5. OUTCOME EVALUATION

THOMAS KELLAGHAN and GEORGE F. MADAUS

The first edition of *Evaluation Models* did not contain a chapter on outcome evaluation. Why is there one in this edition? After all, the idea of measuring outcomes is not new and, as we shall see, outcome evaluation can hardly be regarded as a unitary approach, given the variety of practices encompassed by the term. Nor can it really be considered a *model,* if by model we mean a more or less elaborate representation of the structure and relationships of a range of phenomena. Some would say it is not even evaluation. However, there is still good reason for including in this volume a description of activities that can be broadly categorized as outcome evaluation, since they now account for a considerable amount of program monitoring activities throughout the world, in some cases displacing more traditional approaches to evaluation and research, both in countries with long-established traditions in these disciplines and ones where formal evaluation activities are only being developed. Outcome evaluation has received the backing and financial support of governments as well as of international organizations, such as the European Union, the Organisation for Economic Co-operation and Development, UNESCO, and the World Bank.

In this chapter, we shall first describe the characteristics of outcome evaluation. Following that, we shall outline reasons for its growth and advantages attributed to its use. We shall then identify a number of traditions and developments to which current practice in outcome evaluation is indebted, followed by examples of outcome evaluation at state, national, and international levels. After that, we shall consider approaches used in outcome evaluation. In our concluding remarks, we

shall outline a number of issues raised by outcome evaluation, and consider how it fits among traditional approaches. Most of our illustrative material will come from the field of education, where outcome evaluation probably has had its greatest impact. But such evaluation is by no means confined to education.

WHAT IS OUTCOME EVALUATION?

A number of features of outcome evaluation can be identified. Firstly, it is a term that is applied to activities that are designed primarily to measure the (often presumed) effects or results of programs, rather than their inputs or processes. Second, since more than measurement is required if an activity is to be regarded as evaluative, a judgment as to where a product lies with respect to a standard is often made. Thus, outcomes may be related to a target, standard of service, or achievement. Often the idea of "excellence" is used or implied. The widespread use of the nebulous term "world class standards" by those in the standards-based reform movement in the U.S. is typical of this accent on excellence. Sometimes the judgment of merit or worth is implicit rather than explicit. An implicit judgment is involved when information on outcomes (e.g., the mean achievement level of students in a school) is normative (e.g., indicating where a school stands relative to other schools) and it is left to clients and the public to make the evaluative judgment and, perhaps, to take action.

Third, the range of outcomes that have been used in outcome evaluation is considerable. Within the field of education, academic achievement is the outcome most frequently assessed, and a variety of performance and portfolio modes have been employed with mixed success. Most states now employ writing samples and these have been more successful. Other performance and portfolio assessments, however, have proved to be inefficient, costly, and unreliable. Kentucky had to drop its performance assessment, while Vermont had to rethink its reliance on portfolios (Kortez, 1994; Koretz, Barron, Mitchell, & Stecher, 1996). Other outcomes have also been considered relating to building, educational materials, teaching, attitudes to school, learning motivation, and change in use of a service (student retention rates, absenteeism, and students' post-school destinations). Fourth, the effects or results that are the focus of outcome evaluation may be observed at varying points in a program—during its life, at it completion, or later in time to assess long-term effects. Most frequently, the focus is on outcomes at the completion of a program.

Fifth, it is not usual in outcome evaluation to seek to describe or specify what is actually happening in a program, though the kind of information obtained will obviously, in general terms at least, be chosen to reflect program activities. In many circumstances in which outcome evaluation is used, a description of program activities would be very difficult, if at all possible. This is because many programs are extremely complex and can only be considered programs in the broadest sense of the word (e.g., elementary education). Such programs are perhaps more accurately described as complexes of programs, which are implemented in a variety of ways, and for which the term system might be more appropriate.

Sixth, while outcome evaluations may eschew descriptions of program activities,

efforts may be made to relate outcomes to contextual factors or to presumed relevant antecedent variables. Evaluations vary greatly in the extent to which they attempt to do this, and, later in the paper, we shall refer to analytical techniques used to address the issue. When such techniques are used, the main purpose is to distinguish in outcome data between the gross and net effects of program activity. It is important to do this if outcome information is to be used, as it frequently is, in the management of resources, in control, for quality assurance, or for accountability purposes (e.g., to recognize and attach sanctions to the performance of institutions or individuals with responsibility for the implementation of a program).

Finally, outcome evaluation may be once-off or may involve monitoring (i.e., comparisons of outcomes over time). When integrated into a performance management system, it is likely to be the latter, since it has to fit into an ongoing activity.

