How Credible is the Evidence, and Does It Matter? An Analysis of the Program Assessment Rating Tool

Carolyn J. Heinrich

La Follette School of Public Affairs
University of Wisconsin-Madison
1225 Observatory Drive, Madison, WI 53706
Phone: 608-262-5443, Fax: 608-265-3233
cheinrich@lafollette.wisc.edu

September, 2009

I thank the University of Wisconsin-Madison for support of this research through the Regina Cawley Loughlin Scholar funds, and I thank Maureen Quinn and Samuel Hall for their dedicated research assistance.

Abstract

The Program Assessment Rating Tool (PART) was introduced in 2002 to strengthen the process for assessing public program effectiveness and holding agencies accountable for results by making it more rigorous, systematic and transparent and by directly tying it to the executive budget formulation process and resource allocations. The executive directive was clear in its intent to construct a credible, evidence-based rating tool that would ensure that federal programs receive taxpayer dollars only when they prove that they achieve results. A primary objective of this research is to assess the quality of evidence that agencies have provided to Office of Management and Budget in the PART assessments and to empirically examine the relationship between attributes of the evidence and the PART ratings assigned. This study also explores the relationship between the quality and rigor of evidence provided in PART assessments and the funding subsequently received by programs, as well as the relationship between the assigned performance ratings and funding received. The empirical analysis focuses on the evidence submitted by 95 programs administered by the U.S. Department of Health and Human Services for the PART assessment, using newly constructed measures of the quality of evidence and methods used by the programs and information on characteristics of agencies that might relate to program results and government funding decisions. In general, the study findings suggest that while some aspects of the quality of evidence submitted for PART reviews are significantly and positively associated with the PART ratings, the quality of evidence, PART results scores and overall PART scores had no discernible consequences for program funding over time.

INTRODUCTION

The 1990s are widely recognized as a time in which "results-oriented government" took off as a new way of holding government accountable for how it spends public money, and in particular, for the outcomes or results it produces. The new tools and public management reforms advanced reflected an intentional shift from an emphasis on rules or compliance-oriented accountability toward a focus on performance, or how well an organization does what it does in relation to its organizational goals (Radin, 2000; Heinrich, 2003). Although prior administrations had initiated reforms promoting accountability for results, pay for performance, and performance-based contracting, the National Partnership for Reinventing Government spearheaded by Vice President Al Gore (and drawing on the influential work of Osborne and Gaebler, 1992) brought these themes and principles to broader public attention and transformed them into a movement to improve government performance, complete with reinvention teams (internal and government-wide), reinvention laboratories within agencies, town hall meetings and reinvention summits, and new legislation to mandate performance management at the federal level (Kamensky 1999)[1].

A first fruit of these efforts was the Government Performance and Results Act (GPRA), enacted by Congress and signed into law by President Clinton in 1993 to generate more objective information on government performance and efficiency by measuring progress toward performance goals, providing evidence of performance relative to targets, and holding federal agencies accountable for results in annual reports to the public. In the decade and a half since GPRA, few dispute that there has been a definitive transformation in federal government capacity and infrastructure for managing for results, that is, in its use of outcome-oriented strategic plans, performance measures, and reporting of results (U.S. Government Accountability Office, 2008). Yet some in-depth assessments of the implementation of GPRA have also been

[1] Downloaded from http://govinfo.library.unt.edu/npr/whoweare/history2.html, August 9, 2009.

highly critical. A number of researchers have suggested that overlaying a results-oriented managerial logic on top of an inherently political process where agency goals may be ambiguous or contradictory sets the stage for inevitable problems in implementation, above and beyond the challenges of identifying adequate measures of performance (Frederickson and Frederickson, 2006; Radin, 2000; 2006). Radin (2000) argued that rather than freeing public managers to focus on results, GPRA's performance requirements exacerbated administrative constraints and conflict among program managers and heightened distrust between agencies and legislators.

One of the major goals of Bush administration in introducing the Program Assessment Rating Tool (PART) in 2002 was to strengthen the process for assessing program effectiveness and holding agencies accountable for results by making it more rigorous, systematic and transparent and by directly tying it to the executive budget formulation process and the allocation of resources. The executive directive was clear in its intent to construct a "credible evidence-based rating tool" that would ensure that federal programs receive taxpayer dollars *only* when they prove that they achieve results. As stated in an early policy memo, "A program whose managers fail year after year to put in place measures to test its performance ultimately fails the test just as surely as the program that is demonstrably falling short of success."[2] The Office of Management and Budget (OMB) stanchly asserted that unless the use of performance information was directly linked to budgeting activities that drive policy development, performance management activities would continue to have little impact on policy making and program results (Moynihan, 2008). The emphasis in PART on linking program performance to resource allocations was consistent with Kettl's (2002) observation that despite the ambitious efforts to measure performance outcomes and manage for results, in practice, government control over the flow of funds continues as the primary tool of accountability.

[2] Retrieved from http://www.gpoaccess.gov/usbudget/fy04/pdf/budget/performance.pdf, February 10, 2009.

The Bush administration and OMB were also critical of agency performance measures under GPRA, and indeed, the shift to a focus on assessing the performance of specific programs rather than agencies (organizational units) reflected the aim for a more "evaluative" approach, that is, one based on more rigorous (scientific) methods and evidence. The PART questionnaire administered by OMB asks 25 standard questions in each of four topic areas to rate federal programs: (1) program purpose and design, (2) strategic planning, (3) program management, and (4) program results, and it also includes additional questions tailored to particular program types (e.g., competitive grants, block/formula grants, regulatory programs, etc.). In rating program effectiveness, the OMB accepts historical performance data, GPRA strategic plans, annual performance plans and reports, financial statements, and inspector's general reports as evidence, but it also allegedly accords higher ratings to programs that document their effectiveness through more rigorous methods, such as randomized controlled trials or quasi-experimental methods with longer-term tracking of outcomes. In fact, the recommendations coming from the early OMB PART assessments were more heavily focused on how program assessment methods could be improved to generate better data on performance than on improving program performance itself (U.S. Government Accountability Office, 2005).

Although both GPRA and PART aspire to move beyond "counting beans" in using data to make policy and management decisions, research that has compared GPRA and PART along a number of dimensions points to tradeoffs between an approach that engages political actors from both executive and legislative branches in a broader process of reviewing performance information and setting agency goals and expectations for performance, and a performance assessment tool that emphasizes rigorous, systematic and objectively measured outcomes with direct and unambiguous consequences for resource allocations (Radin, 2006; Moynihan, 2008).

As Breul (2007) conveyed, PART is distinct from GPRA in that it renders a *judgment*, including "results not demonstrated." As the Obama administration considers revamping PART and federal performance management efforts more generally, it is important to have reliable information on how these performance management tools and processes are working in practice. The research presented in this paper focuses on the implementation and use of information produced by the PART, although the findings of this study are considered in the wider context of federal performance management efforts.

