

# **Constructing Effectiveness: The Emergence of the Evaluation Research Industry**

*LBJ Centennial Symposium, LBJ School of Public Affairs, December 4-5, 2008*

*by Peter Frumkin and Kimberly Francis*

## **Abstract**

Evaluation research is today an enormous industry, populated by a limited number of large firms and countless smaller ones. In this chapter, we explore the origins and evolution of this industry and the broader program evaluation profession by looking back at the period of the LBJ presidency when evaluation grew as a response to the enlargement of government and the need to track the effectiveness of new programs. We consider some of early and important evaluations funded by the Office of Economic Opportunity, the Department of Health Education and Welfare, and other agencies, which helped launch and define this field. We also detail the growth and evolution of firms such as Rand, Mathematica, Abt, Westat, and others during these early formative years. Finally, the subsequent professionalization and consolidation of the field of program evaluation is discussed along with the future of the practice of program evaluation.

## **Introduction**

Today, evaluation research is a multi-billion dollar industry focused on answering some variation on the seemingly simple the question: “Did the program work?” Over the past four decades, this enormously complex question has led to the creation of a limited set of large and successful firms – and a massive array of smaller and specialized firms -- that collectively employ a large number of trained experts who spend entire careers searching for evidence of impact and effectiveness. In this chapter, we sketch a brief interpretive history of the evaluation industry, tracking the emergence and expansion of the largest and most visible organizational manifestations of the drive to track effectiveness. Our intent is not to create a comprehensive historical narrative the encompasses all the many actors in this long and intricate story line, but rather to pull out selected moments in the emergence of an increasingly unified and organizational field.

## **Government Programs and the Creation of a New Industry**

So we have seen, in our time, two aspects of intellectual power brought to bear on our Nation’s problems: the power to create, to discover and propose new remedies for what ails us; and the power then to administer complex programs in a rational way. But there is a third aspect of intellectual power that our country urgently needs tonight, and in my judgment it is being supplied sparingly. It is less glamorous...it is less visible and less publicized.... But it is not a bit less critical

to the success or to the failure that we may make in the years that are ahead of us. This is the power to evaluate.”

*-President Lyndon Johnson, September 29, 1966*

When President Johnson made this statement on the occasion of the 50th anniversary of the Brookings Institution, program evaluation was in its infancy as a professional field. At this early stage, major institutional pressures were taking shape in the form of new legislation and policies within the executive and legislative branches, which directly influenced the surge in demand for evaluation during the latter half of the 1960s and throughout the 1970s. Answering the call were hundreds of contract research and evaluation organizations that had either diversified into a new market or had just recently formed.

The most important public policy pressure was the dramatic increase in social spending, exemplified by the Economic Opportunity Act (EOA) in 1964, as well as the Elementary and Secondary Education Act and the Social Security Amendments, both in 1965. Social welfare spending totaled just over \$77 billion in 1965, and increased to almost \$146 billion by 1970, \$290 billion by 1975, and \$493 billion by 1980 (Haveman, 1987). In another example, for a period of time in 1973 the \$110 billion budget of the Department of Health, Education, and Welfare (HEW) was larger than that of the Department of Defense (Staats, 1973). Along with these unprecedented increases in public spending on social programs came a keen interest in which innovative programs were the most effective, sparking a “gold rush” of large-scale quantitative evaluations (House, 1990; Rossi and Wright, 1984). Over time, as federal budget resources became scarce and disillusionment with the effectiveness of social programs grew, accountability, cost-effectiveness, and evaluation became even more critical to the social policy enterprise (Schick, 1971; Shadish, Cook, and Leviton, 1991).

Several important events occurring in the federal government in 1965 and 1966 signaled the birth of modern program evaluation: 1) President Johnson issued an executive order to implement the Planning-Programming-Budgeting-System (PPBS), then in use at the Department of Defense, throughout the agencies of the executive branch; 2) the Office of Economic Opportunity (OEO) launched several national anti-poverty programs, funded 13 evaluations of Head Start, and sponsored the creation of the Institute for Research on Poverty at the University of Wisconsin; 3) the Office of the Assistant Secretary for Planning and Evaluation (ASPE) was established within HEW; and Title 1 of the Elementary and Secondary Education Act (ESEA) included an evaluation reporting requirement, the first major social legislation to do so. While there are doubtless many other signal moments in government’s embrace of evaluation, we focus on these three events because they contributed in ways that almost all of the early leaders of the program evaluation field we interviewed argued were critical to the creation of a new evaluation industry.

### **Planning-Programming-Budgeting-System**

A keystone of the “analytic revolution” in government (O’Connor, 2001), PPBS gave President Johnson a way to centralize control of the major nationwide antipoverty programs, which were receiving weak political support at the local level (Jardini, 1996). It brought a rational decision-making model to the agencies administering these programs, and primed agency staff and legislators to begin thinking in evaluative terms. Rooted in military operations research and the systems analysis framework developed by RAND in the 1950s, PPBS was a management tool that required all agencies to define their program objectives, project the costs of alternative ways to attain these objectives, and improve performance by achieving the highest benefit for the lowest cost (Staats, 1968; Held, 1966).

While PPBS helped move the federal government toward a more analytical approach to program planning and resource allocation, its inherent complexity eventually became too cumbersome for practical implementation across the agencies (Weiss, 1972). Moreover, the system was designed primarily to improve efficiency in the military—cost and benefit projections of alternative courses of action toward a defined military goal (e.g., which weapons system is the most cost-effective at destroying a particular target). Newly minted social policy analysts struggled to define criteria by which to measure effectiveness of social programs (Staats, 1970; O’Connor, 2001; Wholey, Scanlon, Duffy, Fukumoto, and Vogt, 1970), and suffered from a lack of available data on which to draw (Gorham, 1967; Weiss, 1972).

Interest in PPBS eventually flagged and Johnson’s management innovation was abandoned in 1971. But the government-wide mandate had two important byproducts: First, it boosted demand for systematic data and skilled analysts to perform program evaluations, who at the time were principally found in the Department of Defense and RAND (O’Connor, 2001). Second, it set in motion the idea that evaluation and analysis are necessary components of any legitimate social policy design and implementation process.

### **Office of Economic Opportunity**

As PPBS was spreading through the executive branch, the OEO was breaking new evaluation ground with its anti-poverty programs. The original EOA of 1964 did not specifically mandate program evaluation, but was sufficiently broad to allow for evaluation funding at about 1 percent of program funds (Wholey et al., 1970). In large part due to its financial resources and staff designated for evaluation, the Office of Research, Plans, Programs and Evaluation (ORPPE) was an early leader in federal program evaluation. An Urban Institute review of the evaluation practices among four federal agencies described the OEO as having the most highly developed evaluation system compared to HEW and HUD. Not surprisingly, the other agencies were considered to be “grossly underfunded” for evaluation (Wholey et al., 1970). In the late 1960s and early 1970s the ORRPE engaged in ambitious evaluations of Head Start,

Follow Through, Upward Bound, VISTA, Job Corps, and Neighborhood Health Centers, as well as the first large-scale social experiment of the era, the Negative Income Tax experiments. The unfavorable results of the Westinghouse national evaluation of Head Start served as a controversial introduction to assessment for the OEO, but this did not quell the executive branch's interest in expanding its evaluation capacity (O'Connor, 2001).

By the close of the 1960s most federal agencies had planning and evaluation departments to coordinate the evaluations that were now required with most legislation. One estimate found that 800 policy analysts were working across sixteen domestic policy research agencies within the government at this time (Smith, 1991). Nonetheless, staffing was sparse due to remaining funding barriers and the challenge of recruiting skilled researchers to federal evaluation positions at the time (Shadish et al., 1991; Wholey et al., 1970). Facing an enormous demand for evaluation research connected to new public programs, coupled with limited in-house capacity, federal evaluation departments looked outside government to contract research firms as a way of quickly increasing operational capacity. There was an estimated 500 percent increase in federal expenditures on evaluation between 1969 and 1974, with about 60 percent of the 1974 expenditures going toward contract research alone (Rein and White, 1977).