REASONS FOR GROWTH IN OUTCOME EVALUATION

A number of reasons can be identified for growth in outcome evaluation. First, from an historical point of view, the 1966 Equal Educational Opportunity Survey, commonly called the Coleman report, moved the attention of educational policymakers away from a definition of equal educational opportunity in terms of school resources toward a focus on educational outcomes as measured by tests (Coleman et al., 1966). A second reason is the perceived poor record of traditional evaluation approaches in providing direction for policymakers in making decisions about the large number of public programs that have been developed since the 1960s. Short-term readily applicable solutions did not seem to be forthcoming from such evaluation (Radaelli & Dente, 1996), while many evaluations were perceived to be costly, slow, and complex, not paying sufficient attention to outcomes.

A third reason for the growth in outcome evaluation is the development of a corporatist approach to government administration, signaled by a rise in "managerialism."The approach is heavily influenced by ideas from the business world, involving strategic and operational planning, the use of performance indicators, a focus on "deliverables"/results, a growth in incentive and accountability systems based on results (e.g., performance-related pay), and the concept of the citizen as consumer (Davies, 1999). In this situation, "the gentlemanly cult of the amateur administration", as Pollitt (1993) has observed, is being displaced, and its successor is "managerialism, not professional evaluation and analysis" (p. 354). The management consultant is expected to be able to provide the quick, narrow-focused analysis that is needed.

A fourth reason for growth in outcome evaluation is the increasing influence of the accounting and audit community in non-financial areas of public administration. The influence is reflected in "comprehensive audits", "value for money audits", "performance audits", and "environmental audits." In a variety of countries today (e.g., the United States, the United Kingdom, New Zealand, Sweden), audits of performance indicators are carried out, and opinions are issued on the extent to which systems or programs are meeting indicator targets (Davies, 1999).

Fifthly, growth in outcome evaluation reflects increasing use of assessment as a policy tool. In the field of education, this involves a shift from the use of assessment information for localized instructional decision making to centralized high stakes policy making and accountability monitoring (Madaus & Raczek, 1996).

Sixth, the growth of outcome evaluation owes much to a reorganization of the public service in several countries, resulting in the use of relatively autonomous service providers (e.g., National Health Service trusts and grant-maintained schools in Britain). With decentralization of program authority, and the consequent loss of direct control over the implementation of programs, the need arose for new contractual arrangements with service providers and for regulation and compliance monitoring. "Quality" and "standards" are the theme terms, and evaluation arrangements are designed to check that organizations are delivering flexible, cost-effective services to citizen users (Pollitt, 1993).

Finally, a situation in which growth in demand for public services and social program funding (e.g., education, health care, social security) is growing more rapidly than resources can be found for expansion leads to the need for greater efficiency, which in turn calls for selectivity in deciding what programs are to be continued and what new activities are to be launched (Blalock, 1999; Duran, Monnier, & Smith, 1995; Pollitt, 1993).

THE VALUE OF OUTCOME EVALUATION

Several advantages have been attributed to the use of outcome evaluation. One is based on business experience, where well-articulated goals are associated with organizational effectiveness. The situation in schools, which are notorious in lacking such goals, stands in strong contrast to this. It is argued that if schools were to specify outcomes relating to goals, this would identify what is important, and would help focus teachers and students on essential curriculum content (see Schmidt, McKnight, & Raizen, 1996). It is also the position of advocates of outcome evaluation that the specification of outcomes is likely to have a greater impact when aligned with appropriate assessment. This orientation toward specifying outcomes of schooling is at the heart of the standards-based reform movement. Various states have developed curriculum frameworks that mandate, first, academic learning standards by grade and subject area, and second, assessments to measure achievement related to these frameworks. For example, the Massachusetts Comprehensive Assessment System (MCAS), a new assessment program for public schools, "measures the performance of students, schools, and districts on the academic learning standards contained in the Massachusetts Curriculum Frameworks, fulfilling requirements of the Education Reform law of 1993" (Massachusetts Department of Education, 1998, p. 1).

The driving force behind many state reform efforts would appear to be the coupling of rewards or sanctions to performance on the statewide test. Policymakers are aware that testing programs that have the greatest impact on the curriculum, instruction, and learning are ones that students, teachers, administrators, parents, or the general public perceive as having sanctions or high-stakes associated with them

(Madaus & Kellaghan, 1992). In 1999, 33 of the United States had or shortly will have high stakes (e.g., high school graduation, ending social promotion) attached to their tests, while 14 states link moderate stakes (e.g., a special diploma) to their assessment systems. Sanctions may involve financial considerations for districts, schools, or teachers. Sometimes, however, the mere publication of outcome information is considered a sanction. There would appear to be two principles underlying the use of sanctions. First, individuals and institutions that are subject to sanctions will take action to obtain rewards and avoid punishment. Second, if information on outcomes is brought into the public domain, principles of competition will come into operation, and, as in the commercial world, those that do well will thrive, those that do poorly will wither away.

