EVALUATION MODELS
Viewpoints on Educational and Human Services Evaluation
Second Edition

Evaluation in Education and Human Services

Editors:

George F. Madaus, Boston College,
Chestnut Hill, Massachusetts, U.S.A.
Daniel L. Stufflebeam, Western Michigan
University, Kalamazoo, Michigan, U.S.A.

Other books in the series:

Gifford, B. and O'Connor, M.:
Changing Assessments
Gifford, B.:
Policy Perspectives on Educational Testing
Basarab, D. and Root, D.:
The Training Evaluation Process
Haney, W.M., Madaus, G.F. and Lyons, R.:
The Fractured Marketplace for Standardized Testing
Wing, L.C. and Gifford, B.:
Policy Issues in Employment Testing
Gable, R.E.:
Instrument Development in the Affective Domain (2nd Edition)
Kremer-Hayon, L.:
Teacher Self-Evaluation
Payne, David A.:
Designing Educational Project and Program Evaluations
Oakland T. and Hambleton, R.:
International Perspectives on Academic Assessment
Nettles, M.T. and Nettles, A.L.:
Equity and Excellence in Educational Testing and Assessment
Shinkfield, A.J. and Stufflebeam, D.L.:
Teacher Evaluation: Guide to Effective Practice
Birenbaum, M. and Dochy, Filip J.R.C.:
Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge
Mulder, M., Nijhof, W.J. and Brinkerhoff, R.O.:
Corporate Training for Effective Performance
Britton, E.D. and Raizen, S.A.:
Examining the Examinations
Candoli, C., Cullen, K. and Stufflebeam, D.:
Superintendent Performance Evaluation
Brown, S.M. and Seidner, C.J.:
Evaluating Corporate Training: Models and Issues
Osterlind, S.:
Constructing Test Items (Second Edition)
Nettles, M.T. and Nettles, A.L.:
Challenges Minorities Face in Educational Testing and Assessment

EVALUATION MODELS
Viewpoints on Educational and Human Services Evaluation
Second Edition

edited by

Daniel L. Stufflebeam
Western Michigan University

George F. Madaus
Boston College

Thomas Kellaghan
The Educational Research Centre, Dublin

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47559-6
Print ISBN: 0-7923-7884-9

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

CONTENTS

PREFACE

Any attempts to formally evaluate something involves coming to grips with a wide range of concepts such as value, merit, worth, growth, criteria, standards, objectives, needs, norms, client, audience, validity, reliability, objectivity, practical significance, accountability, improvement, inputs, process, product, formative, summative, cost, impact, information, credibility, and, of course, the term evaluation itself. To communicate with colleagues and clients, evaluators need to be clear about what is meant by such concepts. Moreover, it is necessary to integrate the concepts and their meanings into a coherent framework that guides all aspects of their work.

The conceptualization of evaluation is not a once-off activity, nor is it static. Rather, the ideas that guide evaluation work should keep pace with the growth of theory and practice in the field. Further, the design and conduct of any particular study will involve a good deal of thought focused on the job in hand, in which it will be necessary to identify and define audiences and information requirements; the object to be evaluated; the purposes of the evaluation; inquiry procedures; concerns and issues to be examined; variables to be assessed; bases for interpreting findings; and the standards to be invoked in assessing the quality of the work.

It is no small wonder, then, that attempts to conceptualize evaluation have been among the most influential works in the fast-growing literature on the topic, and the contents of this anthology attest to the existence of a rich array of theoretical perspectives. These perspectives vary in many respects, which is not surprising given the complexity of evaluation work; the wide range of situations and political contexts in which it is carried out; its service orientations; and the varied backgrounds

and beliefs of those who write about evaluation. The ways in which evaluation is conceptualized will differ according to the role assigned to objectives in the process; the extent to which it is thought desirable to present convergent or divergent findings; the corollary preference for constructivist or objectivist findings and interpretations; the use or absence of experimental controls; the extent to which theory is used to determine the variables and the interrelationships to be examined; and the role that hard and soft data play in arriving at conclusions. It is understandable that evaluators will sometimes follow one approach in one kind of evaluation assignment, and a quite different approach in another setting. Given the variety of contexts in which evaluations take place and the range of philosophical perspectives reflected in evaluations, it is fortunate that evaluators can find in the literature a variety of ways to conceptualize the evaluation process in their search for the one that best suits a particular context.

