

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 Enough 1

Paul R. Brandon

The Value of Evaluation Standards: A Comparative Assessment 30

Robert Picciotto

The 2004 Claremont Debate: Lipsey vs. Scriven 60

Stewart I. Donaldson and Christina A. Christie

Evaluation Capacity Building and Humanitarian Organization 78

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

PART II

Global Review: Regions and Events

National and Regional Evaluation Networks..... 150

IOCE

A Call to Action: The First International Congress of Qualitative Inquiry..... 155

Chris L. S. Coryn, Daniela C. Schröter, & Michael Scriven

Evaluation in Canada 166

Chris L. S. Coryn

Evaluation in the People's Republic of China 172

<i>Xuejin Lu and Donghai Xie</i>	
Evaluation in Germany: An Overview	180
<i>Gerlinde Struhkamp</i>	
Evaluation—Making it Real in Aotearoa New Zealand: Leading by Example, Leading by Association.....	195
<i>Pam Oliver, Kate McKegg, Geoff Stone, and Maggie Jakob-Hoff</i>	
A Review of the Chinese National Center for Science and Technology Evaluation	197
<i>Laura Pan Luo</i>	
Evaluation in Japan	200
<i>Ryo Sasaki</i>	

PART III

Global Review: Publications

American Journal of Evaluation	205
<i>Lori Wingate</i>	
New Directions for Evaluation	209
<i>Chris L. S. Coryn</i>	
Evaluation: The International Journal of Theory, Research and Practice	214
<i>Daniela C. Schröter</i>	

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.

In the Methods Notes Section, J. Jackson Barnette and Anne Baber Wallis seek to “close one of the few gaps left in the Campbell-Stanley-Cook-Shadish legacy of research designs” (p. 106). They examine how what happens to an intervention between multiple postobservations (e.g., removal, continuation, changes in intensity) in experimental and quasi-experimental evaluation designs impacts validity, data modeling, and analysis. They argue that designs that take these factors into account will produce better inferences.

Donna Mertens’ contribution in the “Historical Record” section provides an account of the “The Inauguration fo the International Organization for Cooperation in Evaluation” (IOCE). IOCE’s mission is “to help legitimate and strengthen evaluation societies, associations or networks so that they can better contribute to good governance, effective decision making, and strengthen the role of civil society” (p. 127). Mertens describes the work done to get the organization off the ground; gives a brief account of the inaugural assembly that took place in Lima, Peru in 2003; and conveys the IOCE’s mission, goals, and current priorities. She concludes by dicussing the organization’s accomplishments, challenges, and opportunities.

In the “Ethical Challenges” section, Gillian Kerr comments on two analyses of “The Steering Committee” ethical challenge in a previous issue of *AJE*. She did not think these analyses paid sufficient attention to “the role of the steering or advisory committee itself and the extent to which membership of such a committee is associated with genuine power” (p. 132) and explains why in “Reflections of ‘The Steering Committee.’”

New Directions for Evaluation

Chris L. S. Coryn

The Spring 2005 issue of *New Directions for Evaluation*, Teaching Evaluation Using the Case Method, edited by Michael Q. Patton and Patricia Patrizi is intended to advance the practice of evaluation teaching using the case method by “providing specially developed cases for teaching and teaching guidelines and discussion points to use in conjunction with the cases” (p. 3). In this issue, chapters 2-4 conclude with “Teaching Guidelines and Questions,” which are intended to provide general case teaching guidance by providing case teaching questions and evaluation points to elicit through questioning.

Chapter 1, Case Teaching and Evaluation, by Michael Q. Patton and Patricia Patrizi, outlines the logic and likely benefits of using and applying cases as a teaching method for students of evaluation. The authors argue that case teaching and training, like the longstanding traditions of using cases for teaching law and medicine, will prepare future evaluators for the practical problems that arise in real-world evaluations (e.g., “professional practice does not lend itself to rules and formulas” and “decisions are rarely routine”, p. 5). The strategies for case teaching strategies presented by the authors in this chapter include (1) facilitating case discussion to provide experiences in evaluative thinking, situational analysis, and practical problem solving for real-world evaluation, (2) set and model norms of civil interaction, (3) emphasizing advanced preparation, (3) setting expectations and creating a learning frame of mind, (4) starting the questioning process by

eliciting the facts of the case, (5) *vive la difference* [e.g., reconciling opposing points of view], (6) adding hypothetical and incorporating role playing, (7) concluding with takeaways and generalized learning, and (8) supporting active, practice-oriented learning. Patton and Patrizi conclude the chapter by stating that

Evaluation as a field of professional practice has long way to go to achieve the prestige of fields like law, medicine, and business, but the challenges we face in supporting the development of skilled practitioners who can analyze unique situations, deal with diverse people, and exercise astute judgment bear striking similarities to these professions.