ORIGINS OF OUTCOME EVALUATION

The rationale for, and practice of, outcome evaluation owe a debt to at least six sources: traditional evaluation, traditions of assessment in education, school effectiveness and education production function research, the performance management movement, accountability concerns, and technical developments.

Traditional Evaluation

A consideration of the outcomes of programs is an integral feature of many traditional approaches to evaluation, and, up to the 1970s, educational evaluations focused primarily on assessing program outcomes. The emphasis on outcomes is most obvious in objectives-oriented evaluation approaches. Tyler (1949), for example, focused on educational objectives and their measurement in the context of curriculum evaluation. Other approaches in the Tylerian tradition also accorded prominence to the specification of objectives and judgments of the extent to which they could be said to have been achieved on the basis of program outcome data (e.g., Provus, 1971). However, these approaches differed from many current outcome evaluation efforts in Unking program objectives to the goals or objectives of individual schools or teachers rather than to statewide curriculum frameworks, while outcomes were not used for high stakes decisions or for accountability purposes.

Traditions of Assessment in Education

Few people would disagree with the view that the outcomes of education are important. However, agreement would not be as widespread on the relative importance of outcomes, since individuals differ in their perceptions of the prominence that should be given to the variety of goals or objectives that have been posited for schooling. Literacy and numeracy skills are usually accorded particular importance, and the use of information on outcomes to make decisions about the effectiveness of schools and teachers, based on students' acquisitions of these skills, reaches back into the last century. Perhaps the best-known examples of this approach are payment-by-results schemes which were introduced into British schools in 1862 to help improve students' literacy and mathematical skills and teacher efficiency, while at the same time saving money. In these schemes, the allocation of funds to schools

was linked to students' achievements as measured by written and oral examinations in reading, writing, and arithmetic. Responsibility for the failure of students was placed on the shoulders of teachers.

Growth in the use of standardized testing in this century, especially in the United States, reflects continuing interest in the outcomes of education. Rice's (1897) work on spelling is an early example of outcome evaluation. Information on outcomes, of course, has been used for a variety of purposes, only some of which related to the evaluation of programs or even of schools. Tests were most frequently used to assess the performance of individual students. On the basis of their value in this context, however, Coleman and Karweit (1972) proposed that they could also be used to provide measures of school performance in evaluating "educational environments."

Over the past three decades, standardized tests have been used increasingly as instruments of national education reform. Their use in diagnosing what is wrong in education, together with the legislative attention which testing has received, reflect a fundamental shift in the official education world, not only in the purpose for which standardized tests are used, but also in perceptions of quality which have moved from a consideration of school facilities, resources, and conditions to the outcomes of schooling (Madaus & Raczek, 1996). A recent illustration of the extent to which outcomes have become a prominent concern of policymakers is to be found in President Bush's America 2000 proposal (US Departments of Education and Labor, 1993) that paved the way for the Educate America Act of 1994. This legislation proposed that new American Achievement Tests should form part of a 15-point accountability package designed to encourage parents, schools, and communities "to measure results, compare results, and insist on change when the results aren't good enough" (Goals 2000: Education America Act, 1994). This legislation was never implemented and the idea of a "voluntary" national test is still on hold. Nonetheless, many states have adopted the central ideas in the legislation in designing their own standards-based reform programs.

School Effectiveness and Education Production Function Research

A large number of studies of school effectiveness and of education production function research has used measures of educational outcomes, usually standardized tests, in their efforts to determine characteristics of effective schools. An input-output representation of schooling was the model most frequently employed: student achievement at a point in time was related to a series of inputs, usually identified as family and background influences, school resources, and school characteristics (e.g., current expenditure, teacher qualifications and experience, pupil-teacher ratio) (see Hanushek, 1997; Madaus, Airasian, & Kellaghan, 1980).

In line with this tradition, several approaches to outcome evaluation collect data on input in an effort to identify factors associated with student achievement. The use of indicators (which might be described as statistics with evaluative relevance) in outcome evaluation fits particularly well with the input-output conceptualization

of schooling. Reflecting the input-output model, indicators used by the National Center for Education Statistics of the U.S. Department of Education now include context and outcome data (Stern, 1986). At the international level, OECD (1997) in describing the education systems of member countries has, during the 1990s, used indicators to describe the demographic, social, and economic context of education, financial and human resources invested in education, the learning environment and the organization of schools, and student achievement.

