National Science Foundation
Math and Science Partnership Program Evaluation (MSP-PE)

Contextualizing Expectations of Success for Federal Programs that Support K-12 Education

Irwin Feller, Ph.D.
American Association for the Advancement of Science
Professor Emeritus, Economics, Pennsylvania State University

December 2007

TABLE OF CONTENTS

PREFACE

This paper is one in a series of briefs for the Math and Science Partnership Program Evaluation (MSP-PE), conducted for the National Science Foundation's Math and Science Partnership Program (NSF-MSP).

The paper's topic originates from discussions held at the MSP-PE's advisory board meeting in October 2006. Advisory board members suggested that "expectations" about the nature and level of outcomes from programs like the NSF MSP Program often go unaddressed. This suggestion led to the commissioning of the present paper. Although the paper does not set specific expectations, it offers a useful framework for others to do so. In this sense, the paper serves as part of the MSP-PE's evaluation design work.

The MSP-PE is led by COSMOS Corporation in partnership with George Mason University (GMU) and Brown University. Robert K. Yin (COSMOS) serves as Principal Investigator (PI), and Jennifer Scherer (COSMOS) serves as one of three Co-Principal Investigators. Additional Co-Principal Investigators and their collaborating institutions (including discipline departments and math centers) are Patricia Moyer-Packenham (GMU) and Kenneth Wong (Brown).

The MSP-PE is conducted under Contract No. EHR-0456995. Since 2007, Bernice Anderson, Ph.D., Senior Advisor for Evaluation, Directorate for Education and Human Resources, has served as the NSF Program Officer. The author is Irwin Feller, Ph.D., of the American Association for the Advancement of Science (AAAS) and Professor Emeritus, Economics, Pennsylvania State University.

ABOUT THE AUTHOR

Irwin Feller is a senior visiting scientist at the American Association for the Advancement of Science (AAAS) and Professor Emeritus of Economics at the Pennsylvania State University, where he was on the faculty between 1963 and 2002. His research interests include science and technology policy, economics of higher education, and program evaluation. He is the author of over 100 referenced journal articles, final research reports, and book chapters, as well as of numerous papers presented to academic, professional, and policy audiences.

His article, "Performance Measurement Redux," American Journal of Evaluation, 23 (2002): 435–452, received the American Society for Public Administration's Joseph S. Wholey Distinguished Scholarship Award, Best Scholarly Article on Performance-based Governance in 2002. His co-authored study, "A Toolkit for Evaluating Public R&D Investment: Models, Methods, and Findings from ATP's First Decade" (with Connie Chang and Rosalie Ruegg) received the American Evaluation Association's 2004 Outstanding Publication Award.

For NSF, he has served as a member and chair of the Advisory Committee to the Assistant Director, Social, Behavioral and Economic Sciences and as a member of the Government Performance and Results Act Review Panel. He also has been a member of several Committee of Visitors panels, numerous proposal selection panels, and numerous NSF-funded project-related advisory panels. At the request of the Director's Office, he organized an NSF-sponsored workshop for the Office of Management and Budget on the design of performance criteria for research and development programs. His published research includes evaluations of NSF's Engineering Research Centers, the EPSCoR program, and NSF's cost-sharing policies.

I. INTRODUCTION

A. Problem Statement

This essay (1) explores the formation of expectations about the projected impacts of federal education programs, and (2) advances a framework for contextualizing these expectations within the policy, scientific, jurisdictional, and diffusion milieus within which federal agencies and programs operate.

Expectations are viewed as a core, albeit frequently subsumed, component of program evaluations directed at assessing the worth of a program: what is expected constitutes a baseline for forming assessments about whether reported results constitute program success, or not. Relatedly, *expectations* also serve as a baseline for determining program accountability: "To hold a public agency accountable for performance, we have to establish expectations for the outcomes that the agency will achieve, the consequences that it will create, or the impact that it will have" (Behn, 2001, p. 8).

An improved analytical understanding of how expectations are formed has immediate utility in providing for a more realistic meshing of performance goals and performance results. But the analysis also has near-term instrumental value: by identifying the contextual factors that condition a program's upper potential, it provides information that program managers can use to design their programs in ways that enable them to reach this potential.

The essay has been prepared as part of the ongoing research design work by the Math and Science Partnership Program Evaluation (MSP-PE). In keeping with a core precept of NSF's approach to the evaluation of its educational programs—that questions

should drive the choice of methodology (Katzenmeyer and Lawrenz, 2006)—as well as to connect its analysis to contemporary events relevant to NSF's consideration of the MSP Program, the essay cites numerous examples from the field of federal policies towards K-12 education and from debates surrounding evaluation methodology in this field.

Particular analytical attention is directed at the "compared-to-what" question frequently encountered in gauging federal agency/program performance, especially to the use of cross-agency or cross-program comparisons. The influence of comparisons is explicit in the Office of Management and Budget (OMB)'s grouping of Performance Assessment Rating Tool (PART) ratings for functionally similar programs across agencies, such as those of the Department of Education's programs with those of the Department of Health and Human Services and the Department of Energy (www.whitehouse.gov/omb/expectmore/topic/Education.html). Comparisons, of course, form the basis of benchmarking, a technique used in the formation of objectives contained with the strategic plans of public sector, private sector, and not-for-profit organizations. Comparisons also permeate the use of analogies, where expectations are expressed in terms of statements about what has happened earlier or elsewhere.

The validity of cross-agency/cross-program comparisons in shaping expectations requires examination, however. Rather than treating them as a starting point or fixed point of reference, the framework presented here starts from the proposition that the environments within which federal programs function are different, even for those with closely aligned objectives. These differences in turn affect the likelihood that programs achieve their stated objectives.

In treating quotidian perspectives and assumptions about meaning and measurement of expectations and success, the essay eclectically distills and integrates theoretical and empirical findings from several research traditions and literatures. Its primary analytical groundings though, are the connections (variables and relationships)—tight in some places, loose in others, absent in yet others—most typically found within the program evaluation and program implementation literatures. These connections are presented as generating more or less hospitable contexts, or environments, for the diffusion, implementation, and impacts, and thus ultimate success, of federal agency R&D programs.

The essay takes frequent note of current methodological and policy debates about the desideratum of evidence-based decision making and the apotheosis of randomized trials in evaluating federal programs, including those directed towards generating scientifically valid education, as called for in the 2002 Education Science Reform Act (Cook, 2002; Lawrenz and Huffman, 2006; Murnane and Nelson, 2007). It does not enter deeply into any of these debates, however, because they are subsidiary to the essay's primary focus on expectations, an activity that generally both precedes and, as contended below, is treated as exogenous to the selection of evaluative techniques.

More specifically, in terms of the Mark, Henry, and Julnes (2000) typology of the four purposes of evaluation, federal program evaluations directed at assessing merit and worth, or at what Chelimsky (2007) has termed cause-and effect questions, are typically designed to determine whether or not a program has produced change in the hypothesized or desired direction and whether the change can be causally linked to its activities after controlling for other probable causes. Among the questions customarily contained in

"impact evaluations," such as for anti-crime programs are: (1) the need for the program; (2) the program conceptualization or design; (3) the program implementation and service delivery; and (4) the program cost and efficiency (National Research Council, 2005). Answers to these questions can indeed influence perceptions about social problems and the selection, and at times discontinuation, of specific social policies (Henry, 2003; Henry and Mark, 2003).

More generally though, questions about whether the magnitude of the change meets/satisfies the expectations/requirements of the program's sponsors are a subject (or question) generally taken as being outside the scope of the evaluation. Definitions, criteria, and measures for assessing whether or not success has occurred are typically treated as exogenous, set by the legislative, judicial, policy, or political environment (Howell and Yemane, 2006).

Interest here is different. The essay is less about what constitutes the "gold standard" of federal government evaluation design and more about whether enough gold from expert evaluation mining is found to meet the expectations of investors. Rather than an exploration of the form of the evaluation design or quality of evidence needed for an intervention to be classified as a model program or the frequency with which lists of model programs are updated (Hallifors, Pankratz, and Hartman, 2006), the question of interest here is whether it is appropriate to use findings on effect size from studies certified by a federal agency, as for example the Substance Abuse and Mental Health Services Administration's National Registry of Evidence-based Programs and Practices (Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, 1999), to set expectations about program outcomes in other

functional areas, say education. Relatedly, debates about which evaluation methodologies are needed to generate reliable, valid, "error-free," and relevant findings in a policy world of evidence-based decision making are cited here selectively, not to referee contending positions than for their utility, as a bridge into discussions about differences between the statistical significance and size of coefficients (McCloskey and Ziliak, 1996; Ziliak and McCloskey, 2004; Boruch, 2007).

Converting the concepts of expectations and success from parameters into variables opens up several lines of inquiry about how to interpret findings from even the most carefully conceived and implemented program evaluations. These questions pervade considerations of evaluation, performance assessment, and accountability, yet are seldom treated explicitly in their own right. Without answers to them, however, links in the analytical, empirical, and policy chains connecting program theory, inputs, outputs, outcomes, and impacts are missing. More saliently, from a public policy/public management perspective that emphasizes deploying evidence-based knowledge to shape the adoption and implementation of best-practice approaches in the provision of public sector services, lack of understanding or clarity about how expectations for success are determined is akin to using a thermometer to assess the health of an individual without knowing that normal body temperature is 98.6 degrees Fahrenheit.

Central to this inquiry is the standing of MSP as both a research and development (R&D) program and an intervention program. MSP is designed to both generate new knowledge and modes of improved practice and promote the adoption and diffusion of this new knowledge into schools. As stated by NSF's 2006 Government Performance and Results Act (GPRA) Advisory Committee, "Simply introducing new pedagogical tools

into the curriculum will not accomplish the goal of improving mathematics and science education unless these tools are adopted by schools and significant progress is detected. Efforts directed at formal assessment of curricular reform initiatives are particularly relevant to NSF's goal to support innovation in research on teaching" (NSF, 2006, p. 27).

It is at the juncture where the three paths of evaluation methodology, adoption and implementation of innovative practices by K-12 schools, and performance management precepts of demonstrated results meet, that issues relating to expectations and what is meant by accomplishment of program goals meet. At this juncture, the performance assessment question changes from how to best generate scientifically valid evidence about the effects of a treatment, to pragmatic questions about how to get a pool of potential adopters to incorporate the findings into their ongoing operations so that it produces the evidence-based beneficial outcomes. In effect, attention shifts, and thus credit or blame for observed outcomes, from the explanatory power of the theory underlying the new program and the quality of the evidence adduced on its behalf to the efficacy of the implementation strategy.

B. Perspective and Thesis

Shaping the essay's tone is longstanding participation in evaluations of federal programs across many domestic agencies and reviews of research findings, including several meta-analyses, of the effects of federally-funded program interventions. Emerging from this prior work and ongoing research (Feller, 2007) is the essay's central thesis: Analytical consideration of how expectations concerning program outcomes are shaped, and thus the establishment of the initial conditions for determining whether reported program accomplishments constitute "success," constitutes a generic gap in the

treatment of federal education (as well as other domestic programs) under the Government Performance Results Act (GPRA), the Office of Science and Technology Policy-Office of Management and Budget (OMB) R&D Investment Criteria, and OMB's PART process.

