Introduction

Thomas Kellaghan
Educational Research Centre, St. Patrick's College, Dublin, Ireland

Daniel L. Stufflebeam
The Evaluation Center, Western Michigan University, MI, USA

Lori A. Wingate
The Evaluation Center, Western Michigan University, MI, USA

Educational evaluation encompasses a wide array of activities, including student assessment, measurement, testing, program evaluation, school personnel evaluation, school accreditation, and curriculum evaluation. It occurs at all levels of education systems, from the individual student evaluations carried out by classroom teachers, to evaluations of schools and districts, to district-wide program evaluations, to national assessments, to cross-national comparisons of student achievement. As in any area of scholarship and practice, the field is constantly evolving, as a result of advances in theory, methodology, and technology; increasing globalization; emerging needs and pressures; and cross-fertilization from other disciplines.

The beginning of a new century would seem an appropriate time to provide a portrait of the current state of the theory and practice of educational evaluation across the globe. It is the purpose of this handbook to attempt to do this, to sketch the international landscape of educational evaluation – its conceptualizations, practice, methodology, and background, and the functions it serves. The book's 43 chapters, grouped in 10 sections, provide detailed accounts of major components of the educational evaluation enterprise. Together, they provide a panoramic view of an evolving field.

Contributing authors from Africa, Australia, Europe, North America, and Latin America demonstrate the importance of the social and political contexts in which evaluation occurs. (Our efforts to obtain a contribution from Asia were unsuccessful.) Although the perspectives presented are to a large extent representative of the general field of evaluation, they are related specifically to education. Evaluation in education provides a context that is of universal interest and importance across the globe; further, as history of the evaluation field shows, the lessons from it are instructive for evaluation work across the disciplines. In

International Handbook of Educational Evaluation, 1–6
T. Kellaghan, D.L. Stufflebeam (eds.)

fact, many advances in evaluation stemmed from the pioneering efforts of educational evaluators in the 1960s and 1970s.

Contemporary educational evaluation is rooted in student assessment and measurement. The distinction between measurement and evaluation, suggested by Ralph Tyler more than 50 years ago and later elaborated on by others, had an enormous influence on the development of evaluation as an integral part of the teaching and learning process. For many years, educational evaluation focused mainly on students' achievements; it concentrated on the use of tests and was immensely influenced by psychometrics. Another major and long-standing influence on educational evaluation is to be found in a variety of programs to accredit schools and colleges. Mainly a U.S. enterprise, accreditation programs began in the late 1800s and are an established reality throughout the U.S. today.

It was only in the mid-1960s and early 1970s, with the increased demand for program evaluation made necessary by various evaluation requirements placed on educational programs and projects by governmental organizations and other agencies, that educational evaluation dramatically expanded and changed in character. While earlier evaluation, as noted above, had focused on student testing and the educational inputs of interest to accrediting organizations, the new thrust began to look at a variety of outcomes, alternative program designs, and the adequacy of operations. To meet new requirements for evaluation, evaluators mainly used their expertise in measurement and psychometrics, though they also took advantage of two other resources: research methodology and administration. Research methodology – mainly quantitative but later also qualitative – provided the guidance for data collection procedures and research designs that could be applied in evaluation. Administration theory and research helped to improve understanding of planning and decision making, which evaluations were designed to service, as well as of the politics of schools.

Most developments in program evaluation took place in the United States and were "exported" to other parts of the world, sometimes only ten or twenty years later. In Europe, for instance, the major concern was – and in some countries still is – testing and student assessment, although tests and other achievement measures have begun to be used for other purposes. Gradually, tests came to be used as outcome measures for other evaluation objects, such as programs, schools, and education systems, sometimes alongside other information regarding the objects' goals and processes. Widely varying applications of evaluation can now be found around the world in many shapes and sizes, reflecting its far-reaching and transdisciplinary nature.

Side by side with all this activity, evaluation has been growing into a fully fledged profession with national and international conferences, journals, and professional associations. It is practiced around the world by professional evaluators in universities, research institutes, government departments, schools, and industry. It is being used to assess programs and services in a variety of areas, such as criminal justice, welfare, health, social work, and education. Each area, while having much in common with evaluation in general, also has its unique features.

Three distinctive features set educational evaluation apart from other types of evaluation. First, it has been strongly shaped by its roots in testing and student assessment, on one hand, and curriculum and program evaluation on the other. In other areas (e.g., health services or criminal justice), evaluation focuses mainly on programs and is usually considered as a type of applied research. Although it took many years for educational evaluation to come to the point where it would not be perceived only as student assessment, such assessment is still an important element of the activity. Second, education is the predominant social service in most societies. Unlike business and industry, or other social services such as health and welfare, education affects, or aspires to affect, almost every member of society. Thus, public involvement and the concerns of evaluation audiences and stakeholders are of special significance in educational evaluation, compared to evaluation in other social services, and even more so when compared to evaluation in business and industry. Third, teachers play very important roles in educational evaluation as evaluators, as evaluation objects, and as stakeholders. They are a unique and extremely large and powerful professional group, with a high stake in evaluation and a long history as practicing evaluators assessing the achievements of their students, and must be taken into account whenever evaluation is being considered.

Education is one of the main pillars of the evaluation field, and thus it is important that those who work in educational evaluation should be part of the general evaluation community, participating in its scientific meetings and publishing their work in its professional journals. There is much that they can share with, and learn from, evaluators in all areas of social service, industry, and business. However, educational evaluators should also be sensitive to the unique features of their own particular area of evaluation and work to develop its capabilities so that they can better serve the needs of education and its constituents. It is our hope that this handbook will aid members of the educational evaluation community in this endeavor.

The handbook is divided into two parts, Perspectives and Practice, each of which is further divided into five sections. While the individual chapters can stand on their own as reference works on a wide array of topics, grouping them under Perspectives and Practice, provides in-depth treatments of related topics within an overall architecture for the evaluation field. In the first part, the perspectives of evaluation are presented in five major domains: theory, methodology, utilization, profession, and the social context in which evaluations are carried out. The second part of the handbook presents and discusses practice in relation to five typical objects of evaluation: students, personnel, programs/projects, schools, and education systems. Chapters in the handbook represent multiple perspectives and practices from around the world. The history of educational evaluation is reviewed, and the unique features that set it apart from other types of evaluation are outlined. Since the chapters in each section are ably introduced by section editors, we will only comment briefly on each section's contents.

