Origins and Development of Program Evaluation as a Discipline and Profession

Orienting Questions

1. What were the characteristics of evaluation before the field began to take shape as a distinct profession and form of inquiry?
2. What were the major periods in the development of professional program evaluation? What events and developments are associated with each period?
3. What events spurred the emergence of modern program evaluation?
4. What other major developments helped advance and shape evaluation as a discipline and profession?

In this chapter, we review the history of evaluation and its progress toward becoming a full-fledged profession and distinct discipline. This history illuminates the forces that have helped shape the field of program evaluation and key advancements in its growth and maturation. This history is United States-centric for two reasons. First, the story of professional program evaluation begins in the United States. Second, as evaluation practitioners and teachers living and working in the U.S., this is the history we (the authors) know best. As we discuss later in the chapter, evaluation is spreading rapidly across the world. We leave the telling of the story of the evaluation’s development outside of the United States to our international colleagues.

/As **Scriven (1996)** i9 noted, “Evaluation is a very young discipline—although it is a very old practice” (p. 395). He pointed out that the formal evaluation of industrial crafts has taken place since prehistoric times: “As long as artifacts have existed—surely hundreds of millennia before any of the highly durable stone tools were made—it is very likely that craft workers have been evaluating their own and their fellow workers’ products as part of a sustained development evaluation process” **(Scriven, 2016 Q,** p. 35). Indeed, ever- advancing technological development by humans has been made possible by our ability to discern the strengths and weaknesses of products and processes and make improvements accordingly.

In the public sector, formal evaluation was evident as early as 2000 BCE, when Chinese officials conducted civil service examinations to measure the proficiency of applicants for government positions. Socrates used verbally mediated evaluations as part of the learning process. Centuries passed before formal evaluations gained a foothold in society to inform decision-making about how best to educate citizens and enhance their well-being.

1800-1940: The Seeds of Modern Program Evaluation are Planted

In this section, we highlight key historical developments that set the stage for the development of program evaluation as a distinct form of inquiry.

##### Empirical Investigation of the Quality of Education Programs and Practices

The seeds for modern program evaluation were planted in the early 1800s with efforts to systematically assess school quality. In England, dissatisfaction with the education system spurred reform movements in which government-appointed royal commissions heard testimony and used other less formal methods to evaluate the respective institutions. This led to still-existing systems of external inspectorates for schools in England and much of Europe (SICI, 2016; **Standaert, 200415).** In the United States, educational evaluation took a slightly different turn. It was influenced by Horace Mann’s comprehensive annual, empirical reports on education in Massachusetts in the 1840s and the Boston School Committee’s 1845 and 1846 use of printed tests in several subjects. This was the first instance of wide-scale student assessment and served as the basis for school comparisons. These two developments in Massachusetts were the first attempts to objectively measure student achievement to assess the quality of a large school system. They set a precedent for today’s widespread use of student test scores as the primary means for judging school effectiveness.

In the late 1800s, liberal education reformer Joseph Rice conducted one of the first comparative studies to assess the quality of instructional methods. His goal was to substantiate his claims that school time was used inefficiently. To do so, he compared schools that varied in the amount of time spent on spelling drills and then examined the students’ spelling ability. He found negligible differences in students’ spelling performance between schools where students spent as much as 100 minutes per week on spelling instruction and those where they spent as little as 10 minutes per week. He used these data to encourage educators to scrutinize their practices empirically. Rice’s study is considered the “first formal educational evaluation in the United States” **(Stufflebeam & Coryn, 2014** [□, p. 30).

A landmark evaluation from the early 1900s was **Flexner’s (1910)** (□ review of medical schools. Backed by the American Medical Association and the Carnegie Foundation, he assessed 155 medical schools operating in the United States and Canada. Following a series of one-day site visits to each school by himself and one colleague, Flexner delivered scathing reviews of the schools’ quality and the state of medical education in general. He was not deterred by lawsuits or death threats due to what the medical schools viewed as his “pitiless exposure” (p. 87) of their medical training practices. He delivered his evaluation findings in scathing terms. He called Chicago’s 15 medical schools “the plague spot of the country in respect to medical education” (p. 84). Soon “schools collapsed to the right and left, usually without a murmur” (p. 87). Flexner’s reports were unambiguously evaluative. His review was a precursor to formal accreditation of academic programs in higher education.

The educational testing movement gained momentum in the early 1900s as measurement technology made rapid advances under E. L. Thorndike and his students. A pioneer in educational testing, Thorndike established norms for student performance in math, reading, handwriting, and other subjects. These norms enabled school administrators to compare their students’ knowledge and abilities with the average achievement of a representative sample of children. By 1918, objective testing was flourishing, pervading the military, private industry, and all levels of education. The 1920s saw the rapid emergence of norm-referenced tests, which were designed to rank students. By the mid-ig3os, more than half of U.S. states had some form of statewide testing. During this period, educators regarded measurement and evaluation as nearly synonymous. The latter was usually thought of as summarizing student test performance and assigning grades.

Although program evaluation as we know it today was still in its infancy, useful measurement tools for evaluation were proliferating. Very few meaningful, formal evaluations of school programs or curricula were published during this period. One notable exception was the ambitious, landmark Eight-Year Study (Smith & Tyler, 1942 l9). The Eight-Year Study set a new standard for educational evaluation with its sophisticated methodology for measuring learning outcomes. With this and later studies Tyler (e.g., 1950) (□ also planted the seeds of standards-based testing—that is, testing to determine students’ mastery of subject matter they were expected to learn, without regard for ranking or comparison with peers as in norm-referenced testing. (In Chapter 613, we discuss Tyler’s profound impact on program evaluation, especially in education.)