The essay's observations about complexity, ambiguity, incompleteness, or inconsistencies are not to be interpreted as a brief for not holding agencies/actors accountable for their activities. Nor do the observations constitute an apologia or rationalization for programs that fail to achieve the objectives for which they were established. Stated more affirmatively, the essay is framed by a perspective that starts from Radin's (2006) statement that federal domestic R&D programs "operate in multiple ways in a world beset by complexity and ambiguity about numbers and data" (p. 29), to which it next adds the refrain increasingly heard in interpreting findings from program evaluations that "context matters," and ends with a conclusion that concepts and measures of success need to derive from the workings of public sector agencies within their specific functional domains.

C. Organization of Essay

Section II first explores the multiple meanings of expectations and success and then tracks through the import of different connotations of these concepts for purposes of program evaluation and performance management. Section III addresses the concept of success for federal R&D programs. Section IV identifies factors held to determine the effects/success of compound federal government R&D/demonstrations programs. Section V sets out the essay's conclusions, including, not surprisingly, suggestions about needed future research.

II. THE IMPORTANCE AND MEANINGS OF EXPECTATIONS AND SUCCESS IN A WORLD OF PERFORMANCE MANAGEMENT

A. Expectations

In a policy and political world dominated by performance-based management systems and requirements for systematic and rigorous, typically quantitative, evidence of demonstrated performance (Kettl, 1997; OECD, 2005), expectations about what constitutes satisfactory performance are a central component of assessing the success of a program, with these assessments in turn affecting decisions about its continuation, expansion, wider promotion, or adoption. Consequently, how success is defined and measured becomes an important policy variable, and likewise a subject warranting critical examination.

For established programs with longstanding, well defined objectives, agreed upon performance criteria, and accepted modes of operation, standards of satisfactory performance may already exist. New requirements for accountability, performance, and evidence for such programs may largely mean that they are expected to attain known or readily identifiable upper limits. OMB's use of the term, "expect more" (www.expectmore.gov) for its PART website carries this interpretation. It suggests that the public has a right to expect more from the performance of federal agencies, implying that these agencies and the programs they administer are not being all that they can, or should, be.[1]

To expect something, an outcome, however can connote something different than expectations about outcomes. To expect something is to regard it as likely to happen or to anticipate its occurrence. An alternative phrasing is to express outcomes in terms of a

high probability, given a known probability distribution. Thus pediatrician's offices frequently contain height-weight charts depicting the "normal," expected development of young children.

In other usage, "expect" may serve as a synonym for "requirement." Legislation that sets forth attainment of minimum standards for student achievement by 2014, or for clean air, clean water, or automobile gas mileage benchmarks, constitute statements of expected performance, with penalties imposed for failure to perform satisfactorily. Similarly, standard business statements of the type, "we expect our investments to yield 'x' percent return," in effect mean that the minimum expected mean value for risk-neutral actors of the distribution of projected returns must be "x" percent in order to induce investment. Subsequent summative judgments about whether or not the investment was successful would be set against these initial requirements. An investment project that yielded 6 percent might be considered to be a failure if the initial minimum requirement (which presumably represents the return from alternative, equally risky investments) was 12 percent.[2]

"Expectations" however has a broader, somewhat different connotation. The words also connotes "the degree of probability of the occurrence of something." Defined thusly, it focuses attention first on the presence of uncertainty surrounding the likelihood of outcomes, and second on the full distribution of probable outcomes, not simply the most likely or average outcome.

This connotation is the appropriate one to use in the context of education programs, basic and/or applied, because uncertainty is their defining essence. The uncertainties relate to whether the idea/approach/concept/thing being explored will work,

work well, work in different settings, or work better in whole or part than existing approaches or other innovations with which it competes in one form or another. Prudent program management seeks to reduce the degree of uncertainty associated with the conversion of fundamentally new, but untested concepts into practical solutions by assessing progress at each of several stages in terms of stage-relevant performance variables (Cohen, Kamienski, and Espino, 1998). Screening approaches are standard parts of the management strategies of federal agencies, as illustrated by the Institute of Education Science's sequential funding of research proposals directed at ideas, development, efficacy, and ramp up, as well as by the Department of Defense's categorization and progression from 6.1 ("Basic Research") to 6.7 ("Operational Systems Development") in its funding of R&D.

For some R&D stages, evidence as to whether to proceed may be readily apparent: schools fail to make adequate yearly progress (AYP); a test rocket crashes; agricultural yields per acre don't improve; an FDA Phase I trial produces negative or statistically insignificant results. But evidence that something is working at a given stage may not be sufficient to induce moving on to the next stage if the new level of performance gained fails to meet stated requirements, e.g., levels of technical speed, strength, durability, etc, or expectations, e.g., projected sales, profit, relative advantage. Alternatively, based on results at initial stages, expectations about next stage results may be adjusted to reflect what now seems the most likely or best-case future outcome (Rosenberg, 1976).

"Expectations" are not givens, however. They arise out of specific historical contexts and contemporary environments. They also may be purposefully managed, as in

beating market expectations about earnings, projected performance in election primaries, and the by now accustomed pre-debate "spin" on who will win presidential debates.

How indeed are expectations for the impacts of federal education programs set? Who sets them? Are program objectives set by systematic distillation of the state-of-knowledge, as typified by use of National Academies-National Research Council panels to define the state of the field in selected scientific areas and the potential for future advances (National Academy of Sciences, 1983); by political actors; by school boards; by superintendents; by others; or in varying combinations of these groups? Are performance goals within each period set so that they are readily achievable in order to readily justify continuation and expansion of the program in subsequent periods? Are program objectives static or do they evolve? If they evolve, what direction do they take over time? Is it to achieve last period's result or last period's result plus some increment; or do they, as Wildavsky (1979) has described, undergo a process of strategic retreat to become what can be achieved rather than what was initially projected or promised?

Is the relationship between those who set program objectives and those who manage and operate programs a command and control relationship such that policymakers can require agents to implement a procedure deemed necessary to attain mandated, higher levels of performance? If a highly directive approach is used, does it contain incentives in which supplemental resources are made available to those agents that follow the recommended course of action but not to others?

Alternatively, is the relationship characterized by indicative planning, such that the principal highlights ("benchmarks") constitute desired or best practice but then defer to the professional standards of the agents to adopt these recommended models? If

derived from benchmarking procedures, where do the benchmarks come from? Are they based on meta-analyses of existing findings or proof-of-concept from an experimental trial followed by directed implementation? Or are they just there?[3]

COSMOS's interim report on the MSP Program, for example, describes use by several grantees of benchmarking as an assessment technique (Yin, forthcoming). The evaluation notes, however, that the MSPs that have chosen to define pre-established benchmarks for later comparison to actual performance usually have not discussed any rationale for selecting their particular numeric benchmark. For instance, the MSPs do not discuss whether such benchmarks as "improving performance by 5 percent each year" might be too conservative or overly ambitious (ibid, p. 24).

How realistic are expectations about the level of performance gain that will be attained if all goes well? Do systematic biases towards overstatement or understatement exist? The period of social experimentation of the late 1960s and 1970s serves as a useful historical example here of the importance and shaping of expectations for public policies connected to federal education and other domestic programs. The period was rife with analogies that the U.S.'s success in landing a man on the moon within a pre-specified time period pointed to the promising prospects for using "hard" and "soft" scientific and technological approaches to achieve comparable successes in K-12 education, urban renewal, elimination of poverty, transportation, housing, etc. As reported in Glickman et al. (1980), high expectations existed for new federal undertakings in federally-subsidized housing, community development block grants, neighborhood revitalization initiatives, and enterprise zones. Instead of success, though, the disjuncture between what was promised or expected and realized accomplishments—the challenge of implementing

innovative ideas and programs emanating from Washington, DC to state and local governments (Pressman and Wildavsky, 1973; MITRE, 1979)—gave rise to the moon-ghetto metaphor, a plangent expression of the federal government's limited capacity to redress, much less solve, complex societal problems, especially when the "impacts of policies depend(ed) in good part on the performance or reaction of people not under the direct control of the policymaker" (Nelson, 1977, p. 34).

The moon-ghetto metaphor has special analytical relevance to the issue of the setting of expectations across federal functional domains. At the metaphor's core is an emphasis on differences in the legal, institutional, knowledge, and resource bases across functional domains, or what is variously termed here contexts, environments, or milieus. To paraphrase Nelson, policy domains vary in the extent to which the "steering wheel" of policy direction is connected to the "rudder" that affects the direction in which the ship moves.

To this point, emphasis has been placed on unrealistic or exaggerated expectations about success as necessary parts of building political coalitions or providing attractive slogans. The possibility also exists that expectations are understated. As suggested above, formal, quantitative statements of expected outputs and outcomes may carry with them the threat of penalties for not reaching stated goals. In such a setting, understatement of what can be accomplished may represent strategic behaviors among players in a multi-period game. In this game, one, both, or several players may have an incentive to understate what can be accomplished in a single period in order to (1) dampen performance requirements in subsequent periods, and (2) produce "surprise" when performance exceeds goals.

But beyond game-theoretic principal-agent views of the world, understated expectations may at times present a failure of imagination or aspiration. Thoreau for example observed that, “In the long run, men hit only what they aim at. Therefore, though they should fail immediately, they had better aim at something high.” And as Shakespeare has Hamlet observe, “There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy.” Might not the same be possible for establishing expectations about program interventions? Success in hitting performance targets may prove of little or modest substantive import if targets are set too close to where the archer stands.

B. Success

As with expectations, success is a concept subject to several definitions and associated with multiple connotations that can influence decisions when used in the context of performance management systems. The *Random House Dictionary of the English Language* defines success in several ways, including “the favorable or prosperous termination of attempts or endeavors,” and “a successful performance or achievement.” To compress what might otherwise require a concordance of nuanced terminological distinctions, “success” as employed here is a portmanteau term denoting positive valences with respect to the criteria and objectives subsumed under performance management systems, such as performance, progress, effectiveness, efficaciousness, and efficiency (Newcomer, 1997; Schweigert, 2006).

Success is performance measured against some standard, objective, or set of expectations. For many activities, the standard—the height of the bar—already exists, as in the form of historical performance records. For others, it is multifaceted, complex, or

loosely defined. At yet other times, it may be opportunistic: recall Senator Aiken's famous suggestion that rather than remain embroiled in a protracted war in Vietnam that seemingly had no clear end, the United States should simply declare victory and then exit.

Criteria for specifying and identifying success of course abound in every field of private and public sector endeavor, ranging from the Bowl Championship Series in college football, the National Research Council's 1995 rankings of graduate programs, the New York Times or the Washington Post's restaurant reviews, to the spectrum of federal programs examined in OMB's PART reviews. As illustrated annually in the tempests surrounding the release of U.S. News and World Report's rating of universities and colleges, debates about the appropriateness, reliability, fairness, and latent unintended consequences of selected criteria in specific settings are widely found.