The opening section deals with perspectives on educational evaluation by examining its theoretical underpinnings. Ernest House introduces the section by

noting that scholars have made substantial progress in developing evaluation theory, but remain far apart in their views of what constitutes sound evaluation. Michael Scriven provides an overview and analysis of theoretical persuasions, which may be grouped and contrasted as objectivist and relativist. Specific evaluation theory perspectives presented in the section include Daniel Stufflebeam's CIPP model, with its decision/accountability and objectivist orientations; Robert Stake's responsive evaluation, that stresses the importance of context and pluralism and advocates a relativist orientation; the constructivist evaluation approach of Egon Guba and Yvonna Lincoln, with its emphasis on participatory process and rejection of objective reality; and the relatively new entry of democratic deliberative evaluation, advanced by Ernest House and Kenneth Howe, which integrates evaluation within a democratic process. Other sections present additional theoretical perspectives, including ones relating to utilization-focused evaluation, participatory evaluation, connoisseurship evaluation, and experimental design. Readers interested in the theory of evaluation will find in these chapters ample material to support dialectical examination of the conceptual, hypothetical, and pragmatic guiding principles of educational evaluation.

Section 2 focuses on evaluation methods. Evaluators, as the section editor Richard Wolf notes, differ in their methodological approaches as much as they differ in their theoretical approaches. Major differences are reflected in the extent to which investigators control and manipulate what is being evaluated. At one extreme, randomized comparative experiments, described by Robert Boruch, are favoured; at the other extreme, completely uncontrolled, naturalistic studies, described by Linda Mabry. Other methods presented in the section include cost-effectiveness analysis, described by Henry Levin and Patrick McEwan, and Elliot Eisner's educational connoisseurship approach. In general, the section reflects the current dominant view that evaluators should employ multiple methods.

The chapters in Section 3 provide in-depth analyses of how evaluators can ensure that their findings will be used. Section editor Marvin Alkin and his colleague Carolyn Huie Hofstetter summarize and examine research on the utilization of evaluation findings. Michael Patton and Bradley Cousins, respectively, present state-of-the-art descriptions of utilization-focused and participatory models of evaluation, and explain how they foster the use of findings.

Section 4 editor Midge Smith acknowledges that the evaluation field has made progress toward professionalization, yet judges that the effort is still immature and in need of much further thought and serious development. The topics treated in the section include Daniel Stufflebeam's report on progress in setting professional standards, Michael Morris's treatise on evaluator ethics, Blaine Worthen's examination of the pros and cons of evaluator certification, Lois-ellin Datta's analysis of the reciprocal influences of government and evaluation, Hallie Preskill's proposal that the evaluation field become a sustainable learning community, and Midge Smith's projection of, and commentary about, the future of evaluation. Overall, contributors to the section characterize evaluation as an emergent profession that has developed significantly but still has far to go.

Section 5 editor Harbans Bhola notes that the practice of evaluation is, and should be, heavily influenced by the local social setting in which the evaluation is carried out, but also characterizes a global context for evaluation. Particular settings for educational evaluation are discussed in chapters by Carl Candoli and Daniel Stufflebeam for the U.S., Ove Karlsson for Europe, Fernando Reimers for Latin America, and Michael Omolewa and Thomas Kellaghan for Africa. Contributions to the section make clear that evaluation practices are heavily influenced by a nation's resources and employment of technology, as well as by local customs, traditions, laws, mores, and ideologies. A clear implication is that national groups need to set their own standards for evaluation.

Section 6 editors Marguerite Clarke and George Madaus introduce chapters on the assessment of student achievement, which has been, and continues to be, a core part of educational evaluation. This is the kind of assessment that impacts most directly on students, often determining how well they learn in the classroom or decisions about graduation and future educational and life chances. It takes many forms. Robert Mislevy, Mark Wilson, Kadriye Ercikan, and Naomi Chudowsky present a highly substantive state-of-the-art report on psychometric principles underlying standardized testing. Peter Airasian and Lisa Abrams describe classroom evaluation practice, which arguably is the form of evaluation which has the greatest impact on the quality of student learning. Caroline Gipps and Gordon Stobart describe concepts and procedures of assessment that have received great attention in recent years in response to dissatisfaction with traditional methods, particularly standardized tests. Thomas Kellaghan and George Madaus provide a description of external (public) examinations and issues that arise in their use in a chapter that 20 years ago would probably have evoked little more than academic interest in the United States. However, having long eschewed the use of public examinations, which have a tradition going back thousands of years in China and form an integral part of education systems today in many parts of the world, the United States over the last decade has accorded a form of these examinations a central role in its standards-based reforms.

Section 7 editor Daniel Stufflebeam argues that educational evaluations must include valid and reliable assessments of teachers and other educators, and that much improvement is needed in this critical area of personnel evaluation. According to Mari Pearlman and Richard Tannenbaum, practices of school-based teacher evaluation have remained poor since 1996 but external programs for assessing teachers, such as teacher induction and national certification assessments, have progressed substantially. According to Naftaly Glasman and Ronald Heck, the evaluation of principals has also remained poor, and shows little sign of improvement. James Stronge ends the section on an optimistic note in his report of the progress that he and his colleagues have made in providing new models and procedures for evaluating educational support personnel. Overall, the section reinforces the message that educational personnel evaluation is a critically important yet deficient part of the educational evaluation enterprise.

James Sanders, the editor of Section 8, drew together authors from diverse national perspectives to address the area of program/project evaluation.

Program evaluation as practiced in developing countries is described by Gila Garaway; in the U.S.A. by Jean King; in Canada by Alice Dignard; and in Australia by John Owen.

Section 9 editor Gary Miron introduces the chapters on evaluation in schools with a discussion of the old and new challenges in this area. Daniel Stufflebeam offers strategies for designing and institutionalizing evaluation systems in schools and school districts; James Sanders and Jane Davidson present a model for school evaluation; while Robert Johnson draws on his work with Richard Jaeger to provide models and exemplars for school profiles and school report cards. Catherine Awsumb Nelson, Jennifer Post, and William Bickel present a framework for assessing the institutionalization of technology in schools.