Beginning in the late 1930s (when positions of influence in federal agencies and higher education were largely held by White men), Black education researchers began to shed light on racial inequities in education (Thomas & Campbell, 202119). Ambrose Caliver, the first Black person to receive a Ph.D. from Columbia University, had an influential role in the U.S. Office of Education beginning in 1932. He spearheaded the collection, analysis, and dissemination of data that revealed gross inequities in the U.S. education system based on race (Thomas & Campbell, 2021|0; Hood, 2001 iQ). Reid Jackson investigated the quality of schools for Black children within the segregated education systems in Kentucky, Florida, and Alabama. Jackson, who held multiple administrator and faculty positions at historically Black colleges and universities, broke new ground by approaching his studies through the lens of culture and race (Thomas & Campbell, 2021 (□; Hopson & Hood, 200519). The launch of The Journal of Negro Education in 1932 provided an important outlet for the publication of evaluative studies that focused on the role of race in education (Hood, 2001 (□; Thomas & Campbell, 2021 [□).

##### Growth of Applied Social Research

In the early 1900s, foundations for evaluation were also being laid in fields beyond education, including health, human services, and business. For example, Cronbach and his colleagues (1980) (□ cited surveys of slum conditions, management and efficiency studies in schools, and investigations of local government corruption. Rossi, Freeman, and Lipsey (2004) (□ noted that at that time evaluations were being conducted in the field of public health, where studies focused on assessing efforts to control infectious diseases. Fredrick Taylor’s influential scientific management theory focused on discovering the most efficient way to perform a task and then training all staff to perform it that way. The emergence of “efficiency experts” in industry soon permeated the business community. As Cronbach et al. (1980) jD noted, “business executives sitting on the governing boards of social service agencies pressed for greater efficiency in those services" (p. 27). Some cities and social service agencies began to develop internal research units. Social scientists began to trickle into government service. They conducted applied social research in areas such as public health, housing, and work productivity. However, these social research precursors to evaluation were small, isolated activities. They had little impact on the lives of the citizenry or the decisions of government agencies.

With the Great Depression in the 1930s came the sudden proliferation of government services and agencies as President Roosevelt’s New Deal programs were implemented to salvage the U.S. economy. This was the first major growth in the U.S. federal government in the 20th century, and its impact was profound. Federal agencies were established to oversee new national programs in welfare, public works, labor management, urban development, health, education, and numerous other human service areas. Increasing numbers of social scientists went to work in these agencies. Applied social research opportunities abounded. Social science academics began to join with their agency-based colleagues to study a variety of variables related to these programs. While some scientists called for explicit evaluation of the new social programs (e.g., Stephan, 1935 (□), most pursued applied research at the intersection of an agency’s needs and their personal interests. Thus, academic sociologists pursued questions that were of interest to both the discipline of sociology and the agency. However, these questions often originated with the academics. The same trend occurred with economists, political scientists, and other academics who conducted research on federal programs. Their projects were considered to be “field research" and provided opportunities to address important questions within their disciplines.

1941-1963: Applied Social and Educational Research Become Commonplace

This period was not especially remarkable in terms of the development of program evaluation. It is notable, however, that applied social research and education research became more commonplace and, in a few instances, began to be institutionalized. However, the limitations of using applied research to answer pressing questions about program quality, value, and effectiveness became apparent. This situation set the stage for the new discipline of evaluation to emerge.

Applied social research expanded during World War II as researchers investigated programs to help military personnel determine how to reduce vulnerability to propaganda, increase morale, and improve the training and job placement of soldiers. In the following decade, studies focused on new government programs for job training, housing, family planning, and community development. As in the past, such studies tended to focus on particular facets of the program in which the researchers happened to be most interested. As these programs expanded, however, social scientists began to broaden their studies to examine entire programs.

With this broader focus came more frequent use of social research methods to investigate and improve specific programs. Rossi et al. (2004) 0 stated that it was typical during this period for social scientists to be “engaged in assessments of delinquency-prevention programs, psychotherapeutic and psychopharmacological treatments, public housing programs, educational activities, community organization initiatives, and numerous other initiatives” (p. 23). Such work also spread to other countries and continents. Many countries in Central America and Africa were the sites of evaluations examining health and nutrition, family planning, and rural community development. Most of these studies relied on existing social research methods and did not extend the conceptual or methodological boundaries of evaluation beyond those already established for behavioral and social research. Such efforts would come later.

Developments in the education sector in the 1940s through early 1960s were unfolding somewhat differently. In this period, earlier developments in educational evaluation were consolidated and refined. School personnel worked to improve standardized testing, quasi-experimental design research, accreditation, and school surveys. In the 1950s and early 1960s, there were also efforts to enhance the Tylerian approach to evaluation (see Chapter 6 0) by teaching educators how to state objectives in explicit, measurable terms so that student progress toward these objectives could be validly and reliably measured.

Black evaluators and researchers continued to call attention to racial inequities in education. Ambrose Caliver’s influence in the U.S. Office of Education continued throughout the 1950s (Thomas & Campbell, 2021 0; Hood, 2001 0). Aaron Brown evaluated the quality of accredited secondary schools for Black children in the South. Brown, a student of Tyler’s, used criteria established by six regional educational associations to evaluate 93 schools (Brown, 1944 0). He identified both “satisfactory features” and “weaknesses" across the schools and discussed possible reasons for notably high and low scores on various criteria (p. 493). Many of the possible causal or interacting factors he described were “peculiar to schools operating for Negroes” (p. 494). Thus, while it would be many decades before culturally responsive practices gained prominence in the evaluation field, Brown’s work called attention to the need to consider culture and context when interpreting the results of educational research and evaluation. Leander Boykin, the first Black person to attain a Ph.D. in education at Stanford University (Hood, 2001 0), studied the differences in financial resources and teacher salaries in schools for Black and White children in the south. Boykin argued for using mixed methods in evaluation, using evaluation to improve education programs, involving stakeholders in the evaluation process, and considering the social and economic context of education programs (Thomas & Campbell, 2021 0; Hood, 2001 0). Like Brown’s, Boykin’s work foreshadowed later developments in the field of program evaluation.