The Performance Management Movement

Sensitivity to the needs of program managers and decision makers is not new in evaluation. Stufflebeam (1983), for example, considered that the decision that had to be made, rather than program objectives, should be the key concern of the evaluator. Current interest in the use of evaluation findings for management decisions has a rather different origin, however: performance management, which has its roots in the 1930s but grew in popularity in the late 1980s and in the 1990s alongside more established evaluation approaches. While the general aims of performance management "to base judgments of the effectiveness of program efforts on more appropriate and trustworthy information, and to improve these efforts" (Blalock, 1999, p. 118) do not differ from the aims of many more traditional evaluation approaches, concepts underlying performance management differ from such approaches in a number of ways.

While traditional evaluation grew out of social science research, adopting its basic concepts and techniques, performance management has its roots in a bureaucratic environment. It is based on planning and management ideas, particularly ones relating to quality assurance, customer satisfaction, and continuous improvement. It involves defining performance in terms of results, setting performance targets, determining the extent to which results are achieved using performance indicators, and basing resource allocation decisions on performance information. Its aim is to provide rapid and continuous feedback on a limited number of outcome measures that are perceived to be of interest to policymakers, administrators, stakeholders, politicians, and customers, and to be of value in making decisions (Blalock, 1999; Davies, 1999). The manager, not the "scientific" policy analyst, is the charismatic figure; efficiency and economy are the main concerns; and the achievement of performance targets is the sign of "administrative health" (Pollitt, 1993).

It was in this context that management information systems (MIS) grew in the 1980s, designed to specify the structures and procedures governing the collection, analysis, presentation, and use of information in organizations. The development was, at least in part, a response to the need to monitor the growth and increasing complexity of systems and to justify decisions about resource allocation. Outcome evaluation fits readily into this picture by providing relatively simple statistical information about a system, program, or activity on a timely basis. While more or less complex analyses may accompany this information in some evaluations, it is not the primary purpose of outcome evaluation to provide them.

Acountability

In recent years, accountability has achieved increasing prominence in government administrations in many countries. Measures to control how stakeholders discharge their obligations have been devised as a mechanism for dealing with issues which arise from a number of phenomena: increasing demand for services coupled with diminishing resources; a multiplication of reform strategies; weak administrative instruments; and competing values and demands in pluralist cultures. These measures, which have been applied to a range of public services, might seem a reasonable way to bring order to complex and poorly understood environments. It is envisaged that information based on the measures would lead to the use of administrative controls over the use of inputs to ensure that specified procedures are complied with. But it might also simply involve the identification of products that meet a specified standard and products that do not. It is regarded as a relatively simple and straightforward task to use data from an outcome evaluation to place the onus for change and adjustment on the person or institution identified as being accountable, and to place one's trust in the operation of a competitive market and the threat or promise of sanctions to bring about the desired effect. In this situation, the onus is not on a manager to identify desirable aspects of implementation or conditions that need to be changed. He or she does not have to try to understand or explain why some individuals or institutions are "effective" and some are not. All that is necessary is to identify the effective and the noneffective, and to have statistical data to support the judgment.

Despite problems associated with outcome evaluation considered below, accountability issues loom large in considerations of school reform today. For example, the Educational Improvement Act adopted in Tennessee in 1991 created the need to specify the means by which teachers, schools, and school systems could be held accountable for meeting objectives set for Tennessee's education systems. Since the focus was on product rather than on process, an outcomes-based assessment system was established and has been embedded in the Tennessee Value Added Assessment System (TVAAS) which forms an integral part of legislation (Sanders & Horn, 1994).

Technical Developments

The availability of relatively low cost technologies with massive computing capabilities has greatly aided the development, not only of large-scale testing programs to obtain outcome data, but also of management information systems in general and logistical planning. Outcome evaluation is greatly facilitated by the capacity to store vast amounts of data, to link data collected at different points in time, and to carry out sophisticated statistical analyses.

THE USE OF OUTCOME EVALUATION

The tendency for governments to take responsibility for quality by setting standards and monitoring scholastic achievement, coupled with an allocation of responsibility

for the use of resources/inputs to providers, can be found in a wide range of countries. This is a change from a situation in which, up to recently, monitoring and evaluation systems were more concerned with resources and implementation than with assessing results. In many countries, aspects of performance measures are now underwritten by legislation.