From this diversity of conceptual approaches to evaluation, however, a consensus has begun to emerge regarding the principles that should undergird all evaluations. The consensus is embodied in the standards issued by the Joint Committee on Standards for Educational Evaluation. Basically, these standards require that evaluations be useful, feasible, ethical, and accurate. The appearance of the standards, and the associated mechanism for regularly reviewing and revising them, signify the maturing of evaluation as a profession. While the standards were developed for use in educational evaluation in North America, they have also been usefully applied, or at least consulted, in fields outside education and in countries around the world.

The present volume is a revision of an anthology that was published in 1983. Two major considerations governed the selection of material for the revision. First, it was decided to retain papers that were regarded as seminal in the history of evaluation as well as ones that described models adequately. Some chapters were dropped because the relevance of their messages had decreased over time. Second, papers which represented developments in evaluation since 1983 were added. We increased the coverage of material that had application outside the field of education and of naturalistic evaluation. These considerations led to the retention of seven papers, the revision of three, and the addition of fifteen.

The result is a book that is an up-to-date reflection of the conceptual development of evaluation, particularly program evaluation, and is divided into five major sections. The first section includes essays on the history of evaluation; models, metaphors, and definitions; and alternative approaches. The second, third, and fourth parts contain articles that represent the current major schools of thought about evaluation, written by leading authors in the field. In Part II, papers are categorized in terms of their questions/methods orientation. They cover objectives-oriented evaluation, outcome evaluation, the role of testing in evaluation, discrepancy evaluation, experimental design, cost analysis, clarification hearings or judicial evaluation, case studies, the technology of criticism, and theory-based evaluation. Papers in Part III address improvement/accountability-oriented approaches: consumer-oriented evaluation, decision-oriented evaluation, and accountability. The entries in Part IV relate to social agenda-directed/advocacy evaluation models, and cover responsive

evaluation, constructivist evaluation, empowerment evaluation, and deliberative democratic evaluation. In the final section, three overarching topics are addressed: utilization-focused evaluation, standards for evaluations, and the methodology of metaevaluation.

The evaluation models described in the book are not models in the sense of mathematical models used to test given theories, but they are models in the sense that each one characterizes its author's idealized view of the main concepts and structure of evaluation work, which form the basis of guidelines which are used to arrive at defensible descriptions and judgments. We are aware that some writers in the field have urged against according alternative perspectives on evaluation the status of models. However, we think the suggestion that they be called something else, such as persuasions or beliefs, might do little more than puzzle readers. We are comfortable in presenting the conceptualizations, not as models of what occurs, but as models for conducting studies according to various authors' beliefs about evaluation. In this sense, they are idealized or "model" views of how to sort out and address the problems encountered in conducting an evaluation.

We wish to emphasize that the presented models should not be considered as discrete options. While they may differ in important aspects, such as in the treatment of objectives and the use of experimental controls, they also overlap. For example, all call for examination of outcomes and most include an examination of process. Clear examples of overlap can be seen in the models proposed by Scriven, Stake, and Stufflebeam when they emphasize the importance of a comprehensive assessment of relevant criteria to illuminate, as well as present judgments of the merit of a program or other object. However, these models also differ in notable ways, such as in the relative importance accorded to an improvement orientation versus a focus on reaching a summative judgment. The practical implication of the concept of overlapping models is that users may combine elements of different models as they design particular evaluations.

We owe an enormous debt to the authors of the articles that appear in the book. We would like to thank those that gave us permission to reprint their publications and those who prepared articles specifically for this volume. We also are grateful to Zachary Rolnik and Michael Williams of Kluwer-Nijhoff Publishing, who consistently supported our effort. Further thanks are extended to Seamus Ó hUallacháin, Brian Carnell, Marguerite Clarke, John Coyle, Ida Holmstedt, Catherine Horn, Diane Joyce, Amandine Passot, Sally Veeder, Hilary Walshe, and Lori Wingate, for their competent editorial, technical, and clerical assistance throughout this project.

