Evaluation in the Context of the Government Market Place: Implications for the Evaluation of Research

Connie Kubo Della-Piana and Gabriel M. Della-Piana
National Science Foundation and independent consultant

The evaluation community has concentrated on examining and explicating implications of the choice of methods for evaluating federal programs, as described in the *New Directions for Evaluation* edited by Julnes and Rog (2007), placing the policy debate in historical and contemporary contexts. In that volume and elsewhere we find that there are several mechanisms described for supporting and/or conducting program evaluation at the federal level. In the Julnes and Rog volume, Chelimsky (2007) describes evaluation activities conducted within the federal government by the Government Accountability Office (GAO). Both grants and contracts supported the work of Yin (Yin and Davis, 2007) in the evaluation of large comprehensive reforms in K-12 science and mathematics education. Other evaluation activities come under the authority of the Office of the Inspector General which conducts performance audits of government programs that draw on program evaluation and its methods (see the *Yellow Book* http://www.gao.gov/govaud/ybk01.htm).

While the current debate focuses heavily on method, we address the challenges faced by the government in the selection of funding mechanisms for supporting program evaluation efforts. The choice of funding mechanism structures the context in which an evaluation is designed and carried out, and in that sense can and does influence the level of specificity involved in describing the requirements of the evaluation. We argue that the type of funding mechanism can limit or enhance the opportunities for the development of theories of and methods for the evaluation of research. We have chosen to explore the use of contracting program evaluations as a way to explicate implications for the evaluation of research.

This paper builds on the work of Biderman and Sharp (1972) and House (1997) on the contracting of social science and program evaluation specifically, and of Kettl (1993) for an overarching perspective on outsourcing of government services. Biderman and Sharp (1972) examine the procurement of evaluation research as social science shifted away from the social scientific research of the academy towards a more intimate connection with the practical concerns of society. House (1997) examines the acquisition of program evaluation in one agency within the context of what he calls an "imperfect market." Kettl (1993) argues that contracting changes the relationship between the public and the private thereby changing the nature of activities, such as evaluation. The goal of the relationship becomes one that is responsive to the requirements of the contract rather than to the citizenry.

In their article, Biderman and Sharp (1972) conclude that the rules and practices that govern government procurement constrain the nature of the relationship between the agency and the contractor and any actions that may be taken before, during, and after the procurement of evaluation services. That is, the relationship

and practice are defined by the contract and the acquisition system. For a detailed description of the acquisition process, see Biderman and Sharp (1972).

House (1997) analyzes the evaluation of federally funded research programs in science and engineering and concludes that evaluation in this context takes place within the context of an "imperfect market." In House's study of evaluation contracting at the National Science Foundation (House, 1997), he finds evidence of an imperfect market on both the supply side (few contractors getting most of contracts with yet a large number of professional evaluators in the field) and the demand side (few buyers, government agencies). House explains that this is so in part due to limited staff and the need for efficiency.

The problem we address is how to develop a set of requirements for an evaluation that takes into account the complexities of contracting evaluations in an imperfect market under the contracting mechanism of "the acquisition of supplies or services for the direct benefit or use of the Federal Government" (Federal Acquisition Regulation, http://www.gsa.gov/far). The means for defining requirements in this context is the statement of work (SOW) which we turn to next. That will be followed by the tensions and obstacles in contracting in the context of a government market, a perspective on normative discourse (the language we use to express evaluations and prescriptions as to what one ought to do or think), the range of inquiry purposes along with strategies and difficulties in matching them to contractor capacity and the requirements of the SOW, the tensions and obstacles in the special case of research and development (R&D) in interdisciplinary contexts, and finally a perspective on management of the complex, ambiguous and unexpected as ways of coping with the demands of contracting outlined the paper.

What is a Statement of Work (SOW) Within a Contractual Request for Quotes (RFQ) for Providing Evaluation Services?

There are three phases of the acquisition process: Acquisition Planning (Pre-Award), Contract Formation (Solicitation and Award), and Contract Administration (Post-Award). The Federal Acquisition Regulation (FAR) regulates the acquisition of products and services, of which evaluation is one. The Guiding Principles of the FAR are: (1) to satisfy the customer in terms of cost, quality and timeliness; (2) to minimize administrative operating costs; (3) to conduct business with integrity, fairness, and openness; and (4) to fulfill public policy objectives" (FAR 1.1.02). The principal customers of the products and services are the users and line managers who act on behalf of the taxpayers (FAR 1.102). In the case of evaluation, the principal customers of a program evaluation are program managers. While there is extensive description of the content and form of a Statement of Work for research and development (FAR 35.005), there is limited guidance on a work statement for products and services requiring a Statement of Work (FAR 8.402a), and evaluation contracting comes under products and services.