While Section 6 deals with the evaluation of individual students, the chapters in Section 10 address the use of aggregated student data to evaluate the performance of whole systems of education (or clearly identified parts of them) in a national, state, or international context. As section editor Thomas Kellaghan points out, the use of this form of evaluation grew rapidly throughout the world in the 1990s as education systems shifted their focus when evaluating their quality from a consideration of inputs to one of outcomes. Two major, and contrasting, forms of national assessment are described in the section. Lyle Jones describes the sample-based National Assessment of Educational Progress in the United States, while Harry Torrance describes the census-based assessment of the national curriculum in England. William Webster, Ted Almaguer and Tim Orsak describe state and school district evaluation in the U.S. Following that, international studies of educational achievement, the results of which have been used on several occasions to raise concern about the state of American education, are described by Tjeerd Plomp, Sarah Howie, and Barry McGaw. William Schmidt and Richard Houang write about a particular aspect of international studies, cross-national curriculum evaluation.

We hope that the handbook will be useful to readers in a variety of ways, helping them to consider alternative views and approaches to evaluation, to think about the role and influence of evaluation in national settings, to gain a better understanding of the complexities of personnel and program evaluation, to gain perspective on how to get evaluation findings used, to look to the future and to the steps that will be needed if evaluation is to mature as a profession, to identify a wide range of resource people, to appreciate the needs for evaluation at different levels, and to identify common themes to ensure integrity for evaluation across national settings.

We are indebted to the authors, section editors, and others who contributed to the handbook. David Nevo assisted the publisher in starting the Handbook project, contributed to the overall plan, and helped select and recruit some of the section editors and chapter authors. We acknowledge the assistance of Michael Williams of Kluwer Academic Publishers, who consistently supported our effort. Hilary Walshe, Regina Klöditz, and Rochelle Henderson at the Educational Research Centre at St Patrick's College in Ireland and Sally Veeder at the Evaluation Center at Western Michigan University provided competent editorial, technical, and clerical assistance.

1
Evaluation Theory and Metatheory[1]

MICHAEL SCRIVEN
Claremont Graduate University, CA, USA

DEFINITIONS

What is evaluation? Synthesizing what the dictionaries and common usage tell us, it is the process of determining the merit, worth, or significance of things (near-synonyms are quality/value/importance). Reports on the results of this process are called *evaluations* if complex, *evaluative claims* if simple sentences, and we here use the term *evaluand* for whatever it is that is evaluated (optionally, we use *evaluee* to indicate that an evaluand is a person).

An evaluation theory (or theory of evaluation) can be of one or the other of two types. Normative theories are about what evaluation should do or be, or how it should be conceived or defined. Descriptive theories are about what evaluations there are, or what evaluations types there are (classificatory theories), and what they in fact do, or have done, or why or how they did or do that (explanatory theories).

A metatheory is a theory about theories, in this case about theories of evaluation. It may be classificatory and/or explanatory. That is, it may suggest ways of grouping evaluation *theories* and/or provide explanations of why *they* are the way that they are. In this essay we provide a classification of evaluation theories, and an explanatory account of their genesis.

DESCRIPTIONS

Of course, the professional practice of evaluation in one of its many *fields*, such as program or personnel evaluation, and in one of its *subject-matter areas*, such as education or public health or social work, involves a great many skills that are not covered directly in the literal or dictionary definition. Determining the merit of beginning reading programs, for example, requires extensive knowledge of the type of evaluand – reading programs – and the methods of the social sciences and often those of the humanities as well. To include these and other related matters in the definition is attractive in the interest of giving a richer notion of serious evaluation, so it's tempting to define evaluation as "whatever evaluators do." But

International Handbook of Educational Evaluation, 15–30
T. Kellaghan, D.L. Stufflebeam (eds.)

this clearly won't do as it stands, since evaluators might all bet on horse races but such betting does not thereby become part of evaluation. In fact, professional evaluators quite properly do various things as part of their professional activities that are not evaluation but which they are individually competent to do, e.g., market research surveys; teaching the theory or practice of evaluation; advising clients on how to write evaluation components into their funding proposals; how to meet legal requirements on information privacy; and when to consider alternative program approaches. Those activities are not part of evaluation as such, merely part of what evaluators often do, just as teaching mathematics is part of what many distinguished mathematicians do, although it's not part of mathematics or of being a mathematician. We often include training in some of these activities as part of practical training in how to be successful in an evaluation career; others are just opportunities to help clients that frequently arise in the course of doing evaluations. Of course, some evaluators are better at, and prefer, some of these activities to others, and tend to emphasize their importance more.

In any case, defining evaluation in terms of what evaluators do presupposes that we have some independent way of identifying *evaluators*. But that is just what the definition of evaluation provides, so we cannot assume we already have it, or, if we do not, that we can avoid circularity through this approach.

It is also true that what evaluators do is to a significant extent driven by the current swings of fashion in the public, professional, or bureaucratic conceptions of what evaluation *should* do, since that determines how much private or public money is spent on evaluation. For example, fashion swings regularly occur about outcome-oriented evaluation – we're in the middle of a favorable one now – and so we currently find many evaluators dedicated to doing mere impact studies. These have many flaws by the standards of good evaluation, e.g., they rarely seek for side effects, which may be more important than intended or desired outcomes; and they rarely look closely at process, of which the same may be said. These are in fact merely examples of incomplete evaluations, or, if you prefer, of evaluation-related activities.