In 1957, the Soviets’ successful launch of Sputnik I sent tremors through the U.S. establishment that were quickly amplified into calls for more effective teaching of math and science to students in U.S. schools. The reaction was immediate. The National Science Foundation started to fund science and math curriculum development initiatives, along with the evaluation of those efforts. According to Cronbach et al. (1980) (□, these studies “were sometimes simple and rather informative, but a few were extensive and conformed to the canons of experimental design” (p. 31). Passage of the National Defense Education Act (NDEA) of 1958 poured millions of dollars into massive, new curriculum development projects, especially in math and science. Only a few projects were funded, but their size and perceived importance led policymakers to fund evaluations of most of them.

Notwithstanding the expanding program evaluation efforts, theoretical work related directly to evaluation did not exist. Therefore, those who conducted evaluation studies were left to utilize what they could from applied social, behavioral, and educational research. Their gleanings were so meager that Cronbach (1963) (□ penned a seminal article in which he sharply criticized past educational evaluations. He called for evaluators to move beyond “comparing score averages" to assessing far- ranging outcomes, including “attitudes, career choices, general understandings and intellectual powers, and aptitude for further learning in the field” (p. 247). His recommendations had little immediate impact. However, they did catch the other education scholars’ attention, helping to spark a greatly expanded conception of evaluation that would emerge in the next decade.

1964-1969: Modern Program Evaluation Emerges

The developments discussed so far were not sufficient in themselves to launch a strong and enduring program evaluation movement. However, they did make conditions ripe for such a development. Much happened to spur the modernization of evaluation between 1964 and 1969—a brief but significant phase in the field’s formation. Suddenly, the need for specialized approaches and professionals to conduct evaluations became acute. Critical developments in this period included (1) massive increases in U.S. federal spending on social programs and (2) a Congressional mandate to evaluate programs funded by the expansive Elementary and Secondary Education Act. Yet, there was a lack of scholarship and specialists in evaluation to address the growing demand.

##### Rapid Expansion of Federally Funded Social Programs

Conditions were ideal for accelerated conceptual and methodological development in evaluation, and a catalyst was found in initiatives spearheaded by U.S.

President Lyndon Johnson, who took office in 1964. His “War on Poverty" legislation sought to equalize and enhance opportunities for all citizens in virtually every sector of society. He aimed to realize his vision for a “Great Society" by pouring millions of dollars into programs in education, health, housing, criminal justice, unemployment, urban renewal, and many other areas. Federal government spending on anti-poverty and other social programs increased by 600 percent after inflation from 1950 to 1979 (Bell, 1983 (□). There was strong interest in learning how programs in areas such as job training, urban development, and housing were working. Managers and policymakers wanted to know how to improve their programs and which strategies worked best to achieve their ambitious goals. Congress wanted information on the types of programs that were worthy of continued funding. Increasingly, evaluations were mandated. In 1969, federal spending on grants and contracts for evaluation was $17 million. By 1972, it had expanded to $100 million (Shadish et al., 1991 [□).

Unlike the private sector, where accountants, management consultants, and research and development experts were readily available to provide feedback on corporate programs’ productivity and profitability, these huge, new social investments had no similar mechanism in place to examine their progress. Some government employees had relevant competence—social scientists and technical specialists in the various federal departments, particularly in the General Accounting Office (GAO) —but they were too few and not sufficiently organized to determine the effectiveness of these vast government innovations. To complicate matters, many inquiry methodologies and management techniques that worked on smaller programs proved inadequate for programs of the size and scope of these sweeping social reforms.

For a time, it appeared that another concept developed and practiced successfully in business and industry might be successfully adapted for evaluating these federal programs: the Planning-Programming-Budgeting System (PPBS). PPBS was used by Ford Motor Company and later brought to the U.S. Department of Defense by Robert McNamara when he became President Kennedy’s secretary of defense in 1961. The PPBS was a variant of the approaches used by many large aerospace, communications, and automotive companies. It was aimed at improving system efficiency, effectiveness, and budget allocation decisions by defining organizational objectives and linking them to system outputs and budgets. Many thought the PPBS would be ideally suited for the federal agencies charged with administering War on Poverty programs, but few bureaucrats heading those agencies were eager to embrace it. The stage was set for the creation of new evaluation approaches and methods, as well as a new kind of professional, with a somewhat different type of training and orientation, to apply them.

##### Elementary and Secondary Education Act (ESEA) of 1965

The single event that arguably was most responsible for the emergence of modern program evaluation is the passage by the U.S. Congress of the Elementary and Secondary Education Act (ESEA) of 1965. This event sent a shock wave through the U.S. education system, awakening both policymakers and practitioners to the importance of systematic evaluation. This bill proposed a considerable increase in federal funding for education, with tens of thousands of federal grants to local schools, state and regional agencies, and universities. The bill’s largest component was Title I, destined to be the costliest federal education program in U.S. history. Wholey and White (1973) |0 argued that Title I was the most important among the array of legislation that influenced evaluation at the time.

When Congress began its deliberations on the proposed ESEA, legislators, especially in the Senate, expressed concerns about a lack of convincing evidence that any federal funding for education had resulted in real educational improvements. Indeed, some members of Congress believed federal funds allocated to education prior to ESEA had sunk like stones into the morass of educational programs with scarcely a ripple to mark their passage. Robert F. Kennedy was the most persuasive voice among these. He insisted that the ESEA require each grant recipient to file an evaluation report showing what had resulted from the expenditure of the federal funds. This Congressional evaluation mandate was ultimately approved for Title I (compensatory education) and Title III (innovative educational projects). The requirements “reflected the state-of-the-art in program evaluation at that time" (Stufflebeam et al., 2000 (□, p. 13). These requirements reflected an astonishing amount of micromanagement at the Congressional level. They also heightened attention to accountability, calling for standardized testing to demonstrate student learning.

Some important milestone evaluation studies occurred at this time. These included the evaluations of Title I, Head Start (Westinghouse, 1969 [□), and the Sesame Street television series (Ball & Bogatz, 1970 (□). The evaluations of Sesame Street demonstrated some of the first uses of formative evaluation, as portions of the program were examined to provide feedback to program developers for improvement.