The Statement of Work (SOW) is the "heart" of the contractual request (solicitation) for quotes (a proposal) and the contract itself. The SOW describes the relationship between the buyer (agency) and the seller (contractor) and explicates the technical requirements of a specified task or set of tasks that is needed by the government. In addition, it serves as the framework for the effort by which the work is to be managed and monitored by the agency. The quote or proposal and modifications requested by the agency become the agreement to which the contractor and the agency are held. In this context, author(s) of an SOW are cautioned not to rely on contractors' commercial descriptions to develop the

requirements, but should tailor "the commercial services performed by the contractor to meet a particular Government need" when using government supply sources (see MOBIS, Mission Oriented Integrated Services, p. 12, available only online at http://www.gsa.gov). The general form and content of the SOW for products and services are specified by the FAR (FAR 8.405-2a), MOBIS provides a description of the form and content of the Statement of Work.

The SOW should include background information about the effort to acquaint the reader with the acquisition situation; description of the scope of the effort; goals and objectives of the work as well as a description of how the results and products will be used; work requirements or tasks that must be completed; description of reporting requirements and what must be delivered; statements describing government-furnished property, security requirements, place of performance; and period of performance. Moreover, the description of the task requirements are dependent on the approach selected to describe the required effort: performance-based (requirements are described in terms of results), level of effort (requirements described in terms of tasks to be performed and hours devoted to each task) or a detailed SOW (requirements described in terms of how the work must be accomplished). The FAR encourages the authors of the SOW to provide a clear and complete work statement, because the "contractor's personnel [are] to perform the service without direct Government supervision." See MOBIS (Mission Oriented Integrated Services, p.12) available only online at http://www.gsa.gov. The General Services Administration (GSA) advises that, requirements should be clear; references should be kept to those sources applicable to the specifications and the standards needed; specifications and other documents should be tailored for the purpose of the effort; and general direction should be separated from information. The key criterion is clarity. There is an assumption that the service rendered can be clearly specified. Yet, clearly describing specifications for the evaluation of research is problematic when the effort and the object of assessment are characterized by complexity, uncertainty, and when the evaluation of research is in the early stages of theory and methodological development.

Is there a clear government need to develop theory and methods for evaluating research? What are the issues that need to be addressed by those developing SOWs and those responding to government requests for evaluation of research? And how will clarification of the issues add value to the efforts and products of requests for services to evaluate research?

The Tensions and Obstacles in Contracting for Evaluation in the Context of a Government Market

The imbalance in expert knowledge between a government agency and an evaluation contractor creates challenges for the government to be a "smart buyer" of evaluation services. As problems, methodology, and substantive areas of work get more complex, it is even more likely that often contractors will have more expertise than government. Furthermore, Kettl (1973) notes that, "Agents…always know more about their qualifications than principals can discover…the government can never be sure that it has hired the best possible contractor" (p. 27).

The normative issue enters into the contracting of evaluation with respect to the balance between the contractor and government expert roles. Government being a servant and representative of the citizenry has an overriding concern for societal impact of evaluation over methodology. The evaluation contractor's overriding concern is typically developing a defensible methodology for completing the evaluation effort. As Kettl (1973) puts it, "…in the search for this balance [between public and private power], seeking the public interest is

paramount. The government is…not just another principal dealing with just another agent…It is a representative of the public and its goals must represent public goals as embodied in law. Pursuing those goals—and the sense of the public interest that lies behind them—is the central task of government" (p. 40). Where ideas of government and contractor conflict, procedural rules are being suggested by some evaluation theorists. For a brief sketch of procedural approaches see Mark (2001, pp. 457-459). That such approaches to dealing with conflict of ideas are difficult even with procedural rules, such as in House & Howe (1999) and in the application of the House and Howe "deliberative democratic evaluation" by Howe & Ashcraft (2005), leaves procedural approaches wide open for research and analysis.