Questions relating to successful performance are inherent in the conduct of federal programs. Abstracting from nomenclature and ideological tints, the consistent theme in the half century progression from program-planning-budgeting, to zero-based budgeting, to management by objective, to reinventing government, to the current emphasis in the United States and many other countries on the new public management has been the need for governments to focus on outputs, outcomes, and impacts, rather than inputs. A shorthand expression is that government expenditures need to be audited less in terms of insuring that funds have been expended for appropriated purposes and more for evidence that they have produced results (Power, 1997).

Within this larger context, three specific questions currently pervade assessments of federal education programs: 1) Is the program successful? 2) What criteria are used in gauging success? 3) What metrics are used to operationalize these criteria?

These questions are standard components of Executive Branch budget reviews, congressional authorizations, appropriations, and oversight hearings, and agency planning, priority setting, and self-assessment activities. They also are routine components of the charges issued by agencies to advisory and review committees, whether part of internal review undertakings, such as the NSF Committee of Visitors (COVs), or external expert reviews, such as may be conducted by National Research Council panels at the request of agencies or congressional committees

Is the Program Successful? Questions about whether or not a federal program has been successful have received new saliency from increased demands for public sector accountability and evidence of demonstrated performance, as represented by GPRA and PART. Federal agencies, for example, are now required to justify their R&D budget requests to OMB in terms of a set of R&D Investment Criteria. These criteria include a set of output or outcome objectives, which a program is expected to achieve. Demonstration that these objectives have been achieved (or at least that a performance management system is in place to document these effects) then becomes part of the input OMB uses in its budget recommendations concerning the future continuation and size of the program (Datta, 2007).

Review of federal agency R&D performance objectives and OMB's responses to them for FY2005 and FY2006 under the PART process however, reveal a diverse array of "technological" and "societal" objectives, and thus implicit measures of success

(www.expectmore.gov). For example, performance of the Department of Education's Early Reading First Program is defined in terms of the number of children from Reading First who enter kindergarten with age appropriate oral language skills; the Department of Energy's hydrogen fuel cell program is defined in terms of specific technical targets (density of hydrogen storage state technologies, in weight percent); performance of the National Institute of Health's AIDS research program is measured in terms of a (deliverable?) vaccine by 2010; FDA's diet-disease program, by "improved consumer understanding," and NSF's nanotechnology research by a "knowledge base."

Successful performance thus has different connotations and contexts across agency programs. In a stylized manner, both the Early Reading First and hydrogen cell programs posit discrete quantitative measure of objectives and outcomes; thus, presumably, readily quantifiable measures of determining whether or not they have been successful exist. The nanotechnology research program objectives also are cast in putatively scientific and technological terms, albeit "knowledge base" is a fuzzy concept, subject to varying interpretations and measures. The two programs, though, share a common feature in that activities directed towards attainment of the stated objectives rest primarily upon the performance of the grantees receiving federal funds. A modicum of control over attainment of scientific goals also may be said to hold for the National Institutes of Health (NIH)'s PART goal of a deliverable AIDS vaccine by 2010, although developing a vaccine and developing a deliverable vaccine are distinct objectives, as even more so is the objective of having a specific population inoculated with the vaccine. But assuming that the goal of a deliverable vaccine is reached, what are the criteria or expectations for success in reducing AIDS? NIH itself is not a provider of health delivery

services. To what extent is its program a success if a vaccine is developed but for whatever reason does not reach its target populations?

What Criteria Are Used in Gauging Success? Although logically antecedent to the first question, answers to this second question tend to be taken as self-evident. But they are not. Is success absolute-total eradication of some disease? Relative-performing, as well as some benchmark or improving over some baseline? Competitive-outstripping some rivals? Something else?

More importantly, what accounts for the use of one standard or another? Are differences between absolute and relative measures substantive or symbolic? Substantive differences relate to major differences in societal outcomes, such that say achieving only an 80 percent inoculation rate for a targeted population still leaves such a large residual untreated population that risk of contagious infection still remains? Are they instead forms of "symbolic politics" (Edelman, 1985) such that the original objective was known to be unattainable but nevertheless set forth because it was a powerful builder of consensus serving to coalesce traditionally different interests and/or so astutely expressed that to oppose it would leave dissenters vulnerable to public, or political, rejection? For example, the goal of No Child Left Behind of having all students tested in reading and math reach grade level by 2014 is increasingly seen less as "lofty" than as unrealistic, but nevertheless rhetorically persuasive (and politically useful).[4]

What Metrics Are Used to Operationalize These Criteria? The third question relates to the reliability and validity of the evidence employed to gauge performance. Are test scores on standardized achievement tests valid measures of the contributions of federal mathematics and science education programs? Are bibliometric measures a valid

indicator of the magnitude of the contribution of federal programs directed at fostering scientific discovery? Are sales an appropriate measure of the success of the Small Business Innovation Research Program? Are patents/venture capital tranches/sales, etc., appropriate indicators, singly or collectively, of the impacts of federal programs directed at fostering increased international economic competitiveness? In each of the above cases, debate exists about the construct validity of the metric cited, the importance of that metric relative to others than might be used to assess program performance, and the potentially dysfunctional aspects of driving program participants to shape their actions to perform well in terms of that metric at the expense of attainment of other program goals (Perrin, 1998; Weingart, 2005).

Opening the question of what constitutes success leads to a daisy chain of connected questions. Consider the several conventional ways in which success may be defined:

- Solution of a problem (removal of cause/symptoms—e.g., traffic congestion; airport delays);
- Statistically significant findings that a program has effects in the predicted direction;
- Measured improvement set against a benchmarking norm, whether in the form of a comparison with one's own baseline performance or against that of comparable, peer, or "stretch" units;
- Benefit-cost ratio greater than 1;
- First or "high" place in competitive races, rankings, or ratings;[5]
- Yes/no checks for selected categories of objectives and actions;[6] or

- Meeting accepted, if conventional, standards of statistical evidence: For example, "for better or worse," the term *statistically significant* has become synonymous with P = 0.05 (http://www.tufts.edu/~gdallal/p05.htm).

But beyond different definitions, yet other complexities arise in gauging program performance. How soon must the impacts of a program be demonstrated for it to be judged a success or not? How much impact must it have to be considered a success? Are the criteria for success based on absolute measures, relative to legislative or judicial requirements, the performance of comparable or bruited exemplars, or to expectations? How long after the experimental phase of an intervention ends must the treatment, or a close equivalent, be retained to permit one to say that success has been continuing? How permanent or stable are program benefits over time? When can one conclude that a demonstration or trial of a particular reform, intervention, or innovation has been shown to be sufficiently effective, efficacious, or effective to warrant ramping up to full scale implementation? When can one say that a program has been sufficiently successful in addressing the need or problem that it was originally intended to satisfy or solve that it is no longer needed? When can one say that a program, after being implemented for some period of time, has been sufficiently successful in achieving its intended objectives that it should shed its designation as an experiment and be incorporated as a routine part of an organization's ongoing operations (Yin, 1981)?

If success is measured in terms of rates of implementation, assuming a conventional logistically shaped diffusion path, what allowances in annual performance measures need to be made for inexorable rates of deceleration? Does "reinvention," using an experimental program to open up possibilities that lead to the rejection or termination

of the intervention but pave the way for ultimately high impact innovations count towards program success or not? And of course there is the normative, distributive, or political question, successful or not successful to whom? One person's wasteful program is another person's income, status, power, etc.

Answers to these questions are not generated by program theories, logic models, or evaluations alone. Rather they are shaped by the interplay among expectations, accommodations for experience, continuing demands from affected, influential and interested groups, the press (and pressure) from other claimants and policy issues for place on the policy agenda, and resources (Baumgartner and Jones, 1993; Kingdon, 1995).

Moreover, leaving aside issues relating to how expectations are formed or success is measured, the opposite of success for federal programs is not necessarily failure. The economic proposition that learning by doing is a source of productivity increase also extends to learning what not to do. If approached from a learning perspective, activities that demonstrably fail to achieve their stated objective may nevertheless produce new knowledge about the state of the world that leads to the more correct direction of future searches. This perspective is contained in Thomas Edison's adage that, rather than failing, his multiple experiments identified 1,000 things that didn't work. It is seen too in empirical findings on the high percentage of R&D projects that fail to achieve either technical or economic success in corporate R&D–"3,000 raw ideas = 1 commercial success," according to Stevens and Burley (1997).

Programmatic if at times opportunistic value also exists for falling short of some stated performance goal (at least in a politically and budgetary benign or neutral

environment). Gaps between goals and accomplishments can serve as justification for continuation of a program; progress has been made, more yet remains to be done. "Science" as an "endless frontier" (Bush, 1945) is perhaps the best example of the resource benefits of setting "unreachable goals."

Of especial importance in examining the iterative linkages between program objectives and assessments of program performance are the consequences of success or failure. Will the programs be terminated? Will those responsible for its conception, operations, oversight, etc. be penalized in some way; or rather will their creative, innovative, risk-taking initiatives be acknowledged and complimented even as the enterprise fails to achieve its intended objectives? If the penalties for failure are high, they may induce risk-averse behaviors and "cautious" setting of objectives, as critics of the new public management have contended, thus producing a program environment characterized by high rates of success but small improvements. If the penalties are small, they may lead to overly optimistic/ambitious form of bureaucratic entrepreneurship that lead to high rates of failure (Yin, 1977; Feller, 1980).

At some point in the evolution of a federal program, differences between the objectives stated in the preamble of legislation or in the goals and objectives advanced by agencies in promulgating internal strategic plans or GPRA documents and actual accomplishments may become an accepted adjustment factor in measuring success. Goals are stated in absolutes; accomplishments in advances from some baseline or in comparative terms.

III. DETERMINING THE SUCCESS OF FEDERAL PROGRAMS THAT SUPPORT K-12 EDUCATION

The analytical, methodological, measurement, and policy complexities associated with uniquely defining and thus measuring success and then using these constructs and measures to answer the three assessment questions described above are compounded (and confounded) when the goal of a federally-supported K-12 program is to generate new interventions that are then to be used to test, validate, or demonstrate a new approach in pursuit of a given educational objective. Success for such a program depends upon satisfactory completion of two logically connected but in fact operationally independent components: first, a research component; and second, an implementation, adoption, or diffusion component (Hallfors and Godette, 2002). Thus, for a program such as MSP to be deemed successful, not only must the program produce findings that contribute in some significant way to an existing body of knowledge or practice and that are relevant to the issues or objectives at hand, but these findings also are expected to contribute to improved performance on the part of K-12 schools systems or post-secondary institutions directly engaged in supplying preservice educations, organizations over which a federal agency frequently has limited control.

The challenge of determining degrees of success for such compound programs is illustrated by Mosteller and Weinstein's (1985) example of an R&D program directed at fostering subsequent changes in clinical practice: "If a new method of diagnosis successfully detects cases of a disease for which we have no effective treatment, how valuable is the technology? It may be useful for counseling or for research, but the effect on health outcome may be negligible" (p. 236).