We believe this book will be of interest and assistance to the full range of persons who are part of any evaluation effort, including the clients who commission evaluation studies and use their results, evaluators, and administrators and staff in the programs that are evaluated. We also believe the book should be useful as a text for courses in program evaluation and for workshops. Further, it should prove to be an invaluable reference source for those who participate in any aspect of formal evaluation work. We hope that it will assist significantly all involved in program evaluation to increase their awareness of the complexity of evaluation; to increase

their appreciation of alternative points of view; to improve their ability to use theoretical suggestions that appear in the literature; to increase their critical appraisal of various approaches; to increase their adherence to the field's professional standards; and, ultimately, to improve the quality and utility of their evaluations.

1. PROGRAM EVALUATION: A HISTORICAL OVERVIEW

GEORGE F. MADAUS and DANIEL L. STUFFLEBEAM

Program evaluation is often mistakenly viewed as a recent phenomenon. Many people date its beginning from the late 1960s with the infusion by the federal government of large sums of money into a wide range of human service programs, including education. However, program evaluation has an interesting history that predates by at least 150 years the explosion of evaluation during the era of President Johnson's Great Society and the emergence of evaluation as a maturing profession since the sixties. A definitive history of program evaluation has yet to be written and in the space available to us we can do little more than offer a modest outline, broad brush strokes of the landscape that constitutes that history. It is important that people interested in the conceptualization of evaluation are aware of the field's roots and origins. Such an awareness of the history of program evaluation should lead to a better understanding of how and why this field has developed as it did.

Where to begin? For convenience we shall describe seven periods in the life of program evaluation. The first is the period prior to 1900, which we call the *Age of Reform;* the second, from 1900 until 1930, we call the *Age of Efficiency and Testing;* the third, from 1930 to 1945, may be called the *Tylerian Age;* the fourth, from 1946 to about 1957, we call the *Age of Innocence;* the fifth, from 1958 to 1972, is the *Age of Development;* the sixth, from 1973 to 1983, the *Age of Professionalization;* and finally the seventh from 1983 to 2000 the *Age of Expansion and Integration.*

D.L. Stufflebeam, C.F. Madaus and T. Kellaghan (eds.). EVALUATION MODELS.

THE AGE OF REFORM 1792–1900

We begin this period in our history of program evaluation in 1792 because that is the year in which William Farish invented the quantitative mark to score examinations (Hoskins, 1968). Replacing qualitative assessments of student performance with a mark for a "correct" answer permitted the ranking of examinees and the averaging and aggregating of scores. This was the first development in the field of psychometrics as we know it today (Madaus & Kellaghan, 1992). In fact Farish revolutionized testing, a technology that plays an important role in the history of program evaluation to the present.

The 19th century was the era of the Industrial Revolution with all of its attendant economic and technological changes. The very structure of society was transformed. Major social changes occurred. There was drastic change in physical and mental health and outlook, in social life and social conscience, and in the structures of social agencies. There was the laissez-faire philosophy of Bentham and the humanitarian philosophy of the philanthropists (Thompson, 1950). There were continued but often drawn out attempts to reform educational and social programs and agencies in both Great Britain and the United States.

In Great Britain there were continuing attempts to reform education, the poor laws, hospitals, orphanages, and public health. Evaluations of these social agencies and functions were informal and impressionistic in nature. Often they took the form of government-appointed commissions set up to investigate aspects of the area under consideration. For example, the Royal Commission of Inquiry into Primary Education in Ireland under the Earl of Powis, after receiving testimony and examining evidence, lamented over the progress of the children in the national schools of Ireland. The Powis Commission recommended the adoption of a scheme known as payment by results, already being used in England, whereby teachers' salaries would be dependent in part on the results of annual examinations in reading, spelling, writing, and arithmetic (Kellaghan & Madaus, 1982; Madaus & Kellaghan, 1992). Another example of this approach to evaluation was the 1882 Royal Commission on Small Pox and Fever Hospitals, which recommended after study that infectious-disease hospitals ought to be open and free to all citizens (Pinker, 1971).