House (1997) notes that Kettl (1993) calls for three actions government can take to be a smart buyer of evaluation services in an imperfect market. We briefly note the three actions and some of the complexity entailed in each. One action is for government to define its goals separately from contractors so that the government knows what it wants to purchase. This is presumably a counter-strategy to reduce the influence of the contractor that grows out of the closeness and thus interdependence of having a few firms dominating the market, whether the domination is for efficiency or otherwise. Another action is the government must come to know which contractors have the capacity to do the job. House suggests that given few contractors one might call upon small contractors (often from universities) to subcontract with large firms. This brings in other complexities associated with the different major missions of the university (to accumulate knowledge) and the government (to acquire services in the public interest). A third action is that the government must be able to judge what it has bought either by judging the proposal, judging the progress of the evaluation as it proceeds, and/or judging the product. Peer review and project monitoring are the typical practices in place for such judgments. A key tension that cuts across all three of Kettl's actions is that the complexity of the knowledge in both disciplinary and interdisciplinary contexts, and the difficulty in specifying precisely what the government wants (since in a complex situation both method and goals evolve and the government may lack expertise in a given arena) plays havoc with implementation of the actions.

In a later report, Kettl (2005) imagines how the political landscape is being transformed and what the next government of the United States might look like. He envisions "imperatives for a new and more effective strategy of government." One example gives the flavor of the issues. Focusing on problems more than structures means that government service may look like a web or network more than a hierarchy. Thus, combining evaluation services across multiple federal divisions, directorates, or even agencies, contracting for evaluation services focuses on the problem (good evaluation services) rather than organizational structure or territory. In such a context, accountability must deal with agency roles in a different way. New mechanisms are needed in such a case for coordinating work among organizations so that they are "interoperable" or so the parts (people, systems, tools) work together more or less seamlessly to solve problems with "different patterns of coordination for different problems." All of this requires information, communication, and performance measures which "transform how the players think and talk about government programs." For legal issues relevant to a major transformation of government to the extent envisioned by Kettl, see reports from the Carnegie Commission on Science, Technology, and Government (http://www.carnegie.org/sub/pubs/ccstfrep.htm).

Up to this point we have illustrated the tensions between the requirements of a Work Statement and the practicalities or realities of

the market. We now turn to the demands of normative discourse.

A Perspective on Normative Discourse Relevant to Contracting

Normative discourse is the language we use to express evaluations, prescribe what one ought to do or think, and give reasons for or against evaluations and prescriptions (Taylor, 1961). As such, understanding the constraints entailed in normative discourse is critical to understanding the work of foundations or other agencies which make evaluative judgments for funding and prescribe, however flexibly, what is appropriate methodology, design, or goals for conducting evaluation of research in a formal statement of work. The perspective outlined here is based largely on Taylor (1961) followed by a statement as to the centrality of the normative in evaluation (Scriven, 1991, 2003, 2004) and some relevant observations by Reddy (2005) and Kelly (2006). The focus is on the normative demands on writing statements of work for letting contracts to conduct evaluations.

Taylor (1961, p.206), specifies conditions under which an "ought" sentence (you ought to do or think or contemplate "X") under "C" conditions (the context, conditions, exceptions) may be taken as a prescription. A prescription is not a command to obey but rather guidance implying choice. Also since it is rational in form, it requires justification with relevant reasons and valid inference that it is the best thing to do in a given situation. Taylor contends that the four conditions under which an "ought" sentence may be taken as a prescription are: (1) The sentence is in earnest (the speaker wants the audience to accept what is said and act accordingly) and affirmed (the speaker has a pro-attitude toward doing it and is not lying or concealing true thoughts; (2) The audience is an agent in a situation or will be in a future one in which doing the act is an alternative of choice, that it is a possibility that will be open to the agent; and (3) The audience is an agent with freedom to choose to do or not to do the act. The presumption is that the agent has the physical ability and intellectual and emotional capacity to do it, is not under external compulsion to do one thing rather than another, and thus will not do X unless she/he chooses to so and will only choose to do so as a result of his/her own deliberations or freely follows another's deliberation). The audience is an agent for whom it is legitimate and proper to demand reasons for doing the prescribed act. This condition follows from the nature of a prescription as the giver of advice, guiding, or making recommendations that imply what it would be rational for a person to do. In the case of evaluation of research, the form of the prescriptive statement (stated or implied) is: You ought to conduct the evaluation in this way for the reason that it will produce valid and useful information that will contribute to social betterment (as defined in the particular project). There are heavy demands for justification and implementation of such prescriptive acts. Reddy (2005) argues that a consequence of normative discourse being prescriptive and thus entailing obligations, as Taylor (1961) outlines, is that successful normative reasoning depends on knowing the context or constraints in a situation so that one does not come to unjustified conclusions about the agents' obligations. In other words it is important to know what is changeable and by whom. Reddy addresses the argument that a constraint may simply be a feature of the situation that is difficult or costly to change by invoking a definition of constraint as "…a feature of the world that can reasonably be judged to have the property that the agent cannot change it without substantial cost or difficulty, if at all" (p. 121). This entails the notion that constraints may be indeed changeable with some difficulty or cost. For thinking about evaluation contracting processes, such constraints play a role in shaping the nature and distribution of obligations. A key question raised by Reddy's analysis is: What are

the respective roles of institutions engaged in a particular contractual process and arrangement (e.g., the contractor, the federal agency, the institutional agencies in the setting of the work, and the professional associations relevant to the expertise or practice)?