In the United States, the Government Performance and Results Act (GPRA) of 1993 was implemented in October 1997 as a response to reports of waste and inefficiency in government spending. To restore public confidence in government, all federal agencies would be held accountable for achieving program results, service quality, customer satisfaction, and for providing Congress with sufficient information to improve decision making. Performance measurement would be required and the resulting data would be made public. A range of publications providing a rationale for, and description of, performance measurement ("managing for results"), as well as experience in its use has been prepared by the U.S. General Accounting Office and other agencies (http://www.reeusda.gov/part/gpra/gpralist.htm).

Major changes have occurred in government agencies following the legislation. For example, the United States Agency for International Development (USAID) has developed for its funded projects a "results framework" which involves specification of goals, objectives, indicators with periodic targets, intermediate results, and long-term net results (representing the effect of the intervention) (Toffolon-Weiss, Bertrand, & Terrell, 1999).

Evaluation activity outside the United States is not well documented. However, it seems reasonable to say that the extent or range of evaluation activities found in the United States is not found elsewhere, despite a recent surge of evaluation activity, or at least a recognition of its need, in many countries. In Spain, for example. government has responded to legislation requiring evaluation following government action in contracting services, creating conditions for competition, and raising the issue of accountability. The response reflects a preference for evaluation approaches that are compatible with the production of management control indicators and are useful in informing decision making in the policy process. For example, the Catalan Health Services Administrative Office monitors populations served, cost, and outputs (e.g., number of visits per inhabitant per day, number and cost of prescriptions) (Ballart, 1998).

Use of evaluation (through usually of a rather old fashioned variety) has also grown rapidly in other countries during the 1980s and 1990s. In Denmark, traditional empirical methodologies (usually surveys) to provide data for political and organizational development, control, monitoring, and modernization are favored (Hansson, 1997). In France, "widespread infatuation with public policy evaluation" as a means of modernizing public service has been reported (Duran, Monnier, & Smith, 1995, p. 45). In Italy, demands to produce an evaluation framework for recent reforms in health services (*aziettda lizzazione della sanita publica*) have resulted in tensions between an approach focused on management and one more oriented to effectiveness and quality assessment. Norway also seems to be showing signs of increasing enthusiasm for evaluation, though issues have not yet developed with the sharpness

of focus observable in Anglo-Saxon countries (see News from the Community, *Evaluation,* 1998, *4,* 373–379). In the Russian Federation, the requirement of a uniform curriculum in schools is being replaced by greater autonomy for regional authorities and schools in conjunction with outcome-based curricula (Bakker, 1999).While the evaluation ambitions of many countries seem less than modest, realization is being hampered by lack of data, expertise, instruments, and the infrastructure required for large-scale data collection and analysis. This point has been made regarding the development of evaluation in the People's Republic of China, where evaluation was unknown up to the early 1980s, but is now seen to be important in the context of national development and economic growth. Many steps are being taken to improve the country's evaluation capacity (Hong & Rist, 1997).

We turn now to descriptions of specific outcome evaluation efforts in education at state level (U.S.), national level, and international level.

Outcome Evaluation at State Level

In the United States, state departments of education are the major players in outcome evaluation, collecting data on student achievement, publishing the data, and allowing comparisons to be made between schools and school districts.

In Texas, for example, outcome data are provided at all grade levels for a range of variables including academic achievement, student promotion rate, student attendance, dropout rate, percentage taking the Scholastic Aptitude Tests, and post-school college enrolment rate. Cash rewards to schools and to individual professional staff are given to schools that provide test data for 95 percent of eligible students and in which at least half its cohorts perform better than a norm group (Webster, Mendro, & Almaguer, 1994).

The Tennessee Value-Added Assessment System (TVAAS) is also an outcomes-based system, in which the focus of accountability is on the product of the educational experience, not the process. The TVAAS has been adopted and legislated for in state law. According to *The Master Plan for Tennessee Schools 1993* of the State Board of Education, "State and local education policies will be focused on results; Tennessee will have assessment and management information systems that provide information on students, schools, and school systems to improve learning and assist policy making" (cited in Sanders & Horn, 1994, p. 301). Testing takes place at all grade levels in reading, mathematics, science, language, and social studies. Judgments are made on the basis of the data that are collected on the effects of school systems, individual schools, and individual teachers. Data on the first two are released to the public.

