Ethnography in educational research: The dynamics of diffusion. *Educational Researcher*, 11(3), 17-29.

Fetterman, D. M. (1986a). *The ethnographic evaluator*. Stanford University. Retrieved January 24, 2005 from <http://www.stanford.edu/~davidf/230class/ethnographic evaluator.html>

Fetterman, D. M. (1986b). Conceptual crossroads: Methods and ethics in ethnographic evaluation. In D. D. Williams (Ed.), *Naturalistic evaluation: New directions for program evaluation*. San Francisco, CA: Jossey-Bass.

Fetterman, D. M. (1998). Ethnography. In L. Bickman & D. J. Rog (Eds.) *Handbook of applied social research methods* (pp. 473-504). Thousand Oaks, CA: Sage Publications, Inc.

Genzuk, M. (2004). *A synthesis of ethnographic research*. University of Southern California-Center for Multilingual, Multicultural Research. Retrieved February 3, 2005 from http://www-ref.usc.edu/~genzuk/Ethnographic_Research.pdf

Guba, E. G. & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage Publications.

Hall, B. (n.d.). *Methods: What is ethnography?* Center for Public Interest Anthropology at University of Pennsylvania. Retrieved January 24, 2005 from <http://www.sas.upenn.edu/anthro/CPIA/METHODS/Ethnography.html>

Hopson, R. K. (2002). Making (more) room at the evaluation table for ethnography: Contributions to the responsive constructivist generation.

Exploring evaluation role and identity (pp. 37-56). Information Age Publishing.

Jessor, R., Colby, A., & Shwedler, R. A. (1996). *Ethnography and human development: Context and meaning in social inquiry*. Chicago, IL: The University of Chicago Press.

Maxwell, J. A. (1998). Designing a Qualitative Study. In L. Bickman & D. J. Rog (Eds.) *Handbook of Applied Social Research Methods* (pp. 69-100). Thousand Oaks, CA: Sage Publications, Inc.

McLean, L. D. (1975). Judging the quality of a school as a place where the alis might thrive.” In R. Stake (Ed.), *Evaluating the arts in education: A responsive approach* (pp.41-58). Columbus, OH: Charles E. Merrill Publishing Company.

Nastasi, B. K. & Berg, M. J (1999). Using ethnography to strengthen and evaluate intervention programs. In J. J. Schensul, M. D. LeCompte, G. A. Hess, B. K. Nastasi, M. J. Berg, L. Williamson, J. Brecher, & R. Glassner (Eds.). *Ethnographers toolkit*. Walnut Creek, CA: Altamira Press.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Sage Publications Newbury Park, CA.

Payne, D. A. (1994). *Designing educational project and program evaluations: A practical overview based on research and experience*. Boston, MA: Kluwer Academic Publishers.

Sanders, J. (1994). *The program evaluation standards* (2nd Ed.). Thousand Oaks, CA: Sage Publications, Inc.

- Scriven, M. (2004). Zen and art of everyday evaluation. *Journal of MultiDisciplinary Evaluation*, 1. Retrieved July 20, 2005 from http://www.wmich.edu/evalctr/jmde/content/JMDE_Num_001_Part_I.htm
- Scriven, M. (1991). *Evaluation thesaurus (4th ed.)*. Newbury Park, CA: Sage Publications, Inc.
- Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed), *School evaluation: The politics and process*. Berkeley, CA: McCutchan Publishing Corporation.
- Scriven, M. (1967). The methodology of evaluation. *AERA Monograph Series on Curriculum Evaluation, Vol. 1* (pp. 39-83). Chicago, IL: Rand McNally.
- Seafeld Research & Development Services. (n.d.). *Fourth generation evaluation*. Retrieved January, 27, 2005 from <http://www.srds.ndirect.co.uk/4th.htm>
- Shadish, W. R. (1994): *The guiding principals of evaluation*. Boston, MA: American Evaluation Association.
- Stake, R. (1975). *Evaluating the arts in education: A responsive approach*. Columbus, OH: Charles E. Merrill Publishing Company.
- United States General Accounting Office. (2003). *Ethnographic studies can inform agencies' actions*. GAO-03-455.
- Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (Eds.) (2004). *Handbook of practical program evaluation* (2nd ed.). San Francisco, CA: Jossey-Bass.
- Wikipedia (n.d.). Hermeneutics. Retrieved March 22, 2005 from <http://en.wikipedia.org/wiki/Hermeneutics.htm>
- Wolcott, H. P. (1980). How to look like an anthropologist without really being one.