Finally, the work of Scriven (1991, 2003, 2004) perhaps more than any other evaluation theorist has made values central to evaluation. His logic of evaluation outlined in various sources is summarized in one sentence in the *Evaluation Thesaurus* (Scriven, 1991, p. 216) as, "The key function of evaluative inference is moving validly to evaluative conclusions from factual (and of course definitional) premises; so the key task of the logic of evaluation is to show how this can be justified." Thus it is justified evaluative conclusions that are central. Scriven also makes clear the centrality of social betterment to evaluation goals. "In my view, one of the most important questions professional evaluators should regularly consider is the extent to which evaluation has made a contribution to the welfare of humankind and, more generally, to the welfare of the planet we inhabit" (Scriven, 2004, p. 183). And elsewhere, Scriven hopes for what he calls the "evaluative social sciences" that "…not only includes the descriptive study of values and those who hold them…but as the home range of *normative* evaluative inquiry, meaning inquiry whose conclusions are directly evaluative, directly about good and bad solutions to social problems, directly about right and wrong approaches, directly about better and worse problems" (Scriven, 2003, p. 21). The normative is thus clearly entailed in both the conduct of evaluation and the writing of statements of work (SOWs) for funding of evaluation of research in science and engineering (S & E). Decisions on evaluation questions and methodology (prescriptions) have the danger of "eliminating knowledge through methodological constraints" (Kelly, 2006, pp. 50-51). In a recent volume on informing federal policy on evaluation methodology (Julnes & Rog, 2007) one key consensus was that method choice should match the needs of specific situations or studies—a task not always or easily accomplished. Both the SOW and the choice of evaluation methodology are normative in intent. They directly or indirectly lead to inferences as to what evaluators ought to think or do. We now turn to the problematics of matching contractor capacity with the methodological requirements embedded in a statement of work.

The Range of Inquiry Purposes and Strategies and Matching Contractor Capacity to the Requirements of a SOW

The program goal oriented Request for Quotes (RFQs) and embedded Statements of Work (SOWs) in the federal sector are designed to get the best product or service for the benefit of society at reasonable cost from the sellers (evaluation contractors) available. One resource to support this effort is the capability of an evaluation contractor to draw on a wide range of evaluation inquiry strategies so as to be better able to match method to significant questions for evaluating individual and interdisciplinary science and technology research. The call for a broad range of inquiry is echoed from governments around the world as indicated in the Perrin (2006) report of a roundtable on moving from outputs to outcomes sponsored by the World Bank and the IBM Center for the Business of Government. Here we illustrate two ways of conceptualizing a range of inquiry purposes and associated strategies. This is followed by drawing implications for the Statement of Work in the context of the government market place and the evaluation of interdisciplinary research. Implications are drawn in part from the perspectives of normative discourse.

Inquiry Purposes from the Natural Sciences

Phillips (2006) outlines ten inquiry purposes characteristic of the natural sciences. The intent

here is to highlight the range of such purposes that might be drawn upon to match contractor capability to the purposes of the evaluation of research. A selected subset of Phillips' listing is noted here:

1. Determining whether an intervention or treatment produces an effect or effects (either intended or unintended)
2. Explaining, or determining the cause of some familiar condition or phenomenon
3. Determining whether a purported effect is a genuine one
4. Determining whether some predicted process or phenomenon actually occurs
5. Noticing, and then describing and investigating, an unexpected phenomenon
6. Testing a widely held explanation for some phenomenon or regularity
7. Determining the structure, architecture, or anatomy of some entity or feature
8. Developing some discovery into a usable product or process

To appreciate this range of purposes the reader is referred to Phillips (2006) for detailed description, context, and examples.