A primary objective of this research is to assess the quality of evidence that agencies have provided to OMB in the PART assessments and to empirically examine the relationship between attributes of the evidence and the PART ratings assigned. In accord with OMB's intentions, it is expected that programs using more rigorous methods of evaluation and producing better quality documentation of their results will achieve higher overall ratings and program results ratings. This study also explores the relationship between the quality and rigor of evidence provided in PART assessments and the funding subsequently received by programs, as well as the relationship between the assigned performance ratings and program budgets and funding received. The empirical analysis focuses on the evidence submitted by 95 programs administered by the U.S. Department of Health and Human Services for the PART assessment, using newly constructed measures of the quality of evidence and methods used by the programs. These data are also merged with information on characteristics of agencies that might relate to program results and government funding decisions.

The following section of this paper reviews the research and information on PART to date, along with related literature on performance management and evidence-based policymaking. The study data, research methods and research hypotheses are described in the

next section, followed by a presentation of the study findings. The paper concludes with a discussion of the findings and their implications for improving ongoing federal performance management efforts and program performance. In general, the study findings suggest that some aspects of the quality of evidence submitted for PART reviews are significantly and positively associated with the PART ratings, but not with changes in program funding received.

REVIEW OF RELATED LITERATURE

The Development of PART

The broad objectives of recent public management reforms to promote more effective, efficient, and responsive government are not unlike those of reforms introduced more than a century ago (and re-introduced over time), although the intended scope, sophistication, and external visibility of performance management activities aimed at holding governments accountable for *outcomes* are indisputably new (Pollitt and Bouckaert, 2000; Radin, 2000; Heinrich, 2003). Active policy and research debates continue along with these waves of reform, suggests Light (1998: 3), because Congress, the executive branch and public management scholars have yet to resolve, in the absence of sufficient evidence one way or another, "when and where government can be trusted to perform well." Indeed, the focus on producing "evidence" of government performance has intensified, in conjunction with the expansion of "evidence-based policymaking"—that is, policies and practices based on scientifically rigorous evidence—beyond its long-time role in the medical field (Sanderson, 2003).

Although major advances in our analytical tools and capacity for assembling performance information and scientific evidence have been achieved, we have yet to realize a consensus, either intellectually or politically, about what should count as evidence, who should produce it and how, and to what extent it should be used in public policy making and program management

(Heinrich, 2007). A 2005 U.S. Government Accountability Office (GAO) study, for example, reported growing friction between OMB and federal agencies regarding the different purposes and timeframes of PART and GPRA and their "conflicting ideas about what to measure, how to measure it, and how to report program results," including disagreement in some cases over the appropriate unit of analysis (or how to define a "program") for both budget analysis and program management (p. 7). Some agency officials saw the program-by-program focus on performance measures as hampering their GPRA planning and reporting processes, and Radin (2006) notes that appropriations committees in Congress have objected to GPRA's concentration on performance outcomes and its de-emphasis of information on processes and outputs. A 2004 GAO report suggested that OMB should communicate earlier in the PART process with congressional appropriators about what performance information is most important to them in evaluating programs, and that Congress should try a more structured approach to communicating with the executive branch about performance goals and outcomes and the oversight agenda.

Some political and policy analysts suggest that the development of PART was largely in response to the perceived failure of GPRA to produce information on "what works" for guiding resource allocations and improving federal program performance (Dull, 2006; Gilmour and Lewis, 2006). Lynn (1998) argues that an unintended effect of GPRA was to focus managers' attention on the procedural requirements (or the paperwork) of the reform—the production of performance plans, development of performance targets, and documentation of performance measures—rather than on using the information to improve results. Others see PART as a performance management tool that builds on the underpinnings of GPRA, including the flow of information that federal agencies have been generating in response to reporting requirements (Breul, 2007). PART aimed to elevate both evaluative capacity and expectations for the rigor

and quality of information produced by agencies, as well as to give some "teeth" to compliance efforts by attaching budgetary consequences to performance ratings (Frederickson and Frederickson, 2006).

In his analysis of the institutional politics associated with budgetary and performance management reforms, Dull (2006) suggests that PART is one more in a succession of president-initiated budget reforms—including Planning, Programming, Budgeting (PPB), Management by Objectives (MBO) and Zero-Based Budgeting (ZBB)—that was probably ill-fated from the start. He asks why the president would expend limited resources on an initiative such as PART, given the dismal record of past initiatives, but also because it binds the president to "a 'transparent' and 'neutral' instrument that, if it works, presumably raises the cost of making political decisions" (p. 188). In effect, the administration ties its hands (politically) in committing to a transparent process of making budget allocations in a neutral manner based on objective program performance evaluations. As Dull points out, this approach is inconsistent with other Bush administration actions that politicized scientific advisory committees, peer review standards for scientific evidence, and other information gathering for policy decision making.

Yet as the Bush administration entered its first full budget cycle (the FY 2003 budget), it backed its commitment to developing a credible performance rating tool with significant staff time and an invitation for wide-ranging public scrutiny of the first PART questionnaire draft. Input and criticism from the GAO, congressional staff, a panel of experts, an internal advisory committee, members of the National Academy of Public Administration, and participants in open forums were taken seriously and followed by modifications to the instrument (Dull, 2006). The first set of PART questions, addressing a program's purpose and design, are generally interpreted as asking whether the government should be doing this activity (or operating this program) at all,

which has led to some objections that PART encroaches on congressional authority (Radin, 2006). The second section concerning strategic planning expands on GPRA in assessing whether the agency sets appropriate annual and long-term goals for programs. The third section rates program management, including financial oversight and program improvement efforts. And the fourth set of questions, the hallmark of PART, is intended to formalize the review of evidence on program performance, with higher standards for accuracy and expectations for longer-term evaluation. The burden of proof is explicitly on the programs to justify a positive ("yes") rating with a superior standard of evidence. In addition, the emphasis on results is reinforced by the weighting of the sections, with accountability for results (the fourth section) contributing 50 percent toward calculation of the overall PART score.

To date (January 2009), the OMB and Federal agencies have assessed the performance of 1,017 federal government programs that represent 98 percent of the federal budget. In a recent survey of senior federal managers (2008), over 90 percent reported that they are held accountable for their results. OMB defines programs as "performing" if they have ratings of 'effective," "moderately effective," or "adequate," and the latest accounting shows that 80 percent of federal programs are performing.[3] Still, a November 2008 poll of the public indicated that only 27 percent of Americans give a positive rating (good or excellent) of the performance of federal government departments and agencies.[4] Is PART really making a difference in how the federal government manages performance, and if so, in what ways?

Is PART Working?

In 2005, OMB was honored with one of the prestigious Innovations in American Government Awards for the development of PART. In announcing OMB as the winner, the award sponsor

[3] See http://www.whitehouse.gov/omb/expectmore/rating/perform.html, accessed August 18, 2009.
[4] "In the Public We Trust", Partnership for Public Service and Gallup Consulting, November, 2008.

described the promising results OMB was achieving through the PART, in particular, in encouraging more programs to focus on results.[5] The announcement noted that in 2004, 50 percent of federal programs reviewed by the PART could not demonstrate whether or not they were having any impact (earning a "results not demonstrated" rating), while only one year later, only 30 percent of programs reviewed fell into this category. In addition, the percentage of programs rated effective or moderately effective increased from 30 percent in 2004 to 40 percent in 2005. In 2009, OMB reports that 49 percent of programs are rated effective or moderately effective, while the number of programs with "results not demonstrated" has dropped to 17 percent. Only 3 percent of current programs are reported to be ineffective, and 2 of these 26 are no longer being funded. In addition, of 127 programs that were initially rated "results not demonstrated," 88 percent improved their scores in a subsequent evaluation (Norcross and Adamson, 2008).