To cast this statement in terms of program evaluation, a difference exists between (1) insuring that an evaluation's design guards against internal threats to validity and accurately measures effect size, and (2) determining whether or not the measured effect size is sufficiently large to meet the requirements or expectations of decisionmakers, program managers, program participants, program recipients, and others whose actions and support influence whether the program is continued, redirected, modified, or terminated.

The further complexity of interest here is the use of cross-program/cross-agency comparisons in setting standards of success for agency-specific programs. "Models," "best practice" lists, or exemplars of what works in specific settings are frequently transferred to other settings with an incomplete understanding of the workings of the model in its original setting, and even less examination of whether the conditions that affected the workings of the program in one setting are to be found in its new one. In agriculture, federal technology transfer programs have long been under the intellectual and programmatic influence of the agricultural extension model with little recognition of the questionable current effectiveness of the program or, more importantly, of the model's limited generalizability to other settings (Feller, 1993). Likewise, attempts at conducting meta-analyses of effect size across program areas has proven difficult because "…standards regarding what are considered 'big,' 'medium,' and 'small' effect sizes vary across fields and contexts" (Gandhi et al., 2007, p. 59).

To illustrate the challenges of determining the success of a federal program directed at improving performance in a specific functional domain, consider the following excerpts from the program evaluation/program implementation fields not as

discrete findings and associated commentaries but as a logically connected sequence from evaluation to measure of success to implementation to impact.

Lipsey (1990) uses the example of the effects of an experimental cancer treatment that in a randomized trial produced a 0.2 standard deviation reduction in death rates from 55 to 45 per 100 within one year to illustrate how the use of specific statistics in meta-analyses, such as the value of a standard deviation, can lead to misleading conclusions about effect size. In fact, according Lipsey, the treatment produces an 18 percent (10/55) reduction in the death rate. He then writes: "When statistical effects are interpreted in this manner, it is difficult to argue that 0.2 is necessarily a trivial effect" (p. 24). The implication is that the program's effects are indeed non-trivial.

Does a non-trivial effect constitute a success, however? One of the basic tenets of diffusion research is that of "relative advantage," namely the "degree to which an innovation is perceived as being better than the idea it supersedes" (Rogers, 1995, p. 212). What impacts on the rate and extent of adoption does a relative advantage of 18 percent produce? Is such an evidence-based impact sufficient to move the program through multiple gates of implementation described in Pressman and Wildavsky's classic study of "How Great Expectations in Washington are Dashed in Oakland?" Is the subsequent rate sufficiently fast and extent of adoption sufficiently large to impact on the societal objective underlying the program, thus leading decision makers or stakeholders to conclude that that the program that led to the initial experiments has succeeded in achieving its stated objectives?

From the perspective of policy makers, program managers, service delivery suppliers, or service delivery recipients engaged in an R&D/demonstration program, the

question might be rephrased as follows: Assuming a statistically significant finding from a well crafted evaluation design of an 18 percent reduction in death rates, is this effect sufficiently high enough to permit one to conclude that the program was successful? What if initial expectations were that the effect would be 30 percent or 50 percent?

More pointedly, assume that the objectives specified in a GPRA or PART submission were 18 percent plus "x." If such were the case, the program would have fallen short of expectations or commitment, however well founded or unfounded these may have been. The "shortfall" might thus be deemed to denote failure and lead to reduction, postponement, or termination of the program. Much the same issue of disparities between expectations and accomplishments can be seen in current controversies about No Child Left Behind's stated goal of having 100 percent of students perform at grade level as determined by standardized tests. This goal may be alternatively perceived as a politically motivated rhetorical flourish or as a legitimate, indeed necessary, response to democratic claims for equitable attention to all. Imagine, for example, stating a goal of 80 percent. Such a goal would immediately raise questions about which populations would not be served. Conversely, if expectations were for a reduction in the 5-10 percent range, reported results of 18 percent might produce a boomlet in budgetary support for the program. (Consider for example the possible boomlet in support for DOE's hydrogen cell R&D program if early findings exceed expectations!)

Posing these different possible adoption/utilization scenarios from the same estimate raises anew the essay's opening question of how expectations or criteria concerning what constitutes success are determined. Thus, why are five-year survival

rates used to gauge the effectiveness of experimental cancer trials, even though "there is no specific biological significance about having survived five years"(Myers and Ries, 1989, p. 21), with evidence accumulating that the rate is an imperfect predictor of mortality rates from various forms of cancer? (As noted by Welch, Schwartz, and Woloshin (2000), "Increased five-year survival for cancer patients is generally inferred to mean that cancer treatment has improved and that fewer patients die of cancer. Increased five-year survival, however, may also reflect changes in diagnosis: finding more people with early-stage cancer, including some who would never have become symptomatic from their cancer")

Assume next that agreement exists that a trial has produced significant results, with the research itself leading to publications and citations, and thus being deemed a scientific success. What can one expect about the rate and extent of its adoption into practice? The safe, empirically justifiable answer is that it may be elongated, that it will vary across innovation, and that it may or may not course the paradigmatic S-shaped time path (Karshenas and Stoneman, 1995).

For example, Balas and Boren (2000), on the basis of their survey of nine clinical procedures, estimated that it would take 15.5 years for a procedure to reach a rate of use of 50 percent, assuming that it had been zero at the time of publication of the landmark study (see Table 1).

But their framing of the issue is even more relevant here: "Is 50 percent utilization rate an acceptable threshold for declaring success in the practical implementation of clinical recommendations?" (ibid, p. 66). Again, what if expectations and requirements were higher? Imagine reporting that a federal program designed to

reach, or impact, all, or most, of a targeted population had reached only 50 percent of its potential users after 15 years!

Table 1

Clinical Procedure	Landmark Trial	Current Rate of Use 2000
Flu vaccination	1968	55%
Thrombolyic therapy	1971	20%
Pneumococcal vaccination	1977	35.6%
Diabetic eye exam	1981	38.4%
Beta blockers after MI	1982	61.9%
Mammography	1982	70.4%
Cholesterol screening	1984	65%
Fecal occult blood test	1986	17%
Diabetic foot care	1993	20%

Source: Balas and Boren, 2000, p. 66.

IV. A FRAMEWORK FOR IMPLEMENTATION

With the likely exception of the functional domains of defense, space, and aspects of homeland security, where federal funds are directed at producing new knowledge for goods and services for which the federal government as direct producer or purchaser can require the utilization of these goods and services, utilization of interventions in most fields of federal domestic activity occurs via the provision of goods and services, either directly or through purchases, by other actors, principally state and local governments, or third sector, not-for-profit organizations.

Getting federally-funded technology, ranging from improved breathing apparatuses for fire fighters to improved methods of K-12 math and science education, into the hands of users, who in turn, actually use the technology in ways that prove productive, is the subject of the literature on research utilization, technology transfer, diffusion of innovations, and related variants. However, not only is this literature extensive and diverse, but on several salient points it is at times both contradictory and competitive, producing in effect a matrix of disciplinary or analytical approaches (rows), and substantive fields (functional fields; e.g., medicine; education; community development) (column) (Downs and Mohr, 1976; Tornatzky and Fleisher, 1990).

No attempt is made here to review the literature or associated debates. Instead, the literature is selectively drawn upon and then matched with selective drawings from the literatures on program evaluation, program implementation, and knowledge utilization to advance an eclectic, reduced form, albeit not simple model of the ability of a federal agency to affect service delivery in state and local government entities through its

programs. This ability in turn produces the performance results that determine whether results meet expectations.

The model is organized about five vectors: (1) federal expenditures, total and as a percent of total expenditures directed at a field; (2) stability and consensus surrounding the field's knowledge base; (3) federal authority to determine program delivery characteristics ("jurisdictional control"); (4) specificity and degree of consensus about overarching objectives; and (5) tightness of the adoption/diffusion process.

Each of these vectors is seen as affecting levels of program expectations or attainment of success, with no single vector alone shaping program success. Also, in what follows, no a priori assumptions are made about the relative importance of the vectors (within and across fields). Instead, the interactive and cumulative effect of these vectors is presented as creating more or less congenial environments within which federal programs operate, and thus their potential for generating required/expected impacts.

Conceptually, these environments may be thought of as constituting a continuum from very hospitable to very inhospitable. At one end would be a federal agency that has a well defined, single objective, a readily observable yes/no measure of success, abundant resources to devote first to basic or discovery research and then to subsequent large-scale randomized trials of new approaches and accompanying independent replications, and high levels of authority to require adoption by end users. A national space program with the objective of landing a person on the moon is an obvious example of such an environment. (Actually, a more complete measure of success is to land the person safely on the moon and then have the individual return safely to earth.)

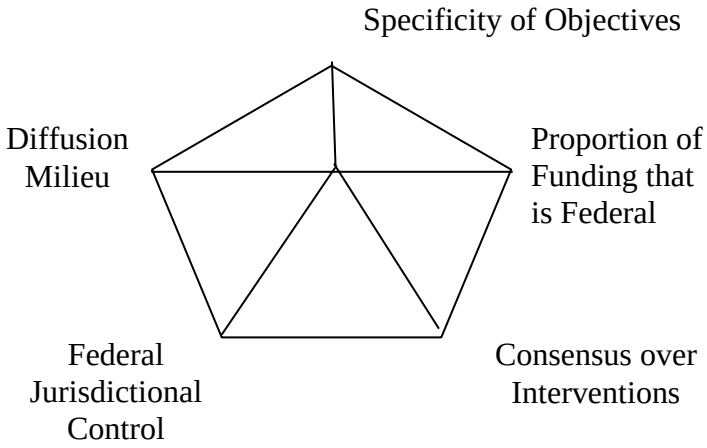
At the other end of the continuum would be a federal agency where the federal investment is small in either or both absolute or relative terms, where multiple, potentially competing objectives exist for the program, where there is a plethora of competing theories and findings concerning the validity and relevance of new interventions, where federal control of the adoption environment is weak, and where the adoption environment is complex (or loosely coupled). The federal role in highway research and technology development has been described in terms that echo these characteristics.[7] Indeed, it is difficult to think of a federal program directed at non-defense, non-space domains where the conditions needed for creating a congenial, welcoming environment hold. Most programs—i.e., education, crime prevention, substance abuse, community development and housing—likely would be placed towards the "inhospitable" end of the continuum.

An alternative way to depict the interplay between federal agencies and these vectors would be to construct a matrix in which the rows represent functional domains and the columns the vectors of influence. Cell entries would represent ordinal measures or assessments, say high, medium, and low. The expectation then would be that federal program interventions in those fields with larger and more highly weighted cell entries would be more likely to achieve success than those with fewer and lower weight entries.

In order though, to highlight variations in the degree to which the values of these vectors can vary across functional domains without having to address their relative weights, the approach presented here depicts functional fields/agencies in terms of their positions on scales associated with each vector. Figure 1, in the form of a pentagon, with

each axis representing a different vector, depicts the general scheme. The lines from the center to the axis represent scales appropriate to the variables contained within the vector.

Figure 1: Positioning the Impact of Federal Programs



As a starting point, Figure 2 uses the approach to depict a stylized version of the moon-ghetto metaphor. In this figure the outer program dominates the inner program in each of the 5 relative vectors. A federal program positioned as the outer program would be expected to show a higher (or faster) rate of success than the inner program.