The passage of the ESEA in 1965 deserves its historical recognition as the birth of contemporary program evaluation. However, it was also marked by significant travail. Social and educational researchers lacked tools and frameworks to evaluate programs effectively. As the evaluation field was still in its infancy, few training programs were in place, and methodologies were largely borrowed from field studies in the social and behavioral sciences.

##### Emergence of Evaluation Specialists and Evaluation Approaches

The need for experts who could conduct useful and rigorous evaluations was sudden and pressing, and the market responded. Congress provided funding for universities to launch new graduate training programs in educational research and evaluation, including fellowship stipends for graduate study in those specializations. Several universities began graduate programs for educating evaluators. Political science programs spawned schools of public administration to train administrators to manage and oversee government programs. Policy analysis emerged as a means to assess different policy options and measure the impacts of implemented policies. Graduate education in the social sciences ballooned. The number of people completing doctoral degrees in economics, education, political science, psychology, and sociology grew from 2,845 to 9463. an increase of 333%, from 1960 to 1970 (Shadish et al., 199119). Many graduates of these programs pursued careers evaluating programs in the public and nonprofit sectors.

Until the late 1960s, there was minimal theoretical and methodological guidance specific to evaluation to inform the work of these new evaluation practitioners. They were left to draw from theories in their primary disciplines and glean what they could from existing social research methods. Such methods included experimental design, psychometrics, survey research, and ethnography. In response to the need for more scholarship on evaluation, some important books and articles were published. **Suchman (1967)** wrote a book reviewing different evaluation methods, and **Campbell (1969)** [□ argued for more social experimentation to examine program effectiveness. **Campbell and Stanley’s book (1966)** (□ on experimental and quasi-experimental designs influenced many working in evaluation to adopt this approach. Scriven (1967). Stake (1967) and Stufflebeam (1968) began to write articles about evaluation practice and theories. Evaluation as a distinct form of inquiry began to take shape.

These conditions led to much excitement among evaluation pioneers about this new area of inquiry and their role in improving social conditions. Donald Campbell, the renowned research methodologist who trained several individuals who later became leaders in evaluation, wrote of the “experimenting society" in his article “Reforms as Experiments." He urged managers to use data collection and experiments to learn how to develop good programs **(Campbell, 1969** (□). He argued that managers should advocate not for their programs but for solutions to the problems their programs were designed to address. He suggested that by advocating for the development and testing of solutions, managers could help policymakers, citizens, and other stakeholders become more patient with the difficult process of solving social problems such as crime, unemployment, and illiteracy. As aptly put by Shadish, “There was this incredible enthusiasm and energy for social problem solving" during this period **(Oral History Project Team, 200315,** p. 271).

1970-1999: Program Evaluation Becomes a Profession

In the last 30 years of the 20th century, program evaluation matured substantially as a professional practice and distinct form of inquiry. There was substantial growth in the volume of evaluation-specific literature. Professional evaluation associations formed. Standards of practice were established to help shape the professional identity of evaluators. During this time, the contexts and approaches to evaluation diversified, with evaluation increasingly leveraged to enhance organizational learning.

##### Growth of Evaluation Literature and Professional Evaluation Associations

In the absence of any comprehensive textbooks on evaluation, Caro (1971) i5 published a collection of readings on evaluation. Soon after, program evaluation textbooks began to be published. Examples include Evaluation Research: Methods of Assessing Program Effectiveness by Weiss (1972) i5, Educational Evaluation: Theory and Practice by Worthen and Sanders (1973) i9, and Evaluations Systematic Approach by Rossi, Freeman, and Rosenbaum (1979)15. Many more followed, including subsequent editions of those early texts. Articles about evaluation began to appear with increasing frequency in academic journals. These publications featured new evaluation models and approaches to respond to the needs of specific types of evaluation (e.g., ESEATitle III evaluations, and evaluations of mental health programs).

The number of journals that focused on evaluation grew dramatically in this period. These included American Journal of Evaluation: Canadian Journal of Program Evaluation: Educational Evaluation and Policy Analysis: Evaluation: The International Journal of Theory, Research, and Practice: Evaluation and Program Planning: Evaluation and the Health Professions: Evaluation Practice: Evaluation Studies Review Annual: Evaluation Quarterly: Evaluation Review: ITEA Journal of Tests and Evaluation: New Directions for Program Evaluation: Performance Improvement Quarterly: Practical Assessment, Research, and Evaluation: Research Evaluation: and Studies in Educational Evaluation. Some journals omitted an explicit reference to evaluation in their titles but highlighted it in their contents. These included Journal of Policy Analysis and Management, Performance Improvement Quarterly, and Policy Studies Review. In addition, the Journal of Negro Education continued to publish evaluative studies related to the education of Black children and racial disparities in the education system. Most originated in North America, with a few in Europe.

In the late 1970s and throughout the 1980s, the publication of evaluation books, including textbooks, reference books, and even compendia and encyclopedias of evaluation, increased markedly. In response to the demands for guidance and the collective experience gained from practicing evaluation in the field, a body of specialized evaluation literature developed and expanded.

Simultaneously, professional associations and related organizations were formed. The American Educational Research Association’s Division H was an initial focal point for professional activity in evaluation. Two national professional associations were founded in the United States that focused exclusively on evaluation: the Evaluation Network in 1975 and the Evaluation Research Society (ERS) in 1976. In 1986, these organizations merged to form the American Evaluation Association (AEA). Many local and regional AEA affiliates were created to offer options for more localized professional exchange related to evaluation.

With a growing literature, professional associations, and conferences where evaluators could exchange ideas with colleagues engaged in similar work, evaluation began to take shape as a distinct discipline and profession.

##### Development of Standards for Evaluation

In 1975,12 professional associations concerned with evaluation in education came together to form the Joint Committee on Standards for Educational Evaluation. The Committee’s charge was to develop standards that evaluators, evaluation clients, and evaluation consumers could use to judge the quality of program evaluations in education settings. In 1981, the Joint Committee published Standards for Evaluations of Educational Programs, Projects, and Materials. The initial 30 standards for evaluation were organized under the headings of utility, feasibility, propriety, and accuracy. That is, they called for evaluations to be useful, practical, ethical, and valid. The primacy of the standards for the utility of evaluations is notable. It signaled the profession’s commitment to providing useful and relevant service and information to evaluation stakeholders. A second edition of the standards was published in 1994 and a third in 2011; the latest edition introduced a new domain of standards focused on evaluation accountability.