Royal commissions are still used today in Great Britain to evaluate areas of concern. Rough counterparts in the United States to these commissions are presidential commissions (for example, the President's Commission on School Finance), White House panels (e.g., the White House Panel on Non Public Education), and congressional hearings. Throughout their history royal commissions, presidential commissions, and congressional hearings have served as a means of evaluating human services programs of various kinds through the examination of evidence either gathered by the Commission or presented to it in testimony by concerned parties. However, this approach to evaluation was often only emblematic or symbolic. N. J. Crisp (1982) captures the pseudo nature of such evaluations in a work of fiction. One of his characters discusses a royal commission this way: "Appoint it, feel that

you've accomplished something, and forget about it, in the hope that by the time it's reported, the problem will have disappeared or been overtaken by events" (p. 148).

In Great Britain during this period when reform programs were put in place, it was not unusual to demand yearly evaluations through a system of annual reports submitted by an inspectorate. For example, in education there were school inspectors that visited each school annually and submitted reports on their condition and on pupil attainments (Kellaghan & Madaus, 1982; Madaus & Kellaghan, 1992). Similarly the Poor Law commissioners had a small, paid inspectorate to oversee compliance with the Poor Law Amendment Act of 1834 (Pinker, 1971). The system of maintaining external inspectorates to examine and evaluate the work of the schools exists today in Great Britain and Ireland. In the United States, external inspectors are employed by some state and federal agencies. For example, the Occupational Safety and Health Administration (OSHA) employs inspectors to monitor health hazards in the workplace. Interestingly, the system of external inspectors as a model for evaluation has received scant attention in the evaluation literature.

Two other developments in Great Britain during this period are worthy of note. First, during the middle of the nineteenth century a number of associations dedicated to social inquiry came into existence. These societies conducted and publicized findings on a number of social problems that were very influential in stimulating discussion (for example, Chadwick's 1842 Report on the Sanitary Condition of the Laboring Population of Great Britain [Pinker, 1971]). Second, often in response to these private reports, bureaucracies established to manage the programs sometimes set up committees of enquiry. These were official, government-sponsored investigations of various social programs, such as provincial workhouses (Pinker, 1971). Both these examples are important in that they constitute the beginnings of an empirical approach to the evaluation of programs.

In the United States perhaps the earliest formal evaluation was in 1815 when the Army Ordnance Department drew up a system of regulations for the "uniformity of manufacture of all arms ordnance" (Smith, 1987, p. 42). To accomplish this it became clear that the engineering of people was as important as the engineering of materials. The idiosyncrasy of the skilled craftsman had to yield to uniformity. Over several decades the Ordnance Department developed the administrative, communication, inspection, accounting, bureaucratic, and mechanical techniques that fostered conformity and resulted in the technology of interchangeable parts and the eventual manufacture of a host of mass-produced products in the 20th century (Smith, 1987).These early efforts by the Ordnance Department foreshadowed Frederick Taylor's Scientific Management movement discussed below.

The first formal attempt to evaluate the performance of schools took place in Boston in 1845.This event is important in the history of evaluation because it began a long tradition of using pupil test scores as a principal source of data to evaluate the effectiveness of a school or instructional program. Then, at the urging of Samuel Gridley Howe, written essay examinations were introduced into the Boston

grammar schools by Horace Mann and the Board of Education. Ostensibly the essay exam, modeled after those used in Europe at the time, was introduced to replace the viva voce or oral examinations. The latter mode of examination had become administratively awkward with increased numbers of pupils and was also seen as unfair because it could not be standardized for all pupils. The interesting point in terms of program evaluation was the hidden policy agenda behind the move to written examinations; namely, it was the gathering of data for inter-school comparisons that could be used in decisions concerning the annual appointment of headmasters. Howe and Mann attempted to establish differential school effects and used these data to eliminate headmasters who opposed them on the abolition of corporal punishment. This is an interesting early example of politicization of evaluation data.