Note, second, that evaluation is not just the process of determining facts about things (including their effects), which, roughly speaking, we call research if it's difficult and observation if it's easy. An evaluation must, by definition, lead to *a particular type* of conclusion – one about merit, worth, or significance – usually expressed in the language of good/bad, better/worse, well/ill, elegantly/poorly etc. This constraint requires that evaluations – in everyday life as well as in scientific practice – involve three components: (i) the empirical study (i.e., determining brute facts about things and their effects and perhaps their causes); (ii) collecting the set of perceived as well as defensible values that are substantially relevant to the results of the empirical study, e.g., via a needs assessment, or a legal opinion; and (iii) integrating the two into a report with an evaluative claim as its conclusion. For example, in an evaluation of a program aimed to reduce the use of illegal drugs, the empirical study may show (i) that children increased their knowledge of illegal drugs as a result of the drug education part of the program,

which is (we document by means of a survey) widely thought to be a good outcome; and (ii) that they consequently increased their level of use of those drugs, widely thought to be a bad outcome. A professional evaluator, according to the definition, should do more than just report those facts. While reporting such facts is useful research, it is purely empirical research, partly about effects and partly about opinions. First, a further effort must be made to critique the values, e.g., for consistency with others that are held to be equally or more important, for the validity of any assumptions on which they are built, and for practicality, given our relevant knowledge. Second, we must synthesize all these results, mere facts and refined values, with any other relevant facts and values. Only these further steps can get us to an overall evaluative conclusion about the merit of the program. The values-search and values-critique part of this, and the synthesis of the facts with the values, are what distinguish the evaluator from the empirical researcher. As someone has pithily remarked, while the applied psychologist or sociologist or economist only needs to answer the question, "What's So?", the evaluator must go on to answer the question, "So What?"

In this case, the reason the knowledge about illegal drugs is thought to be good is usually that it is expected to lead to reduced use (a fact extracted from interviews, surveys, or focus groups with parents, lawmakers, police, and others). Hence the second part of the factual results here trumps the first part, since it shows that the reverse effect is the one that actually occurred, and hence the synthesis leads (at first sight) to an overall negative conclusion about the program. However, more thorough studies will look at whether *all* the consequences of the use of *all* illegal drugs are bad, or whether this is just the conventional, politically correct view. The social science literature does indeed contain a few good books written on that subject, which is scientifically-based values-critique; but the significance of these for the evaluation of drug education programs was not recognized. The significance was that they showed that it was perfectly possible to combine a scientific approach with critique of program goals and processes; but this was contrary to the existing paradigm and hence just ignored.

There's a further complication. It's arguable that good and bad should be taken to be implicitly defined by what the society *does* rather than what it *says*. That means, for example, that good should properly be defined so that alcohol and nicotine and morphine are acceptable for at least some adults in some situations, perhaps even good (in moderation) in a subset of these circumstances. With that approach, the overall evaluative conclusion of our program evaluation example may be different, depending on exactly what drugs are being taken by what subjects in what circumstances in the evaluation study. If we are to draw any serious conclusions from such studies, it is essential to decide and apply a defensible definition of social good and to answer the deeper questions as illustrated above. These are the hardest tasks of evaluation. Of course, these challenges doesn't come up most of the time since there is usually little controversy about the values involved, e.g., in learning to read, or in providing shelters for battered women and children. But it's crucial in many of the most important social interventions and policies. Avoidance of this obligation of evaluation vitiated or

rendered trivial or immoral the research of many hundreds, perhaps thousands, of social scientists who did not question the common assumptions on these matters, for example in the notorious case of social science support of dictators in South America.

These further steps into the domain of values, beyond the results of the empirical part of the study, i.e., going beyond the study of what people do value into the domain of what the evidence suggests they should value, were long held (and are still held by many academics) to be illicit – the kind of claims that could not be made with the kind of objectivity that science demands and achieves. This skeptical view, known as the doctrine of value-free science, was dominant throughout the twentieth century – especially in the social sciences. This view, although also strongly endorsed by extraneous parties – for example, most religious and political organizations, who wanted that turf for themselves – was completely ignored by two of the great applied disciplines, medicine and the law. For example, no doctor felt incapable of concluding that a patient was seriously ill from the results of tests and observations, although that is of course an evaluative conclusion that goes beyond the bare facts (it is a fact in its own domain, of course, but an evaluative fact). This legal/medical model (partly adopted in education and social work as well) would have been a better model for the social sciences, whose chosen theory about such matters, the value-free doctrine, rendered them incapable of addressing matters of poverty, corruption, and injustice because, it was said, the mere identification of those things, since the terms are value-impregnated, represented a non-scientific act. Consequently, many areas languished where social science could have made huge contributions to guiding interventions and improving interpretations, and people suffered and died more than was necessary. Ironically, many of those who despised excursions into the logic or philosophy of a domain, thinking of themselves as more practical for that choice, had in fact rested their efforts on one of the worst logical/philosophical blunders of the century, and thereby had huge and highly undesirable practical effects.

Great care is indeed needed in addressing the validity of value judgments, but science is no stranger to great care; nor, as we'll see in a moment, is it any stranger to objectively made value judgments. So neither of these considerations is fatal to scientific evaluation. The real problem appears to have been the desire to "keep science's nose clean", i.e., to avoid becoming embroiled in political, theological, and moral controversies. (It is clear that this is what motivated Max Weber, the originator of the value-free doctrine in the social sciences.) But getting embroiled in those issues is what it takes to apply science to real world problems, and balking at that challenge led to a century of failed service in causes that society desperately needed to press. To the credit of educational researchers, some of them followed the medical/legal model, which led to half a century of pretty serious evaluation of various aspects of education. In fact, to put it bluntly, educational research was several decades ahead of the rest of social science in the search for useful models of evaluation, and still is, to judge by most of the evaluation texts of the new millennium (see references). Sadly

enough, although this seems clear enough from a glance at the literature, it is rarely acknowledged by mainstream social scientists who have gotten into serious evaluation, a shoddy example of misplaced arrogance about the relative intellectual importance of education and the mainline social sciences.

Part of the explanation for the avant garde role of education in improving evaluation approaches may be due to three factors. First, a typical premier school of education, with its historians, philosophers, statisticians, and qualitative researchers, is remarkably interdisciplinary and less controlled by a single paradigm than the typical social science (or other science) department. Second, education is heavily committed to the application of its research, in this respect it is like engineering, medicine, and the law. And thirdly, it is usually an autonomous college, not so easily driven by the fashions espoused by fellow high-prestige fellow departments.