(p. 13)

In Chapter 2, *Evaluation of the Fighting Back Initiative*, by Kay E. Sherwood, presents the case of the Robert Wood Johnson Foundation's Fighting Back initiative, an \$88 million dollar investment by the foundation for developing community-generated strategies for reducing use and abuse of alcohol and illegal drugs. This investment included \$14 million for an independent evaluation of the foundation's initiative. In the case, Sherwood provides all of the necessary background and contextual information for making the case a usable teaching tool. Also presented in the case are early efforts at evaluating the initiative, beginning in 1990, where the evaluation floundered as the research team was "unable to manage the complexity and comprehensiveness of the design" (p. 23). This team purportedly wasted \$4.6 million, 4 years, baseline for future efforts, and credibility for the overall effort. Eventually the evaluation was rescued by a new research team, which conducted the 1994-2000 evaluation of the initiative. All in all, the case of the Fighting Back Initiative provides a rich, complex teaching example.

In Chapter 3, *Evaluation of the Central Valley Partnership of the James Irvine Foundation*, by Martha S. Campbell, Michael Q. Patton, and Patricia Patrizi, the case presented was initiated by the foundation as a “partnership for citizenship” (p. 39). Thus, the purpose of the Central Valley Partnership (CVP) was to engage low income, immigrant, and disenfranchised residents in civic action. In this example, the authors present a case where the role of the evaluator shifts from pure evaluation to “an organizational development resource” (p. 46). In this sense, the case illustrates the various roles and responsibilities that evaluators are often required or requested to perform. The case concludes with comments from Martha Campbell, now the vice president for programs at the Irvine Foundation, in which she states

Irvine’s experience with CVP and its other evaluations has reinforced, as well as tempered, its view of the role and potential of evaluation...As such, Irvine currently adopts an approach to evaluation that has a strong focus on improving program delivery and documenting program innovations or practices for the larger field.

(p. 54)

Chapter 4, *Evaluating Home Visitation: A Case Study of Evaluation at the David and Lucile Packard Foundation*, by Kay E. Sherwood, presents a case where the foundation used an evaluation-focused strategy to making grants for child development projects. Through this strategy, the foundation’s evaluation efforts frequently emphasized results-based evidence to support project effectiveness, primarily in the form of experimental designs. Unfortunately, as the case presents, these effects were generally “mixed” or “non-significant” (p. 67). Much of the case involves the publication of these poor, disappointing results and the subsequent

fallout generated by them, including efforts for damage control by the foundation and other stakeholders.

In Chapter 5, *Evaluation Case Teaching from a Participant Perspective*, by John Bare, the author describes the benefits of the case teaching method from the view of a learner. Most interesting in Bare's chapter is the "surfacing of values," wherein the author argues that values are pervasive and shape both program planning and evaluation. Moreover, the author notes that "cases help reveal these" (p. 89).

The issue concludes with Chapter 6, *Diverse and Creative Uses of Cases for Teaching*, by Michael Q. Patton. In this chapter Patton presents suggestions for using the cases presented in the issue, and other cases, for the "broader context of evaluation teaching and training" (p. 91). First, the author provides issues for exploring cross-case comparisons including (1) connecting parts into a whole, (2) the personal factor, (3) evaluator roles and purposes, (4) complex relationships and institutional arrangements, (5) controversies and politics, and (6) what is missing? Second, Patton explores additional teaching uses for cases. These uses could include (1) insights into evaluator competencies, (2) learning to write executive summaries, (3) practicing qualitative analysis and extracting lessons learned, (4) stakeholder analysis and stakeholder mapping, (5) developing ethical commitments and sensitivities, (6) metaevaluation training, and (7) applying model, theorists, and conceptual distinctions. Patton summarizes the issue by stating that

This volume on using cases for teaching evaluation aspires to contribute to professional excellence in evaluation by grounding training real-world experiences captured and presented in detailed cases. Case teaching and the additional practice-oriented teaching ideas presented in this chapter seek to bridge the gap between knowing and doing.