Between 1887 and 1898, Joseph Rice conducted what is generally recognized as the first formal educational program evaluation in America. He carried out a comparative study on the value of drill in spelling instruction across a number of school districts. Rice, like Mann and Howe before him, used test scores as his criteria measures in his evaluation of spelling instruction. He found no significant learning gains between systems which spent up to 200 minutes a week studying spelling and those which spent as little as ten minutes per week. Rice's results led educators to reexamine and eventually revise their approach to the teaching of spelling. More important from the point of view of this history of program evaluation is his argument that educators had to become experimentalists and quantitative thinkers and his use of comparative research design to study student achievement (Rice, 1914; 1897). Rice was a harbinger of the experimental design approach to evaluation first advanced by Lindquist (1953) and extended and championed by Campbell (Campbell & Stanley, 1963; Campbell, 1969) and others in the 1960s and 1970s and by Mosteller and his colleagues in the mid 1990s (see Chapter 8).

Before leaving this very brief treatment of the age of reform, another development should be mentioned. The foundation of the accreditation or professional judgement approach to evaluation can be traced directly to the establishment of the North Central Association of Colleges and Secondary Schools in the late 1800s. The accreditation movement did not, however, gain great stature until the 1930s when six additional regional accrediting associations were established across the U.S. Since then the accrediting movement has expanded tremendously and gained great strength and credibility as a major means of evaluating the adequacy of educational institutions. (See Floden, 1983 for a treatment of the accreditation approach to evaluation.)

THE AGE OF EFFICIENCY AND TESTING 1900–1930

During the early part of the twentieth century the seminal work by Fredrick Taylor launched the scientific management movement, an early form of personnel evaluation. Taylorism continues to affect almost all aspects of American life to this day. (For a detailed treatment of Taylor's impact on society see Doray, 1988 and Banta, 1993.) Taylor's ideas became a powerful force in administrative theory in educational

and industrial circles (Biddle & Ellena, 1964; Callahan, 1962; Cremin, 1962). The emphasis of this movement was on systemization, standardization, and, most importantly, efficiency. Typifying this emphasis on efficiency were the tides of the fourteenth and fifteenth yearbooks of the National Society for the Study of Education (NSSE), which were, respectively, *Methods for Measuring Teachers' Efficiency* and the *Standards and Tests for the Measurement of the Efficiency Of Schools and School Systems.*

Surveys done in a number of large school systems during this period focused on school and/or teacher efficiency using various criteria (for example, expenditures, pupil dropout rate, promotion rates, etc.). By 1915, thirty to forty large school systems had completed or were working on comprehensive surveys on all phases of educational life (Kendall, 1915; Smith & Judd, 1914). A number of these surveys employed the newly developed "objective" tests in arithmetic, spelling, handwriting, and English composition to determine the quality of teaching. (For a detailed treatment of the history of mathematics and arithmetic tests during this time see Madaus, Clarke & O'Leary, in press.) These tests were often developed in large districts by a bureau or department set up specifically to improve the efficiency of the district. For example, the Department of Educational Investigation and Measurement in the Boston public schools developed a number of tests that today would be described as objective referenced (Ballou, 1916). Eventually tests like those in Boston took on a norm-referenced character as the percentage of students passing became a standard by which teachers could judge whether their classes were above or below the general standard for the city (Ballou, 1916). In addition to these locally developed tests there were a number of tests developed by researchers like Courtis, Ayers, Thorndike, and others, which were geared to measuring a very precise set of instructional objectives. These tests by famous researchers of the day had normative data that enabled one system to compare itself with another (Tyack & Hansot, 1982).

Many of these early twentieth-century surveys were classic examples of muckraking, "often initiated by a few local people who invited outside experts to expose defects and propose remedies" (Tyack & Hansot, 1982, p. 161). Another problem associated with these early surveys—a problem not unknown to evaluators today—was that the "objective" results obtained were often used as propaganda "to build dikes of data against rising tides of public criticism" (Tyack & Hansot, 1982, p. 155). However, researchers at the time did recognize that such surveys could and should avoid muckraking and public relations use. Many of them were indeed constructive, done in cooperation with local advisors, and designed to produce public support for unrecognized but needed change (Tyack & Hansot, 1982).