Outcome Evaluation at National Level

The most obvious exemplars of outcome evaluation at national level are "national assessments", which have operated in the United Kingdom in one form or another since 1948, in the United States since 1969, and in France since 1979. The United States National Assessment of Educational Progress (NAEP) is the most widely reported assessment model in the literature. It is an ongoing survey, mandated by

the U.S. Congress and implemented by trained field staff, usually school or district personnel. The survey is designed to measure students' educational achievements at specified ages and grades and reports the percentage of students scoring in the three controversial performance categories: "basic", "proficient", and "advanced". It also examines achievements of subpopulations defined by demographic characteristics and by specific background experience. Over the years, details of the administration of NAEP have changed; for example, in the frequency of assessment and in the grade level targeted. At present, assessments are conducted every second year on samples of students in grades 4, 8, and 12. Eleven instructional areas have been assessed periodically. Most recent reports have focused on reading and writing, mathematics and science, history, geography, and civics. Data have been reported by state, gender, ethnicity, type of community, and region.

National assessments are now a feature of many other education systems throughout the world, not only in industrialized countries (e.g., Australia, Canada, Finland, France, Ireland, the Netherlands, Norway, Sweden, New Zealand, United Kingdom) but also in developing countries (see Chinapah, 1997; Greaney & Kellaghan, 1996). An assessment of students' first language and mathematics at the elementary school level is included in all national assessments. Science is included in some, and a second language, art, music, and social studies in a small number. In most countries, data are collected for a sample of students at a particular age or grade level, but in some countries, all students at the relevant age or grade level are assessed (Kellaghan & Grisay, 1995).

Outcome Evaluation at International Level

International assessments differ from national assessments in that they involve measurement of the outcomes of education systems in several countries, usually simultaneously. Representatives from many countries (usually from research organizations) agree on an instrument to assess achievement in a curriculum area, the instrument is administered to a representative sample of students at a particular age or grade level in each country, and comparative analyses of the data are carried out (Kellaghan & Grisay, 1995). The main advantage of international studies over national assessments is the comparative framework they provide in assessing student achievement and curricular provision. International assessments give some indication of where the students in a country stand relative to students in other countries. They also show the extent to which the treatment of common curriculum areas differs across countries, and, in particular, the extent to which the approach in a given country may be idiosyncratic. This information may lead a country to reassess its curriculum policy.

The International Association for the Evaluation of Educational Achievement (IEA) has pioneered international assessment studies and has carried out a series of studies of school achievement, attitudes, and curricula in a variety of countries since 1959. Although one of lEA's primary functions is to conduct research designed to improve understanding of the educational process, studies were also intended to have a more practical and applied purpose: to obtain information relevant to policy-

making and educational planning in the interest of improving education systems (Husén, 1967; Postlethwaite, 1987).

To date, the IEA has conducted studies of mathematics achievement, science achievement, reading literacy, written composition, English as a foreign language, French as a foreign language, civic education, computers in education, and preprimary childcare. Levels and patterns of achievement have been described and compared across countries. So also have differences in intended and implemented curricula and in the course-taking patterns of students. A variety of correlates of achievement has been identified, including students' opportunity to learn, the amount of time a subject is studied, the use of computers, and resources in the homes of students.

APPROACHES IN OUTCOME EVALUATION

A variety of approaches, depending on the outcome to be assessed, has been used in outcome evaluation. In evaluations in the field of education, assessments of student achievement usually involve the administration of tests or examinations. The performances of individual students may then be aggregated to the level of the teacher, school, district, state, or even nation to allow judgments to be made about achievement at the desired level.

Judgments may be made on the basis of unadjusted results. In British league tables, the percentages of students in schools awarded varying grades on public examinations ("performance tables") have been published since 1992. In the United States, most state accountability systems in the past compared schools and school districts on the basis of unadjusted outcome measures (Guskey & Kifer, 1990). Similarly, in international comparative studies, countries are ranked on the basis of unadjusted mean scores.

This procedure is perhaps not surprising if outcome evaluation is concerned primarily with description, not explanation, with the product of the educational experience, not the process by which it was achieved. There is, however, concern about the extent to which such comparisons are fair, particularly if evaluation results are used for accountability purposes. The issue at stake is that of distinguishing between the "net" impact of a program which represents outcomes that are directly attributable to the program, and "gross" impact which reflects, in addition to net impact, influences other than the program being monitored. The distinction is readily illustrated in the case of student achievements, which are generally recognized as reflecting a variety of influences, including genetic endowment, achievement on entering school, and the support and assistance that students receive at home and in the community, all of which may be independent of school and teacher influences (Sanders & Horn, 1994; Webster et al., 1994). If students differ from school to school in their levels of achievement when entering a school, measures of absolute levels of student achievement at a later date may not adequately reflect a school's success in moving students from their initial entry levels. However, it seems reasonable to say that schools and institutions should be held accountable only for things that they can be

expected to influence, not for the characteristics students bring with them when they come to school (Woodhouse & Goldstein, 1996).