Inquiry Strategies for Evaluating R & D Programs

A directory and overview of evaluation methods for R & D programs focused on technology development has been provided by Ruegg & Jordan (2007) for the U.S. Department of Energy as a resource for program managers and Federal agencies. The directory includes a logic model of seven sequential steps in R & D in phases from design to diffusion and ultimate outcomes. Methods are listed with descriptions, limitations, and examples of use. Here is a brief listing of the methods:

1. Peer review/expert judgment
2. Monitoring
3. Data compilation and use of indicators
4. Bibliometric methods, including counts and citation analyses, data mining, and hotspot patent analyses
5. Network analyses
6. Case study methods
7. Survey method
8. Benchmarking method
9. Technology commercialization tracking method
10. Benefit-cost case study
11. Econometric methods
12. Historical tracing method
13. Spillover analysis

The Range of Inquiry Strategies, the Statement of Work, and the Demands of Normative Discourse

The Statement of Work (SOW) embedded in the Request for Quotes for purchases of services to conduct an evaluation project, includes goals and objectives of the evaluation and may provide guidance suggesting the range of capabilities needed for the tasks. In effect the SOW takes on the form of a prescription for producing, in the given context, valid and useful information that contributes to the goals of the program to be evaluated and the public good. Being rational in form, as noted above, such statements are subject to justification. From the point of view of "normative discourse" what is to be justified by the agency is the conditions under which an "ought" sentence (which is not a command but guidance as to what one ought to do or think) may be taken as a prescription. In the current context, the ought sentence is that the evaluation contractor ought to call on a wide range of inquiry purposes and strategies to find the best alternatives open for this time, place, manner, and circumstances. However, from the perspective of the federal agency issuing the prescription, there are demands on justification for the four conditions noted earlier.

First, is that the prescription be in earnest. This would entail expecting the contractor to

conduct considerable front-end analysis and reflection throughout the study to draw on the range of inquiry purposes and strategies for an appropriate match to the requirements of the project—a heavy time and cost burden. Second, is that drawing on a broad range of inquiry purposes and strategies entails a government agency judgment that the contractor is in a situation (or will be in one) where the prescribed act is an open possibility. This may be difficult for the federal agency to assume, for example, in the common cases where data is known to not be generally available or the context is not likely to support accessing the data. Such support may not be forthcoming due to time and cost demands on the persons and institutions being studied and the sometimes expected negative consequences of the transparency of the findings. Third, is the entailment that the contractor (or the person within the contracting agency expected to conduct the study) is indeed free to choose to do the act (presupposing intellectual and emotional capacity, craft skill and substantive and contextual know-how). Finding out whether a contractor has such capacity is often done through examining previous work and resumes, though clearly this cannot easily do the job for all capacities. It also entails finding the "best" contractor and in the current imperfect market for such work, one of the obstacles is the limited number of contractors to draw upon. Fourth is that the contractor is an agent for whom it is legitimate for the government agency to imply what it would be rational for a contractor (or persons) to do. This runs counter to the notion of expertise. Contractors know their expertise better than the federal agency knows or can find out. Also, when the evaluation contractor is an academic researcher, policies of the academic institution support academic freedom which may conflict with goals and objectives of a SOW. The above demands, from the perspective of normative discourse, are difficult though not impossible to address, but become even more difficult and more complex when one encounters the case of interdisciplinary research where this paper now turns.

Tensions and Obstacles Illuminated in Attempts to Assess R&D in Interdisciplinary Contexts

Evaluation of Science and Engineering increasingly entails assessment of interdisciplinary work, even within a discipline. This is a technically and normatively complex domain for evaluation. Mark (2003) contends that the government as "gatekeeper" in this context has an opportunity to contribute to the shaping of the discourse around which evaluation approaches are most appropriate for this complex domain. Here we present a brief sketch of some of the issues and obstacles drawn from empirical studies aimed at assessing the quality of interdisciplinary research in science and engineering.

Impact measures are complicated by the time delay of impact, spread of effects in multiple directions, and diverse citation practices across disciplines (Boix-Mansilla, 2006). Peers for reviews can not easily be identified since, by definition, the lack of peers is a consequence of interdisciplinary research (IR) being a new kind of synthesis of expertise. Even within disciplines, finding peers is problematic since the increase of specialization justifies thinking of interdisciplinarity applying to any combination of knowledge that goes beyond the specialization of a single researcher (Boix-Mansilla, 2006; Laudel, 2006).

What makes it difficult for assessment based on knowledge production is that epistemic principles for mathematics, physics, physiology, molecular biology, nanophysics and other scientific disciplines vary with their respective disciplinary aims (Boix-Mansilla, 2006).

The quality and extent of collaborative inquiry is in part a function of degree of organization in a research group and partly on the extent of cognitive coupling. There is no

one standard for collaboration, but rather the field should work toward an awareness of mechanisms, potentials and problems of each depending on organizational and epistemic conditions. Following that aim takes one into a multi-method descriptive and quantitative assessment to capture the richness of the collaboration (Lengwiler, 2006).