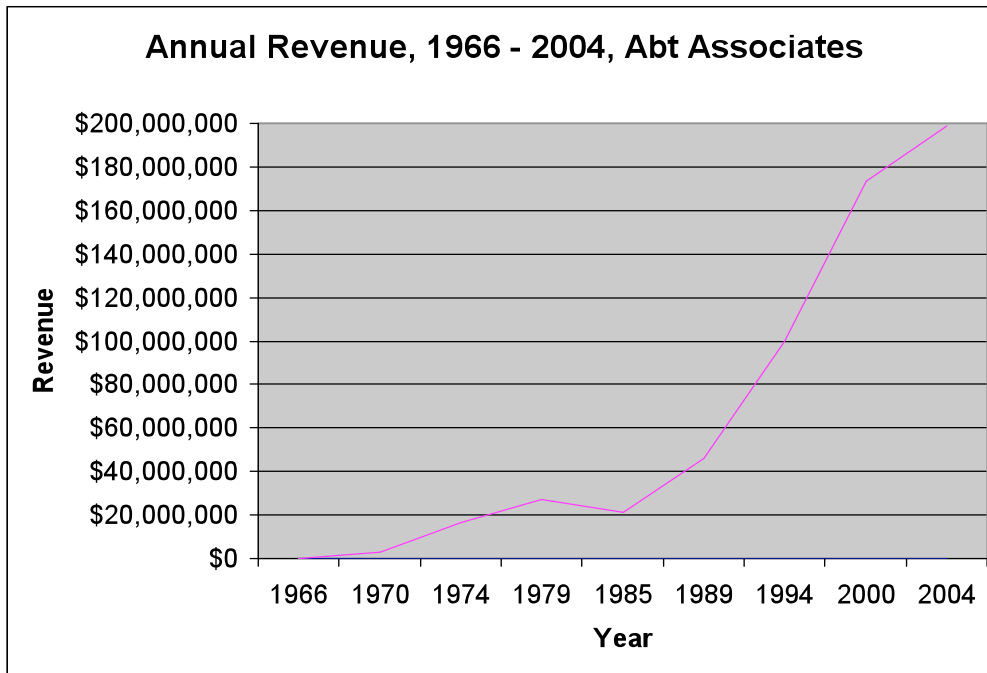
Among the beneficiaries of these new federal dollars were the entrepreneurs who developed research organizations to serve the growing appetite for social research and evaluation. In 1970, there was an estimated 300 firms (both for profit and not-for-profit) qualified to receive Requests for Proposals from the OEO (Wholey et al., 1970). Some of these firms existed prior to the advent of modern program evaluation in the mid-1960s, such as RAND, Westat, and American Institutes for Research (AIR), and some were initiated in response to or anticipation of the new market. For existing government contractors working in allied research fields, the increased demand for program evaluation in the federal sphere provided an opportunity to diversify and strengthen their industry.

A couple of years before President Johnson's 1966 address to the Brookings Institution, Clark Abt assembled an interdisciplinary collaborative of social scientists and engineers in Cambridge, Massachusetts, to begin offering planning, research, and evaluation services to federal government programs. Clark Abt's training and experience was rooted in the defense research industry and the tools of systems analysis and cost-benefit analysis—indeed, Abt Associate's first contract was to design a game for the Department of Defense to teach counter-insurgency strategy to military trainees—but during his years at defense contractor Raytheon, he had seen the opportunity to apply those analytical skills to examining and solving problems in education, housing, and social welfare. Abt's interest in instructional game design led to further contracts for elementary school educational games, and the eventual publication of *Serious Games* (Abt, 1970), which argued that simulations could be used to guide decision-making in the business, government, and education sectors.

In the late 1960s the newly created Office of Economic Opportunity asked Abt Associates to evaluate several of its War on Poverty programs, and in 1972 the company began its first large-scale social experiment for HUD—the Housing Allowance Demand Experiment. Also in 1972, Abt Associates took over from Stanford Research Institute the responsibility for the national evaluation of Follow Through, a massive educational experiment aimed at finding ways to break the cycle of poverty through better education. Follow Through started in 1967 as part of President Johnson’s War on Poverty, lasted almost two decades, costing hundreds of millions of dollars.

The noble intent of the fledgling Department of Education (DOE) and the Office of Economic Opportunity was to break the cycle of poverty through better education. Poor academic performance was known to correlate directly with poverty. Poor education then led to less economic opportunity for those children when they became adults, thus ensuring poverty for the next generation. Follow Through planned to evaluate whether the poorest schools in America, both economically and academically impoverished, could be brought up to a level comparable with mainstream America. The actual achievement of the children would be used to determine success. By this time, Abt Associates was well on its way to becoming one of the top-performing contract research firms for the federal government, specializing in social and economic programs (see Figure 1). Its early work on these education programs allowed it to build its practice on the national scene and develop the capacity to carry out very large projects.

**Figure 1.**



While several of the large scale early evaluations were taking place under the direction of the executive branch for the purposes of finding out what worked and even how program implementation might be improved, conservatives were interested in using evaluation to find out which programs did not work and could be terminated. As a senior executive at Abt Associates put it: “Conservatives like Nixon after his 1968 victory wanted to save money on social programs by evaluating their ineffectiveness and cutting away what they felt was liberal waste, and liberals wanted to do just the opposite, to prove that the social programs were having some productive effect.” Evaluation research—while in principle operating independently of politics—was thus early on seen by some as a potentially powerful instrument of policy change. In terms of driving the behavior of policymakers, evidence of program failure could prove to be as potent, or even more potent, than evidence of impact.

The performance of Johnson’s anti-poverty programs came under congressional scrutiny in 1967, and the hearings surrounding the re-authorization of the EOA resulted in the passing of the so-called Prouty Amendment named for Senator Winston Prouty (R-Vermont). This legislation required the General Accounting Office (GAO) to review the effectiveness of several anti-poverty programs and the rigor of the evaluations administered by the OEO (Sperry, 1981). The subsequent favorable review of the OEO’s evaluation practices in 1969 further legitimated and justified expanded funding for program evaluation (O’Connor, 2001), and this requirement became an essential element in most social program legislation of the period.

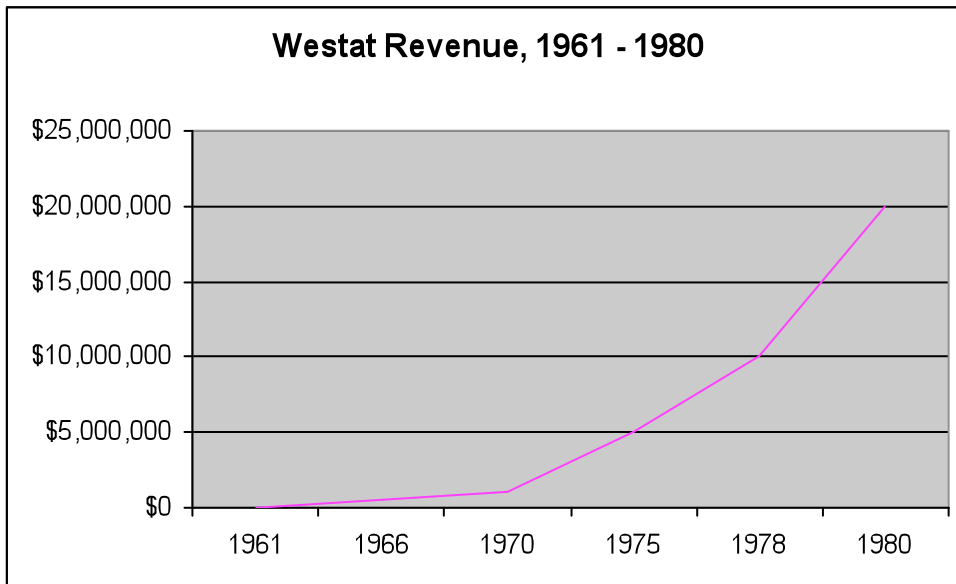
Aside from spurring the development of evaluation start-up firms, new evaluation legislation bolstered the organizations in existence well before the drastic increase in domestic spending. One such firm was Westat, a statistical and survey research consulting firm that became involved in a national evaluation of daycare for the OEO in 1970. Westat emerged in 1961 as a partnership among three statisticians from the University of Wyoming: Dr. Edward C. Bryant, who was chair of the statistics department at the time, and Donald W. King and James Daley, both of whom had graduated from Wyoming with masters degrees in statistics in 1960. In 1961, both King and Daley were looking for jobs, and Bryant was leaving his academic post for health reasons. Daley, aware of Bryant’s health constraints, suggested forming a consulting company to serve the statistical needs of government, business, and industry. The idea appealed to Bryant, and with the addition of King, the three formed the partnership that became Westat (Bryant, 1981). Morris Hansen, a Wyoming graduate and survey statistics pioneer from the U.S. Census Bureau, later joined as Senior Vice President in 1968.