In 1982, the Evaluation Research Society (ERS) published its own standards for program evaluation. They distinguished their standards from the Joint Committee’s by noting they were for program evaluation in any context, not just for educational programs (Evaluation Research Society Standards Committee, 198213). The 55 ERS standards were organized around five domains of evaluation activity, including (1) formulation and negotiation, (2) structure and design, (3) data collection and interpretation, (4) communication and disclosure, and (5) utilization. When ERS merged with the Evaluation Network (ENet) in 1986 to form the American Evaluation Association, the new organization determined that it would formulate new guidance to replace the ERS standards.

These activities to develop and communicate a shared understanding of what constituted quality in program evaluation were critical in shaping evaluation as a profession and distinguishing it from applied social and educational research. (In Chapter 313, we discuss the Joint Committee standards, the AEA guiding principles, and other sets of standards, principles, criteria, and competencies for program evaluation in more detail.)

##### Shifts in the Market for Evaluation and Role of Evaluators

While the infrastructure for professional evaluation was being formed, the markets for evaluation were changing dramatically. Ronald Reagan’s election as the U.S. president in 1980 led to a sharp decline in the number of federally funded evaluations. Instead, the federal government awarded block grants to states. States made their own decisions about spending and their own choices about evaluation requirements. However, the decline in evaluation at the federal level resulted in a healthy diversification of evaluation settings and approaches (Shadish et al., 1991 |D). Reflecting on the contraction of funding at the national level, Worthen (1995) Ö observed that evaluation became more commonplace among state agencies and local organizations, with a more authentic commitment to using the results. He noted, “Evaluation plays an increasingly important informational role as the level of the evaluation becomes more local, while evaluations at national levels typically continue to serve symbolic, non-informational functions” (p. 29).

Many state and local agencies began conducting their own evaluations, with less reliance on external experts. Foundations and other nonprofit organizations increased their attention to evaluation. Senge’s (1990) (□ book, The Fifth Discipline, spurred thinking about how organizations learn and change. This topic was highly relevant to evaluators since they were concerned with getting stakeholders in organizations to use evaluation information. In the education sector, evaluation continued to focus on student outcomes, measured by standardized testing. In other fields such as public administration, adult learning, and organizational management and change, there was a growing interest in leveraging evaluation to enhance organizational learning. Evaluative Inquiry for Learning in Organizations (Preskill & Torres, 1998 l5) brought these concepts to the attention of evaluators. Evaluators began thinking more broadly about the role of evaluation in organizations and tasks evaluators should perform. Reichardt (1994) published an article reflecting on what the field had learned from evaluation practice, suggesting that evaluators should become more involved in the planning stages of programs. He argued that evaluators’ skills might be more useful in a program’s beginning stages rather than after it ended. Evaluators increasingly used logic models (see Chapter 6 [□) to determine an evaluation’s focus and put that focus in an appropriate context. Engaging in logic model development helped program stakeholders to think about their programs evaluatively (Rogers & Williams, 200613).

In 1996, the United Way of America introduced a strategy for outcome measurement for the nonprofit agencies it funded. Their approach represented a convergence of the traditional focus on outcome evaluation with organizational learning. Measuring Program Outcomes: A Practical Approach (United Way of America, 199610) offered nonprofits a framework that included developing logic models to link inputs, activities, outputs, and outcomes; employing quantitative and repeated measures of outcomes; and using results for program improvement. The activities were not labeled as “evaluation" by United Way. Instead, they were characterized as “a modest effort simply to track outcomes” (Hendricks et al., 2008 (□, p. 16). United Way’s outcome focus was different from the outcome focus in the education sector in that (1) accountability was considered secondary to the purpose of program improvement and (2) expectations for measuring outcomes were generally more realistic than requirements for public-sector agencies. The United Way recognized that many nonprofit human service organizations lacked resources to conduct sophisticated evaluations of all outcomes. It also understood that engaging in evaluation could catalyze organizational learning and improvement. Nonprofit organizations, like many public-sector organizations, had typically reported inputs and activities to funders, with minimal attention to outcomes. The move to assess and monitor program outcomes was a notable shift to more comprehensive evaluation of nonprofit programs. Furthermore, it demonstrated that evaluation of outcomes and evaluation for organizational learning were indeed compatible.

##### Diversification of Evaluation Approaches and Methods

During this period, several prominent writers in the field proposed new and divergent approaches to evaluation. Evaluation moved beyond simply measuring whether objectives were attained. Evaluators started to consider program managers’ information needs. They realized that evaluation should address unintended outcomes as well as those that were intended. The importance of making judgments about merit and worth, not just goal achievement, became apparent. The role of values and standards in reaching those judgments gained attention among evaluation scholars and practitioners.

As evaluation funders and practitioners diversified, the nature and methods of evaluation adapted and changed. Formative evaluations provided feedback for incremental change and improvement. Evaluators helped programs theorize and measure links between program actions and outcomes. Patton developed his utilization-focused evaluation approach, which emphasized the importance of identifying intended evaluation users and adapting questions and methods to those users’ needs (Patton, 1975 (□, 1978 (□, 1986 [□)• Guba and Lincoln (1981) (□ urged evaluators to make greater use of qualitative methods to develop “thick descriptions” of programs, providing more authentic portrayals of the nature of programs in action. Fetterman (1984) (□ advocated for the use of ethnographic methods for educational evaluation. As different types of organizations funded more evaluations and expressed different needs, evaluators who had previously focused on policymakers (e.g., Congress, cabinet-level departments, legislators) as their primary audience began to consider multiple stakeholders and use more qualitative methods. Participatory methods for involving many different stakeholders, including those often detached from decision making, became commonplace. Although the dramatic decline in federal funding for evaluation caused anxiety among evaluators at the time, it actually increased the number, depth, and breadth of approaches to evaluation.