Practicing Anthropology, 3(2), 56-59.

Wolcott, H. F. (1982). *Ethnographers sans ethnography: The evaluation compromise*. Bloomington, IN: Agency for Instructional Television.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines* (2nd ed.). White Plains, NY: Longman Publishers.

About the Author

Brandon W. Youker obtained his Bachelor of Arts in Social Work from Michigan State University, Master of Science in Social Work from Columbia University in the City of New York, and is a former post-graduate advanced clinical social work fellow at Yale Child Study Center-Yale School of Medicine. Currently, Brandon Youker is a doctoral student in Interdisciplinary Evaluation at Western Michigan University and an evaluator at Western Michigan University's Evaluation Center. His academic interests include evaluation theory, methodology and design; international program evaluation; evaluation of social work practice and programming; the evaluation of human service programs; and the evaluation of arts programs.

Book Reviews

Revisiting Realistic Evaluation

Chris L. S. Coryn

Albeit some might argue that this review is a little late in coming, it is worth revisiting Pawson and Tilley's 1997 book, *Realistic Evaluation* (reprinted in 1998, 2000, 2001, and 2002) as the debate about causation and evidence-based research and evaluation continues to be a topic of debate and concern in the evaluation and research communities (see *A Call to Action: The First International Congress of Qualitative Inquiry and The Claremont Debate*, in this issue of *JMDE*). *Realistic Evaluation* is rooted in the tradition of scientific realism, which is said to be one of the "dominant axes in modern European thinking" (p. 55). In the most general of terms scientific realism concerns "the nature and operation of causal forces" (p. 55). The essential ingredients for assessing these causal forces are C-M-O configurations—where C represents context, M represents mechanisms, and O represents outcomes. Context refers to the "spatial and institutional locations of social situations, together, crucially, with the norms, values, and interrelationships found in them" (p. 216). Mechanisms are the "choices and capacities which lead to regular patterns of social behavior" and the causal mechanisms which generate these patterns of behavior are "deemed 'social problems' and which are the rationale for a program" (p. 216). Outcomes "provide the key evidence for the

realist evaluator in any recommendation to mount, monitor, modify, or mothball a program” (p. 217). From the C-M-O configuration, the authors argue that the way in which causation in the “social world should be constructed” and that the “basic realist formula” is “mechanism + context = outcome” (p. xv).

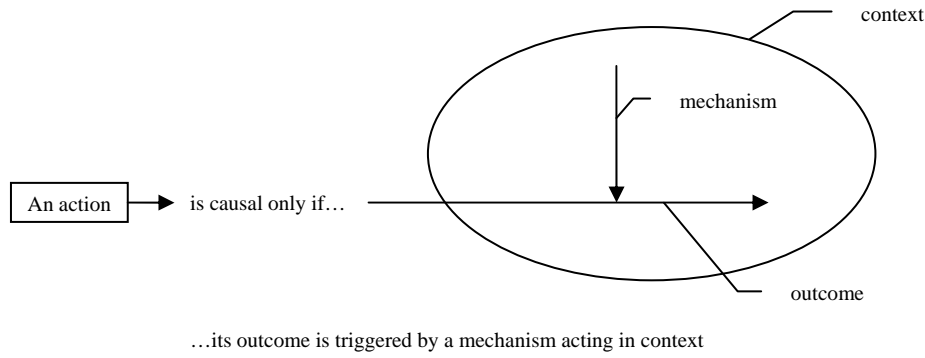
Chapter 1, A History of Evaluation in 28 ½ Pages, presents Pawson and Tilley’s version of the history of evaluation. The authors begin this history with the experimental evaluations of the 1960s of the “great social programs of the ‘great society’ [the U.S.]” (p. 2) brought about by the work of Stanley and Campbell, among others. In short, the experimental paradigm is described as a failure for a variety of reasons, including the lack of external validity (i.e., generalizability) brought about by experiments’ inability to reproduce results in the ‘real world.’ Somewhat out of place, but next in the short history of evaluation are the utilization-focused approaches. These approaches are criticized on the grounds that “he who pays the researcher calls the methodological tune” (p. 14). Finally, the emergence of constructivism in the 1970s is reviewed and also described as a disappointment because of the “inability to grasp those structural and institutional features of society which are in some respects independent of individuals’ reasoning and desires” (p. 23). All in all, the authors paint a bleak picture of evaluation’s past and contend that if the future is to be brighter then evaluators had better take theory seriously, although the authors also find serious flaws in the various theory-driven approaches of Chen, Weiss, and others. These faults are described as the lack of attention given to context and the emphasis on experimental methods, for example. This 28 ½ page history of evaluation is intended to set the stage and substantiate the authors approach to and purposes for evaluation: determining not only ‘if’ a program works, but also ‘how’ and for ‘whom.’