One must concede, however, that it was difficult to conceptualize what was going on, since most educational researchers are, appropriately enough in most respects, strongly influenced by social scientists as role models, and there was no help to be found from them. Not surprisingly, there emerged from this confused situation a remarkably diverse zoo of models, or theories of evaluation, or, we might equally well say, conceptions of evaluation. And since many people who came to do the evaluations that government funded had been brought up on the value-free doctrine, it is not surprising that that conception – it's really a denial of all models rather than a model in its own right – was very popular. This negative view was reconciled with actually doing evaluation, as many of the value-free doctrine's supporters did, by saying that evaluators simply gathered data that was relevant to decisions, but did not draw or try to draw any evaluative conclusions from it. This was "evaluation-free evaluation", perhaps the most bizarre inhabitant in the evaluation-models zoo. Let's now look in slightly more detail at this and some other evaluation models.

MODELS OF EVALUATION: EIGHT SIMPLIFIED ACCOUNTS

Evaluators play many roles in the course of doing what is loosely said to be evaluation, and, like actors, they sometimes fall into the trap of thinking that their most common role represents the whole of reality – or at least its essential core. There seems to be about eight usefully distinguishable cases in the history of evaluation in which this has happened. I list them here, calling them models in order to bypass complaints that they are mostly lacking in the detailed apparatus of a theory, and I append a note or two on each explaining why I see it as providing a highly distorted image of the real nature of evaluation. In most cases, the conception is simply a portrayal of one activity that evaluators often perform, one function that evaluation can serve – no more an account of evaluation's essential nature than playing the prince of Denmark provides the essence of all acting. Then I go on to develop the general theory that is implicit in these criticisms, one that I suggest is a more essential part of the truth than the others,

2
The CIPP Model for Evaluation

DANIEL L. STUFFLEBEAM
The Evaluation Center, Western Michigan University, MI, USA

This chapter presents the CIPP Evaluation Model, a comprehensive framework for guiding evaluations of programs, projects, personnel, products, institutions, and evaluation systems. This model was developed in the late 1960s to help improve and achieve accountability for U.S. school programs, especially those keyed to improving teaching and learning in urban, inner city school districts. Over the years, the model has been further developed and applied to educational programs both inside and outside the U.S. Also, the model has been adapted and employed in philanthropy, social programs, health professions, business, construction, and the military. It has been employed internally by schools, school districts, universities, charitable foundations, businesses, government agencies, and other organizations; by contracted external evaluators; and by individual teachers, educational administrators, and other professionals desiring to assess and improve their services.[1] This chapter is designed to help educators around the world grasp the model's main concepts, appreciate its wide-ranging applicability, and particularly consider how they can apply it in schools and systems of schools. The model's underlying theme is that evaluation's most important purpose is not to prove, but to improve.

Corresponding to the letters in the acronym CIPP, this model's core concepts are context, input, process, and product evaluation. By employing the four types of evaluation, the evaluator serves several important functions. Context evaluations assess needs, problems, and opportunities within a defined environment; they aid evaluation users to define and assess goals and later reference assessed needs of targeted beneficiaries to judge a school program, course of instruction, counseling service, teacher evaluation system, or other enterprise. Input evaluations assess competing strategies and the work plans and budgets of approaches chosen for implementation; they aid evaluation users to design improvement efforts, develop defensible funding proposals, detail action plans, record the alternative plans that were considered, and record the basis for choosing one approach over the others. Process evaluations monitor, document, and assess activities; they help evaluation users carry out improvement efforts and maintain accountability records of their execution of action plans. Product evaluations

International Handbook of Educational Evaluation, 31–62
T. Kellaghan, D.L. Stufflebeam (eds.)

identify and assess short-term, long-term, intended, and unintended outcomes. They help evaluation users maintain their focus on meeting the needs of students or other beneficiaries; assess and record their level of success in reaching and meeting the beneficiaries' targeted needs; identify intended and unintended side effects; and make informed decisions to continue, stop, or improve the effort.

According to the CIPP Model, evaluations should serve administrators, policy boards, military officers, and other clients; teachers, physicians, counselors, clinicians, engineers, social workers, and other service providers; students, parents, patients, and other beneficiaries; and funding organizations, regulatory bodies, and society at large. Evaluators should present their audiences with evaluations that help develop high quality, needed services and products; help identify and assess alternative improvement options; help assure high quality and ongoing improvement of services; certify the effectiveness of services and products; expose deficient, unneeded, and/or unsafe services and products; and help clarify the factors that influenced an enterprise's success or failure. Thus, the CIPP Model is oriented to administration, development, effective service, prevention of harm, accountability, dissemination, and research.

This chapter introduces the CIPP Model by presenting a general scheme to show relationships among the model's key components. Next, evaluation is defined. The chapter subsequently delineates the CIPP Model's improvement/formative and accountability/summative roles. It follows with a brief discussion of self-evaluation applications of the model. Following discussion of the model's use for improvement purposes, general guidance and an example checklist are provided for using the model for accountability purposes. Context, input, process, and product evaluation are next explained in some detail as applied mainly to group efforts; these explanations include a few cogent examples and a range of relevant techniques. The chapter is concluded with guidelines for designing the four types of evaluation. The Evaluation Center's[2] experiences in applying the model are referenced throughout the chapter.

A GENERAL SCHEMA

Figure 1 portrays the basic elements of the CIPP Model in three concentric circles. The inner circle represents the core values that provide the foundation for one's evaluations. The wheel surrounding the values is divided into four evaluative foci associated with any program or other endeavor: goals, plans, actions, and outcomes. The outer wheel denotes the type of evaluation that serves each of the four evaluative foci. These are context, input, process, and product evaluation.