The attempt at mutual learning and deliberation (as for example when program staff and evaluators attempt to include or be accountable to extra-academic stakeholders outside the disciplinary community) reduces three types of complexity: the factual, in deciding about knowledge; the temporal, by reducing what is done and what stays on the agenda; and the social, by interaction of heterogeneous actors drawing on their own respective competencies (Maasen & Lieven, 2006). But for an example of diversity decreasing simplification see Weick and Sutcliffe (2001, pp. 59f) who argue for diversity as one way for an organization to enable seeing different things in the same event. They provide an example of how people working with nuclear power build in resistance to simplification as a way of coping with the unexpected in a complex technology that is important to control. And for how culture is interwoven or should be interwoven with research see Gordon (1999).

One heuristic for brainstorming issues in evaluation of science activities is to contemplate the intersection of three domains: issues (quality, effects, appropriateness, process improvement, and so on), system variables (policy, resources, structure/process, outputs, outcomes, impacts), and current evaluation tools (for example, case studies, surveys, cost-benefit and all the methods listed above). What can be learned with each approach and where is each weak? (Arnold & Balazs, 1998).

Peer review is for many still the method of choice for review of science and engineering proposals and publications. There is a large literature some of which is summarized in the NRC report on the subject (National Research Council, 2004). The classic objections are the lack of reliability and predictive accuracy. Some highlights from Langfeldt (2006) point up critical issues. For example, the rejection paper that 34 years later won a Nobel Prize; the finding that, if the independent follow up of panel peer reviews of proposals were the conclusive ratings, 24% to 30% of the proposals would get a reverse outcome; and the finding that reviewers favored fields they were familiar with and proposed that interdisciplinary projects should not be reviewed in the same way. There is evidence and argument that goes the other way, but the point is that there are continuing attempts to understand and refine peer review processes, to minimize limitations.

Bozeman, Dietz, & Gaughan (2001) propose a scientific and technical human capital model for capturing what "enables researchers to create and transform knowledge and ideas in ways that would not be possible without these resources". The model combines scientists' human capital (for example, cognitive and tacit knowledge, substantive knowledge of the craft, and know-how based on understanding context beyond the construction of research designs and the carrying out of implementation) and productive social capital (including networks and whatever combines to contribute to career trajectories and to enhancing or generating their own capabilities). Bozeman and others (2001) argue that the model, perhaps combined with conventional product oriented approaches, provides policy makers with demonstrable improvements in capacity while waiting for long term effects.

A Congressional Budget Office (2005) background paper on R & D productivity growth concluded that while it is quite likely that R & D positively impacts productivity, does so with a rate of return at least equal to other investments, and privately funded R & D benefits from basic science research carried out with government support, there are knowledge gaps that are worthy of attention. For example, available data makes it difficult to estimate the

size of the R & D influence with any precision. Also, spillovers, or benefits to firms, industries, and other nations than the one doing the research, are difficult to measure.

Cozzens (2002) reports on a workshop to review research assessment as part of a larger project to connect the United States research assessment community to world discussions of the craft. Two troublesome or at least challenging issues that arose include the time scale required for evaluation and the chilling effect of performance metrics on teamwork and creativity. On the time scale for example, one study reported 14% of bliomedical research knowledge eventually influences clinical practice and then on an average of 17 years after publication. Challenges included "political pressure, limitations on data, and unrealistic expectations from sponsors for simple indicators and answers." A key methodology suggested is tracing studies looking backward from innovation and forward from certain points in development of science.

Tracing studies are also called for in the assessment of interdisciplinary research (IR) by Laudel and Origgi (2006) in a special journal issue on assessment of IR noting "how patchy our knowledge of this subject is" and calling for "access to decision-making bodies which allow [the] study [of] interdisciplinary assessment procedures in vivo, that is by observation and interviews." Tracing studies have a long history. Some sources tapping that history are cited by Della-Piana and Della-Piana (2005), including Dunbar (1999), Feist & Gorman (1998), Klahr & Simon (1999), Shadish & Fuller (1994), and Thagard (1997). More directly relevant to the current context is the call for tracing studies in assessment of research by Ruegg & Jordan (2007), Cozzens (2002), and Bozeman, Dietz, and Gaughan (2001).

Chelimsky (2007) draws on years of experience in running PEMD (the Program Evaluation and Methodology Division in the U.S. Government Accounting Office) to inform government policy in assessment of research. Perhaps the key sentence summary of preparing for critiques of evaluations was, " [PEMD] developed all [of its] evaluation designs with an eye toward defending them and especially, our method of analysis choices later" (p. 28).