Figure 2: Positioning the Moon-Ghetto Metaphor

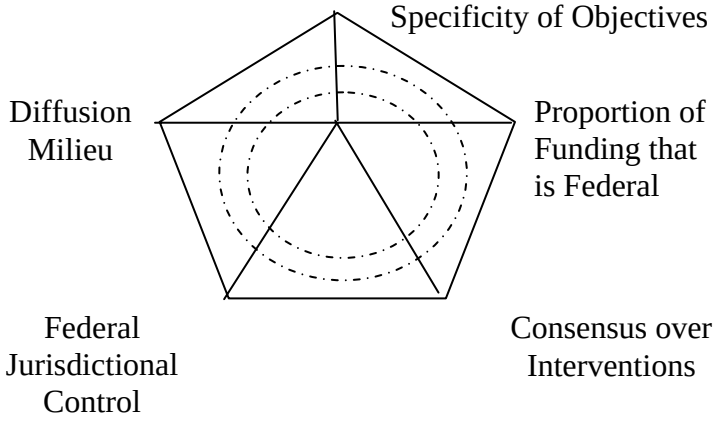
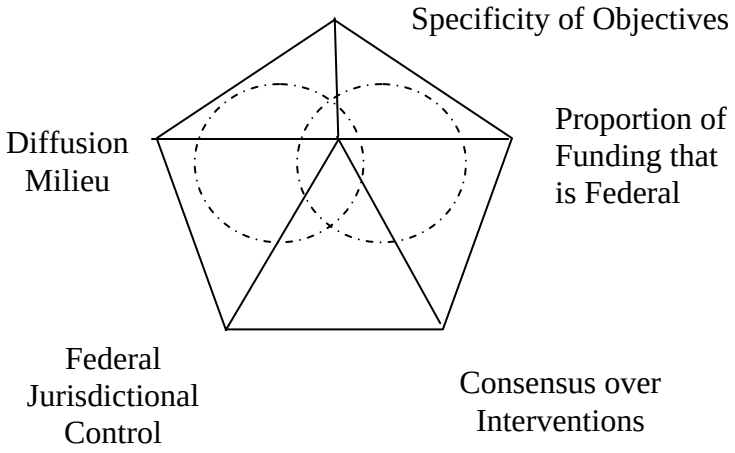


Figure 3 illustrates a case where one program provides a more hospitable environment for program success in some but not all of the vectors, say in being better funded or having a more systematic, evidence-base for recommending new approaches but being confronted by a more complex, resistant diffusion/implementation environment. Comparisons between the two programs in terms of expected or realized levels of success are more problematic here, and would depend on the relative weight/importance of the five vectors in shaping program performance.

Figure 3: Positioning Different Domestic Programs



Brief accounts of how each of these vectors can affect program success and the interactions among them are presented in the sections below.

A. Proportion of Federal Investment

The absolute and relative size of federal investment in a functional domain has several different impacts on shaping the environment for the acceptance of findings from experimental or innovative programs. In absolute terms, expenditure size affects the number of embryonic ideas that can be nurtured to the point where they can be transitioned from concept to demonstration stage, the number that can be advanced to

experimental, demonstration, or field tests, the feasibility of alternative evaluation designs to assess efficacy or effectiveness under different contextual designs, and the feasibility (and number) of replications of even the most carefully designed and implemented evaluations. As Cronbach (1982) has suggested, a "fleet of studies" provides both for critiques of single studies and cumulative advances in methodology.

Multiple projects, assuming multiple, independent investigators, increase the number of independent perspectives and the likelihood that limitations or flaws in even the most well crafted studies will be detected, and reported. Multiple studies also provide for triangulation of findings from closely related but not identical studies. They serve to generate ranges of estimates that in effect produce confidence intervals for gauging the plausibility of findings from any single study (as well as contributing to the formation of expectations about the impacts of comparable future programmatic investments).

Consider, for example, the interpretation given to the summary of selected estimates on returns to agricultural R&D and R&D spillovers offered by Griliches (1995). Table 2 follows an extended discussion by Griliches of the conceptual and measurement problems associated with estimating returns from science and technology. Based though on the number, quality and convergence of findings, he is able to conclude as follows:

"In spite all these difficulties, there have been a significant number of reasonably well done studies all pointing in the same direction: R&D spillovers are present, their magnitude may be quite large, and social rates of return remain significantly above private rates" (p. 72).

Relative expenditures impact the likelihood that findings from federally-funded evaluations will be dominant or hegemonic in determining what is known about a program's effectiveness. Multiple sources of funding increase the prospects of alternative perspectives, methodologies, assessments of existing programs, and a range of policy alternatives. In all, having multiple projects strengthens the prospects for a field's acceptance of the validity and significance of findings, and in turn, possibly although not necessarily as suggested by continuing resistance to findings about global warming, to the increased acceptance and utilization of interventions findings by policy makers.

Table 2: Selected Estimates of Returns to R&D and R&D spillovers

I. Agriculture	*Rates of return to public R&D*	
Griliches (1958) Hybrid corn	35-40	
Hybrid sorghum	20	
Peterson (1967) Poultry	21-25	
Schmitz-Seckler (1970) Tomato harvester	37-46	
Griliches (1964) Aggregate	35-40	
Evenson (1968) Aggregate	41-50	
Knutson-Tweeten (1979) Aggregate	28-47	
Huffman-Evenson (1993) Crops	45-62	
Livestock	11-83	
Aggregate	43-67	
II. Industry	*Rates of return to R&D*	
Case Studies	Within	From outside
Mansfield et al. (1977)	25	56
I-O Weighted		
Terleckj (1974) total	28	48
private	29	78
Sveikauskas (1981)	10 to 23	50
Goto-Suzuki (1989)	26	80
R&D Weighted (patent flows)		
Griches-Lichtenberg (1984)	46 to 69	11 to 62
Mohen-Lepine (1988)	56	28
Proximity (technological distance)		
Jaffe (1986)		30% of within
Cost Functions		
Bernstein-Nadiri (1988, 1989)		20% of within

differs by industry	9 to 27	10 to 160
Bernstein-Nadiri (1991)	14 to 28	Median: 56% of within

Source: Grilches, op. cit., p. 72

The size and proportion of federal investments in different sectors vary in a widely recognized manner. Federal R&D expenditures currently represent approximately 30 percent of total U.S. R&D expenditures, with the largest share, 64 percent, coming from industry, and the balance from colleges and universities, state and local governments, foundations, and other not-for-profit organizations (AAAS, 2007). The federal government obviously dominates total funding for defense. At a lower end of the expenditure continuum, federal outlays for K-12 education represented only about 5.7 percent of the total K-12 expenditures in 1990-91; the proportion has risen since then, but only to 8.3 percent in 2004-05 (U.S. Department of Education, 2005).

B. Consensus Over Interventions

Federal support of mission-oriented programs, including associated evaluations of demonstration projects, as in K-12 education, is predicated on the two-part proposition that public sector programs (1) should be a theory-based and empirically proven set of cause-effect relationships, and (2) articulating the scientific, evidence-based basis of an intervention will foster user receptivity to the program (Coalition for Evidence-Based Policy, 2002). These premises underlie the heightened emphasis on the scientific basis of federal initiatives in fields such as education and substance abuse prevention, and help account for the attachment of the word "science" to recent initiatives in education (e.g., Education Sciences Reform Act; Institute of Education Sciences), and the emergence of the new field, prevention science.

The two premises are accepted in the following discussion.[8] Of interest here, rather, is the identification of generic issues relating to the degree of consensus within a field concerning the validity or import of new scientific findings, the frequency and impact of replications, and the jostling for relative position of evidence-based knowledge and ordinary knowledge. With theory-based programs and well-crafted evaluations, these issues can affect the confidence that potential users have about the applicability and benefits of innovative programs being offered to them, and thus their willingness to accept them. For each of these generic issues, differences may exist across functional domains.

Two further introductory comments are needed here because the issues discussed below are entangled with other aspects of treatments of what is termed "knowledge utilization." First, the discussion here is limited to consideration of differences in the scientific basis and stability of knowledge among functional fields; it is not a general overview of the conditions determining how, if at all, social and behavioral science research is used in policy making (e.g., Hunter and Schmidt, 1996). Second, it abstracts from political or ideological acceptance, resistance, or rejection, to evidence-based findings about the success or failure of specific programs.[9]

Consensus. New research findings seldom quickly and totally displace existing bodies of knowledge. At any point in time, disagreement can exist among experts concerning the truthfulness of specific findings, or perhaps more importantly the generality of one finding as contrasted with another as each seeks to explain the same phenomena (Cole, 1992). Even where there is a clearly articulated behavioral or social science program theory underlying a federal domestic R&D program followed by well

crafted studies or evaluations, the scientific community's acceptance of findings may not be incomplete.

Indeed, the very term "consensus" currently used to describe various ongoing meta-analyses endeavors should give one pause. The very term implies continuing differences. Procedures directed at fostering or distilling consensus, such as meta-analysis, may be helpful in separating the wheat from the chaff, especially in fields reported to be beset by poorly done research—as indeed is suggested by the 2004 NRC report *Advancing Scientific Research in Education*. Not all wheat strains are the same, however; consensus is not unanimity. Consensus may be viewed as a part-full/part-empty glass, with the ratio loosely defined.

The formation of consensus is an integral component of the processes leading to the acceptance of new scientific findings (Bowler and Marcus, 2005). Frequently, it is only after extended periods of time that agreement is reached about what is true, and what works, and then not always. As expressed by Planck (2002) in an oft-cited statement, "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation growing up is familiar with it" (as quoted in Farmelo, 2002, p. 26).

Examples of continuing disagreements among experts are readily found for K-12 education. Abstracting from the controversies surrounding possible conflicts of interest in the administration of the Department of Education's Reading First Program, substantive disagreements exist among reading specialists about the findings from the National Academies report, *Preventing Reading Difficulties in Young Children*, and the subsequent report of the National Reading Panel, *Teaching Children to Read*, that emphasized

instruction in phonics as the preferred scientifically based basis for teaching students to read (Schemo, 2007).

Different schools of thought may continue for long periods of time. These schools may offer empirically tested, proven, but competing theories of how the world operates, from which they deduce, at times, contradictory policy recommendations. Rule, for example, has argued that "although social science research produces many 'findings,'" one must strain to identify what could legitimately be called "social science" discoveries or empirical observations by any name that decisively settles theoretical controversies (Rule, 1997, p. 3). A similar view on the problem of too many findings yielding too little consensus is contained in Smelser's (2005) observation, that the "history of the social sciences has been one of intellectual ferment that produces increasing numbers of perspectives, schools, and approaches, few of which die, and all of which provide frameworks for identifying what counts as appropriate social science knowledge…The result of this growth and diversification of perspectives is that consensus about the nature of discoveries and findings grows less likely over time" (p. 246).