In line with this thinking, several attempts have been made to develop statistical methodologies that will permit an assessment of the contributions of schools to student development in situations in which the nonrandom assignment of students is assumed. These methodologies are based on two concepts. One relates to "normal" academic progression, which is the average progression that students make from a given starting point over a particular period in the school system (described as "expected" progress).The other is related to the extent to which individual students or groups of students (e.g., in a class or school) exceed or fall below that average progress in the specified time period. The difference is regarded as representing the value which a particular class or school has "added" to students' progress.

Statistical procedures are usually based on multiple-regression analysis and involve comparing actual student outcomes with expectations or predictions determined empirically on the basis of relevant inputs (attendance, gender, ethnicity, earlier achievement). The most sophisticated of these approaches use longitudinal student data, in which individual students' earlier achievement scores are matched with their later achievement scores. In the Tennessee Value-Added Assessment System, for example, estimated student gain scores are aggregated to the levels of teacher, school, and system and are compared with national norm gains, which each school is expected to achieve. Schools with scores less than two standard deviations below the norm must show positive progress or risk intervention by the State (Sanders & Horn, 1994).

Problems associated with the use of value-added measures include inadequate coverage of the achievements of schools, which may vary by curriculum area, grade level, and teacher; incomplete data for students arising from absenteeism or student turnover rate; regression to the mean in statistical analyses; problems with reliability of measures when the number of students in a school is small; and how to factor in the contextual effect on achievement created by the ability level of students in a school or class (Sanders & Horn, 1994; School Curriculum and Assessment Authority, 1994; Tymms, 1995; Webster et al., 1994; Woodhouse & Goldstein, 1996).

ISSUES IN OUTCOME EVALUATION

Despite its popularity, the use of outcome evaluation gives rise to a series of issues. First, since outcome evaluation rests primarily on assumptions related to planning, incentives, accountability, and consumerism, it is not likely to lead to greater understanding of what goes on in programs, or to an identification of the factors that affect outcomes (e.g., the relative contributions of teachers, schools, and a variety of other influences, within a program or outside it). However, many would regard progress in understanding "how" and "why" programs have an impact as important for real improvement. Second, and related to the first point, is the issue of identification and specification of the responsibility of providers and clients, particularly

in situations in which roles may be ambiguous and not clearly separated. How does one establish that a particular outcome was, even in part, amenable to the influence of a person to whom responsibility for it may have been assigned? For example, while it is reasonable to assume that a school and teachers bear some responsibility for student achievement, do not students and parents also bear responsibility? If this is so, how should responsibility between the parties be apportioned? And should the apportionment be the same for all students, in all circumstances, at all age levels?

Third, performance indicators may be used, recorded, and interpreted in varying ways, thus giving rise to problems of comparability. For example, a core set of measures developed by a Federal Interagency Task Force to monitor market programs in the United States was designed to form the basis of state-level management information systems supporting performance monitoring. However, since no state operates a fully integrated data system serving multiple programs, and since choice of performance measures differ from one program to another, data are not directly comparable (Blalock, 1999).

Fourth, since many outcome evaluations focus on a limited range of outcomes, the data that are obtained may not adequately reflect system or program goals and objectives. The temptation, of course, is to focus on what is easy to measure, but this may be to the detriment of important objectives. Perrin (1998) reminds us that "many activities in the public policy realm, by their very nature, are complex and intangible and cannot be reduced to a numerical figure . . . What is measured, or even measurable, often bears little resemblance to what is relevant" (pp. 373–373). However, focusing on a limited set of outcomes is likely to mean that other outcomes will be neglected in program implementation.

Fifth, when outcome evaluation is associated with high stakes, meeting the requirements of measuring and reporting may become more important than what a program was designed to achieve, resulting in goal displacement. In education, for example, when assessment results become the goal of instruction, the true purpose of the instructional process may be subverted as goals are reoriented to meet or exceed "standards." Further, efforts to improve performance on the measure do not necessarily result in improvement in the areas that programs were designed to achieve. When meeting standards becomes the basis for budgetary decisions, there is the further consequence that programs that meet standards, rather than program goals, may be continued, while programs that meet goals, but not standards, may be discontinued.