Each double arrow denotes a two-way relationship between a particular evaluative focus and a type of evaluation. The task of setting goals raises questions for a context evaluation, which in turn provides information for validating or improving goals. Planning improvement efforts generates questions for an input evaluation, which correspondingly provides judgments of plans and

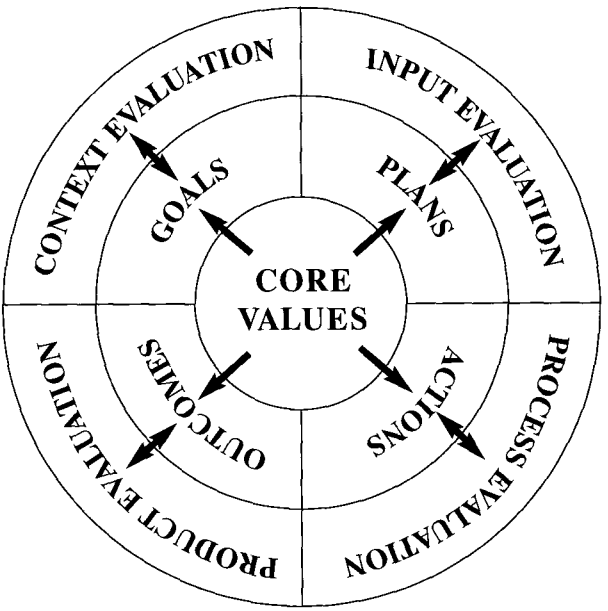

Figure 1: Key Components of the CIPP Evaluation Model and Associated Relationships with Programs

direction for strengthening plans. Improvement activities bring up questions for a process evaluation, which in turn provides judgments of actions and feedback for strengthening them. Accomplishments, lack of accomplishments, and side effects command the attention of product evaluations, which ultimately judge the outcomes and identify needs for achieving better results.

These reciprocal relationships are made functional by grounding evaluations in core values, as denoted by the scheme's inner circle. The root term in evaluation is *value*. This term refers to any of a range of ideals held by a society, group, or individual. Example values – applied in evaluations of U.S. public school programs – are students' meeting of state-defined academic standards, equality of opportunity, human rights, technical excellence, efficient use of resources, safety of products and procedures, and innovative progress. Essentially, evaluators assess the services of an institution, program, or person against a pertinent set of societal, institutional, program, and professional/technical values. The values provide the foundation for deriving the particular evaluative criteria. The criteria, along with questions of stakeholders, lead to clarification of information needs. These, in turn, provide the basis for selecting/constructing the evaluation instruments and procedures and interpreting standards. Evaluators and their clients must regularly employ values clarification as the foundation of their evaluation activities.

A FORMAL DEFINITION OF EVALUATION

The formal definition of evaluation underlying the CIPP Model is as follows:

> Evaluation is the process of delineating, obtaining, providing, and applying descriptive and judgmental information about the merit and worth of some object's goals, design, implementation, and outcomes to guide improvement decisions, provide accountability reports, inform institutionalization/ dissemination decisions, and improve understanding of the involved phenomena.

This definition summarizes the key ideas in the CIPP Model. The definition posits four purposes for evaluation: guiding decisions; providing records for accountability; informing decisions about installing and/or disseminating developed products, programs, and services; and promoting understanding of the dynamics of the examined phenomena. It says the process of evaluation includes four main tasks: delineating, obtaining, providing, and applying information. Hence, trainers should educate evaluators in such areas as systems thinking, group process, decision making, conflict resolution, consensus building, writing reports, communicating findings, and fostering utilization of evaluation results. To fully implement the evaluation process, evaluators also need technical training in collecting, processing, and analyzing information and in developing judgmental conclusions. The definition also notes that evaluators should collect both descriptive and judgmental information; this requires employment of both quantitative and qualitative methods. According to the definition, evaluations should assess goals, designs, implementation, and outcomes, giving rise to the needs, respectively, for context, input, process, and product evaluations. Also highlighted is the fundamental premise that evaluators should invoke the criteria of merit (the evaluand's quality) and worth (its costs and effectiveness in addressing the needs of students or other beneficiaries).

The CIPP Model also posits that evaluators should subject their evaluations and evaluation systems to evaluations and that such metaevaluations should invoke appropriate standards. The standards for judging evaluations that employ the CIPP Model go beyond the traditional standards of internal and external validity employed to judge research studies. The standards employed to judge CIPP evaluations of North American public school programs and personnel include utility, feasibility, propriety, and accuracy (Joint Committee, 1981; 1988; 1994). These standards are targeted to educational evaluations in the U.S. and Canada, but they provide examples that other countries can consider as they develop their own standards for educational evaluations.

THE CIPP MODEL'S IMPROVEMENT/FORMATIVE AND ACCOUNTABILITY/SUMMATIVE ORIENTATIONS

The CIPP Model is designed to serve needs for both formative and summative evaluations. CIPP evaluations are formative when they proactively key the

collection and reporting of information to improvement. They are summative when they look back on completed project or program activities or performances of services, pull together and sum up the value meanings of relevant information, and focus on accountability.

The relationships of improvement/formative and accountability/summative roles of evaluation to context, input, process, and product evaluations are represented in Table 1. This table shows that evaluators may use context, input, process, and product evaluations both to guide development and improvement of programs, projects, or materials – the formative role – and to supply information for accountability – the summative role. Based on this scheme, the evaluator would design and conduct an evaluation to help the responsible teachers, principals, or other service providers plan and carry out a program, project, or service. They would also organize and store pertinent information from the formative evaluation for later use in compiling an accountability/summative evaluation report.

While improvement/formative-oriented information might not answer all the questions of accountability/summative evaluation, it would help answer many of them. In fact, external evaluators who arrive at a program's end often cannot produce an informative accountability/summative evaluation if the project has no evaluative record from the developmental period. A full implementation of the CIPP approach includes documentation of the gathered formative evaluation evidence and how the service providers used it for improvement.

This record helps the external summative evaluator address the following questions:

1. What student or other beneficiary needs were targeted, how pervasive and important were they, how varied were they, how validly were they assessed,

Table 1. The Relevance of Four Evaluation Types to Improvement and Accountability

	Context	*Input*	*Process*	*Product*
Improvement/ Formative orientation	Guidance for choosing goals and assigning priorities	Guidance for choosing a program/ service strategy Input for specifying the procedural design, schedule, and budget	Guidance for implementation	Guidance for termination, continuation, modification, or installation
Accountability/ Summative orientation	Record of goals and priorities and bases for their choice along with a record of assessed needs, opportunities, and problems	Record of chosen strategy and design and reasons for their choice over other alternatives	Record of the actual process and its costs	Record of achievements, assessments compared with needs and costs, and recycling decisions

3
Responsive Evaluation

ROBERT STAKE

Center for Instructional Research and Curriculum Evaluation, University of Illinois at Champaign-Urbana, IL, USA

Responsive evaluation is an approach, a predisposition, to the evaluation of educational and other programs. Compared to most other approaches it draws attention to program activity, to program uniqueness, and to the social plurality of its people. This same predisposition toward merit and shortcoming can be built into or can be recognized in other approaches, such as a stakeholder evaluation or connoisseurship evaluation.