Economics offers another example here, with its well-known schism between "freshwater" (Harvard/MIT) and "lake water" (Chicago, Rochester, Minnesota) growth theorists, and Nobel laureates holding dissimilar views on monetary and fiscal policy as Paul Samuelson and Milton Friedman (Warsh, 2006).

Indeed, stepping back from specific controversies, disagreements, or even differences within fields, one may see contemporary efforts such as meta-analysis, Campbell Collaborations, and NIH Consensus Conferences, and like efforts, to distill "evidence based findings" from a cacophony of "reported conflicting and contradictory

findings and conclusions" from different research studies on the same question as an attempt to forge, and legitimate "consensus," about what "research tells us" about a field (Hunter and Schmidt, op. cit., p. 32).

Replication. The importance of replication in validating findings from large-scale program interventions has already been noted. In practice, few replications exist for many major federal program interventions. Moreover, a goodly number of existing replications, at least in the fields of education and substance abuse prevention, have been conducted by the developers of the interventions being tested. Systematic biases for developers of programs to report larger effect sizes than independent researchers have been reported (Borman, Hewes, Overman and Brown, 2003). Additionally, in some recent cases, independent replications have been conducted of widely disseminated, research-based interventions, finding little or no support for the intervention's initially reported efficacy (St. Pierre et al., 2005; Mandel, Bialous, and Glantz, 2006).

Ordinary Knowledge. Orthogonal to contemporary emphasis on formal theory construction and testing as the basis for designing and evaluating public sector programs is a perspective rooted in practitioner, field-based experience, or what Lindblom and Coehn (1979) have termed "usable" or "ordinary" knowledge. Ordinary knowledge is "knowledge that does not owe its origin, testing, degree of verification, truth status, or currency to distinctive PSI (professional social inquiry) professional techniques, but rather to common sense, casual empiricism, or thoughtful speculation and analysis" (p. 12). This perspective emphasizes the importance of the contribution of practical, but systematic knowledge that can be used in real-world settings (contexts) over empirically-validated exemplary models (Machlup, 1962).

Ordinary knowledge would appear to be especially important in practitioner assessments of the applicability of innovative approaches to their specific contexts. Without questioning the validity of the findings underlying the newly proposed model, even accepting without question the good intentions and professional standards of the organization sponsoring a list of models or certified programs, practitioners may simply not see the model as relevant to their setting. Absent positive or negative incentives, funds for adopting it, or loss of accreditation for not adopting it, for example, they may choose to retain existing approaches or adopt others not on the certified list.

C. Jurisdictional Control

In the U.S.'s federal system of government, authority for the conduct of different public sector functions is variously distributed among the federal government, state governments, local governments, as well a myriad of derivative interstate and intrastate regional and special purpose governmental organizations (e.g., water districts) (Wright, 1988). The result is what has been described as a "marble cake" rather than a "layer cake" of jurisdictional intermingling. Initiation of new governmental policies or innovative program approaches may emerge from state or local government action, becoming the basis for federal government legislation mandating implementation across the nation—the oft-cited "laboratories of democracy" metaphor. New polices and programs instead may flow directly from federal action that imposes new requirements on subnational jurisdictions or preempts them from acting on their own.

Jurisdictional boundaries within functional areas between and among governmental units are often blurred, and subject to both structural and episodic change. As illustrated both by ongoing disagreements between the Department of Education and

several state educational agencies about the Department's emphasis on the use of phonics in granting awards under the Reading First Program, and between the Department and state higher education offices (as well as universities) about the Commission on the Future of Higher Education (Spelling Commission) concerning proposed new federal rules governing accreditation, conflicts over jurisdictional authority can dominate consideration of the merits or effectiveness of program intervention under review.

Within the boundaries of its jurisdictional control, the federal government employs several carrot-and-stick techniques, singly and in combination to induce or require state and local governments to adopt interventions flowing from its programs. The principal ones are grants (to encourage adoption), regulations (that mandate specific action to be eligible for project specific grants, or more importantly, affect eligibility for ongoing programmatic appropriations), and public education programs (to foster awareness and interest). These techniques are frequently packaged as follows: federal legislation mandates that state and local governments achieve specified objectives; federal appropriations are made available to these units to implement policies and programs directed at achieving these objectives, subject to scrutiny and approval by the federal agency charged with overseeing compliance with the legislation; and technical assistance is provided, including launching new interventions directed at generating a sounder knowledge base for achieving the objective.

A direct testable hypothesis is that the tighter the degree of federal government control of a federally-distributed functional domain, the greater is its ability to set *standards* about what constitutes satisfactory performance, either in terms of specific levels of output or achievement, or in terms of the mandated use of specific inputs or

production techniques. An agency's degree of control over standards, in turn, would therefore determine its ability to foster the adoption of findings emerging from its R&D programs, including here evaluations that attest to the efficacy of these findings.

These patterns are readily seen in the trend towards increased federal government involvement in the performance of America's elementary schools. Historically, the province primarily of state and local government responsibility and control, federal government standard setting and regulatory oversight have increased markedly since the Elementary and Secondary Education Act of 1965, reaching new heights with the standard setting requirements specified by the No Child Left Behind Act of 2001 (Manna, 2007). NCLB in effect requires schools to show measurable student progress on mandated standardized tests if they want federal money.

The seeming logic, and possible tempting interpretation, of the above discussion is that the fostering of adoption and implementation of federal interventions is best accomplished when it is accompanied by increased jurisdictional control. Indeed, historically a plausible case could be made that earlier broad-based federal efforts, circa 1970s, that promoted the development and adoption of what were termed new public technologies largely failed because they represented technology pushes in functional domains where federal jurisdiction was weak.

There are two downsides contained in this logic, however, that also need to be considered. Standard setting is a risky business, latent with both positive and negative consequences. Viewed positively, standard setting holds the promise of insuring (high) minimum levels of performance, improving interoperability among component parts of larger systems, and reducing search and transactions costs in deciding which of multiple

preferred approaches fulfills new requirements. However, in a dynamic termed by David as "narrow windows, blind giants," federal agencies are likely to have the greatest influence on practice when they set standards at the beginning of the implementation of a new policy. This is the time, however, when they have the least information about the best technology today and in the future. Setting the wrong standard or setting a standard too early in the development of a new science-based technology may lock a system into an inefficient mode of production/service delivery (Stoneman, 1987).

Second, it is a basic principle of the economics of regulation that once a standard is imposed (and enforced), incentives are created to cheat in order to qualify for payments tied to meeting the standard and to bribe those responsible for monitoring compliance with the standard. These dynamics have already been observed in the case of NCLB, where teachers have been found to have inflated student test scores to qualify for bonus pay based on student achievements (Levitt and Dubner, 2005). The current imbroglio over conflicts-of-interest in the determination of state eligibility for awards under the Reading First Program is another example of rent-seeking behavior made possible by the search for scientifically-based standards.

C. Specificity and Degree of Consensus About Objectives

Federal programs frequently contain multiple objectives; these objectives may be reinforcing, or confounding and contradictory.[10] The programs may be comprised of multiple elements, each connected to one another by a program theory and a structured logic model, but each containing the prospects for subgoal or subunit optimization in ways not necessarily consistent with, or conducive to, the attainment of overarching objectives. Spliced together, specificity and degree of consensus about overarching goals

and subgoals affect the degree of hospitality of state and local units to findings from federal programs, how these findings are used, and thus the program's measured impact and success.

Whenever multiple goals exist, opportunities for trade-offs are present. Where programs are based on multi-component, non-linear relationships, the logic of logic models is best described as fuzzy logic: successes or failures at single nodes may or may not be critical determinants of successfully meeting overarching performance objectives. Thus, a distinctive feature of the U.S. national science policy is to situate the larger part of federally-funded basic research in universities. Part of the rationale for this policy is that research conducted in universities, in comparison say, to that conducted in a government or private sector research laboratory, is viewed as yielding a joint product: new research findings, and the training of students, the future generation of scientists and technically trained individuals. In practice though, principal investigators operating within the constraints of given project budgets must decide how many students (and at which degree levels) to support. Since rewards to the faculty researcher tend to be based on research output rather than student output, substitution of lower cost or more reliably available personnel for students becomes a rational subunit decision, albeit one not necessarily consistent with overarching agency missions.

The settings into which federal initiatives to improve student performance in K-12 education enter offer additional examples of the importance of specificity and consensus about objectives. Improved student performance, as measured by standardized tests, is a core objective of NCLB and MSP. However, agreement that improved performance on standardized tests is the overarching objective of K-12 education, or indeed of any single

aspect of elementary education is far from universal (Shulman, 2007). As Gardner (2004) has observed:

"Members of a society can reach agreement with relative ease about the purpose of medicine—to deliver high-quality health care to all citizens; nor need the purposes of the military or the monetary system be perennially disputed. However, except for certain fundamentals, the purposes of education, and the notion of what it means to be an educated person, are subjects about which individuals—both professionals and lay—hold distinctive and often conflicting views" (p. 236).

Leaving aside the accuracy of this statement with respect to its examples of purported consensus—significant disagreement frequently exists about both the purposes and the techniques of monetary policy—the relevance here of Gardner's statement is its highlighting of differences about agreement of purposes across functional fields.

E. Diffusion/Implementation

As described above, for federal programs to be successful, their findings must be adopted and used by the entities that directly provide services to the populations whose improved performance or well-being is the ultimate objective. Also as noted, the literature on the adoption, use, and diffusion of innovations is extensive, diverse, and at times contradictory. The conclusion that follows from this assessment is that any distillation of this literature emphasizes the likelihood of different combinations and importance of variables across functional domains. Thus variables and relationships that are simultaneously both important and susceptible to external manipulation are likely to be different across the illustrative fields of K-12 education, agriculture, highways, coastal zone management, and health care delivery.

Studies of the diffusion of educational innovations indeed abound, although "education is less important to the theoretical understanding of the diffusion of innovations" (Rogers, 1995, p. 63). These studies are characterized by the same complexity, or diffuseness, that besets diffusion research in general, namely the challenge of sorting through the relative importance of the characteristics of the innovation, the characteristics of the innovator or organizational setting, and the larger environment (or milieu) on rates and extent and adoption. Perhaps the most stable themes to emerge from these studies are that diffusion of educational innovations occurs only after an extended period and varies among innovations. [11]

Assessments then of the impacts of federal programs directed at improved academic performance must be predicated on assumptions or findings related to how quickly and widely findings course through the multiple gates of adoption and diffusion. Given differences in the characteristics of diffusion milieus across fields (Feller and Menzel, 1977), the use of comparisons or benchmarks would appear problematic, which is probably a good thing since several such comparisons, as suggested by the Boren and Balas study (2000) cited above, point to elongated adoption curves.

The passive voice is inappropriate here, however. Federal agencies seek to actively transfer program findings or to have these findings implemented. They provide funds to adopt and implement new practices; they publish lists of model or exemplary practices directed at inducing and guiding adopters to select validated new approaches; they foster partnerships among key organizations; and more comprehensively now, as with NCLB, they impose requirements on school districts to attain specific performance standards, support and validate new approaches directed at enhancing the capacity of

schools to meet these standards, and then provide funds and technical assistance to adopt these approaches, but not necessarily others which have not been similarly certified.