Sixth, when evaluations are based on predetermined objectives or standards, it is unlikely that unintended or unanticipated consequences will be detected. Seventh, the interpretation of data in outcome evaluations may not adequately acknowledge diversities in the environment in which programs were implemented. It may well be that a program is "successful" in one context, but not in another. Finally, the cost of outcome evaluation may divert funds from other needs, a not unimportant consideration at a time of resource constraints (Battistich et al., 1999; Blalock, 1999; Davies, 1999; Natriello, 1996; Perrin, 1998).

OUTCOME EVALUATION AND OTHER FORMS OF EVALUATION

In conclusion, we may ask: Where does outcome evaluation fit among traditional approaches to program evaluation? The question may be addressed from three not entirely mutually exclusive points of view: the context in which an evaluation is carried out, its methodology, and its relationship to the policy process and decision making.

Context

As far as context is concerned, outcome evaluation, as it has recently developed, differs from traditional approaches in a number of ways, fitting more comfortably with its managerial antecedents that with any program evaluation approach. First, it tends to be part of a bureaucratic routine, providing knowledge that, in theory at any rate, is relevant to policy. Second, it frequently involves accountability considerations, relating to the scrutinization of programs and reporting of performance indicators. Third, the most common use of such evaluation is in the context of very broad and complex programs (represented in, for example, all the efforts made by a school or school system over a number of years) rather than more discrete and more clearly specified programs. Fourth, outcome evaluation, as most commonly practised, relates to on-going practice rather than to innovative or experimental programs designed to address social or economic problems. Thus, it is not normally associated with trial runs of new programs, as traditional program evaluation is, nor is it normally combined with qualitative approaches to assess program implementation and impact.

Methodology

The methodologies of outcome evaluation have some affinity with early (1960s) evaluation approaches, which were largely based on Popperian logical positivism, employing quantitative measures, deductive chains, and aspirations towards generalization. While outcome indicators in themselves will not provide valid causal knowledge, interest in causality associated with their use is evidenced in efforts to identify correlates of achievement and in the assumptions underlying the use of added value techniques.

While these aspects of outcome evaluation may point to an affinity with traditional views of evaluation and indeed of research, there are also indications that outcome evaluation is perceived as a genre that is distinct from traditional evaluation (see Blalock, 1999; Pollitt, 1993). This conclusion seems warranted when one considers that outcome monitoring (represented in national assessments and international comparative studies) is being promoted by governments and international agencies at the same time as, and independently of, more traditional approaches to evaluation (see, e.g., European Commission, 1997).

Policy and Decision Making

At this stage, there is little documentation available on the use of outcome evaluations in a policy context. The extent to which information derived from such eval-

uations enters the policy arena will no doubt differ from country to country, depending on a country's traditions of government and of policy and decision making, as well as on the relationships which have already been established between policymakers, decision makers, and evaluators. Insofar as the methodology of outcome evaluations seems close to that involved in empirical quantitative approaches, with their rational view of the policy process, one might expect outcome information to be considered exogenous to the process, providing "objective", "neutral", and apolitical information to be used instrumentally in policy and decision making. In this view, as in early evaluation efforts, the evaluator has a role to play in resolving policy issues, but not as a player in the actual policy process (Radaelli & Dente, 1996). This conclusion is reinforced when we consider the number of outcome evaluation projects in which there often is no identifiable "evaluator." Indeed, the term evaluation often does not have a prominent place in discourses on the activities of what we are calling outcome evaluation.

This should not surprise us, given the limited number of goals of information production that are considered relevant to outcome evaluation. Of the six goals identified by Blalock (1999) that more conventional methods of evaluation strive to meet, outcome evaluation is likely to address only one: determining if a program's outcomes for clients (and perhaps its net impact) are consistent with desired outcomes and to improve these outcomes. Outcome evaluation is not likely to provide information on Blalock's five other goals: whether or not a program's interventions are as intended; whether a program is being delivered to the intended target population; whether a program is being implemented as intended; identification of the major influences shaping a program's outcomes; or the appropriateness, utility, and societal value of policies on which a program is based.

The way in which outcome evaluation information is predicted to work in some systems suggests that the effort to accommodate the information in policy will be slight. If, for example, the prime purpose of providing outcome information on school performance is to attach to it rewards or punishments for school districts, schools, or teachers, then there would seem to be little need to reflect on, or try to understand, how schools function, or what it is about programs that facilitates student growth. Perhaps, the questions raised by these issues are too demanding and challenging for a busy administrator. The easier course is to import market models and leave it to competition and consumer choice to bring about desired reform. However, as long as this approach is followed, many questions that have traditionally occupied evaluators will remain unanswered: does a program contribute to improvement, is it equitable, what are the unintended consequences, and at what cost is change achieved?