Primarily a statistical consulting firm (rather than a social science or defense consulting firm), Westat’s first projects included expert testimony in a lawsuit about the value of a uranium mine, quality control of crushed rock for a construction company, an assessment of the efficiency of the use of live animals in research for the Humane Society of the

United States, and the first Westat survey, focused on 80 customers of a bank in Golden, Colorado. In 1962, Westat was awarded its first major contract that helped to establish the firm as a major federal contractor: a five-year project to help the U.S. Patent Office improve information retrieval processes. In the late 1960s, Westat's leadership decided that adding survey research expertise to their statistical capabilities would increase their marketability (Bryant, 1981) as by this time more large scale social program evaluations were taking shape across the federal landscape. Westat's early survey work included the national survey of day care for the Westinghouse Learning Corporation (under an OEO contract), and a two-year longitudinal evaluation of the Public Employment Program for the DOL. From 1970 to 1980, Westat grew from \$1 million in revenue to \$20 million (see Figure 2).

Westat now provides a wide range of research and evaluation services in the areas of health, clinical trials, social services, employment/national service, housing, education, substance use, energy and environment, science and technology, transportation, military human resources, and marketing research, with 2006 revenue in excess of \$425 million.

**Figure 2.**



Source: Bryant, 1981.

### **The Department of Health, Education, and Welfare and the Elementary and Secondary Education Act/Title I**

HEW was also progressively becoming involved in evaluation. The ASPE office was created in 1965 primarily to coordinate the implementation of PPBS throughout the

agency, and William Gorham, former RAND and defense department analyst, was brought in to head the effort (two years later he became the founding president of the Urban Institute). He soon found that the areas of health, education, and welfare suffered from a profound lack of data suitable for evaluation and the cost-benefit analyses of PPBS. That same year, Gorham recommended to Charles Schultze, Bureau of the Budget Director, that 1 percent of all appropriations to HEW be designated for program evaluation (Gorham, personal communication, 2007). This suggestion eventually became law in 1967 and 1968 through 11 pieces of legislation, though in practice evaluation funding among Department of Labor (DOL), HEW, OEO, and HUD averaged about 0.4 percent of program funds in 1969 (Wholey et al., 1970).

One of ASPE's tasks was to oversee the landmark evaluation mandate of Title I of the ESEA. Originally pushed by Senator Robert Kennedy as a way for local schools to be accountable to parents and communities for how the money was used, ASPE saw the evaluation requirement as an opportunity to test different education strategies targeted to disadvantaged children using the input-output model of the PPBS (McLaughlin, 1975). Either way, this was an example of direct institutional pressure helping to create an industry. HEW enlisted the services of American Institutes for Research (AIR), a contract research firm in existence since 1946, to write case studies of effective Title I programs and to review the evaluation reporting practices over the seven-year period of 1965 to 1972.

AIR was a product of the post-World War II research boom that also produced RAND and the Stanford Research Institute. After John C. Flanagan, an industrial psychologist for the U.S. Army Air Corps in World War II, joined the psychology department at the University of Pittsburgh, he started AIR to focus on workforce, personnel, and education research, which are still emphasized by the firm today. Previously, Flanagan had developed aptitude tests for the Aviation Psychology Program using the "critical incident technique" to evaluate Army Air Corps candidates (University Times, University of Pittsburgh, 1996).

AIR's first major educational research effort was the 1957 Project TALENT, a longitudinal survey of high school students that measured the aptitudes and interests of a national sample of 440,000 students. The database became a national resource for improving education through vocational guidance and curriculum development (AIR News, Winter 2007). In the 1970s AIR leveraged their expertise in educational evaluation and began conducting evaluations of domestic social programs, including delinquency prevention programs for Office of Juvenile Justice and Delinquency Prevention in 1976. Currently, one of AIR's high-profile contracts is to support the National Center for Education Statistics for the U.S. Department of Education.

The initial attempt to mandate evaluation reporting from the school districts receiving Title I funding was unsuccessful, for reasons chronicled in other studies (McLaughlin 1975). These included the incompatibility of the cost-benefit evaluation design with the messy reality of school systems, the guarded resistance of school administrators toward

collecting the data needed by ASPE, and the resultant lack of usable information for management or accountability. Despite the dismal implementation of the ESEA's evaluation activities, federal agencies proceeded to engage in even more evaluation, "...information gathering has become a necessary activity (qua activity) in the policy system, and faith in the science of systems analysis remains undiminished at the higher echelons of the federal government" (McLaughlin, 1975, p. 118). An almost ritualistic and unreflective embrace of the idea of evaluation (later identified by Carol Weiss as "symbolic" evaluation) seems to have helped perpetuate the early diffusion of a culture of evaluation, no matter if the state of practice remained imperfect (Meyer and Rowan, 1991).

The culture of evaluation had been present in Washington for years and related very much to the early needs of HEW. The Urban Institute is an example of an industry stalwart that did not emerge to meet government contracting trends and to seize a new market. Rather it was formed directly and consciously by government interests. The idea for an urban research institute started all the way back with President Johnson's 1964 Task Force on Cities, which recommended a "national Institute of Urban Development" to be part of a cabinet-level Department of Housing and Urban Development (Bassett, 1969). Three years later, in the context of the proliferation of somewhat uncoordinated Great Society programs and a tightening domestic budget, President Johnson's White House staff began to plan the new institute, but as an entity separate from any federal agency (Smith, 1991). Special Assistant to President Johnson Joseph Califano cited two problems that jumpstarted the planning process for the new research institute: a severe lack of data appropriate for policy decision-making, and the lack of objective program analysis and evaluation. In one example, Califano found that no one within HEW could tell him who was receiving welfare benefits except that it was approximately 7.3 percent of the population (Bassett, 1969).

The plan was to create a nonprofit research institution that could provide nonpartisan analysis of the nation's urban problems and use these data to advise government on appropriate programs and policies. Initial start-up funds were provided primarily by HUD and the Ford Foundation, with additional contributions from DOL, Department of Transportation, HEW, and OEO. The model for this institute was RAND, a think tank designed to serve the exclusive research needs of the defense department until it started to diversify into domestic research in the late 1960s (Bassett, 1969; Hayes and Japha, 1978; Smith, 1991). In fact, RAND tried for a year to land the federal contract to locate the Urban Institute within its purview (Jardini, 1996).