The above-described trends in PART ratings appear to suggest that federal government performance is improving, as is the capability of federal programs to marshal evidence in support of their effectiveness. Of course, this is predicated on one's belief that the rating tool is "credible" and that the evidence presented by programs during the reviews is of high quality and reflects the achievement of federal program goals. Norcross and Adamson (2008) suggest that programs could also be getting better at responding to procedural requirements associated with providing information to examiners or that OMB could be relaxing its criteria. The public's significantly less positive view of the performance of federal programs likely suggests either ignorance on their part of the performance information generated by PART, or possibly their doubts of its veracity.

[5] See http://www.innovations.harvard.edu/awards.html?id=7496, accessed August 6, 2009.

Studies of PART to date have described some of the challenges inherent in assigning ratings to programs in a consistent way, including some subjective terminology in the questions, the restrictive yes/no format, multiple goals of programs, and a continuing lack of credible evidence on program results. A GAO (2004) study of the 2004 PART process reported that "OMB staff were not fully consistent in interpreting the guidance for complex PART questions and in defining acceptable measures," and that the staff were constrained by the limited evidence on program results provided by the programs (p. 6). Among its recommendations, the GAO suggested that OMB needed to clarify its expectations for the acceptability of output versus outcome measures and the timing of evaluation information, and to better define what counts as an "independent, quality evaluation." The OMB has since generated a number of supporting materials to aid programs and agencies and PART examiners in the implementation of PART, including a document titled "What Constitutes Strong Evidence of a Program's Effectiveness?" that describes different methods for producing credible evidence and the hierarchy among them in terms of their rigor.[6]

In the same 2004 GAO report, PART was lauded for introducing greater structure to a previously informal process of performance review by asking performance-related questions in a systematic way. In its interviews with OMB managers and staff and agency officials, the GAO heard that the PART was also contributing to richer discussions of what a program should be achieving and how it could be achieved, as it brought together program, planning, and budget staffs, including those outside of the performance management area, to complete the questionnaire. Still, contrary to the intent of the PART, some federal programs appeared in practice to largely ignore the requests for more scientifically rigorous evidence and quantitative information on performance outcomes. Gilmour and Lewis' (2006) analysis suggested that in

[6] See http://www.whitehouse.gov/omb/assets/omb/performance/2004_program_eval.pdf, accessed August 19, 2009.

the absence of acceptable performance information, OMB made decisions on the basis of what they could rate, and in other cases, the fact that programs had high-quality measures did not appear to influence budget decisions. In general, they found that the ‘‘results’’ component of PART scores was relatively inconsequential for budget decisions, suggesting that this information is not being used in performance budgeting to redirect resources to programs that produce results. Norcross and Adamson (2008) analyzed data from the fifth year of PART and likewise saw little evidence that Congress uses PART scores in making funding decisions, although they did suggest that there might be a tendency for Congress to award higher budget increases to effective and moderately effective programs.

The most recent GAO (2008) report also suggested that there has been little progress in getting federal managers' to use performance information in decision making. Based on the supposition that "buy-in" by Congress of PART is essential to its sustainability, Stalebrink and Frisco (2009) conducted a recent analysis of nearly 7,000 hearing reports from both chambers of Congress between 2003 and 2008 to assess changes in congressional exposure to PART information and members' use of the information over time. They tracked trends of rising congressional exposure to PART information between 2003 and 2006 followed by declining exposure and interest. Based on their assessment of the content of the hearing report comments, they concluded that PART information is rarely applied in making congressional allocation decisions.

This discussion motivates the central focus of this study: if credible, high-quality information is being generated that accurately reflects agency and program performance and progress toward program goals, it should be influential in decision making about the allocation of budgetary resources. On the contrary, in the absence of solid evidence (and supporting

documentation) on program performance, other criteria for allocating resources will and probably should drive the process. In other words, if the quality of information on which results are judged is weak, it is presumably not in the public interest for there to be a tight link between program performance ratings and resource allocations. In the next section, the data and methods used to test hypotheses about the relationships between the quality of evidence provided by agencies, their PART ratings, and the funding subsequently received by programs are described.

STUDY DATA, METHODS AND RESEARCH HYPOTHESES

This study focuses on the information submitted by 95 programs administered by the U.S. Department of Health and Human Services for the PART process in years 2002-2007. Although OMB currently reports 115 PART reviews for the Department of Health and Human Services (DHHS) at ExpectMore.gov, this number includes the recent (2008) assessment of four programs[7] that came after this study sample was constructed, as well as reassessments of some programs. The Department of Health and Human Services was selected for this study in part because of some of the additional challenges that are well-noted in the performance management and evidence-based policy making literature on measuring the outcomes of social programs (see Radin, 2006 and Heinrich, 2007 for a discussion of these issues). Indeed, DHHS is second only to the Department of Education in the total number of programs rated as not performing (ineffective or results not demonstrated), and the percentage of programs currently not performing (27%) is above the average for all programs (20%). In addition, the substantive experience of the researchers involved in the assembly and coding of the data for this project lies in the area of social program evaluation. Because of the intensive nature of the work undertaken

[7] This information is current as of August 2009. The programs newly rated in 2008 include: CDC Division of Global Migration and Quarantine, Health Information Technology Research, Office of Medicare Hearings and Appeals, and Substance Abuse Drug Courts.

to construct new measures for the analysis, limited time and resources precluded its expansion to additional agencies.

The PART data for these programs, including the overall program scores, the four section ratings, the ratings/responses to each question asked in each of the four sections of the PART questionnaire, and the weights assigned to the individual PART questions, were downloaded from publicly available files or extracted directly from the published PART reports. The OMB maintains a website titled "Assessing Program Performance"[8] where completed PART assessments, assessment details and program funding levels can be accessed, along with all supporting documentation, including technical guidance letters that provided essential information for the construction of new measures for this research. The OMB website also includes sample PART questions, and the exact set of questions asked in the review of each program can be viewed in the "view assessment details" link for each program. Thus, the core data for this project can all be readily accessed electronically without restrictions.

The assessment details from the OMB review of programs were used to construct new measures of the quality of methods and evidence supplied for the PART assessments by the sample of DHHS programs included in this study. Specifically, this information was analyzed by three researchers[9] to code and develop measures of:

- the types of information employed and reported by the programs—quantitative, both quantitative and qualitative/subjective information, qualitative/descriptive only, or none;
- whether the programs were externally evaluated and/or if internal data collection was used, and if one or more comparisons were made to other programs;

[8] See http://www.whitehouse.gov/omb/performance/.
[9] The other two researchers, besides the author, who coded the data were Maureen Quinn**,** a legislative analyst at the Wisconsin Legislative Audit Bureau and Sam Hall, a researcher at the Urban Institute.

- documentation/evidence of the types of performance measures used by the agencies—long-term, annual, and whether a baseline and/or targets were established; and
- if actual performance outcomes were reported.