This burgeoning body of evaluation literature revealed sharp differences in the authors’ philosophical and methodological preferences. It also underscored a fact about which there was much agreement: Evaluation was a multidimensional technical and political enterprise that called for new theories and methods. Shadish and his colleagues (1991) i9 said it well when they noted, “As evaluation matured, its theory took on its own special character that resulted from the interplay among problems uncovered by practitioners, the solutions they tried, and traditions of the academic discipline of each evaluator" (p. 31).

21st Century Program Evaluation: 2000-Present

Today, evaluations are conducted in diverse settings using a variety of approaches and methods. The continued development of the field in the 21s1 century has involved rapid growth of evaluation around the globe, diffusion and mainstreaming of evaluation, proliferation of opportunities to learn about evaluation, development of evaluation competency frameworks, creation of systems for credentialing evaluators in Canada and Japan, issuance of policy statements by the American Evaluation Association, and increasing attention to culture, race, and social justice in evaluation.

##### Rapid Global Expansion

Evaluation has grown rapidly worldwide since the turn of the century, spurred by international collaborative efforts to enhance the practice and highlight the importance of evaluation for improving the human condition.

Rist’s (1990) (□ study of differences in evaluation across countries identified the United States, Canada, Germany, and Sweden as countries in the “first wave of evaluation development." These were countries where modern program evaluation originated in the 1960s and 1970s. In this first wave, evaluation was oriented to improving social programs and interventions. Countries included in the “second wave” are the United Kingdom, the Netherlands, Denmark, and France. In these countries, evaluation began as an effort to control federal budgets and reduce government spending. Evaluation was oriented more to accountability and identifying unproductive programs than to social experimentation and program improvement.

A follow-up study to Rist’s has not yet been conducted to identify specific countries in the third or subsequent waves of evaluation development, but there has been rapid growth of evaluation in Africa, Asia, and South America. Based on casual observation rather than systematic study, it appears that the expansion of evaluation in the global south was spurred in part by the demand for evaluation of international development efforts. Donors and multilateral agencies that were funding development projects wanted evidence of the value and impact of their investments. This led to a growing need for professionals in those locations who could conduct evaluations. Rather than relying solely on evaluators from North America and Europe, organizations and countries in the global south recognized and responded to a need to develop local expertise in evaluation.

In 2003, the International Organization for Cooperation in Evaluation (IOCE) was created. It began as a coalition of 24 voluntary organizations for professional evaluation (VOPEs) to “build and strengthen the global network of relationships between existing and emerging VOPEs.” (IOCE, n.d., n.p. i5). By 2016, the official count of VOPEs grew to 173, representing about 52,000 individual members worldwide (IOCE, 2016 (□).

Less than a decade after its formation, IOCE, along with UNICEF, helped facilitate the creation of EvalPartners in 2012. EvalPartners describes itself as an “innovative partnership whose members are Civil Society Organizations (CSOs) and Voluntary Organizations for Professional Evaluation (VOPEs)” (EvalPartners, 2021 [□, para. 1). It works to enhance the capacity of organizations within civil society to engage in “national evaluation processes, contributing to improved country-led evaluation systems and policies that are equity-focused and gender equality responsive” (para. 8). In addition to creating the first-ever global forum for knowledge sharing about evaluation, a major achievement of EvalPartners was its work to designate 2015 as the International Year of Evaluation. EvalYear 2015, as it came to be known, was endorsed by the United Nations General Assembly in Resolution 69/237: Building Capacity for the Evaluation of Development Activities at the Country Level (UN,

2015a). A UN press release described this resolution as follows:

For the first time in the history of the United Nations, a landmark, stand-alone UN Resolution on national evaluation capacity development has been adopted by the Second Committee of the United Nations General Assembly. This resolution was presented to the Committee by Fiji, and supported at global and country level by the United Nations Evaluation Group (UNEG) and national evaluation partners around the world. It received a very strong cross-regional sponsorship from more than 42 countries and a general consensus recognizing evaluation capacity as a country-level tool to strengthen evidence-based policymaking (UN Web TV, 2014).

The UN sees evaluation as playing a critical role in achieving its sustainable development goals, which span 17 areas that need to be addressed in order to “end poverty, protect the planet, and ensure prosperity for all” **(UN, n.d. (□).**

Along with practice, evaluation scholarship has expanded far beyond evaluation’s origins in Canada and the U.S. Examples of evaluation-focused journals that started in the 2000s include the *Evaluation Journal of Australasia,* which started in 2001; *Evidence Base,* launched by the Australia and New Zealand School of Government in 2012; and the *African Evaluation Journal* and the *International Journal of Evaluation and Research in Education,* both of which started in 2013.

##### Diffusion of Evaluation Practice

As evaluation has spread geographically, the practice has also become more widespread in terms of the individuals engaged in evaluation and evaluation-related tasks. In the nonprofit sector, it is common for in-house evaluators to have responsibility for major components of data collection and evaluation in their organizations. Individuals charged with evaluation are typically program managers and staff who have other program responsibilities. In 2003, Christie found that many of the evaluators she surveyed in California were internal and held other responsibilities, which were mostly management-related. Many had little or no training in evaluation and were unfamiliar with evaluation theories and approaches. Although we don’t know of other, more recent studies that reveal how pervasive this situation is, the plethora of basic evaluation guides and toolkits online suggests that many foundations and nonprofit organizations expect their grantees and employees to engage in evaluation with little or no formal training.

For some, the diffusion of evaluation responsibilities within organizations may raise concerns about the quality of evaluation studies. We believe the involvement of more individuals in evaluation endeavors has great advantages. In his American Evaluation Association presidential address in 2002, James Sanders (a coauthor of this book) advocated for mainstreaming evaluation, which he described as “the process of making evaluation an integral part of an organization’s everyday operations.” He pointed out the value of evaluation becoming “a routine part of the organization’s work ethic if it is mainstreamed. It is part of the culture and job responsibilities at all levels of the organization” (2002, p. 254). A study of evaluation practice and use within the federal government conducted by the U.S. Government Accountability Office (GAO) confirmed the value of evaluation mainstreaming. It concluded that “involving program managers and staff in evaluation studies and priority goal reviews offer promise for building capacity in a constrained budget environment” **(GAO, 2014 |D,** n.p.).