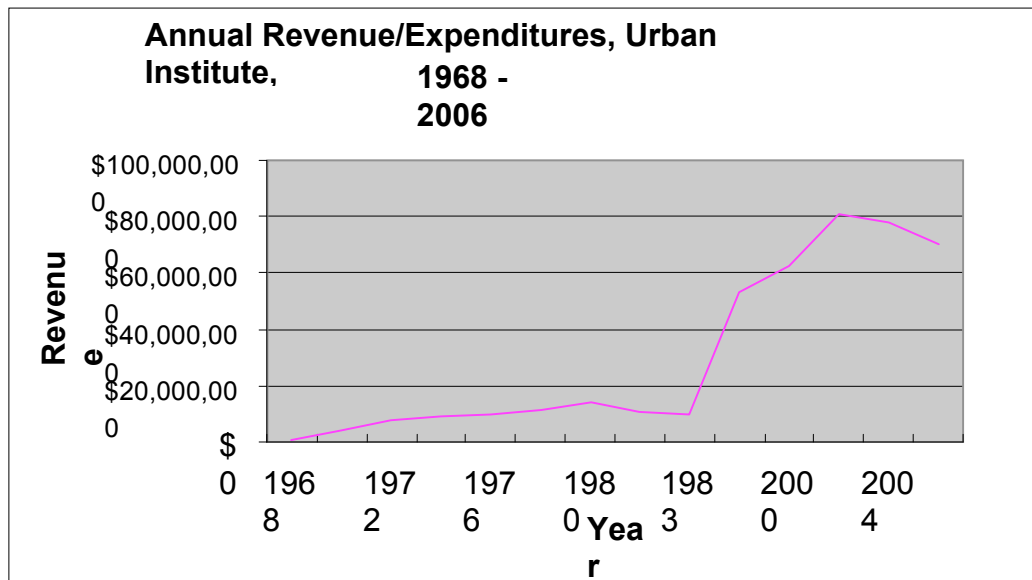
The Urban Institute was different from the RAND model, which has started off focused on the needs of the Air Force, in that the Urban Institute would serve any federal agency involved in urban programs, rather than the exclusive needs of HUD. Its first major project was to examine how four government agencies (HUD, OEO, HEW, and Labor) evaluated the social programs they sponsored, culminating in the classic study by Wholey (1970) and his associates, "Federal Evaluation Policy: Analyzing the Effects of Public Programs." Other milestones included the design of a major social experiment, the

Experimental Housing Allowance Program, and the development of computer models to simulate how changes in food stamp and welfare programs would affect family incomes.

The environment in Washington would hardly stay stable. With the new Nixon administration in 1969 came greater control over the research activities of contractors. The funding relationship with HUD survived in part because the contract was amended to give HUD the authority to determine the research projects to be under. The unrestricted funds that were briefly enjoyed by the Urban Institute were thus no more, though staff were still able to propose research ideas to HUD (Hayes and Japha, 1978). Overall, this move to add controls to the nascent evaluation field further consolidated the industry, with research organizations soon conforming and mirroring the standards and practices of government funders at HEW, HUD, and DOL—on whom all the early players were dependent for financial support. The power to control funding and to define projects would be crucial to the establishment of a new industry that would be independent of government but responsive to its needs.

By the 1980s, funding for the Urban Institute shifted away from the exclusive focus on federal grants and contracts and began to include more foundation funding (Smith, 1991). Over time, the Urban Institute turned into a multi-dimensional policy research institute, working on a wide array of projects, funded by an equally diverse array of clients. The Urban Institute was led to find a way to enhance its viability and diversify its revenue mix in the face of declining federal resources in the 1980s. This move also had the salient effect of reducing the level of government influence on its organizational focus, structure, and activities. Figure 3 chronicles revenue growth during this period of diversification and expansion, from \$10 million in the 1980s to \$80 million 20 years later.

Figure 3.



Source: Bovbjerg, 1983; Hayes and Japha, 1978. Note: Data from 1983-1996 are missing.

The Urban Institute's chosen path aside, in the nascent stages of the evaluation field, and for several firms and institutes this holds true today, contract research and evaluation firms were heavily dependent on the same sources of federal government evaluation dollars, mainly from HEW, OEO, and HUD. What we know about organizational fields (DiMaggio and Powell, 1991) suggests that this places pressure on competing firms to offer similar services and to structure themselves in similar ways, in order to respond to industry standards expected by the federal funders. As one CEO of a leading federal evaluation firm confirmed,

The dominant clients of most of the key players...are either federal government, state and local governments...those clients have requirements, have modes of doing business, have expectations that all of the players have to be responsive to. And so you get lots of similarities [among firms] driven primarily by client requirements, client behaviors, client characteristics.

In this way, the state can consciously or unconsciously shape industry standards by asking for certain approaches and methods of evaluation, as well as determining what is to be evaluated.

The Nixon administration eventually dismantled the OEO and most of its programs were absorbed into existing bureaucracies at HEW, HUD, and DOL. Social spending, including program evaluation and applied social research, continued to increase until 1981 when the Reagan administration drastically cut many of the programs built over the preceding twenty years. The evaluation industry, as illustrated by the emergence of several contract research firms and the Urban Institute, emerged in the context of governmental pressures and constraints in the form of funding mandates and the expectations of a new federal evaluation culture. We argue that these "coercive" pressures act on the industry to create a loosely consolidated field.

### **Diffusion of Existing Evaluation Models**

Rapidly unfolding legislative change in the 1960s led to a period of uncertainty surrounding the new evaluation mandates. The uncertainty rested in the growing expectations for data-driven decision making throughout government coupled with a social policy arena that had little or no capacity to carry this out. One response to institutional conditions marked by ambiguous goals and lack of specific technical expertise is to look at what has been successful in neighboring organizational sectors, and adopt or adapt those models in the new setting (DiMaggio and Powell, 1991). President Johnson's executive order for widespread implementation of PPBS is an example of this sort of response; a tried and tested management system that had worked well for the DOD was a logical next step for HEW and other agencies. Soon, analysts who were trained to evaluate the efficiency and effectiveness of military programs were being asked

to apply evaluation and cost-benefit analysis worked in the context of a social program, where the outcomes were more difficult to operationalize.

The PPBS was rooted in systems analysis, an approach created by RAND for the military in their efforts to develop a scientific approach to managing war. When Robert McNamara decided to use PPBS as a management tool, he brought its chief architects from RAND to the DOD. With the expansion into the domestic arena, the RAND model was diffused even more widely, which gave this federal contractor remarkable organizational influence. RAND emerged from the post-World War II recognition of the increased military success attributed to technology research and development, and was sustained by burgeoning defense contracts during the Cold War years. A collaboration of the War Department, the Office of Scientific Research and Development, and private industry, in 1948 RAND separated from the Douglas Aircraft Company, established its nonprofit status and by 1950 secured a \$1 million interest-free loan from the Ford Foundation for operating expenses (Jardini, 1996).