Chapter 2, Weaknesses in Experimental Evaluation, presents Pawson and Tilley's expose on the inherent problems with the experimental tradition; namely, the experimentalists' "epistemological assumptions about causation and their lack of fit with the nature of social programs" (p. 30). Essentially, the authors argue that more often than not that change cannot be captured in *OXO* terminology. All in all, it is asserted that "by its very logic, experimental evaluation either ignores these underlying process [causal mechanisms], or treats them incorrectly as inputs, outputs or confounding variables, or deals with them in a *post hoc* and thus arbitrary fashion" (p. 54).

In Chapter 3, In With the New: Scientific Realism, the authors present the principles and practice of scientific realism. As previously mentioned, the realist view (generative) of causation can be described thusly (as illustrated by the explosion of gunpowder):

Our basic concern is still, of course, the *outcome* (the spark causing the explosion). But what does the explanatory work is first of all the *mechanism* (the chemical composition of the substance which allows the reaction), and secondly the *context* (the physical conditions which allow the mechanism to come into operation). This proposition—causal outcomes follow mechanisms acting in contexts—is the axiomatic base upon which all realist explanations build.

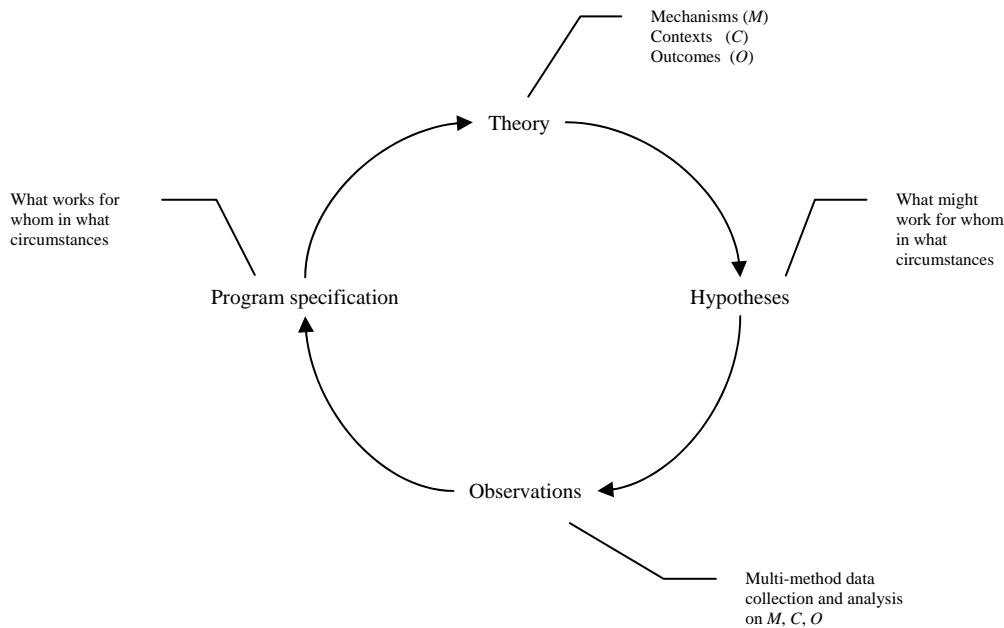
(Pawson & Tilley, 1997, p. 58)



Source: Pawson, R. and Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

Figure 1. Generative Causation

Chapter 4, How to Design a Realistic Evaluation, presents the realist evaluation cycle (see Figure 2) and three case studies which apply realist evaluation principles to varying degrees.