A responsive evaluation is a search and documentation of program quality. The essential feature of the approach is a responsiveness to key issues or problems, especially those recognized by people at the site. It is not particularly responsive to program theory or stated goals but more to stakeholder concerns. Its design usually develops slowly, with continuing adaptation of evaluation goal-setting and data-gathering in pace with the evaluators becoming well acquainted with the program and its contexts.

Issues are often taken as the "conceptual organizers" for the inquiry, rather than needs, objectives, hypotheses, or social and economic equations. Issues are organizational perplexities or complex social problems, regularly attendant to unexpected costs or side effects of program efforts. The term "issue" draws thinking toward the interactivity, particularity, and subjective valuing already felt by persons associated with the program. (Examples of issue questions: Are the eligibility criteria appropriate? Do these simulation exercises confuse the students about authoritative sources of information?) People involved in the program are concerned about one thing or another (or likely to become concerned). The evaluators inquire, negotiate, and select a few issues around which to organize the study.

The evaluators look for troubles and coping behavior. To become acquainted with a program's issues, the evaluators usually observe its activities, interview those who have some role or stake in the program, and examine relevant documents. These are not necessarily the data-gathering methods for informing the interpretation of issues; but are needed for the initial planning and progressive focusing of the study. And even later, management of the study as a whole usually remains flexible – whether quantitative or qualitative data are gathered.

International Handbook of Educational Evaluation, 63–68
T. Kellaghan, D.L. Stufflebeam (eds.)

OBSERVATIONS AND JUDGMENTS

Directed toward discovery of merit and shortcoming in the program, responsive evaluation study recognizes multiple sources of valuing as well as multiple grounds. It is respectful of multiple, even sometimes contradictory, standards held by different individuals and groups.

Ultimately the evaluators describe the program activity, its issues, and make summary statements of program worth. But first they exchange descriptive data and interpretations with data givers, surrogate readers, and other evaluation specialists for recognizing misunderstanding and misrepresentation. In their reports they provide ample description of activities over time and personal viewing so that, with the reservations and best judgments of the evaluators, the report readers can make up their own minds about program quality.

There is a common misunderstanding that responsive evaluation requires naturalistic inquiry, case study, or qualitative methods. Not so. With the program staff, evaluation sponsors and others, the evaluators discuss alternative methods. Often the clients will want more emphasis on outcomes, and responsive evaluators press for more attention on the quality of processes. They negotiate. But knowing more about what different methods can accomplish, and what methods this evaluation "team" can do well, and being the ones to carry them out, the evaluators ultimately directly or indirectly decide what the methods will be. Preliminary emphasis often is on becoming acquainted with activity, especially for external evaluators, but also the history and social context of the program. The program philosophy may be phenomenological, participatory, instrumental, or in pursuit of accountability. Method depends partly on the situation. For it to be a good responsive evaluation the methods must fit the "here and now", having potential for serving the evaluation needs of the various parties concerned.

Even so, it has been uncommon for a responsive evaluation study to emphasize the testing of students or other indicators of successful attainment of stated objectives. This is because such instrumentation has so often been found simplistic and inattentive to local circumstances. Available tests seldom provide comprehensive measures of the outcomes intended, even when stakeholders have grown used to using them. And even when possible, developing new tests and questionnaires right is very expensive. For good evaluation, test results have too often been disappointing – with educators, for example, probably justifiably believing that more was learned than showed up on the tests. With the responsive approach, tests often are used, but in a subordinate role. They are needed when it is clear that they actually can serve to inform about the quality of the program.

In most responsive evaluations, people are used more as sociological informants than as subjects. They are asked what they saw as well as what they felt. They are questioned not so much to see how they have changed but to indicate the changes they see.

SUBJECTIVITY AND PLURALISM

My first thoughts about how to evaluate programs were extensions of empirical social science and psychometrics, where depersonalization and objectivity were esteemed. As I have described elsewhere (Stake, 1998), in my efforts to evaluate curriculum reform efforts in the 1960s, I quickly found that neither those designs nor tests were getting data that answered enough of the important questions. Responsive evaluation was my response to "preordinate evaluation", prior selection and final measurement of a few outcome criteria. Over the years I came to be comfortable with the idea that disciplining impressions and personal experience led to better understanding of merit and worth than using needs to identify improvement with strict controls on bias (Stake et al., 1997).

Case study, with the program as the case, became my preferred way of portraying the activity, the issues, and the personal relationships that reveal program quality. Not all who have a predilection for responsive evaluation use a case study format. Many evaluators do their work responsively without calling it that and some who do call their work responsive are not responsive to the same phenomena I am. There is no single meaning to the term.

Those who object to the responsive approach often do so on the ground that too much attention is given to subjective data, e.g., the testimony of participants or the judgments of students. For description of what is happening, the evaluators try (through triangulation and review panels) to show the credibility of observations and soundness of interpretations. Part of the program description, of course especially that about the worth of the program, is revealed in how people subjectively perceive what is going on. Placing value on the program is not seen as an act separate from experiencing it.

The researchers' own perceptions too are recognized as subjective, in choosing what to observe, in observing, and in reporting the observations. One tries in responsive evaluation to make those value commitments more recognizable. Issues, e.g., the importance of a professional development ethic, are not avoided because they are inextricably subjective. When reporting, care is taken to illuminate the subjectivity of data and interpretations.