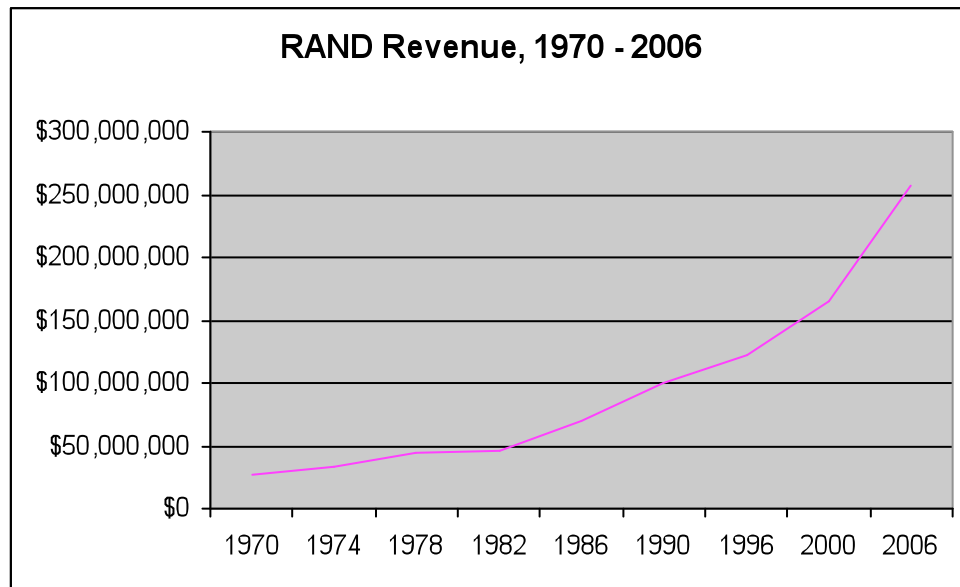
Though RAND's original purpose was narrowly defined by military defense issues (specifically Air Force contracts), over the years it expanded into other sectors as defense contracts became less lucrative and the country's interests turned to domestic social policies in the mid-1960s. RAND made its first overture into non-defense research in 1958 with a \$35,000 Ford Foundation study to assess the applicability of systems analysis to elementary and secondary education. Two years later, Ford paid RAND \$500,000 for a three-year analysis of urban transportation systems. Then, with the widespread implementation of PPBS throughout the executive branch in 1965, RAND analysts were in high demand in the social policy arena. Henry Rowen, a former RAND economist who had recently been in charge of implementing PPBS across federal social welfare agencies at the Bureau of the Budget, returned to RAND as president in 1966—the same year its board of trustees terminated the exclusive focus on military research and moved into the social policy arena (Jardini, 1996). Other RAND alumni brought in to high-level government positions included Charles Hitch, Alain Enthoven, and William Gorham.

Though RAND lost a year-long battle to house the Urban Institute, the New York City-RAND Institute was established in 1968 with help from the Ford Foundation and became a springboard for RAND's diversification into social policy research (Jardini, 1996). Throughout the 1970s, RAND's income from the domestic policy arena grew to equal its income from defense concerns; but the political climate of the 1980s reversed this trend and social policy research shrank to about 20 percent of RAND's total revenue (RAND annual reports, 1970-1990). Overall, RAND's income grew to over \$250 million by 2006 (see Figure 4).

The story of RAND's diversification into social research underscores the influence of a changing policy environment on the evolution of the evaluation industry. In response both to government need for management tools and RAND's need to generate income and remain competitive, they replicated their innovative systems analysis approach to management in the uncharted waters of health, education, and welfare administration. In

this way, diversifying into new policy sectors was at once an innovative and mimetic response to an uncertain contracting environment, where sole dependence on defense resources was no longer viable.

Figure 4.



Source: RAND Annual Reports, 1970- 2006.

Research Triangle Institute (RTI), a contract research and development institute founded in 1959, had a similar experience. Grounded in chemistry, physics, statistics, engineering, semiconductors, and civil defense research, the late 1960s brought a profound shift in RTI's research docket. From 1966 to 1969 RTI's revenue from health, education, population, environment, and transportation sectors rose from 21 percent to 65 percent of the total (Larrabee, 1991). This growth coincided with RTI's desire to double the size of the Institute in five years (Larrabee, 1991), and reflected the changing research priorities of federal government and private industry.

Another example of diffusion within the industry is the practice of modeling new research institutes, firms, and departments after existing prototypes. As mentioned above, RAND was offered as a prototype for the Urban Institute, though Urban's founders were quite clear on how the two differed (Hayes and Japha, 1978). Prior to that, the Stanford Research Institute (now known as SRI International), one of the first research and development institutes founded after World War II, was a model for RTI. SRI was explicitly mentioned in a 1959 memorandum from the vice president of Duke University announcing the establishment of RTI: "It will be similar to Stanford Research Institute in California, Southern Research Institute in Birmingham, Armour Research Institute Foundation in Chicago and several others" (from Larrabee, 1991).

Incidentally, SRI's eventual diversification of its research activities followed a similar course to RTI and RAND, starting with the six-year national evaluation of Follow Through for the OEO. While the majority of SRI's current revenue still comes from the

Department of Defense (data from [www.sri.com](http://www.sri.com)), it is also known for its contributions to education research and evaluation.

The ASPE office and the OEO's ORPPE were conceived as replicas of the offices within DOD that had parallel functions (Haveman, 1987). "McNamarism" was not just a new way of planning and assessing programs, then, but it affected the way new departments were envisioned and organized. The ORPPE attracted early champions of systems analysis, including economist Joseph Kershaw, who co-authored a RAND publication on systems analysis in 1959, became provost at Williams College in 1962, and then the first director of the ORPPE. Robert Levine, also of RAND, succeeded Kershaw at OEO. And Sargent Shriver, director of the OEO, reportedly wanted the ORPPE to mimic the Systems Analysis Office at McNamara's Department of Defense: "Systems analysis had the reputation at the time of being *the* solution to all planning and some administrative ills, and Shriver wanted some of that" (Levine, 1970, p. 59).

So far we have seen two examples of how existing cost-benefit management techniques and organizational forms were diffused throughout the emerging evaluation landscape. But the department of defense and its attendant contracting industry were not the only exemplars; the field also adopted the social experiment, at the time a leading-edge methodology practiced in several social science disciplines.

### **The Age of Social Experiments**

The use of random assignment and control groups dates back to the early 1900s, when psychologists like Edward Thorndike conducted educational experiments and sociologist F. Stuart Chapin studied issues like the effect of public housing and programs for delinquent boys (Oakley, 1998). The watershed moment linking this established social science tradition with the evaluation field was the publication of Campbell and Stanley's "Experimental and Quasi-experimental Designs for Research" (1963). Campbell and Stanley provided an explicit framework that linked social research to applied settings at a time when social policymakers were uncertain about how to effectively design interventions for social problems (Haveman, 1987). Campbell's vision of the "experimenting society" converged with this uncertain context and produced the "golden age" of social experiments (Campbell, 1969; Oakely, 1998; Rossi and Wright, 1984).

Though scattered social experiments were conducted prior to the 1960s, social policy evaluators in the late 1960s adopted the experimental design during a period of uncertainty surrounding how best to conduct an evaluation:

I think very clearly, obviously there was a point in time in which the experience base was not very deep, and so it had to evolve. And I think that firms like Abt and Mathematica and a few of the others that place a lot of emphasis on sophisticated applications of these analytical tools and data collection tools...helped to develop the field (CEO of large evaluation firm).

There were few existing research models to draw upon for evaluation designs, but a few of us were widely read and accessed to social scientists and learned about the applications of experimental designs in agricultural field experiments and medical drug clinical trials and testing (founder of large evaluation firm).

In this environment of intense political interest (both liberal and conservative) in the outcomes of the War on Poverty, key anti-poverty programs were repeatedly evaluated without being able to provide conclusive evidence of their effectiveness or ineffectiveness. The impetus for experimental evaluations of social policy came from the social sciences, specifically economics, and was adopted first by the OEO and later HUD and DOL. The first large-scale social experiment in the Great Society era usually is credited to the Negative Income Tax experiments (Greenberg, Shroder, and Onstott, 1999; Rossi and Wright, 1984). These experiments began as a dissertation funding proposal by Heather Ross, an MIT graduate student, who wanted to know whether or not guaranteed income payments to low-income families would result in a work disincentive (Greenberg et al. 1999). The OEO accepted the idea and turned it into four large experiments spanning six years.<sup>1</sup>

About two years prior to the first income maintenance experiment, the OEO had seen an immediate need for a pool of researchers that would focus exclusively on poverty and inform their efforts (Haveman, 1987). The Institute for Research on Poverty at the University of Wisconsin was established in 1966, and one of its first projects was the design and analysis of the first two negative income tax experiments (Haveman, 1987). Mathematica Policy Research, at the time a fledgling department within a mathematical consulting firm in Princeton, was chosen to collect and analyze data from the NIT experiments.