Source: Pawson, R. and Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

Figure 2. The Realist Evaluation Cycle

The first case study presented is an evaluation of property marking and described by the authors as *testing theory*, the second is an evaluation of a housing project and described as *theory formation and development*, and the third is an evaluation of a prison-delivered higher education program also described as *theory formation and development*. These case studies are described in some detail and are intended to represent exemplars of realistic evaluation.

I have opted to exclude a review of the remaining chapters (5-9) as these merely focus on collecting realist data and the methodological procedures involved in conducting evaluation as prescribed by Pawson and Tilley.

Despite the book's title, the true underlying premise of Pawson and Tilley's *Realistic Evaluation* is not merely a proposition of how to conduct evaluation, but rather a treatise on the nature of causation and science. While the author's notion of causation (scientific realism) is compelling, I am not entirely convinced that it is the "final solution" to the causation debate. Neither is it a dramatic improvement over either successionist or other traditions. In their haste to prescribe generative explanations they fail to recognize or acknowledge that numerous experimentalists (and non-experimentalists) give considerable attention to context in their accounts of causation (e.g., moderators, mediators, interaction effects), often to a greater degree than the examples provided throughout the book suggest. Moreover, these causal accounts (i.e., realist accounts) seem little more than explanations of program effectiveness for different groups or consumers, which can be accomplished without the use of realist principles.

Prior reviews (Patton, 1999; Rogers, 1999) of *Realistic Evaluation* have been mixed. For example, Rogers (1999) stated that "this is one of those rare books that has the potential to permanently change one's perspective on program evaluation"

(p. 381). Patton (1999), on the other hand, was not entirely convinced of the credibility of Pawson and Tilley's contribution and responded to their criticisms of utilization-focused evaluation (p. 14) thusly:

I rarely respond to attacks on or distortions of my views, especially when they're based on the twenty-year-old first edition of the book (Patton, 1978) and don't take into account subsequent revisions and elaborations (Patton, 1986, 1997) that I hope have corrected at least some earlier weaknesses, and have benefited from well-deserved and well-meaning critiques. I have learned that responding to a distortion risks reinforcing the very thing I want to correct by calling attention to it. However, the distortions in the opening chapter of Pawson and Tilley, in which they sarcastically and disparagingly review (and bemoan) the history of evaluation, are anything but innocent or trivial. The irony is that, in the introduction, the authors claim the mantle of "detachment," "objectivity," and "scientific evaluation" (p. xiii). Their mocking review of evaluation's history has one primary purpose: positioning themselves as saviors of the profession by redirecting us to be scientists first and foremost.

(Patton, 1999, p. 387)

While *Realistic Evaluation* has spurred serious interest and debate, and even spawned an issue of *New Directions for Evaluation* (Henry, Julnes, & Mark, 1998), the approach has not quite received the attention in North America that it has in the United Kingdom and Europe. A search of the American and Canadian evaluation journals did not turn-up any publications related to the approach (with the exception of Patton and Roger's reviews of the book). While a search of the major European evaluation journal (*Evaluation: The International Journal of Theory, Research, and Practice*) returned 56 articles which focused on, or emphasized, the realistic evaluation approach.

References

- Henry, G. T., Julnes, G. & Mark, M. M. (Eds.) (1998). Realist evaluation: An emerging theory in support of practice. *New Directions for Evaluation*, 78.
- Patton, M. Q. (1999). Realistic evaluation [review of the book *Realistic evaluation*]. *American Journal of Evaluation*, 20(2), 385-388.
- Pawson, R. and Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.
- Rogers, P. J. (1999). Realistic evaluation [review of the book *Realistic evaluation*]. *American Journal of Evaluation*, 20(2), 381-383.