Objection to a responsive approach is also expressed in the belief that a single authority, e.g., the program staff, the funding agency or the research community, should specify the key questions. Those questions often are worthy of study, but in program evaluation for public use, never exclusively. There is general expectation that if a program is evaluated, a wide array of important concerns will be considered. Embezzlement, racial discrimination, inconsistency in philosophy, and thwarting of creativity may be unmentioned in the contract and not found in the evaluators' expertise, but some sensitivity to all such shortcomings belong within the evaluation expectation, and the responsive evaluator at least tries not to be blind to them.

Further, it is recognized that evaluation studies are administratively prescribed, not only to gain understanding and inform decision-making but also to legitimatize and protect administrative and program operations from criticism,

especially during the evaluation period. And still further, evaluation requirements are sometimes made more for the purpose of promulgating hoped-for standards than for seeing if they are being attained. Responsive evaluators expect to be working in political, competitive, and self-serving situations and the better ones expose the meanness they find.

By seeking out stakeholder issues, responsive evaluators try to see how political and commercial efforts extend control over education and social service. They are not automatically in favor of activist and legitimate reform efforts, but they tend to feature the issues they raise. Responsive evaluation was not conceived as an instrument of reform. Some activists find it democratic; others find it too conservative (Shadish, Cook, & Leviton, 1991). It has been used to serve the diverse people most affected personally and culturally by the program at hand – though it regularly produces some findings they do not like.

ORGANIZING AND REPORTING

The feedback from responsive evaluation studies is expected to be in format and language attractive and comprehensible to the various groups, responsive to their needs. Thus, even at the risk of catering, different reports or presentations may be prepared for different groups. Narrative portrayals, story telling, and verbatim testimony will be appropriate for some; data banks and regression analyses for others. Obviously the budget will not allow everything, so these different communications have to be considered early in the work.

Responsive evaluation is not participatory evaluation, but it is organized partly around stakeholder concerns and it is not uncommon for responsive evaluation feedback to occur early and throughout the evaluation period. Representatives of the prospective audience of readers should have directly or indirectly helped shape the list of issues to be pursued. Along the way, the evaluator may ask, "Is this pertinent?" and "And is this evidence of success?" and might, based on the answer, change priorities of inquiry.

Responsive evaluation has been useful during formative evaluation when the staff needs more formal ways of monitoring the program, when no one is sure what the next problems will be. It has been useful in summative evaluation when audiences want an understanding of a program's activities, its strengths and shortcomings and when the evaluators feel that it is their responsibility to provide a vicarious experience. Such experience is seen as important if the readers of the report are to be able to determine the relevance of the findings to their own sense of program worth.

As analyzed by Ernest House (1980, p. 60) responsive evaluation will sometimes be found to be "intuitive" or indeed subjective, closer sometimes to literary criticism, Elliot Eisner's connnoisseurship, or Michael Scriven's *modus operandi* evaluation than to the more traditional social science designs. When the public is seen as the client, responsive evaluation may be seen as "client centered", as did Daniel Stufflebeam and Anthony Shinkfield (1985, p. 290). But usually it

differs from those approaches in the most essential feature, that of responding to the issues, language, contexts, and standards of an array of stakeholder groups.

When I proposed this "responsive evaluation" approach (at an evaluation conference at the Pedagogical Institute in Göteborg, Sweden, in 1974) I drew particularly upon the writings of Mike Atkin (1963); Lee Cronbach (1963); Jack Easley (1966); Stephen Kemmis (1976); Barry MacDonald (1976); and Malcolm Parlett and David Hamilton (1977). They spoke of the necessity of organizing the evaluation of programs around what was happening in classrooms, drawing more attention to what educators were doing and less attention to what students were doing. Later I reworked some of my ideas as I read Ernest House (1980); Egon Guba and Yvonna Lincoln (1985); Tom Schwandt (1989); and Linda Mabry (1998). Of course I was influenced by many who proposed other ways of evaluating programs.

It is difficult to tell from an evaluation report whether or not the study itself was "responsive." A final report seldom reveals how issues were negotiated and how audiences were served. Examples of studies which were clearly intentionally responsive were those by Barry MacDonald (1982); Saville Kushner (1992); Anne McKee and Michael Watts (2000); Lou Smith and Paul Pohland (1974); and Robert Stake and Jack Easley (1979), indicated in the references below. My meta-evaluation, *Quieting Reform* (1986), also took the responsive approach.

REFERENCES

Atkin, J.M. (1963). Some evaluation problems in a course content improvement project. *Journal of Research in Science Teaching*, **1**, 129–132.

Cronbach, L.J. (1963). Course improvement through evaluation. *Teachers College Record*, **64**, 672–683.

Easley, J.A., Jr. (1966). *Evaluation problems of the UICSM curriculum project*. Paper presented at the National Seminar for Research in Vocational Education. Champaign, IL: University of Illinois.

Greene, J.C. (1997). Evaluation as advocacy. *Evaluation Practice*, **18**, 25–35.

Guba, E., & Lincoln, Y. (1981). *Effective evaluation*. San Francisco: Jossey-Bass.

Hamilton, D., Jenkins, D., King, C., MacDonald, B., & Parlett, M. (Eds.). (1977). *Beyond the numbers game*. London: Macmillan.

House, E.R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage.

Kushner, S. (1992). *A musical education: Innovation in the Conservatoire*. Victoria, Australia: Deakin University Press.

Mabry, L. (1998). *Portofios plus: A critical guide to alternative assessment*. Newbury Park, CA: Corwin Press.

MacDonald, B. (1976). Evaluation and the control of education. In D.A. Tawney (Ed.) *Curriculum evaluation today: Trends and implications*. London: Falmer.

MacDonald, B., & Kushner, S. (Eds.). (1982). *Bread and dreams, CARE*. Norwich, England: University of East Anglia.

McKee, A., & Watts, M. (2000). *Protecting Space? The Case of Practice and Professional Development Plans*. Norwich, England: Centre of Applied Research in Education, University of East Anglia.

Parlett, M., & Hamilton, D. (1977). Evaluation as illumination: A new approach to the study of innovatory programmes. In D. Hamilton, D. Jenkins, C. King, B. MacDonald, & M. Parlett (Eds.), *Beyond the numbers game*. London: Macmillan.

Schwandt, T.A. (1989). Recapturing moral discourse in evaluation. *Educational Researcher*, **18**(8), 11–16.

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