Mathematica, Inc., was founded in 1959 as a division of the Market Research Corporation of America by Princeton economist Oskar Morgenstern and several of his university colleagues, including Tibor Fabian, who joined in 1961. Mathematica, Inc. specialized in constructing mathematical models, cost-benefit analyses, and applying the use of computers to help solve economic problems. Its initial projects were wide-ranging, including several for the Department of Defense, development of the lottery system for the state of New Jersey, and cost-benefit analyses of a transportation corridor, the performing arts, and the space shuttle program.

The Institute for Research on Poverty had decided upon New Jersey for the site of the first Negative Income Tax experiment because the welfare laws in that state were conducive to a control-group design. That led to the need for an organization in New Jersey to manage the random assignment, data collection, and income payments. The subcontract was given to a newly formed division of Mathematica, called the Urban Opinion Surveys Group. In 1975 the division incorporated as Mathematica Policy Research, and became an independent company in 1986. MPR was involved in several

---

<sup>1</sup> The exception is the Denver-Seattle experiment, which lasted from 1970 until 1991.

seminal social experiments from the late 1960s through the 1970s, and is one of the “big three” evaluation firms that dominate the social experiment market today (Greenberg et al. 1999).

Another landmark social experiment, the National Supported Work Demonstration, was launched in 1974 by the newly formed Manpower Demonstration Research Corporation (MDRC). Supported Work was a demonstration project aimed at increasing the employability of former offenders, out-of-school youth, substance abusers, and long-time public assistance recipients. MDRC was formed in New York City for the express purpose of centralizing management of the experiment. A collaborative non-profit venture of the Ford Foundation and the U.S. Depts. of Labor (Manpower Administration), HEW, Justice, Housing and Urban Development, and the Special Action Office for Drug Abuse Prevention, some of MDRC’s founders envisioned that if successful at this first task, the organization could manage future demonstrations as well and ultimately build a body of evidence on the effectiveness of anti-poverty programs (Brecher, 1978). The Ford Foundation prompted the idea for the national demonstration as an extension of a New York City supported work program operated by the Vera Institute of Justice since 1969.

In 1973, Mitchell Sviridoff, Ford Foundation’s Vice President for National Affairs, sought advice about the potential project from Eli Ginzberg, a manpower expert at Columbia University’s Graduate School of Business. Ginzberg was supportive and agreed to provide research advice. By framing the project as a research and development effort, Sviridoff hoped to avoid the mistakes of Headstart, widely seen as having expanded prematurely (Brecher, 1978). With \$6 million for the first year of a five-year effort, MDRC contracted Mathematica Policy Research and the Institute for Research on Poverty to design, collect and analyze the evaluation data. The other bidders were an Urban Institute/Westat team and a RAND/NORC team (Brecher, 1978).

Shortly after launching the Supported Work project, MDRC began the National Tenant Management Demonstration Program at the request of the Ford Foundation and HUD. Brecher (1978) recounts that after some initial concern that it would interfere with the Supported Work demonstration, the MDRC board agreed to take on the second project with the expansion of staff and promises that it would be separated from MDRC if it interfered. This action marked the evolution of MDRC into a “general purpose research and development corporation, likely to prove long-lived and likely...to tackle other social issues...arising along the borders between the governmental and the private nonprofit sectors of the American economy” (p. 83).

The above examples, while illustrating the application of the social science model of randomized experiments to evaluation questions, also reveal how evaluation firms and institutes were both competitors and collaborators from very early on in the life of the industry. Collaboration among research and evaluation firms and institutes is routine. Other high-profile joint projects from the early years include the Westinghouse and Westat daycare study of 1970, and the RAND/ Mathematica health insurance experiment

of 1972. A preliminary scan of the websites of some of today's largest evaluation organizations reveal many more examples of recent collaborations: Westat subcontracted to Abt Associates for the design of a web survey for mental health outpatient program staff, the Urban Institute hired Abt Associates to implement a survey of public housing residents, Mathematica Policy Research subcontracted with the Urban Institute to help evaluate children's health insurance programs and with MDRC to assist with the evaluation of disability programs for youth, and Abt Associates and MDRC are leading the evaluation of the Reading First Program for the U.S. Department of Education.

The main reason for embarking on a joint project is when one firm alone does not have the capacity that it would take to win a particular contract. The need to partner with other firms is determined by the scope of the project, which is in turn controlled by the client (in this case, federal government). In this way, government clients further consolidate the evaluation industry by creating requests for proposals and scopes of work that require the collaboration of competing firms. Collaborative projects between competing firms also helps generate the context in which professional network ties are forged, cementing an interdependence within the industry. Collectively, these network ties provide the backdrop for the development of professional norms.

### **Professional Networks and the Search for Legitimacy**

The interdependence born out of the experience with collaboration on joint projects reinforces the growing peer networks in the field, which contribute to the overall professionalization of the industry. The peer networks no doubt influence hiring practices, as evidenced by the common strategy of firms hiring personnel from each other's ranks:

You've got a thriving private sector research community but with a limited number of firms. And if you need to develop additional senior people and more rapidly than you can grow them yourself, you've got to get them from somewhere...if you need particular talents your competitors are the most fertile place to get them (CEO of major evaluation firm).

From the firm's perspective, the motivation for hiring from a competitor is the simple need for expertise not found in-house. Many organizations within an industry collectively engaging in this practice results in filtering of personnel (DiMaggio and Powell, 1991). Filtering happens when firms look for similar attributes in hiring and promoting staff, and hire from within the same industry and handful of graduate programs in economics, public policy and sociology. The implications of this practice include senior staff who "tend to view problems in a similar fashion...and approach decisions in much the same way" (p. 72), and, at the field level, an industry that begins to coalesce as a field made up of organizations that are more similar than they are different.

A cursory review of recent announcements at Abt Associates provides several examples. At the senior level, Abt Associates has hired a vice president with ten years of experience

at RAND, a senior associate who whose career moves included RAND and COSMOS Corporation, a principal associate who left Abt in 1996 to work at the Urban Institute and was re-hired in 2005, a principal associate who had been at RTI for 29 years, and a senior vice president with 26 years of experience at RTI and the National Opinion Research Center (NORC). Similarly, the Urban Institute in 2002 hired a former Abt Associates vice president and RAND researcher to direct its Justice Policy Center.

It helps for firms to be organizationally and technically similar for these personnel transitions to work. If unique firms tend to converge as a result of pursuing similar projects from a handful of federal funders, this also facilitates the flow of personnel across the industry. The tendency for the larger firms to resemble each other structurally (e.g., similar career paths and job titles) reinforces their ability to network and collaborate on joint projects.

### **Professional Associations and the Maturation of a Field**

The professionalization of evaluation has been a perennial topic for evaluators (Bickman, 1997; Connor and Dickman, 1979; Flaherty and Morell, 1978; Freeman and Solomon, 1979; Morell and Flaherty, 1978; Morell, 1990; Smith, 2001). Peer networks support the development of professional associations and subsequent development of standards, norms, and ethical codes. The establishment of professional associations is a key marker in the professionalization process (CITE). In the evaluation field, three associations began about ten years after the evaluation “boom”: the Evaluation Network (E-Net), the Council for Applied Social Research, and the Evaluation Research Society (ERS).

E-Net was founded in 1974 “to bring together individuals dedicated to improving theories, practices, programs, and education in evaluation” (Evaluation News, February 1981). Of the two or three associations at the time, E-Net catered more to evaluators who worked on local school and health systems evaluation (King, Mark, and Miller, 2004). Clark Abt organized the Council for Applied Social Research in the early 1970s, with the intent to connect top quantitative social scientists and the most competitive research organizations such as Mathematica Policy Research, RAND, RTI, Westat, SRI, and NORC with officials from the Office of Management and Budget, Office of Education, DOL, HUD, and HEW. These evaluators were involved in large-scale evaluations of federal programs, but most federal programs were not being evaluated at the time. The hope was to take stock of what had been learned from evaluation studies thus far, and promote more widespread use of sophisticated quantitative evaluation techniques (Abt, 1976).