(p. 98)

As a student of evaluation I found “Teaching Evaluation Using the Case Method” a compelling, logical approach to teaching and learning evaluation. Each of the cases presented in Chapters 2-4 offer a unique series of problems and possibilities. Furthermore, I found Patton’s presentations of teaching guidelines and questions at the end of these chapters useful and relevant to the cases presented. While I agree with Patton that evaluation teaching and training needs to “bridge the gap between knowing and doing” (p. 98), there are alternatives to cases which should be considered as well. For example, cases may in fact be “real-world,” but the use of the case is still “hypothetical.” That is, learners are not really evaluating the programs or projects presented in the cases. They may be confronted with the complexities and problems of real-world evaluation, but real-world practice should include “real” evaluation as opposed to merely practicing on cases. Although cases are an invaluable teaching tool, I would argue that what many professional programs of study call “field or professional experience” would be the real, real-world equivalent of cases.

References

Patton, M. Q. & Patrizi, P. (Eds.) (2005). Teaching evaluation using the case method. *New Directions for Evaluation*, 105.

Evaluation: The International Journal of Theory, Research and Practice

Daniela C. Schröter

In a time of results-based management and budgeting, the question whether or not the *inputs* have been in line with the policies of donors and partner countries is not longer really relevant. The real question is whether the *results* of our actions are in line with the policies and the problems that these policies tried to address.

(van den Berg, p. 35)

The first 2005 issue of *Evaluation* (Volume 11(1), January 2005) begins with two contributions to *A Visit to the World of Practice*, both of which focus on results-based evaluation and impact assessment within the context of the Millennium Development Goals (MDGs). Please visit <http://ddp-ext.worldbank.org/ext/MDG/-home.do> for information on the MDGs.

First, Kusek, Rist, and White discuss how the shift from implementation-focused monitoring and evaluation (M&E) to results-based M&E is taking place in various developed and developing countries, which challenges are being faced in this transition, and what strategies should be considered when introducing results-based M&E, including readiness assessments, political and organizational issues, and potential challenges with implementation, reliable data collection and analysis.

Second, van den Berg discusses some methodological issues in the assessment of development cooperation. Monitoring, for example, would not assess if the right

things are done in development, but only whether things are done right. Impact assessments, in contrast to monitoring, would be complicated and expensive, because impacts occurs over long terms, require increased scope of research, and rely on baseline data often unavailable. Moreover, counterfactuals have to be considered to indicate that observed outcomes in fact resulted from the intervention under investigation. Causality as the key to the establishment of impact would be reduced in the social science context to “specific causality”, because there are no general laws as in the natural sciences. To proof linkages between outcomes and impact, methods such as lab research, RCTs, and double-blind studies with comparison groups are commonly utilized by social scientist. Van den Berg argues for the methodological inclusion of historical analysis to ascertain causality, utilizing triangulation “par excellence” to insure reliability and reasoning for validity. Current evaluation practice employs triangulation only methodologically. However, using historical triangulation eliminates the need for counterfactuals to establish causality. Moreover, linear causality as established through statistical techniques is often thwarted by societal complexities. Therefore, discussions in social sciences should shift toward “conditionalities” (p. 34). Van den Berg believes “that the development community should move from causality or plausibility to contribution, and from direct linkages to necessary but not sufficient conditions for change” (p. 34).

Four articles follow. First, Saunders, Charlier, and Bonamy discuss how evaluation can be used to support change, exemplified in two international higher education case examples. Second, Kautto and Similä provide an account of evaluating “recently introduced policy instruments (RIPs)” (p. 55) supported by intervention theories and recommend (1) the utilization of theory-based approaches, (2) the selection of criteria and establishment of causal links between evaluation criteria,

(3) the selection of causal linkages for which information can readily be ascertained, (4) determination of procedures for proceeding with the criteria for which information is not readily available, and (5) consideration of potential for theory failure. Third, Byng, Norman, and Redfern provide a case example within a mental health context, utilizing realistic evaluation as coined by Pawson and Tilley in combination with analytic induction. Fourth, Shadish, Chacón-Moscoso, and Sánchez-Meca describe how meta-analysis and systematic reviews have been developed historically, utilized in Europe, and contributed to policy making and practice.

In the *Review* section of *Evaluation 11*(1), Kushner looks at a current UK Cabinet Publication entitled “Quality in qualitative evaluation: a framework for assessing research and evidence.”

The final section, *News from the Community*, discusses the fifth annual Japanese Evaluation Society (JES) and third annual African Evaluation Association conferences. The section also introduces the International Organization for Cooperation in Evaluation (IOCE; also see this issue of JMDE). The final news from the community is the Univation/German Evaluation Society conference, which focused on network evaluation.