A basic description of the coding of the PART data to construct the above measures is included in the appendix, as well as a listing of the PART questions and descriptive statistics for the variables used in the analysis. The reviews were time-intensive in that they involved reading all of the detailed comments, descriptions of measures, explanations of the ratings, and other documentation included in the PART reports. The review and coding of this information was completed for each question asked in each of the four major sections of the PART. The review and coding was conducted by multiple researchers to allow for checks of inter-rater reliability.

Inter-rater reliability was very high in the data coding. Given that the coding to generate these new variables primarily involved assigning 0 or 1 values, a simple measure of joint probability of agreement (taking into account the number of ratings but not accounting for the possibility of chance agreement) was computed. In the researcher coding of information for 25 questions and 95 programs, there were only three discrepancies (in coding an evaluation as external or internal), implying an inter-rater reliability rate of more than 99 percent.

In addition, the PART information and these newly-constructed measures were merged with a dataset assembled by Jung Wook Lee, Hal Rainey and Young Chun (2009) that provides additional information on characteristics of the agencies that might be relevant to program results and government funding decisions, including: directive, goal and evaluative ambiguity; congressional, presidential and media salience; agency size, age, and financial publicness; measures of professional staffing, managerial capacity and other aspects of program governance; and policy and regulatory responsibilities. Descriptive statistics of variables from this dataset

that were used as control variables in the analysis are also presented in Appendix A. A limitation is that they are measured at the agency rather than the program level (with 15 different agencies represented). In addition, a number of these variables were highly intercorrelated, and thus, their inclusion in models was determined in part by tests for multicollinearity. For example, we included only the congressional salience measure in the analysis, as variance inflation factors exceeded acceptable levels when presidential salience was also included.

The primary method of empirical analysis employed in this study is multiple regression, with appropriate corrections for potential violations of basic model assumptions, (e.g., clustered robust standard errors to account for correlated errors due to programs grouped within agencies, changes in model specification to correct for multicollinearity). Multiple regression models are estimated to test for positive associations between the rigor of evidence provided by programs and their PART scores, and between the rigor of evidence and the funding received by the programs, as well as their results scores and funding received by the programs, holding constant other program and agency characteristics.

The dependent variables in the analyses include: (1) the program results (section 4) PART score, (2) the overall PART score assigned to the programs, and (3) the change in funding received from one fiscal year before the PART assessments to fiscal year 2008. Gilmour and Lewis (2006) argue that the first three sections of PART, which are not concerned with results, but rather with purpose, planning and management, measure the extent to which federal programs produce the required paperwork under GPRA. They also note that some of the questions regarding program purpose are open to politicization, which is contrary to an objective focus on results. Thus, I do not necessarily expect the same or as strong of a relationship

between the quality of evidence and overall PART scores, as between the rigor of evidence and the results (section 4) PART score.

The two core sets of explanatory variables include: the newly constructed measures that describe the nature and quality of evidence provided by programs in the PART review, and the measures of other program and agency characteristics that are used as controls in the analysis. In accord with OMB guidelines on what constitutes strong evidence,[10] I expect higher ratings for programs that provide quantitative evidence, are independently evaluated, and that report longer-term measures of outcomes and establish explicit performance targets.

The specific research hypotheses tested (and corollary hypotheses) are as follows:

H1: The PART results (section 4) score will be higher when more rigorous evidence is provided in support of the responses to the five questions about program results.

C1a: Results scores will be positively related to the number of question responses backed by quantitative evidence.

C1b: Results scores will be positively related to the number of question responses backed by external (or independent) evaluations.

C1c: Results scores will be negatively related to the number of question responses backed by qualitative evidence only.

C1d: Results scores will be negatively related to the number of question responses with no supporting evidence.

C1e: Results scores will be negatively related to the number of question responses backed by internal evaluations only.

H2: The overall program PART score will be higher when more rigorous evidence in support of program results is provided and when higher quality measures and evidence are provided in support of question responses throughout the questionnaire.

The same five corollary hypotheses as above apply in relation to overall PART scores, in addition to the following:

[10] See http://www.whitehouse.gov/omb/assets/omb/performance/2004_program_eval.pdf, accessed Aug. 20, 2009. This document identifies methods such as random assignment experiments and quasi-experimental or nonexperimental methods with matched comparison groups as more rigorous methods.

C2a: Overall PART scores will be higher for programs that use long-term measures of program outcomes.

C2b: Overall PART scores will be higher for programs that annually measure progress toward long-term goals.

C2c: Overall PART scores will be higher for programs that establish baseline measures and targets for assessing performance.

In accord with the statement of former President Bush suggesting that better documentation of results is as important as higher performance in and of itself, I also expect that measures reflecting higher quality evidence provided to demonstrate program results to be positively related to funding received by the programs.

H3: Increases in federal funding received (from the fiscal year prior to the PART assessments to fiscal year 2008) will be positively related to the quality and rigor of evidence provided in support of the responses to program results questions and in support of questions responses throughout the questionnaire.

DATA ANALYSIS AND FINDINGS

The descriptive statistics in Appendix A provide some indication of what types of evidence programs are more and less likely to offer in response to the PART questions concerning program results. For example, it is clear that programs are least likely to provide externally-generated evidence of their performance relative to long-term and annual performance goals (just 9% and 7%, respectively). Only in response to the fifth results question, asking whether independent evaluations of sufficient scope and quality indicate the program is effective, do a majority offer external evidence; at the same time, however, for 46 percent of the programs (most of those providing evidence), this evidence is only qualitative (no quantitative measures). In addition, close to half of the programs provided no evidence of how their performance compares to other government or private programs with similar purposes and goals. On the positive side, over 90 percent of programs report regularly collecting timely performance

information, and more than 80 percent of programs have identified specific long-term performance measures focused on outcomes, although this does not imply anything about the quality of those measures or the evidence produced for evaluating program performance.

In coding the information supplied by programs for the PART assessment, information that described trends in outcomes without any specific numbers or data in support, such as "reduces incidences by around half," was coded as "qualitative." In addition, supporting evidence that was contained in reports and documents from jointly-funded programs, grantees or sub-grantees, contractors, cost-sharing partners and other government partners was coded as internal. The findings that nearly half of the programs do not make comparisons to other programs with similar purposes and that the independent evaluations conducted generated mostly qualitative evidence on performance were not surprising, given that rigorous, large-scale evaluations of national programs are a costly undertaking. Some programs were awarded points for these questions by OMB examiners if they cited prior GAO reports, university studies or evaluations by organizations such as the Urban Institute. However, these studies are frequently not initiated by the programs themselves, and thus, some programs may have an advantage in these performance reviews that does not reflect organizational efforts to improve performance evaluation (and may even reflect larger concerns about the effectiveness of programs).

Results of hypothesis testing

The first hypothesis (and its corollaries), stating that the PART results score will be higher when more rigorous evidence is provided in support of the responses to the program results questions, was tested with the multiple regression model shown in Table 1, column 1. The results of this regression, with the results (section 4) score as the dependent variable, generally confirm the hypothesized relationships. The average results score is 0.419. For each results question for

which no evidence is provided, the results score is reduced by 0.175 (p=0.001), and for each question for which only qualitative evidence is offered in response, the results score is reduced by 0.047 (p=0.027); the reference category in this model is the provision of some quantitative evidence. And although statistically significant only at $\alpha < 0.10$, there is also a negative relationship between the reporting of internal evaluations as evidence and the results score. The relationship between the provision of external evidence and the results score is positive, as expected, but not statistically significant.