With the growth of standardized achievement tests after World War I, school districts used these tests to make inferences about program effectiveness. For example, May (1971) in an unpublished paper described the history of standardized testing in Philadelphia from 1916 to 1938. He found that commercially available achievement tests, along with tests built by research bureaus of large school districts, were used to evaluate the curriculum and overall system performance, in addition to being

used to make decisions about individuals. Throughout its history, the field of evaluation has been closely linked to the field of testing. Test data have often been the principal data source in evaluations; this use of tests has been a mixed blessing as we shall see presently.

It is important to point out that studies of efficiency and testing were for the most part initiated by, and confined to, local school districts. In contrast to the national curriculum development projects of the late 1950s and early 1960s, curriculum development before the 1930s was largely in the hands of a teacher or committee of teachers. It was natural, therefore, that evaluations of that period were addressed to localized questions. This focus or emphasis on local evaluation questions continued into the 1960s despite the fact that the audience for the evaluations was state-wide or nation-wide; this resulted in many useless educational evaluations being carried out during the 1960s. It was only in the 1970s that educators and evaluators recognized and began to deal with this problem of generalizability. And, it wasn't until the 90s with the advent of standards based reform that the focus shifted from local to state level control over many aspects of the curriculum.

During the late 1920s and 1930s, university institutes specializing in field studies were formed and conducted surveys for local districts. The most famous of these institutes was the one headed by George Strayer at Teachers College (Tyack & Hansot, 1982). These institutes could be considered the precursors of the university centers dedicated to evaluation that grew up in the 1960s and 1970s.

THE TYLERIAN AGE 1930–1945

Ralph W. Tyler has had enormous influence on education in general and educational evaluation and testing in particular. He is often referred to, quite properly we feel, as the father of educational evaluation. Tyler began by conceptualizing a broad and innovative view of both curriculum and evaluation. (Cf. Chapter 4.) This view saw curriculum as a set of broadly planned school experiences designed and implemented to help students achieve specified behavioral outcomes. Tyler coined the term "educational evaluation" which meant assessing the extent that valued objectives had been achieved as part of an instructional program. (This development is the foundation of today's outcome evaluation described in Chapter 4). During the early and mid-1930s, he applied his conceptualization of evaluation to helping instructors at Ohio State University improve their courses and the tests that they used in their courses.

During the depths of the Great Depression, schools, as well as other public institutions, had stagnated from a lack of resources and, perhaps just as importantly, from a lack of optimism. Just as Roosevelt tried through his New Deal programs to lead the economy out of the abyss, so too John Dewey and others tried to renew education. The renewal in education came to be known as the Progressive Education Movement, and it reflected the philosophy of pragmatism and employed tools from behavioristic psychology.

Tyler became directly involved in the Progressive Education Movement when he was called upon to direct the research component of the now-famous Eight Year Study (Smith & Tyler, 1942a). The Eight-Year Study (1932–1940), funded by the Carnegie Corporation, was the first and last large study of the differential effectiveness of various types of schooling until well after World War II. The study came about when questions were asked in the early 1930s about the efficacy of the traditional high school experience relative to the progressive secondary school experience. As a result of these questions, leading colleges began to refuse progressive school graduates admittance because they lacked credits in certain specific subjects. To settle the debate, an experiment was proposed in 1932 in which over 300 colleges agreed to waive their traditional entrance requirements for graduates from about 30 progressive secondary schools. The high school and college performance of students from these secondary schools would be compared to the high school and college performance of students from a group of traditional secondary schools.

The Eight-Year Study introduced educators throughout America to a new and broader view of educational evaluation than that which had been in vogue during the age of efficiency and testing. Tyler conceptualized evaluation as a comparison of intended outcomes with actual outcomes. His view of evaluation was seen by advocates as having a clear-cut advantage over previous approaches. Since a Tylerian evaluation involves internal comparisons of outcomes with objectives, it need not provide for costly and disruptive comparisons between experimental and control groups, as were required in the comparative experimental approach that Rice had used. Since the approach calls for the measurement of behaviorally defined objectives, it concentrates on learning *outcomes* instead of organizational and teaching *inputs,* thereby avoiding the subjectivity of the professional judgment or accreditation approach; and, since its measures reflect defined objectives, there was no need to be heavily concerned with the reliability of differences between the scores of individual students. Further, the measures typically cover a much wider range of outcome variables than those assessed by standardized norm-referenced tests.