Professional evaluators are playing a crucial role in helping organizations develop their capacity for evaluation. A substantial amount of scholarship has been dedicated to this topic in recent years (e.g., Carden, 2017 (□; Edwards et al., 2016 (□; Grack-Nelson et al., 2018 |ö; Morkel & Ramasobama, 2017 (□; Ngwabi et al., 2020 iÇ; Mapitsa et al., 2019 (□; Nichols et al., 2018 (□; Zhao et al., 2017 (□). Fierro and Christie’s (2016) (□ study found that organizational capacity for evaluation and evaluation mainstreaming are enabled by several factors. Two salient factors are as follows:

* organizational infrastructure and policies such as a shared framework to guide practice, access to information about evaluation and completed evaluation studies, and supervisors who use and support evaluation
* personnel who have the requisite evaluation knowledge and skills and actually engage in conducting and using evaluation.

The latter point indicates a need for more opportunities for professionals to develop their evaluation skills. Fortunately, along with the diffusion of evaluation practice, the 21st century has also seen an expansion in terms of how individuals can build their evaluation knowledge and skills, described next.

##### Proliferation of Opportunities to Learn about Evaluation

According to **LaVelle (2018)** (□, as of 2017, there were 42 evaluation-specific certificate programs and 71 programs that offered master’s or doctoral degrees in evaluation in the United States. (Comparable data are not available for programs outside the U.S.) As noted above, however, many people involved in conducting evaluations within their organizations have primary professional identifications other than as evaluators, and they are not necessarily interested in becoming full-time evaluators. Therefore, university-based education is not the best option for these individuals to develop their evaluation capacity. In response to the growing demand for evaluation training, several nonacademic professional development opportunities have emerged.

Many in-person and online training options are available to practitioners who wish to develop or advance their evaluation skills through continuing education. Several voluntary organizations for professional evaluation around the world offer free or relatively low-cost opportunities to learn about evaluation. The International Organization for Cooperation in Evaluation maintains a directory of VOPEs on its website (<http://ioce.net>). The International Program for Development Evaluation Training, located in Switzerland and supported by the World Bank, offers advanced training for evaluators working in international development contexts. In the U.S., The Evaluators’ Institute is a well-respected evaluation training program. In addition to training, a plethora of evaluation workshops, webinars, guides, and toolkits have been developed by organizations specifically for their constituents. In short, there are numerous options for professionals to develop their evaluation knowledge and skills outside of formal degree programs.

##### Development of Evaluation Competency Frameworks

In 2005, Stevahn et al. proposed a set of competencies for evaluators. They argued that defining the competencies needed for evaluation practice would benefit the field by informing evaluation training programs, enhancing reflective practice by evaluators, helping to advance research on evaluation, and furthering the field’s professionalization **(Stevahn et al., 2005** (□). This competency framework defined the necessary skills, knowledge, behaviors, and attitudes for professional evaluators in the areas of (1) professional practice, (2) systematic inquiry, (3) situational analysis, (4) project management, and (5) reflective practice.

The framework developed by Stevahn et al. was adopted and adapted by the Canadian Evaluation Society as the foundation for its evaluator credentialing system, initiated in 2010 (see next section). Within a few years, various professional evaluation organizations and other groups substantially involved in evaluation began putting forth their own sets of competencies. These included the Aotearoa New Zealand Evaluation Association, Australasian Evaluation Society, European Evaluation Society, German Evaluation Society, International Development Evaluation Association, South African Department of Performance Monitoring and Evaluation, United Kingdom Evaluation Society, and United Nations Evaluation Group.

In 2015, the American Evaluation Association formed a task force to revisit the competencies published by Stevahn et al. The task force engaged AEA members as they reviewed and revised the competencies in advance of formal adoption by the organization. This process culminated in 2018 when the revised competencies were approved by the AEA board, following a vote by AEA members.

The development and endorsement of competencies for professional evaluators has been an important step in the field’s development. The competencies provide a common frame of reference for those involved in evaluation education and help cultivate a shared professional identity.

##### Creation of Evaluator Credentialing Programs

As mentioned above, a significant development in the professionalization of evaluation occurred in Canada. In 2010, the Canadian Evaluation Society (CES) launched its Credentialed Evaluator designation program. The program is voluntary and does not carry the weight of certification or licensure, which do not exist for evaluators. Instead, it indicates “the holder provided evidence of the education and experience required by the CES to be a competent evaluator” (Kuji-Shikatani, 2015 (□, p. 71).

Applicants for the Canadian credential prepare a portfolio for review by the CES credentialing board, which is made up of “senior evaluation professionals with at least 25 years of evaluation experience" (CES, 2014a i3, para. 2). The portfolio must demonstrate qualifications in three areas:

• a graduate degree or equivalent work experiences and professional development

• at least 2 years of evaluation work experience in the past 10 years

• capabilities related to at least 70% of the CES competencies (CES, 2014b (□).

For the first two qualifications, documentation must be provided as evidence of education and work experiences. To demonstrate the third qualification, applicants submit brief written explanations of how they achieved each selected competency. To maintain the credential, evaluators with the designation must obtain and document a minimum of 40 hours of continuing education credits over three years.

As suggested by these requirements, the Competencies for Canadian Evaluation Practice are the cornerstone of the CES credentialing program. These competencies define the knowledge, skills, and attitudes required for professional evaluation practice in five areas: reflective practice, technical practice, situational practice, management practice, and interpersonal practice. The Canadian list of competencies was based on the one developed by Stevahn et al. (2005) (□ in the U.S.

Before the CES credentialing program started, the Japan Evaluation Society (JES) initiated an evaluator certification program. The JES program has three levels: certified professional evaluator, certified specialty evaluator (for educational evaluators), and certified advanced evaluator. To obtain the JES certifications, an evaluator must attend a multi-day training, then pass a multiple-choice exam (Wilcox & King, 2013 (□).