According to an oral history interview with Lois-Ellen Datta, Marcia Guttentag organized the ERS in 1976 as an alternative to the Council for Applied Research with more attention paid to developing a diverse membership (King et al., 2004). In 1981, a merger between the ERS and the Council for Applied Social Research was approved (Evaluation News, February 1981). ERS and E-Net merged in 1986 under the name American Evaluation Association (AEA).

The consolidation of the professional associations occurred during a time of slow growth in the federal evaluation industry, as the Reagan administration cut back many social programs and attendant evaluation efforts. So, as the federal funding forces that jump-started the industry weakened, the normative and professional forces continued to develop as the field sought its identity. But despite the convergence of professional associations and the organizational similarities within the federal evaluation industry, the evaluation field remains highly varied with regard to what evaluation is, what training is required to practice as an evaluator, and what methods and techniques constitute competent practice (Flaherty and Morell, 1978; Smith, 1999; Worthen, 1999). Indeed, in 1996, facing declining membership, the board of the AEA agreed that the association was “not a strong and unified organization” (Bickman, 1997). That same year, the association began to explore the possibility of certifying evaluators as one way to strengthen the field.

**Figure 5.**



Source: AEA; AEA News, 1986; Summary of AEA Membership, *Evaluation Practice*, 1988 vol. 9, 84-86; Bickman, 1997.

Note: AEA did not systematically collect membership data until 2001; data missing for 1997-2000.

The subsequent debate over the certification of evaluators illustrates how the evaluation field struggles with professionalization. One characteristic of a fully developed profession is that its members exercise strict control over who is allowed to practice and the competencies required to do so (Morell and Flaherty, 1978; Wilensky, 1961). Thus, attempting to credential evaluators is an attempt to control and shape the field to a particular set of norms and expectations. While certification programs have surfaced,

such as the Evaluator's Institute and the Certificate of Advanced Study in Evaluation at Claremont Graduate School, formal credentialing has not transpired. The main reason is best summarized by Worthen (1999):

Evaluation is at present so splintered, rooted as it is in so many disciplines, with today's evaluators trained in so many diverse specializations and through such diverse means, that it seems rather optimistic to presume that any agreement can be forged within AEA about what constitutes essential evaluation competencies. Indeed, we evaluators no longer even agree on what evaluation is.

The development of professional standards and guidelines is another way an emerging field tries to consolidate its base and create an identity among its practitioners (Wilensky, 1964). The AEA's Guiding Principles for Evaluators, first published in 1994, is a code of ethics; the statements are intentionally general to encompass the wide range of methods and epistemologies living under the umbrella of evaluation. Indeed, its function is more to socialize evaluators around the identity of being an evaluator, rather than provide regulations and standards for how to do evaluation (Bustelo, 2006). The preface to the guiding principles illustrates this intent:

Based on differences in training, experience, and work settings, the profession of evaluation encompasses diverse perceptions about the primary purpose of evaluation....Despite that diversity, the common ground is that evaluators aspire to construct and provide the best possible information that might bear on the value of whatever is being evaluated. The principles are intended to foster that primary aim (AEA Guiding Principles, 2004).

Nonetheless, critics such as Rossi (1995) observed that the vagueness of the guidelines, while indicating a diversity of perspectives, weakens the field. Indeed, if the base of knowledge for a field is too general or vague, the achievement of professional status is less likely (Wilensky, 1964). This is distinct from the purpose of one of the earlier professional associations, CASR, which expressly promoted the use of experimental designs for evaluation.

A major characteristic of the evaluation field is thus its heterogeneity, approximating what Morell (1990) called a "loose coalition" of evaluation. However, conflicts over such areas as mission, priorities, and methodology often characterize the development of a profession (Bucher and Strauss, 1961), and are a sign of healthy growth (Connor and Dickman, 1979). There are several contested areas within the field that vary over time (Smith, 2001), but one issue, concerning whether or not randomized controlled trials should be the "gold standard" for evaluation, persists. The realization that these designs did not provide timely information to decision-makers, apply realistically to smaller, localized programs or programs with shifting strategies, and the uncertainty of the results coupled with high cost, led to the resurgence of alternative evaluation methodologies (House, 1990; Maynard, 2000; Rossi and Wright, 1984). While experimental designs were still used for selected federally-sponsored evaluations, the focus for many

evaluators turned toward how to increase use of evaluation findings, which in turn led to the popularity of participatory and collaborative evaluation practices (Smith, 2001).

Consequences of this debate include Peter Rossi resigning his membership when AEA took a position against the U.S. Department of Education's preference for randomized experiments in educational evaluations in 2003 (Lipsey, 2007), and perhaps the migration to associations like the American Economics Association, American Public Health Association, and the Association for Public Policy Analysis and Management (APPAM) which more readily accept experimental designs as the gold standard. The AEA, on the other hand, has purposefully decided not to prioritize particular methodologies in recognition of the diversity of perspectives and purposes of evaluation—and its membership has increased steadily from 3,000 to 5,000 in the last five years. Intra-field differences were seen as a necessary sign of progress in the professionalization of evaluation by its practitioners in the late 1970s, and such debates continue today.

The number of professional associations in which evaluators take part exemplifies the breadth and diversity of the field: in addition to the organizations mentioned above, one could add the American Psychological Association and the American Educational Research Association as well. The fact that the primary disciplines of evaluators include psychology, public administration, sociology, economics, public health, and many others (Morell, 1990), makes evaluation inherently interdisciplinary and thus loosely consolidated. Indeed, evaluation has been called a “transdiscipline” (Scriven, 1993) because it provides a set of tools and methods for use by the primary disciplines—in this way it takes root across disciplinary boundaries (e.g., public health, education, criminal justice, welfare, personnel and workforce). Others have classified evaluation as a “meta-discipline,” one that encompasses most social science research (Picciotto, 1999), while still others lament the lack of professional identity and failure to develop core practices and methods (Sechrest, 1994).

In sum, the evaluation industry was distinguished early on by its interdependence, which set the stage for networking and professionalization. This in turn gave rise to a diversity of perspectives in the broader field of evaluation, as discontentment with large-scale experimentally designed evaluations grew. Professional associations that began in the mid 1970s representing divergent approaches to evaluation were eventually consolidated until the profession had a single organization, the AEA, by 1986. Nonetheless, the way the field evolved with its origins in a diverse range of “home” disciplines, the variety of avenues available to become an evaluator, and the contested areas within the field signifies its status as a “loose coalition” (Morell, 1990).

## **Conclusion**

The idea of program evaluation can be traced back very far into the history of American public life. Here, we have focused on its emergence and expansion over the past half century. During this time, the field has witnessed the growth of a sophisticated industry and the emergence of a profession aimed at discerning effectiveness. While the field of

program evaluation took major strides in the 1960s and 1970s to define its agenda and ambitions, it took decades for practices and systems to be built to support the goal of measuring the impact of social programs. The explosive growth of evaluation firm revenues, from the 1970s through the 1990s is evidence that not only has the demand for evidence of program impact been great, the supply of evaluation services has risen to meet this demand.

While we have focused here on the early history of government's move toward measurement, it is important to note that the quest for good evaluation research now extends deeply into the nonprofit sector. Long funded by government and exposed to its mandates and regulations, nonprofit service providers have also contributed to the growth of the evaluation industry by contracting out to research firms to meet their assessment needs. In this sense the move to evaluate social programs may have started in government, but today it has been fully absorbed into the practices and priorities of the nonprofit sector, whose role in social service delivery has only expanded through the trend toward contracting out and the resulting hollowing out of the state (Milward and Provan).