Clearly by the middle of the 1940s Tyler had, through his work and writing, laid the foundation for his enormous influence on the educational scene in general and on testing and evaluation in particular during the next 25 years.

THE AGE OF INNOCENCE 1946–1957

We have labeled the period 1946–1957 as the *Age of Innocence,* although we might just as well have called it the *Age of Ignorance.* It was a time of poverty and despair in the inner cities and in rural areas, but almost no one except the victims seemed to notice. It was a period of extreme racial prejudice and segregation, to which most white people seemed oblivious. There was exorbitant consumption and widespread waste of natural resources with little apparent concern about the depletion of these resources. It was a period of vast development of industry and military capabilities with little provision for safeguards against the many negative side effects.

More to the point of this review, there was expansion of educational offerings, personnel, and facilities. New buildings were erected. New kinds of educational institutions, such as experimental colleges and community colleges emerged. Small school districts consolidated with others to be able to provide the wide range of educational services that were common in the larger school systems, including mental and physical health services, guidance, food services, music instruction, expanded sports programs, business and technical education, and community education. College enrolments increased dramatically and enrolments in teacher-education programs ballooned. Throughout American society, the late 1940s and 1950s were a time to forget the war, leave the depression behind, build and expand capabilities, acquire resources, and engineer and enjoy a "good life."

This general scene in society and education was reflected in educational evaluation. While there was great expansion of education there was no particular interest on the part of society in solving social and education problems and holding educators accountable. There was little call for educators to demonstrate the efficiency and effectiveness of any of the many developmental efforts. Educators did talk and write about evaluation, and they did collect considerable amounts of data (usually to justify the need for expansion or for broad, new programs). However, there is little evidence that these data were used to judge and improve the quality of programs or that the data could have been used for such a purpose.

We have labeled the period 1946 to 1947 The Age of Innocence, not because work in evaluation did not proceed but because the work seemingly had no social purpose. The great deal of technical development in evaluation was just that. It was not geared to identifying beneficiaries' needs and critically examining society's response to the needs.

During this period there was considerable development of some of the technical aspects of evaluation; this was consistent with the then-prevalent expansion of all sorts of technologies. Chief among these developments was the growth in standardized testing. Many new nationally standardized tests were published during this period. Schools purchased these tests by the thousands and also subscribed heavily to machine scoring and analysis services that the new technology made available. The testing movement received another boost in 1947 with the establishment of the Educational Testing Service.

By the 1950s, the standardized testing business had expanded tremendously, and the professional organizations concerned with testing initiated a series of steps designed to regulate the test-related activities of their members. In 1954, a committee of the American Psychological Association prepared *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954). In 1955, committees of the American Educational Research Association and the National Council on Measurements Used in Education prepared *Technical Recommendations for Achievement Tests* (AERA and NCMUE, 1955). These two reports provided the basis for the 1966 edition of the joint AERA/APA/NCME *Standards for Educational and Psychological Tests and Manuals* (APA, 1966) and the 1974 revision entitled, *Standards for Educational and Psychological Tests* (APA, 1974). The latter report recognized the need for

separate standards dealing with program evaluation. A revision of the Standards in 1985 contained a chapter on the use of tests in program evaluation, as did a further revision in 2000.

The rapid expansion of testing was not the only technical development related to program evaluation during this period. Lindquist (1953) extended and delineated the statistical principles of experimental design. Years later, many evaluators and educators found that the problems of trying to meet simultaneously all of the required assumptions of experimental design (for example, constant treatment, uncontaminated treatment, randomly assigned subjects, stable study samples, and unitary success criteria) in the school setting were insurmountable.