Even though the effort to define competencies originated in the U.S., the topics of evaluator credentialing and certification are controversial among U.S. evaluators. To clarify, in the U.S., credentialing usually involves a person completing a specified training program but does not require the person to demonstrate competency. In contrast, certification requires the person to demonstrate mastery of skills through formal assessment managed by an external body such as a professional organization. Some professions (such as accounting, cosmetology, and law) require professional to have a license, which is typically issued by a governmental body (Altschuld, 200519). Some see credentialing as a necessary and inevitable step in the process of professionalizing evaluation. Others regard it as unmanageable in a field where there is such diversity of perspectives, approaches, and contexts for practice. The fact that evaluations are often conducted by teams whose members have complementary skills is also a complicating factor in terms of credentialing or certifying individual evaluators in the U.S.

##### Publication of AEA Policy Statements

As described previously, the evaluation discipline has been shaped by the market’s demands and needs for evaluation. Federal policies and different organizations’ interest in evaluation led to increased numbers of evaluators and training programs. Those same forces shaped the nature of the work. As the profession grew, the flagship professional evaluation association, AEA, turned the tables with attempts to shape evaluation policy and practice. Having reached a critical mass in terms of size and collective expertise, it began issuing policy statements in the early 2000s.

In 2002, AEA issued a position statement on high-stakes testing in PreK-12 education (AEA, 2002 (□). The statement summarized research on the risks and benefits of high-stakes student testing and concluded that “evidence of the impact of high stakes testing shows it to be an evaluative practice where the harm outweighs the benefits" (2003 [□, p. 1). The AEA task force wrote:

Although used for more than two decades, state mandated high stakes testing has not improved the quality of schools; nor diminished disparities in academic achievement along gender, race, or class lines; nor moved the country forward in moral, social, or economic terms. The American Evaluation Association (AEA) is a staunch supporter of accountability, but not test driven accountability. AEA joins many other professional associations in opposing the inappropriate use of tests to make high stakes decisions. (2003 1P.P.1)

The task force made several recommendations for improved evaluation practice. They recommended better test validation, better alignment between tests and intended uses, use of multiple measures, and consideration of a wide range of perspectives when determining what is in the best interests of students.

In 2006, AEA issued its second policy statement, on the subject of educational accountability (AEA, 2006 [□). This statement expressed concerns about three significant issues:

• Overreliance on standardized test scores that are not necessarily accurate measures of student learning, especially for very young and for historically underserved students, and that do not capture complex educational processes or achievements;

• Definitions of success that require test score increases that are higher or faster than historical evidence suggests is possible; and

• A one-size-fits-all approach that may be insensitive to local contextual variables or to local educational efforts, (p. 1)

This AEA policy statement encouraged use of multiple measures, measures of individual student progress over time, context-sensitive reporting, use of data to inform decisions about resource allocation, accessible appeals processes, and public participation and access.

AEA’s most proactive attempt to influence evaluation at the federal level to date was its issuance of An Evaluation Roadmap fora More Effective Government (AEA, 2016 (□). The organization declared that “evaluation is an essential function of government" (p. 2) and offered the document as a “roadmap for improving government through evaluation” (p. 2). The comprehensive document included

• recommendations for specific ways the government should use evaluation

• a framework that listed 17 activities to support effective evaluation, addressing evaluation, management, quality and independence, and transparency

• principles to guide practice with regard to evaluation scope, analytic approaches and methods, resources, professional competence, evaluation plans, dissemination of results, follow-up and tracking, and policies and procedures

• distinct strategies for institutionalizing evaluation to be adopted by the executive and legislative branches.

Although many factors influence evaluation policies and practices at the federal level, there are hints that AEA’s roadmap has caught the attention of those working in government. In the U.S. Office of Management and Budget’s (2018) federal budget publication Analytical Perspectives, Chapter 6 [□, titled “Building and Using Evidence to Improve Government Effectiveness," called on federal agencies to build and use portfolios of evidence, develop learning agendas, create evidence infrastructure, and maximize use of administrative data. The specific strategies recommended to advance this agenda are highly consistent with the AEA roadmap’s recommendations for institutionalizing evaluation.

##### Increasing Attention to Culture, Diversity, and Social Justice in Evaluation

In her synopsis of the history of evaluation, McBride (2018) [□ pointed out increasing attention to culture and social justice in evaluation since 2000. Several developments demonstrate this heightened focus in the field:

• Formation of the Center for Culturally Responsive Evaluation and Assessment (CREA): Founded in 2003 by noted evaluation scholar Stafford Flood, this center at the University of Illinois promotes “a culturally responsive stance in all forms of systematic inquiry including evaluation, assessment, policy analysis, applied research, and action research” (CREA, n.d. (□). Since 2013, CREA has held an annual conference to bring together scholars and practitioners interested in advancing cultural responsiveness and attending to power and privilege in evaluation.

• heightened attention to culture in the Program Evaluation Standards: In the third edition of the Program Evaluation Standards (Joint Committee, 2011), 4 of 30 standards mention culture—one in each of the domains of utility, feasibility, propriety, and accuracy. In contrast, none of the standards in the second edition published in 1994 mentioned culture.

• Graduate Education Diversity Internship (GEDI) program: The GEDI program provides evaluation training and paid internships to graduate students from historically underrepresented groups. It started in 2004 to diversify the evaluation profession, improve the capacity of evaluators to work competently with diverse populations, and deepen thinking about cultural responsiveness in evaluation (AEA, n.d. e>-

• American Evaluation Association’s public statement on cultural competence in evaluation: In 2011, AEA issued an 11-page document describing the role and importance of culture in evaluation and how evaluators can conduct their work with cultural competence. (Chapter 3 (□ includes more details on this statement.)

The body of literature on culture, diversity, and social justice in evaluation has also grown dramatically since 2000. In Chapter 10 (□, we discuss the role of culture in evaluation in more detail.