These findings thus support the main hypothesis that the rigor of the evidence (or having at least some quantitative evidence) is positively associated with PART scores on the results section. That said, less than a third of the variation in the results scores is explained by these variables. In Model 2 in Table 1, variables measuring program and agency characteristics are added to the model, and robust, clustered standard errors are estimated to adjust for the grouping of programs within agencies. The percentage of total variation in the PART results scores explained by this model almost doubles (to approximately 58 percent), and the statistically significant, negative effects of having no evidence or only qualitative evidence in support of program performance hold. In addition, after removing the variable indicating the number of questions for which internal evidence was provided due to multicollinearity, the measure of the number of questions for which *external* evidence was provided is now also statistically significant, suggesting that the results score increases by 0.045 for each question in this section that is supported by external evidence.

The results in Model 2 also show that research and development (R&D) and capital asset programs receive significantly higher PART results scores.[11] Congressional salience (measured in Z-scores) is significantly and positively associated with PART results scores as well, while the relationship of the age of the agency (the year in which it was established) to the results score is negative (and statistically significant). Although one might readily construct plausible arguments for why R&D programs such as the National Center for Health Statistics and those administered by the Agency for Healthcare Research and Quality (AHRQ) and National Institutes for Health (NIH) might be better equipped to "demonstrate" progress toward achieving annual or long-term goals and cost-effectiveness, as well as capital asset programs administered by agencies such as the Center for Disease Control and the NIH, it is more difficult to see the logic for a relationship between congressional salience and results scores. If it is the case that more politically salient programs and those with a higher percentage of financial resources from government are more likely to have external (independent) evaluations mandated, then this might contribute to higher results scores. However, regressions employing as dependent variables the raw scores from questions 1-3 in the results section (on demonstrating results), and separately, the raw scores from questions 4 and 5 (on comparing the program to other similar programs and the scope and quality of independent evaluations) both showed statistically significant associations between congressional salience and the results scores.

The second hypothesis and its corollaries ask whether providing more rigorous evidence in support of program results and in response to other questions throughout the questionnaire is positively related to the *overall* PART scores. As noted above, the first three sections of the PART questionnaire are less directly concerned with program results and are more focused on

[11] The number of observations in the models with agency characteristics drops to 89 from 95, due to missing information for some agencies. Sensitivity tests indicated that the results of models including all 95 observations did not change substantively when estimated with the subset of 89 programs.

the process of measuring performance, and thus, it is possible that the relationship between the quality or rigor of evidence and overall PART scores may be weaker. The same model as shown in Table 1, column 1 (with measures of the rigor of evidence on results) was estimated with the overall PART score as the dependent variable, and the results are presented as Model 3 in this table. The results of this estimation are comparable to those for the results section scores (Model 1), although the magnitude of the coefficients differs because the scale of scores is different. The average overall PART score is 1.87 (between ineffective and adequate), and for each results question for which there was no evidence provided, the overall PART score is reduced by 0.567 (or about 30% of the average overall PART score). The provision of only qualitative results evidence is also negatively and statistically significantly related to the overall PART score. About the same proportion of total variation in overall PART scores is explained by these variables (as in the model with the results scores as the dependent variable).

In the next model (Model 4 in Table 1, other measures of the program's performance management efforts (based on PART questionnaire responses) were added to this model with the overall score as the dependent variable. These include indicator variables for whether the programs were recorded as using long-term measures of program outcomes, annually measuring progress toward long-term goals, establishing baseline measures and targets for assessing performance, regularly collecting timely measures of performance, tying budget requests explicitly to their accomplishment of program goals, and holding federal managers and program partners accountable for performance results. After including these measures of performance management efforts, the effects of the other measures characterizing the rigor of the results evidence are slightly weaker but still statistically significant.[12] In addition, there are negative,

[12] Again, the measure of the number of questions for which internal evidence was provided was excluded due to multicollinearity problems that emerged after adding additional variables to the model.

statistically significant relationships between overall PART scores and programs' reports of having no long-term measures, no baseline measures or targets, and no independent evaluations, while programs that report holding federal managers accountable and tying budget requests explicitly to results have significantly higher overall PART scores. Not having long-term measures is most strongly (negatively) associated with overall PART scores (reducing the score by 0.803). That said, only 12 percent of programs presented externally-generated evidence of these measures, and the percentages are even smaller for the other indicators of program performance management efforts. Thus, although PART scores are probably based more on what programs report they have done to measure results than on objective reviews of the nature and rigor of the evidence supplied (as suggested by Gilmour and Lewis, 2006), the associations are at least consistent with the intent and expectations of the performance rating exercise.

In the fifth model in Table 1, agency characteristics are added to this same model as explanatory variables, and robust clustered standard errors are again estimated. Not having evidence in support of results or having only qualitative evidence are still negative and statistically significant predictors of overall PART scores, as is the indicator for no long-term measures. In addition, holding federal managers accountable for program performance is still positively and significantly related to overall PART scores. Among agency characteristics, the only variable statistically significant at the $\alpha<0.05$ level is the measure of congressional salience, which is again positively to the PART ratings. Contrary to what one would expect, the percentage of agency positions that are classified as management or program analysis—what Lee, Rainey and Chun (2009) characterize as a measure of managerial capacity—is negatively related to overall PART scores, although only weakly.

As discussed earlier, the "teeth" of PART were supposed to be the budgetary consequences attached to the performance ratings. Furthermore, the Bush administration emphasized that better documentation of results would be as important as high performance ratings themselves for program funding. This motivates the third hypothesis that changes in program funding following the PART assessments (FY 2008) will be related to the quality and rigor of evidence provided in support of program results and throughout the PART questionnaire. The same explanatory variables included in the fifth model in Table 1—measures of the rigor of the results evidence, program performance management efforts, and program and agency characteristics—were also included the model predicting the change in program funding (in millions of dollars) from the year prior to the program's PART assessment to FY 2008. The results, shown in column 1 of Table 2, are contrary to expectations. Not a single measure of the quality of the results evidence or the indicator variables measuring the use of long-term and annual measures, baseline measures and other program performance management efforts are statistically significant predictors of changes in program funding. The only observed relationship consistent with the aims of PART is the negative relationship of agency evaluative ambiguity (the percentage of subjective or workload-oriented performance indicators, as opposed to objective and results-oriented performance indicators) to increases in funding. The other two statistically significant associations with funding increases are agency size (the log of the number of full-time employees), which is negatively related to funding changes, and congressional salience, which is a strong, positive predictor of post-PART funding changes.

The final two regression models explore an even more basic relationship: are changes in funding related to program performance, as measured by the PART ratings (holding other agency and program characteristics constant)? The model in column 2 of Table 2 adds the program's

score on section four of the PART (results) to the model that includes measures of the rigor of results evidence, program performance management efforts, and agency characteristics, and the model in column 3 adds the overall PART score to this same base model. Strikingly (or perhaps not), the results of these regressions show no relationship between either the PART results (section four) score and funding changes, or between the overall PART score and funding changes. In each of these models, about 56-58 percent of the total variation in program funding changes is explained, apparently primarily by agency characteristics (agency size, evaluative ambiguity and congressional salience).