Many obstacles still remain before the field. There remains a significant difference of opinion on the validity of program evaluation in cases where experimental designs are not possible. Perhaps the most prominent challenges lies in discovering and communicating the limits of evaluation to its consumers. While the work of the major evaluation firms are typically communicated with certitude and scientific heft, the field of program evaluation remains more of an art than a science.

The rise of an entire multi-billion dollar industry in evaluation research is testament to the power of the question, "Did the program work?" This simple question lies at the origin of many of the largest and most complex publicly funded evaluations. As government has sought to become more effective and to maximize the impact of its spending programs, evaluation has been a critical means to that end. Not only does it lead to information about operational effectiveness, but it also can have powerful uses for both securing and terminating future funding. Critical to evaluation data's wise use are an industry and a profession that are both technically sophisticated and ethical. When it comes to evaluation research, the last half century's experience demonstrates, that knowledge is indeed power.

## References

- AIR News. 2007. "AIR celebrates six decades of success."
- Bassett, G. 1969. *The Urban Institute: A History of its Organization*. Washington, D.C.: The Urban Institute.
- Brecher, E. M. 1978. *MDRC: Origin and Early Operations*: Ford Foundation.
- Bryant, E. C. 1981. *Twenty Years and Counting: A Personal History of Westat*: Westat.
- Bucher, R., and Strauss, A. 1961. "Professions in process." *American Journal of Sociology*, 66 (4), pp. 325-334.
- Bustelo, M. 2006. "The potential role of standards and guidelines in the development of an evaluation culture in Spain." *Evaluation*, 12 (4), pp. 437-453.
- Campbell, D. T. 1969. "Reforms as experiments." *American Psychologist*, 24, pp. 409-429.
- Campbell, D. T., and Stanley, J. C. 1963. *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.
- Conner, R.F., and Dickman, F.B. 1979. "Professionalization of evaluative research: conflict as a sign of health." *Evaluation and Program Planning*, 2, pp. 103-109.
- DiMaggio, P. J., and Powell, W. W. 1991. The iron cage revisited: institutional isomorphism and collective rationality. In W. W. Powell and P. J. DiMaggio (eds.), *The New Institutionalism in Organizational Analysis*. Chicago: University of Chicago Press.
- Flaherty, E. W., and Morell, J. A. 1978. Evaluation: manifestations of a new field. *Evaluation and Program Planning*, 1(1), pp. 1-10.
- Freeman, H.E. and Solomon, M.A. 1979. "The next decade in evaluation research." *Evaluation and Program Planning*, 2, pp. 255-262.
- Greenberg, D. H., Shroder, M., and Onstott, M. 1999. The social experiment market. *The Journal of Economic Perspectives*, 13(3), pp. 157-172.
- Gorham, W. 1967. Notes of a practitioner. *The Public Interest*, 8 (Summer), pp. 4-8.
- Haveman. 1987. *Poverty Policy and Poverty Research: The Great Society and the Social Sciences*. Madison: University of Wisconsin Press.
- Hayes, F., and Japha, A. 1978. *The Urban Institute, 1968-1978: An Evaluation of its Performance, Prospects and Financial Problems*: The Ford Foundation.

- Held, V. 1966. "PPBS comes to Washington." *Public Interest*, 4 (Summer), pp. 102-115.
- House, E.R. 1990. "Trends in evaluation." *Educational Researcher*, 1(3), pp. 24-28.
- Jardini, D. R. 1996. *Out of the Blue Yonder: The RAND Corporation's Diversification into Social Welfare Research, 1946-1968*. Carnegie Mellon University.
- King, J., Mark, M., and Miller, R. 2004. "The oral history of evaluation, part 2: An interview with Lois-ellin Datta." *American Journal of Evaluation*, 25(2), pp. 243-253.
- Larrabee, C. X. 1991. *Many Missions: Research Triangle Institute's First 31 Years*. Research Triangle Park, North Carolina: Research Triangle Institute.
- Levine, R. A. 1970. *The Poor Ye Need Not Have With You: Lessons from the War on Poverty*. Cambridge: M.I.T. Press.
- Lipsey, M. W. 2007. "Peter H. Rossi: formative for program evaluation." *American Journal of Evaluation*, 28(2), pp. 199-202.
- Maynard, R.A. 2000. "Whether a sociologist, economist, psychologist or simply a skilled evaluator." *Evaluation*, 6(4), pp. 471-480.
- McLaughlin, M. W. 1975. *Evaluation and Reform: The Elementary and Secondary Education Act of 1965/Title I*. Cambridge, MA: Ballinger.
- Morell, J. A. 1990. "Evaluation: status of a loose coalition." *Evaluation Practice*, 11(3), pp. 213-219.
- Morell, J. A., and Flaherty, E. W. 1978. "The development of evaluation as a profession: Current status and some predictions." *Evaluation and Program Planning*, 1(1), pp. 11-17.
- Oakley, A. 1998. "Experimentation and social interventions: a forgotten but important history." *BMJ (British Medical Journal)*, 317, pp. 1239-1242.
- O'Connor, A. 2001. *Poverty Knowledge: Social Science, Social Policy, and the Poor in Twentieth-Century History*. Princeton: Princeton University Press.
- Picciotto, R. 1999. "Towards an economics of evaluation." *Evaluation*, 5(1), pp. 7-22.
- Rein, M., and White, S. H. 1977. Can policy research help policy? *Public Interest*, 49, pp. 119-136.
- Rossi, M. and Wright, James D. 1984. "Evaluation research: an assessment." *Annual Review of Sociology*, 10, pp. 331-352.

DO NOT CITE WITHOUT PERMISSION OF AUTHORS

- Rossi, P. H. 1995. "Doing good and getting it right." *New Directions for Program Evaluation*, pp. 55-60.
- Scriven, M. 1993. "Hard-won lessons in program evaluation." *New Directions for Program Evaluation*, 58.
- Schick, Allen. 1971. "From analysis to evaluation." *Annals of the American Academy of Political and Social Science*, 394, pp. 57-71.
- Sechrest, L. 1994. "Program evaluation: oh what it seemed to be!" *Evaluation Practice*, 15(3), pp. 359-365.
- Shadish, W. R., Cook, T. D., and Leviton, L. C. 1991. *Foundations of Program Evaluation*. Newbury Park: Sage.
- Smith, J.A. 1991. *The Idea Brokers: Think Tanks and the Rise of the New Policy Elite*. New York: The Free Press.
- Smith, M. F. 1999. "Should AEA begin a process for restricting membership in the profession of evaluation?" *American Journal of Evaluation*, 20(3), pp. 521-532.
- Smith, M.F. 2001. Evaluation: preview of the future #2. *American Journal of Evaluation*, 22(3), pp. 281-300.
- Sperry, R. L., Desmond, T. D., McGraw, K. F., and Schmidt, B. 1981. *GAO 1966-1981: An Administrative History*. Washington, D.C.: U.S. General Accounting Office.
- Staats, E.B. 1968. Perspective on planning-programming-budgeting. *The GAO Review* (Summer), pp. 3-2.
- Staats, E. B. 1970. "The relationship of budgeting, program planning, and evaluation." *The GAO Review* (Winter), pp. 3-10.
- Staats, E. B. 1973. *Challenges and problems in the evaluation of governmental programs*. Unpublished manuscript, Pittsburgh, PA.
- University Times. May 9, 1996. John C. Flanagan Obituary.
- Weiss, C. H. 1972. *Evaluation Research: Methods of Assessing Program Effectiveness*. Englewood Cliffs, NJ: Prentice-Hall.
- Wholey, J. S., Scanlon, J.W., Duffy, H.G., Fukumoto, J.S. and Vogt, L.M. 1970. *Federal Evaluation Policy: Analyzing the Effects of Public Programs*. Washington, D.C.: Urban Institute.
- Worthen, B. R. 1999. "Critical challenges confronting the certification of evaluators." *American Journal of Evaluation*, 20(3), pp. 533-556.