During the 1950s and early 1960s there was also considerable technical development related to the Tylerian view of evaluation. Since implementing the Tyler approach in an evaluation required that objectives be stated explicitly, there was a need to help educators and other professionals to do a better job articulating their objectives. Techniques to help program staffs make their objectives explicit, along with taxonomies of possible educational objectives (Bloom et al., 1956; Krathwohl, 1964), were developed to fill this need. The Tyler rationale was also used extensively during this period to train teachers in test development.

During this period evaluations were, as before, primarily within the purview of local agencies. Federal and state agencies had not yet become deeply involved in the evaluation of programs. Funds for evaluation that were done came from local coffers, foundations, voluntary associations such as the community chest, or professional organizations. This lack of dependence on taxpayer money for evaluation would end with the dawn of the next period in the history of evaluation.

THE AGE OF DEVELOPMENT 1958–1972

The age of innocence in evaluation came to an abrupt end with the call in the late 1950s and early 1960s for evaluations of large-scale curriculum development projects funded by federal monies. This marked the end of an era in evaluation and the beginning of profound changes that would see evaluation expand as an industry and into a profession, focused on helping meet society's needs and dependent on taxpayer monies for support.

As a result of the Russian launch of Sputnik in 1957, the federal government enacted the National Defense Education Act of 1958. Among other things, this act provided for new educational programs in mathematics, science, and foreign language; and expanded counseling and guidance services and testing programs in school districts. A number of new national curriculum development projects, especially in the areas of science and mathematics, were established. Eventually funds were made available to evaluate these curriculum development efforts.

All four of the approaches to evaluation discussed so far were represented in the evaluations done during this period. First, the Tyler approach was used to help define objectives for the new curricula and to assess the degree to which the objectives were later realized. Second, new nationally standardized tests were created to better

reflect the objectives and content of the new curricula. Third, the professional-judgment approach was used to rate proposals and to check periodically on the efforts of contractors. Finally, many evaluators evaluated curriculum development efforts through the use of field experiments.

The best and the brightest of the educational evaluation community were involved in efforts to evaluate these new curricula; they were adequately financed, and they carefully applied the technology that had been developed during the past decade or more. Nonetheless, by the early 1960s it became apparent to some leaders in educational evaluation that their work and their results were neither particularly helpful to curriculum developers nor responsive to the questions being raised by those who wanted to know about the programs "effectiveness."

This negative assessment was reflected best in a landmark article by Cronbach (1963; cf. Chapter 14). In looking at the evaluation efforts of the recent past, he sharply criticized the guiding conceptualizations of evaluation for their lack of relevance and utility, and advised evaluators to turn away from their penchant for post hoc evaluations based on comparisons of the norm-referenced test scores of experimental and control groups. Instead, Cronbach counseled evaluators to reconceptualize evaluation—not in terms of a horse race between competing programs but as a process of gathering and reporting information that could help guide curriculum development. Cronbach was the first person to argue that analysis and reporting of test item scores would be likely to prove more useful to teachers than the reporting of average total scores. When first published, Cronbach's counsel and recommendations went largely unnoticed, except by a small circle of evaluation specialists. Nonetheless, the article was seminal, containing hypotheses about the conceptualization and conduct of evaluations that were to be tested and found valid within a few years.

In 1965, guided by the vision of Senator Hubert Humphrey, the charismatic leadership of President John Kennedy, and the great political skill of President Lyndon Johnson, the War on Poverty was launched. These programs poured billions of dollars into reforms aimed at equalizing and upgrading opportunities for all citizens across a broad array of health, social, and educational services. The expanding economy enabled the federal government to finance these programs, and there was widespread national support for developing what President Johnson termed the Great Society.

Accompanying this massive effort to help the needy came concern in some quarters that the money invested in these programs might be wasted if appropriate accountability requirements were not imposed. In response to this concern, Senator Robert Kennedy and some of his colleagues in the Congress amended the Elementary and Secondary Education Act of 1964 (ESEA) to include specific evaluation requirements. As a result, Title I of that Act, which was aimed at providing compensatory education to disadvantaged children, specifically required each school district receiving funds under its terms to evaluate annually—using appropriate standardized test data—the extent to which its Tide I projects had achieved their objectives. This requirement, with its specific references to standardized test data and an