The lack of an observed relationship between PART performance ratings and program funding changes and between the rigor of the evidence offered in the PART assessments and program funding changes did not reflect insufficient variation in funding from one fiscal year to another to detect these relationships. Fourteen percent of the DHHS programs saw funding declines (by up to 100%, or a loss of all funding), and the others saw increases ranging from one-half percent to 44 percent. Approximately one-third of the programs realized funding changes (measured in millions of dollars) of plus or minus 10 percent or more. Simple descriptive statistics did show a positive correlation between PART results scores and overall PART scores and program rankings in terms of the size of funding increases they received, but these relationships were not statistically significant. The general finding that performance (as assessed by the PART) and the quality of the evidence supplied by programs (as evaluated in this study) bear little relationship to program funding changes following the PART exercise suggest that at least for DHHS programs, the PART is not working as intended.

These study findings are consistent with those of Gilmour and Lewis (2006) and Norcross and Adamson (2008), both who found little evidence that Congress uses the PART

results scores or overall PART scores in making funding decisions. In addition, based on their assessment of hearing report comments, Stalebrink and Frisco concluded that PART information is rarely applied in making congressional allocation decisions. On the other hand, these findings differ from those of Blanchard's (2008) study, which concluded that higher-performing programs are being rewarded with larger approved funding increases by OMB and Congress. Blanchard's study differs from this one in that he uses data from a larger number of departments and programs; he constructs different measures of program performance and funding changes and includes different control variables in his models, and he draws some conclusions from descriptive analyses of trends in results and budget proposals. He acknowledges that his study would have benefited from better measures of political factors, which may not be fully accounted for in his models, although there are too many differences in study design and methods to attempt to reconcile the findings of these studies.

As Moynihan (2008) notes, OMB contends that partisan preferences do not affect PART assessments, but they do influence resource allocations, acknowledges the OMB. This is expected, adds Moynihan, given that the PART "explicitly feeds into the highly political budget process" (p. 134). In the analysis in this study of what predicts PART results and overall scores, as well as changes in federal funding, congressional salience was a consistent, statistically significant predictor across the models. Even if intended to be objective rather than partisan, the political nature of the PART appears to be inherent in its implementation.

CONCLUSION AND IMPLICATIONS

The Bush administration's Program Assessment Rating Tool (PART) sought to advance the promises of "results-oriented" government by strengthening the performance assessment process—that is, making it more rigorous, systematic and transparent—and by directly linking

the allocation of budgetary resources to the program PART ratings. The findings of this study thus present a mixed review of the PART's effectiveness. The empirical analysis using data on 95 DHHS programs with newly constructed measures to evaluate the quality/rigor of evidence supplied by the programs did show some statistically significant relationships between the nature and rigor of the evidence and PART ratings, confirming that programs that supplied only qualitative evidence (or no evidence) and that did not identify long-term or baseline measures for use in performance assessments were rated lower. At the same time, the ratings of program results and the overall PART scores had no discernible consequences for program funding over time. Furthermore, the agencies' congressional salience was the strongest predictor of subsequent program funding increases or decreases than program performance, suggesting that the PART has a long way to go toward fostering evaluative capacity and an emphasis on results that the focus on programs (rather than agencies) was intended to promote.

Going forward, how should policymakers and the public see the future of program performance rating tools such as the PART, when the overall trends in PART ratings and the findings of this study suggest that more programs are providing stronger evidence of their performance (i.e., the number with "results not demonstrated" has steadily declined), and yet, this does not appear to also be shaping how the federal government allocates its resources? Since the departure of President George W. Bush, PART has continued as the main policy tool used by the executive branch to assess program results, and President Obama, who was a co-author of the Federal Funding Accountability and Transparency Act of 2006 (S. 2590) that requires full disclosure to the public of all entities or organizations receiving federal funds, has pledged unprecedented openness and transparency in reporting of government results to the public.

One policy direction would be to continue efforts to strengthen program evaluative capacity along the lines of current government guidelines that call for more experimental evaluations and other rigorous methods of performance analysis. A problem noted earlier, however, is that these expectations for producing more credible evidence and reliable knowledge of program outcomes are frequently incompatible with the requirements of other performance management initiatives such as GPRA to produce timely information for decision-making. In the context of increasing public demands for accountability that include high-stakes pressures to demonstrate performance improvements, policymakers frequently have little choice but to consider and use a mix of different types of information and methods in producing annual performance reports. Along these same lines, a bolder action would be to end GPRA and shift the effort and resources that are currently used in developing annual strategic plans and performance reports to more rigorous, longer-term program evaluation efforts under the PART or a similar program evaluation tool. This could help to reduce conflicts over what to measure and how to measure it, as well as the paperwork burdens associated with complying with both PART and GPRA performance measurement requirements. It might also help to promote a genuine longer-term focus on government performance improvements and reduce the extent to which political considerations play a role in performance assessment processes.

Indeed, the fact that some aspects of the quality of evidence supplied by programs are influencing their PART ratings is a real advance. Still, if the PART ratings do not matter for policy and program decision making, it is a small step forward. In addition, given that the quality of evidence is still very uneven across programs and that this appears to be overlooked in some PART assessments, it might be in the interest of the public to place less weight (officially) on evidence of program performance in decisions about resource allocations until evaluative

capacity and the evidence produced are stronger. (Unofficially, there appears to be very little weight placed on evidence of performance right now).

This, of course, begs another question: what other objective factors, besides program performance, should influence funding allocations? For example, are there substantial, measurable factors that contribute to congressional salience that could be made more explicit or transparent in funding allocations? For the purposes of accountability, the public should, at a minimum, understand how much influence program performance really has on funding decisions and what the tradeoffs are between emphasizing performance and other goals, such as equity in access to public services. Although there are mechanisms such as televised hearings and other channels through which the public can ascertain some of this information, making the decision making processes more transparent (that is, in an explicit, concrete and accessible way) and allowing for full disclosure as President Obama has promised would be another step forward.

Dull's analysis (2006) suggests, however, that this would be undesirable for the Obama administration, which would sacrifice important political flexibility by committing to greater transparency in making budget allocations, particularly if objective program performance evaluations were actually given the weight that the PART intends. The Bush administration designed a bold policy tool, but it remains to be seen whether another presidential administration will be willing to give it the "teeth" it needs to ensure that federal programs that receive funding are doing all that they can to prove they achieve results. And if PART is not continued, Light's (1998) analysis suggests that another similar policy tool or reform will surely follow, as the public interest in seeing evidence of government performance is unlikely to abate anytime soon.

REFERENCES

Blanchard, Lloyd A. 2008. Part and Performance Budgeting Effectiveness. In *Performance Management and Budgeting: How Governments Can Learn from Experience*, Steve Redburn, Robert Shea, and Terry Buss, (eds.), M.E. Sharpe, Inc., pp. 67-91.

Breul, Jonathan. 2007. Three Bush Administration Management Reform Initiatives. *Public Administration Review* 67(1): 21-26.

Frederickson, David G., and George H. Frederickson. 2006. *Measuring the performance of the hollow state*. Washington, DC: Georgetown University Press.