A Perspective on Managing the Unexpected

Managing the unexpected under conditions of complexity, ambiguity, and sometimes lack of control has become a common challenge in an information society. Bringing disciplinary and other stakeholders into various stages of evaluation, while engaged in the complexity of selecting appropriate evaluation designs for research that is often interdisciplinary, creates ambiguity and the unexpected. For guidance on key processes to consider under such conditions, we go to Weick and Sutcliffe (2001) who draw on the experience of high reliability organizations (HROs) such as aircraft carriers, nuclear power plants, and firefighting crews to specify five dimensions of mindfulness. We adapt those dimensions with slightly different language as a framework for considering how they might be relevant to constructing Statements of Work (SOWs) for evaluation of science and engineering (S & E) research.

A Counterforce to the Tendency to Focus on Success

Success breeds complacency, drift into automatic processing, and routinization or trivialization of adaptive technique. This is often the posture that "we followed the procedure," albeit in a check-off mode. The counterforce is to look for failure.

A Counterforce to Simplification

Consensus is often achieved by ignoring nuances and diverse views. The counterforce is to know the situation is complex and to work

on reconciling differences in ways that do not destroy the nuances that diverse views illuminate.

A Counterforce to Direct Reduction of Error

Error can be disabling and cause dropping out attempts to solve a problem. The counterforce is to focus on resilience or the detection, containment, and especially bouncing back from inevitable error in complex contexts loaded with the unexpected.

A Counterforce to "One Expert"

Having one expert based on hierarchical position or one dimension relevant to the problem can lead to missing key elements in complex environments. The counterforce is to defer to the relevant expertise, to cultivate diversity and push for input and decisions to migrate in directions of where the expertise is for a given context.

A Counterforce to a Working System

When the total system "works", small gaps might be missed. The counterforce is to have someone attentive to or sensitive to the total system to look for small failures, gaps, loopholes, or short-term underperformance that has potential for longer-term useful outcomes.

Conclusion

A perspective on managing the complex, unexpected or ambiguous is central to writing an appropriate SOW for evaluating science and technology research. As sketched above, the difficulties rest not only in the problems of an imperfect market for buying evaluation services, but also in the complexities of contracting to achieve government goals while providing flexibility for the contractor to use its capabilities, the demands of justifying normative claims, the difficulty of anticipating appropriate designs to match research contexts, and the difficulty of finding "peers" for review as interdisciplinarity increases and there are no established peers. Given these complexities it is perhaps well to be reminded of Kenneth Prewitt's 1980 testimony before the Subcommittee on Science, Research, and Technology, House of Representatives as quoted in Cronbach (1982), "The complexities of the problems for which the social and behavioral sciences might be helpful are always going to be one step ahead of the problem-solving abilities of those sciences…They are sciences whose progress is marked, and whose usefulness is measured, less by the achievement of consensus or the solving of problems than by a refinement of debate and a sharpening of the intelligence upon which collective management of human affairs depends" (p. 82).

Author Note

This article was prepared in a personal capacity. The views expressed in this article do not necessarily represent the views of the National Science Foundation (NSF) of the United States. The article draws on work the second author has conducted supported in part as a consultant on a contract the Guardians of Honor (GOH) has with the National Science Foundation. The paper does not necessarily represent the views of GOH or NSF.

References

Arnold , E., & Balazs, K. (1998). *Methods in the evaluation of publicly funded basic research: A review for OECD*. Brighton, UK: Technopolis Ltd.

Biderman, A. D., & Sharp, L. M. (1972). Evaluation research: Procurement and method. *Social Science Information*, *11*(3/4), 141-170.

Boix-Mansilla, V. (2006). Assessing expert interdisciplinary work at the frontier: An

empirical exploration. *Research Evaluation*, *15*(1), 17-29.

Bozeman, B., Dietz, J. S., & Gaughan, M. (2001). Scientific and technical human capital: An alternative model for research evaluation. *International Journal of Technology Management*, *22*(7/8), 716-740.

Chelimsky, E. (2007). Factors influencing the choice of methods in federal evaluation practice. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation* (pp. 13-34). *New Directions for Evaluation* (No. 113). San Francisco: Jossey-Bass.

Congressional Budget Office. (2005, June). *R & D productivity and growth* (Background paper). Washington, DC: Author.

Cozzens, S. E. (2002). Research Assessment: What's Next? Final report on a workshop. *Research Evaluation, 11*(2), 65-79.