As both a practical and theoretical matter though, implementation is a mixed bag in many sectors. Thus, commenting on the state-of-the-art implementation processes in the private sector, Thomas and Strickland (2000) have observed that, "Unfortunately, there are no 10-step checklists, no proven paths, and few concrete guidelines for tackling the job. The best evidence on dos and don'ts comes from the reported experiences and 'lessons learned' of managers and companies—and the wisdom they yield is 'inconsistent'" (p. 269-270). Likewise, in higher education, considerable divergences have been found between the boldly announced initiatives to promote interdisciplinary research, and actual steps towards implementation (Feller, 2007b). Thus it should not be surprising, nor is it a singling out, to observe that few evidence-based approaches exist on how to implement programmatic innovations in K-12 education, a system that at best can only be described as loosely coupled (Weick, 1969). Moreover, it is not simply that few sure guides exist about how to implement new approaches; rather it is that many of the very features above that would appear to contribute to a tightly linked research-dissemination-adoption system may also create environments inimical to improved performance.

The current controversy about the Department of Education's procedures for awarding grants to school districts under the Reading First Program for example illustrates how the combination of differences among experts (or the lack of consensus, much less unanimity, in what constitutes evidence) and between experts and

practitioners, the extension of federal government influence on curricular decisions in K-12 education, and the provision of federal grants can create the potential for self-serving behaviors, that, aside from issues of improper legal or ethical behavior—judgments which are beyond the scope of this essay—can adversely impact on sought after educational outcomes.

V. CONCLUSION

The essay is seen as generating value along four dimensions. First, conceived of as an exploratory study located at the intersections of policy analysis, program evaluation, and program implementation, it fills a gap in current analysis and public policy debates about the impacts of the new public management, especially with respect to education programs. Second, it identifies a research agenda for subsequent elaboration and testing of specific hypotheses about the determinants of the impacts of federally-supported education programs. Third, even in its current exploratory form, it offers a heuristic—a way of thinking—about the expected benefits of an education program that may assist program managers, policy makers, and program participants to shape program goals that more realistically approximate what the best and the brightest can attain, given the nature of the societal problem being addressed, the existing and projected new knowledge base from funded interventions and available resources. Fourth, it offers an analytical framework that permits decision makers and program managers to frame expectations about program impacts within the context of impacts generated by similar programs. Most importantly here, the framework can be employed by program managers as a diagnostic to identify and distill the contextual factors that have shaped favorable program impacts elsewhere, with a view towards designing and implementing specific program approaches that increase the predicted impact of the innovative program.

Thus, returning to the earlier hypothesized positioning of federal programs in terms of the pentagonal, a more refined, but still hypothesized positioning, would be to group programs into three broad, non-intersecting circles: program areas where the federal impact is likely to be (relatively) high—defense, space; medium—agriculture;

and low—education, community development. Ideally, a federal program would like to be able to improve its projected impacts by moving outward along all five of the pentagon's rays simultaneously or in a coordinated manner; that is, to represent a larger share of total resources; have fewer and more explicit, universally shared objectives; operate with a fecund and agreed upon knowledge base; strengthen jurisdictional control; and operate within a tightly coupled adoption-diffusion-routinization system. Not all these moves may be feasible at a point in time; trade-offs may exist in agency and program strategy between and among outward moves.

Contextualizing expectations in this manner lays the groundwork for the next stages of the current research. These stages include data collection to construct the scales embedded in the pentagonal model, and then to use the model to identify ways in which federal programs might position themselves to have greater impacts in the K-12 educational arena.

REFERENCES

American Association for the Advancement of Science (2007). *R&D FY 2008* (Washington, DC: AAAS).

Balas, E. and S. Boren (2000). "Managing Clinical Knowledge for Health Care Improvement," *Yearbook of Medical Informatics*, 65-70.

Baumgartner, F. and B. Jones (1993). *Agendas and Instability in American Politics* (Chicago, IL: University of Chicago Press).

Behn, R. (2001). "Rethinking Democratic Accountability," (Washington DC: Brookings Institution), 8.

Borman, G., G. Hewes, L. Overman, and S. Brown (2003). "Comprehensive School Reform and Achievement: A Meta-Analysis," *Review of Educational Research*, 73: 125-230.

Boruch, R. (2007). "Encouraging the Flight from Error: Ethical Standards, Evidence Standards, and Randomized Trials," in Julnes and Roeg, op. cit., 55-73.

(2005). "Better Evaluation for Evidence-Based Policy: Place Randomized Trials in Education, Criminology, Welfare and Health," *Annals of the American Academy of Political and Social Science*, 599: 6-18.

Bowler, P. and I. Morus (2005). *Making Modern Science* (Chicago, IL: Chicago University Press).

Bush, V. (1945). Science-The Endless Frontier, A Report to the President on Program for Postwar Scientific Research. Reprinted 1960 (Washington, DC: National Science Foundation).

Chelimsky, E. (2007). "Factors Influencing the Choice of Methods in Federal Evaluation Practice," in Julnes and Roeg, op. cit., 13-31.

Coalition for Evidence-Based Policy (2002). *Bringing Evidence-Driven Progress to Education* (Washington, DC).

Cohen, L., P. Kamienski, and R. Espino (1998). "Gate System Focuses Industrial Basic Research," *Measuring and Improving the Performance and Return on R&D* (Washington, DC: Industrial Research Institute), 221-224.

Cole, S. (1992). *Making Science* (Cambridge, MA: Harvard University Press).

Cook, T. (2002). "Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for Not Doing Them," *Educational Evaluation and Policy Analysis*, 24: 175-199.

COSMOS Corporation (2006). *The Math and Science Partnership Program Evaluation* (MSP-PE), Second Annual Report.

Cousins, B. and L Shulha (2006). "A Comparative Analysis of Evaluation Utilization and Its Cognate Fields of Inquiry: Current Issues and Trends," in *SAGE Handbook of Evaluation*, edited by I. Shaw, J. Greene, and M. Mark, 266-291.

Cronbach, L. (1982). *Designing Evaluations of Educational and Social Programs* (San Francisco, CA: Jossey-Bass).

Datta, L. (2007). "Looking at the Evidence: What Variations in Practice Might Indicate," in Julnes and Rog, op. cit., 35-54.

Department of Health and Human Services, Substance Abuse and Mental Health Services Administration (1999). *Understanding Substance Abuse Prevention-Toward the 21st Century*.

Downs, G. and L. Mohr (1976). "Conceptual Issues in Innovation," *Administrative Science Quarterly*, 21(4): 700-714.

Edel, M. (1980). "'People' versus 'Places' in Urban Impact Analysis," Glickman, op. cit., 175-191.

Edelman, M. (1985). *The Symbolic Use of Politics* (University of Illinois Press).

Farmelo, G. (2002). "A Revolution with No Revolutionaries," *It Must be Beautiful* (London: Granta Books), 1-27.

Feller, I. (2007). "Mapping the Frontiers of Evaluation Research," Science and Public Policy (forthcoming).

(2007b). "Interdisciplinarity: Paths Taken and Not Taken", *Change*, November/December, 46-51.

(1982). "Innovation Processes: A Comparison in Public Schools and Other Public Sector Organizations," *Knowledge,* 4: 271-291.

(1980). "Public Sector Innovation as 'Conspicuous Production,'" *Policy Analysis,* 6: 1–20.

(1993). "What Agricultural Extension Has to Offer as a Model for Manufacturing Modernization," *Journal of Policy Analysis and Management,* 12(3): 574–581.

and. D. Menzel (1977). "Diffusion Milieus as a Focus of Research on Innovation in the Public Sector," *Policy Sciences,* 8(March): 49–68.

Flay, B. and L. Collins (2005). "Historical Review of School-Based Randomized Trials for Evaluating Problem Behavior Prevention Programs," Annals, op. cit., 115-146.

Gandhi, A., E. Murphy-Graham, A. Petrosino, S. Chrimer, and C. Weiss (2007). "The Devil is in the Details: Examining the Evidence for 'Proven' School-Based Drug Abuse Prevention Programs," *Evaluation Review*, 31: 43-74.

Gardner, H. (2004). "How Education Changes: Considerations of History, Science and Values," in M. Suarez-Orozco and D. Qin-Hilliard (Eds.), *Globalization: Culture and Education in the New Millennium* (Berkeley: University of California Press), 235-256.

Glickman, N. (1980). (ed.) *The Urban Impacts of Federal Policies* (Baltimore, MD: Johns Hopkins Press).

Griliches, Z. (1995). "R&D and Productivity: Econometric Results and Measurement Issues," in *Handbook of the Economics of Innovation and Technological Change*, edited by Paul Stoneman (Oxford, UK: Blackwell), 52-89.

Hallfors, D. and D. Godette (2002). "Will the 'Principles of Effectiveness' Improve Prevention Practices? Early Findings from a Diffusion Study," *Health Education Research*, 17: 461-470.

Hallfors, D., M. Pankratz, and S. Hartman (2006). "Does Federal Policy Support the Use of Scientific Evidence in School-Based Prevention Programs?" *Prevention Science*.

Hamermesh, D. (2007). "Replication in Economics," National Bureau of Economic Research Working Paper 13026 (Cambridge, MA: National Bureau of Economic Research).

Henry, G. (2003). "Influential Evaluations," *American Journal of Evaluation*, 24: 515-524.

and Mel Mark (2003). "Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions," *American Journal of Evaluation*, 24: 293-314.

Howell, E. and A. Yemane (2006). "An Assessment of Evaluation Designs: Case Studies of 12 Large Federal Evaluations," *American Journal of Evaluation*, 27: 219-26.

Hunter, J. and F. Schmidt (1996). "Cumulative Research Knowledge and Social Policy Formulation: The Critical Role of Meta-Analysis," *Psychology, Public Policy and Law*, 2: 324-347.

Julnes, G. and D. Rog (2007). (eds.), *Informing Federal Policies on Evaluation Methodology: Building the Evidence Base for Method Choice in Government Sponsored Evaluation* (San Francisco, CA: Jossey-Bass).

Karshenas, M. and P. Stoneman (1995). "Technological Diffusion," *Handbook of Economics of Innovation and Technological Innovation*, op. cit., 265-297.

Katzenmeyer, C. and F. Lawrenz (2006). "National Science Foundation Perspectives on the Nature of STEM Program Evaluation," in *Critical Issues in STEM Evaluation*, edited by D. Huffman and F. Lawrenz, (Minneapolis, MN: University of Minnesota Press), 7-18.

Kettl, D. (1997). "The Global Revolution in Public Management: Driving Themes, Missing Links," *Journal of Policy Analysis and Management*, 16: 446-462.

Kingdon, J. (1995). *Agendas, Alternatives, and Public Policies*, Second Edition (New York: Harper Collins).

Lawrenz, F. and D. Huffman (2006). "Methodological Pluralism: The Gold Standard of STEM Evaluation," in Huffman and Lawrenz, op. cit., 19-34.

Levitt, S. and S. Dubner (2005). *Freakonomics* (New York: Harper Collins).