Dull, Matthew. 2006. Why PART? The Institutional Politics of Presidential Budget Reform. *Journal of Public Administration Research and Theory* 16(2): 187-215.

Gilmour, John B. and David E. Lewis. 2006. "Does Performance Measurement Work? An Examination of OMB's PART Scores." *Public Administration Review* 66(5): 742 – 752.

Heinrich, Carolyn J. 2003. "Measuring Public Sector Performance and Effectiveness." In *Handbook of Public Administration*, Guy Peters and Jon Pierre, (eds.), London: Sage Publications, Ltd., pp. 25-37.

Heinrich, Carolyn J. 2007. "Evidence-based policy and performance management: Challenges and Prospects in Two Parallel Movements." *The American Review of Public Administration* 37(3): 255-277.

Kamensky, John. 1999. National Partnership for Reinventing Government: A Brief Review. Washington, DC: National Partnership for Reinventing Government.

Kettl, Donald. 2002. *The transformation of governance: Public administration for twenty-first century America*. Baltimore, MD: The Johns Hopkins University Press.

Lee, Jung Wook, Hal G. Rainey and Young Han Chun. 2009. Of Politics and Purpose: Political Salience and Goal Ambiguity of US Federal Agencies. *Public Administration* 87(3): 457 – 484.

Light, Paul C. 1998. *The Tides of Reform Making Government Work, 1945-1995*. New Haven, CT: Yale University Press.

Lynn, Laurence E., Jr. 1998. The *New Public Management***:** How to Transform a Theme into a Legacy. *Public Administration Review* 58(3): 231-7.

Moynihan, Donald P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform.* Washington D.C.: Georgetown University Press.

Norcross, Eileen and Joseph Adamson. 2008. An Analysis of the Office of Management and

Budget's Program Assessment Rating Tool (PART) for Fiscal Year 2008. Government Accountability Project Working Paper, Mercatus Center, George Mason University.

Osborne, David, and Ted Gaebler. 1992. *Reinventing government*. New York, NY: Penguin Press.

Pollitt, Christopher and Geert Bouckaert. 2000. *Public Management Reform: A Comparative Analysis*. Oxford: Oxford University Press.

Radin, Beryl A. 2000. The Government Performance and Results Act and the Tradition of Federal Management Reform: Square Pegs in Round Holes?' *Journal of Public Administration Research and Theory* 10(1): 11-35.

Radin, Beryl A. 2006. *Challenging the Performance Movement*. Washington, D.C.: Georgetown University Press.

Sanderson, Ian. 2003. Is it "what works" that matters? Evaluation and evidence-based policy-making. *Research Papers in Education* 18(4): 331-345.

"Senior Executive Services Survey Results," U.S. Office of Personnel Management, May 2008. United States Government Accountability Office, 2004. Performance Budgeting: PART Focuses Attention on Program Performance, but More Can Be Done to engage Congress. Washington, DC. GAO Report #04-174.

Stalebrink, Odd J. and Velda Frisco. 2009. PART of the Future: A Look at Congressional Trends. Unpublished Manuscript, School of Public Affairs, The Pennsylvania State University, Harrisburg, PA..

United States Government Accountability Office. 2005. Performance budgeting: PART focuses attention on program performance, but more can be done to engage Congress. GAO Report #06-28.

United States Government Accountability Office. 2008. Government Performance: Lessons Learned for the Next Administration on Using Performance Information to Improve Results. GAO Report #08-1026T.

Table 1: Relationship of Evidence Quality to PART Scores

Dependent variable:	**PART results score**				**Overall PART score**					
Explanatory variables	*Model 1*		*Model 2*		*Model 3*		*Model 4*		*Model 5*	
External evidence-results (#)	0.022		0.045	**	0.019		0.051		0.039	
	(0.031)		(0.018)		(0.158)		(0.125)		(0.116)	
Internal evidence-results (#)	-0.083	*			-0.044					
	(0.046)				(0.236)					
No evidence-results (#)	-0.175	**	-0.073	**	-0.567	**	-0.245	**	-0.211	**
	(0.052)		(0.026)		(0.266)		(0.110)		(0.068)	
Qualitative evidence only-results (#)	-0.047	**	-0.049	*	-0.449	**	-0.187	*	-0.294	**
	(0.021)		(0.024)		(0.109)		(0.109)		(0.070)	
Block grant			0.060						-0.147	
			(0.077)						(0.418)	
R&D			0.213	**					0.458	
			(0.084)						(0.372)	
Capital assets			0.208	*					0.455	
			(0.112)						(0.356)	
Direct federal			-0.006						-0.542	
			(0.085)						(0.504)	
Regulatory			0.028						-0.523	
			(0.090)						(0.462)	
Evaluative ambiguity			-0.004						-0.014	
			(0.004)						(0.011)	
% management/ analysis positions			-0.003						-0.029	*
			(0.004)						(0.014)	
Log # full-time employees			0.034						-0.187	*
			(0.021)						(0.092)	
Age of agency			-0.0019	**					-0.00008	
			(0.0008)						(0.005)	
Congressional salience			0.178	*					1.235	**
			(0.092)						(0.312)	
No long-term measures							-0.803	**	-0.488	**
							(0.371)		(0.195)	
No annual measures							-0.100		-0.527	*
							(0.408)		(0.262)	
No baseline/targets							-0.606	**	-0.307	
							(0.282)		(0.344)	
No independent evaluation							-0.506	**	-0.157	
							(0.239)		(0.357)	
No regular performance info							-0.251		-0.511	
							(0.450)		(0.358)	
Performance budgeting							0.498	**	0.242	
							(0.229)		-0.283	
Managers held accountable							0.633	**	0.609	**
							(0.259)		-0.241	
Constant	0.951	**	0.603	*	3.251	**	2.267	**	5.233	**
	(0.241)		(0.289)		(1.234)		(0.380)		(1.420)	
R-squared	*31.6%*		*58.3%*		*31.9%*		*53.8%*		*62.1%*	

Standard errors in parentheses; * coefficient statistically significant at α<0.10; ** coefficient statistically significant at α<0.05