Cronbach, L. J. (1982). Prudent aspirations for social inquiry. In W. H. Kruskal (Ed.), *The social sciences: Their nature and uses* (pp. 61-81). Chicago: University of Chicago Press.

Della-Piana, G. & Della-Piana, C. K. (2005, April). *Approaches to discovery of how and why evaluations deviate from design and differ from reports*. In B. Olds (Chair), Some National Science Foundation responses to difficulties in linking evaluation and program improvement. Symposium conducted in the meeting of the American Educational Research Association, Montreal, Canada.

Dunbar, K. (1999). How scientists build models: In Vivo science as a window on the scientific mind. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 89-98). New York: Plenum Press.

Feist, G. J., & Gorman, M.E. (1998). The psychology of science: Review and integration of a nascent discipline. *Review of General Psychology*, *2*(1), 3-47.

Gordon, E. W. (1999). *Education and justice: A view from the back of the bus*. New York: Teachers College Press.

House, E. R.(1997). Evaluation in the government marketplace. *Evaluation Practice*, *8*(1), 37-48.

House, E. R. & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage. 1999.

Howe, K.R. & Ashcroft, C. (2005). Deliberative democratic evaluation: Successes and limitations of an evaluation of school choice. *Teachers College Record*, *107*(10), 2275-2296.

Julnes, G., & Rog, D. J. (Eds.). (2007). *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation*. *New Directions for Evaluation* (No. 113). San Francisco: Jossey-Bass.

Kelly, G. J. (2006). Epistemology and educational research. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research*. Mahwah, NJ: Lawrence Erlbaum.

Kettl, D. F. (2005). *The next government of the United States: Challenges for performance in the 21st century* (Transformation of Organization Series). Washington, DC: IBM Center for the Business of Government.

Kettl, D. F. (1993). *Sharing power: Public governance and private markets*. Washington, DC: The Brookings Institution.

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, *125*(5), 524-543.

Langfeldt, L. (2006). The policy challenges of peer review: Managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation*, *15*(1), 31-41.

Lengwiler, M. (2006). Between charisma and heuristics: Four styles of interdisciplinarity. *Science and Public Policy*, *33*(6), 423-434.

Maasen, S., & Liven, O. (2005). Transdisciplinarity: A new mode of

Governing Science? *Science and Public Policy*, *33*(6), 399-410.

Maasen, S., Lengwiler, M., & Guggenheim, M. (2006). Practices of transdisciplinary research: Close(r) encounters of science and society. *Science and Public Policy*, *33*(6), 394-398.

Mark, M. M. (2003). Toward an integrative view of the theory and practice of program and policy evaluation. In S. I. Donaldson & M. Scriven, (Eds.), *Evaluating social programs and problems: Visions for the new millennium* (pp. 183-204). Mahwah, NJ: Lawrence Erlbaum Associates.

Mark, M. M. (2005). Evaluation's future: Furor, futile, or fertile? *American Journal of Evaluation*, *22*(3), 457-479.

National Research Council (2004). *Strengthening peer review in federal agencies that support education research*. Washington, DC: The National Academies Press.

Perrin, B. (2006). *Moving from outputs to outcomes: Practical advice from governments around the world* (Managing for Performance and Results Series). Washington, DC: IBM Center for the Business of Government.

Phillips, D. C. (2006). Muddying the waters: The many purposes of educational inquiry. In C. F. Conrad & R. O. Serlin (Eds.), *The Sage handbook of research in education*. Thousand Oaks, CA: Sage.

Reddy, S. (2005). The role of apparent constraints in normative reasoning: A methodological statement and applications to global justice. *The Journal of Ethics*, *9*, 119-125.

Ruegg, R., & Jordan, G. (2007). *Overview of evaluation methods for R & D programs: A directory of evaluation methods relevant to technology development programs*. Washington, DC: U.S. Department of Energy. Office of Energy Efficiency and Renewable Energy.

Scriven, M. (2004). Reflections. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 183-195). Thousand Oaks, CA: Sage.

Scriven, M. (2003). Evaluation in the new millennium: The transdisciplinary vision. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social problems: Visions for the new millennium* (pp. 19-41). Mahwah, NJ: Lawrence Erlbaum.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park: Sage.

Shadish, W. R., & Fuller, S. (1994). *The social psychology of science*. New York: The Guilford Press.

Taylor, P. (1961). *Normative discourse*. Englewood Cliffs, NJ: Prentice-Hall.

Weick, K. E., & Sutcliffe, K. M. (2001). *Managing the unexpected: Assuring high performance in an age of complexity*. San Francisco, CA: Jossey Bass.