Leibenstein, H. (1966). "Allocative Efficiency vs. "X-Efficiency," *American Economic Review*.

Lipsey, M. (1990). *Design Sensitivity* (Newbury Parks: SAGE).

Lindblom, D. and D. Cohen (1979). *Usable Knowledge* (New Haven, CT: Yale University Press).

Machlup, F. (1962). *The Production and Distribution of Knowledge in the United States* (Princeton, NJ: Princeton University Press).

Mandel, L., S. Bialous, and S. Glantz (2006). "Avoiding 'Truth': Tobacco Industry Promotion of Life Skills Training," *Journal of Adolescent Health*, 39: 868-879.

Manna, P. (2007). *School's In* (Washington, DC: Georgetown University Press).

Mark, M., G. Henry, and G. Julnes, G. (2000). *Evaluation: An Integrated Framework for Understanding, Guiding and Improving Public and Nonprofit Policies and Programs* (San Francisco, CA: Jossey-Bass).

McCloskey, D. and S. Ziliak (1996). "The Standard Error of Regression," *Journal of Economic Literature*, 34: 97-114.

Mitre Corporation (1979). *Institutionalization of Federal Programs at the Local Level*, edited by E. Chelimsky, M78-80 (McLean, VA).

Mosteller, F. and Weinstein, M. (1985). "Toward Evaluating the Cost-Effectiveness of Medical and Social Experiments," in *Social Experimentation*, edited by J.

Hausman and D. Wise (Chicago, IL: National Bureau of Economic Research), 221-246.

Murnane, R. and R. Nelson (2007). "Improving the Performance of the Education Sector: The Valuable, Challenging, and Limited Role of Random Assignment Evaluations," Economics of Innovation and New Technology, 16:307-322.

Myers, M. and L. Ries (1989). "Cancer Patient Survival Rates: SEER Program Results for 10 Years of Follow-up," *CA: A Cancer Journal for Clinicians*, 39: 21-32.

National Academy of Science (1983). Frontiers in Science and Technology (Washington, DC), Report to the National Science Foundation under contract No. PRM-8206308.

National Reading Panel (2000). *Teaching Children to Read.*

National Research Council (2005). *Improving Evaluation of Anticrime Programs* (Washington, DC: National Academies Press).

(2004). *Advancing Scientific Research in Education* (Washington, DC: National Academies Press).

National Research Council, Transportation Research Board (2001). *The Federal Role in Highway Research and Technology*, Special Report 261 (Washington, DC: National Academy Press).

National Science Foundation (2006). *Report of the Advisory Committee for GPRA Performance Assessment* (Arlington, VA: National Science Foundation).

Newcomer, K (1997). "Using Performance Measurement to Improve Programs," in *Using Performance Measurement to Improve Public and Nonprofit Programs*, edited by K. Newcomer, New Directions for Evaluation, (San Francisco, CA: Jossey-Bass Publishers) 75.

Nelson, R. (1977). *The Moon and the Ghetto* (New York: W.W. Norton & Company).

OECD (2005). *Modernizing Government* (Paris: OECD).

Paley, A. (2007). ""No Child" Target is Called Out of Reach," www.washingtonpost.com, March 13, 2007.

Perrin, B. (1998). "Effective Use and Misuse of Performance Measurement," *American Journal of Evaluation*, 19: 367-379.

Power, T. (1997). *The Audit Society* (New York: Oxford University Press).

Pressman, J. and A. Wildavsky (1973). *Implementation: How Great Expectations in Washington are Dashed in Oakland; Or, Why It's Amazing that Federal Programs Work at All*.

Radin, R. (2006). *Challenging the Performance Movement* (Washington, DC: Georgetown University Press).

Rosenberg, N. (1976). "On Technological Expectations," *Economic Journal*, 86: 523-535.

Rogers, E. (1995). *Diffusion of Innovations*, Fourth Edition (New York: Free Press).

Rule, J. (1997). *Theory and Progress in Social Science* (Cambridge, UK: Cambridge University Press).

Schweigert, F. (2006). "The Meaning of Effectiveness in Assessing Community Initiatives," *American Journal of Evaluation*, 27: 416-436.

Schemo, D. (2007). "In War Over Teaching, A U.S.-Local Clash," *New York Times*, March 9, 2007, p. 1ff.

Shulman, L. (2007). "Counting and Recounting: Assessment and the Quest for Accountability," *Change Magazine*, 39.

Smelser, N. (2005). "The Questionable Logic of 'Mistakes' in the Dynamics of Growth in the Social Sciences," *Social Research*, 72: 237-262.

St. Pierre, T., D. Osgood, C. Mincemoyer, D.K. Kaltreider, and T. Kauh (2005). "Results on an Independent Evaluation of Project Alert Delivered in Schools by Cooperative Extension," *Prevention Science*, 6: 305-317.

Stevens, G. and J. Burley (1997). "3000 Raw Ideas = 1 Commercial Success," Research Management Review, May-June, *Measuring and Improving the Performance and Return on R&D*, op. cit., (Washington, DC: Industrial Research Institute), 4-15.

Tornatzky, L. and M. Fleischer (1990). *The Processes of Technological Innovation* (Lexington, MA: D.C. Heath & Company).

Thomas and Strickland (1998). Crafting and Implementing Strategy, Tenth Edition (Boston, MA: Irwin-McGraw Hill).

Warsh, D. (2006). *Knowledge and the Wealth of Nations* (New York: W.W. Norton).

Weick, K. (1976). "Educational Organizations as Loosely Coupled Systems," *Administrative Science Quarterly*, 21(1), 1-19.

Weingart, P. (2005). "Impact of bibliometrics upon the science system: Inadvertent consequences?" *Scientometrics*, 62(1), 117-131.

Welch, H., G. L. Schwartz, and S. Woloshin (2000). "Are Increasing 5-Year Survival Rates Evidence of Success Against Cancer?" *JAMA.* 283: 2975-2978.

Wildavsky, A. (1979). "Strategic Retreat from Objectives: Learning from Failure in American Public Policy," *Speaking Truth to Power* (Boston: Little Brown).

Wolf, C (1990). *Markets or Governments* (Cambridge, MA: MIT Press).

Wright, D. (1988). *Understanding Intergovernmental Relations*, Third Edition (New York: Harcourt).

Yin, R. K. (1977). "Production Efficiency versus Bureaucratic Self-Interest: Two Innovative Processes," *Policy Sciences*, 8:381-399.

(1981). "Life Histories of Innovations: How New Practices Become Routinized," *Administrative Science Quarterly*, 41: 21-28.

(forthcoming). "The Math and Science Partnership Program Evaluation (MSP-PE). Overview of the First Two Years," *Peabody Journal of Education*.

Ziliak, S. and D. McCloskey (2004). "Size Matters: The Standard Error of Regressions in the American Economic Review," *Journal of Socio-Economics*, 33: 527-546.

FOOTNOTES

[1] In economic terms, the objective of enactments and administrative procedures such as GPRA and PART is to move agencies/programs from presumed positions of X-efficiency or satisfying behavior to those of allocative efficiency and maximizing behavior (Leibenstein, 1966).

[2] The famous flagship message of Admiral Horatio Nelson at the Battle of Trafalgar that "England expects that every man will do his duty," highlights how the two meanings of high likelihood and requirement can segue into one another. Nelson's initial instruction to his signal officer was to send the message, "England confides (is confident) that every man will do his duty." The signal officer suggested that "expects" be substituted for "confides" since the former word was in the signal book, whereas the latter would have to be spelt out letter-by-letter (http:en.wikipeidia.org/wiki/England).

[3] For example, the report notes that "…..a common approach has been for an MSP to establish a target or benchmark for the expected change, and then to determine whether such a benchmark has been met. As one example, an MSP had 40 participating districts and had set an initial expectation that, by the end of the MSP's five-year period, 90 percent of these districts would exceed two benchmarks: that at least 75 percent of each district's students would have scored "advanced" or "proficient" on the state assessment in mathematics, and that 10 percent or fewer of the students would have scored "below basic." The MSP interpreted the trends for its first three years as suggesting that the districts were making progress toward these benchmarks in the 5th and 8th grades, but not in the 11th grade" (Yin, forthcoming, p.19).

[4] "There is zero percent chance that we will ever reach a 100 percent target. But because the title of the law is so rhetorically brilliant, politicians are afraid to change this completely unrealistic standard. They don't want to be accused of leaving some children behind," Robert Linn, co-director, National Center for Research on Evaluation, Standards and Testing, UCLA (as quoted in Paley, 2007, p. 1).

[5] In races, meets, or matches, success is synonymous with coming in first, or in the money, with "odds" serving as a priori probabilities of success. Public sector organizations generally do not function in competitive environments, but rather are described as having a monopoly on the provision of services, coupled with an uncertain technology (Wolf, 1990). In effect, the privatization/voucher movement represents efforts to introduce the prod of competition, with related metrics of success, into K-12 education.

[6] NSF, for example, organizes its GRPA performance assessment report around a matrix in which the rows represent the Foundation's strategic goals (and component sub goals) of Ideas, Tools, People, and Organizational Excellence, and the columns represent "Significant Achievement," "Quality," and "Relevance" (NSF, 2006). In terms of NSF's 2006 GRPA report, STEM programs fall under sub goal P3 (Develop the Nation's capability to provide K-12 and higher education faculty with opportunities for continuous learning and career development in science, technology, engineering, and mathematics.

[7] The Federal Highway Administration is only one source of funding for highway R&D. It also is described as confronting many barriers to innovation: "Highway innovation is difficult because the highway industry is so decentralized, its procurement practices at times provide little incentive to innovate, and there is considerable aversion to risk in the public sector. *Achieving widespread implementation of innovations often requires a great deal of proactive technology transfer"* (National Research Council, Transportation Research Board, 2001, p. 5; emphasis added).

[8] An important but tangential topic for this essay is the "tactical" political conservatism of calls for evidence-based decision-making and randomized trials as a programmatic threshold requirement. The cost and complexity of conducting such trials and the challenges of isolating effect sizes may tend to or reinforce prior positions that government intervention is ineffective. Compare, however, the contrary view of Flay and Collins (2005): "In general, the ethical arguments for conducting randomized studies are stronger than the arguments against them. In many cases, it would be more unethical to provide intervention or education that has unknown or possibly negative effects than it would be to assign people or schools randomly to find out which programs are most effective" (p. 123).

[9] As I have paraphrased Tolstoy on another occasion, in reference to NIST's Advanced Technology Program, a favorable evaluation may not save a program; an unfavorable one may not kill it.

[10] "Policy-making and analysis have often confused individual or group 'people' impacts with 'place impacts…'(S)ome tendency exists to confuse poor people with poor areas, with the result that regional development and antipoverty objectives have often been substituted for each other….At times this confusion has been benign. But these objectives cannot be considered perfect substitutes" (Edel, 1980, p.176).

[11] Rogers, 1994, cites Mort et al.'s 1953 conclusion to the effect that "The average American school lags twenty-five years behind the best practice" (p. 64), and then notes that "There is a wide range in the adoption of educational innovations."