Table 2: Relationship of Evidence Quality and Performance to Funding

Dependent variable	**Change in federal funding, before PART to 2008 ($ mill.)**					
Explanatory variables	*Model 1*		*Model 2*		*Model 3*	
	272		752		296	
External evidence-results (#)	(436)		(690)		(414)	
	614		248		422	
No evidence-results (#)	(853)		(604)		(680)	
	-815		-896		-1060	
Qualitative evidence only-results (#)	(1272)		(1206)		(1223)	
Block grant	-1226		-435		-1301	
	(2198)		(2447)		(2376)	
R&D	399		2033		808	
	(1727)		(2195)		(1718)	
Capital assets	4638		6116		5089	
	(5125)		(6050)		(5535)	
Direct federal	30506		30752		30220	
	(25019)		(24722)		(25015)	
Regulatory	-10529		-11208		-10926	
	(6302)		(6826)		(6638)	
Evaluative ambiguity	-258	**	-289	**	-268	**
	(90)		(96)		(92)	
	-133		-143		-151	
% management/ analysis positions	(229)		(228)		(225)	
	-3710	*	-3244	*	-3860	*
Log # full-time employees	(2009)		(1620)		(2079)	
Age of agency	74		61		74	
	(69)		(62)		(69)	
	15769	**	17086	**	16799	**
Congressional salience	(4479)		(4861)		(5169)	
	-2867		-3992		-3366	
No long-term measures	(3855)		(4539)		(4239)	
No annual measures	-257		-745		-662	
	(1850)		(1965)		(2005)	
No baseline/targets	2041		1282		1798	
	(2576)		(1888)		(2273)	
No independent evaluation	-1394		-1995		-1521	
	(2373)		(2579)		(2420)	
No regular performance info	-6888		-7444		-7361	
	(6650)		(6905)		(7046)	
Performance budgeting	-867		-128		-674	
	(2305)		(1993)		(2234)	
Managers held accountable	-6011		-5961		-5574	
	(5183)		(5153)		(4941)	
Results section score			-9603			
			(6631)			
Overall PART score					-855	
					(726)	
Constant	51884	**	55307	**	56176	**
	(19218)		(20356)		(21642)	
R-squared	*55.7%*		*57.6%*		*56.1%*	
Standard errors in parentheses; * coefficient statistically significant at α<0.10; ** coefficient statistically significant at α<0.05						

APPENDIX: Description of PART Questions, Newly Constructed Measures and Variable Descriptives

Basic questions in PART

Section I. Program Purpose and Design

1.1: Is the program purpose clear?

1.2: Does the program address a specific and existing problem, interest, or need?

1.3: Is the program designed so that it is not redundant or duplicative of any other Federal, State, local or private effort?

1.4: Is the program design free of major flaws that would limit the program's effectiveness or efficiency?

1.5: Is the program design effectively targeted so that resources will address the program's purpose directly and will reach intended beneficiaries?

Section II. Strategic Planning

2.1: Does the program have a limited number of specific long-term performance measures that focus on outcomes and meaningfully reflect the purpose of the program?

2.2: Does the program have ambitious targets and timeframes for its long-term measures?

2.3: Does the program have a limited number of specific annual performance measures that can demonstrate progress toward achieving the program's long-term goals?

2.4: Does the program have baselines and ambitious targets for its annual measures?

2.5: Do all partners (including grantees, sub-grantees, contractors, cost-sharing partners, and other government partners) commit to and work toward the annual and/or long-term goals of the program?

2.6: Are independent evaluations of sufficient scope and quality conducted on a regular basis or as needed to support program improvements and evaluate effectiveness and relevance to the problem, interest, or need?

2.7: Are Budget requests explicitly tied to accomplishment of the annual and long-term performance goals, and are the resource needs presented in a complete and transparent manner in the program's budget?

2.8: Has the program taken meaningful steps to correct its strategic planning deficiencies?

Section III. Program Management

3.1: Does the agency regularly collect timely and credible performance information, including information from key program partners, and use it to manage the program and improve performance?

3.2: Are Federal managers and program partners (including grantees, sub-grantees, contractors, cost-sharing partners, and other government partners) held accountable for cost, schedule and performance results?

3.3: Are funds (Federal and partners') obligated in a timely manner, spent for the intended purpose, and accurately reported?

3.4: Does the program have procedures to measure and achieve efficiencies and cost effectiveness in program execution?

3.5: Does the program collaborate and coordinate effectively with related programs?

3.6: Does the program use strong financial management practices?

3.7: Has the program taken meaningful steps to address its management deficiencies?

Section IV. Program Results/Accountability

4.1: Has the program demonstrated adequate progress in achieving its long-term performance goals?

4.2: Does the program (including program partners) achieve its annual performance goals?

4.3: Does the program demonstrate improved efficiencies or cost effectiveness in achieving program goals each year?

4.4: Does the performance of this program compare favorably to other programs, including government, private, etc., with similar purpose and goals?

4.5: Do independent evaluations of sufficient scope and quality indicate that the program is effective and achieving results?

Information recorded for each PART question

Weight: enter percentage

Answer:
4 – yes
3 – large extent
2 – small extent
1 – no
0 – NA

New measures characterizing the evidence

Quantitative data/measures: 1 – yes, 0 – no

Both quantitative and qualitative/subjective information: 1 – yes, 0 – no

Qualitative (descriptive information, reports with no indication of data or empirical measures included): 1 – yes, 0 – no

No evidence: 1 – yes, 0 – no

Externally provided or evaluated: 1 – yes, 0 – no

Internal data collection/reporting: 1 – yes, 0 – no

Enter measure: (character field)

Term of measure is long-term: 1 – yes, 0 – no

Term of measure is annual: 1 – yes, 0 – no

Baseline established: 1 – yes, 0 – no

Target established: 1 – yes, 0 – no

Actual measure reported: 1 – yes, 0 – no

Year of actual measure: enter year or range (e.g. 1999-2002)

Descriptive Measures of Study Variables			
Variable	**N**	**Mean**	**Std. Dev.**
Q 4_1 external	95	0.094	0.293
Q 4_1 internal	95	0.865	0.344
Q 4_2 external	95	0.073	0.261
Q 4_2 internal	95	0.844	0.365
Q 4_3 external	95	0.104	0.307
Q 4_3 internal	95	0.802	0.401
Q 4_4 external	95	0.198	0.401
Q 4_4 internal	95	0.469	0.502
Q 4_5 external	95	0.573	0.497
Q 4_5 internal	95	0.594	0.494
Q 4_1 only qualitative	95	0.208	0.408
Q 4_2 only qualitative	95	0.198	0.401
Q 4_3 only qualitative	95	0.229	0.423
Q 4_4 only qualitative	95	0.281	0.452
Q 4_5 only qualitative	95	0.458	0.501
No evidence Q 4_1	95	0.104	0.307
No evidence Q 4_2	95	0.146	0.355
No evidence Q 4_3	95	0.188	0.392
No evidence Q 4_4	95	0.490	0.503
No evidence Q 4_5	95	0.177	0.384
Overall PART score	95	1.874	1.378
Results score	95	0.419	0.268
# results questions-external	95	1.042	0.988
# results questions-internal	95	3.589	1.317
# results questions-qualitative	95	3.021	1.487
# results questions-quantitative	95	2.505	1.570
# results questions-only qualitative	95	1.379	1.213
# results questions-no evidence	95	1.095	1.272
% externally evaluated	95	0.199	0.128
No long-term measures	95	0.168	0.376
No annual measures	95	0.137	0.346
No baseline/targets	95	0.263	0.443
No independent evaluation	95	0.453	0.500
No regular performance info	95	0.063	0.245
No comparison program	95	0.537	0.501
Managers held accountable	95	0.726	0.448
Performance budgeting	95	0.347	0.479
Competitive grant	95	0.400	0.492
Block grant	95	0.337	0.475
R&D grant	95	0.084	0.279
Capital assets	95	0.084	0.279
Direct federal	95	0.063	0.245
Regulatory	95	0.032	0.176
Evaluative ambiguity	89	56.093	10.905
Congressional salience	89	-0.339	0.396
Presidential salience	89	-0.310	0.201
Log # full-time employees	89	7.664	1.106
Age of agency	89	51.843	25.617
% management/analysis positions	89	9.430	5.841
Federal funding change	93	1129	10509
Percent of funding change	93	-0.064	0.293