PAR PUBLIC ADMINISTRATION REVIEW

Carolyn J. Heinrich
University of Texas at Austin

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

This research empirically assesses the quality of evidence that agencies provided to the Office of Management and Budget in the application of the Program Assessment Rating Tool (PART), introduced in 2002 to more rigorously, systematically, and transparently assess public program effectiveness and hold agencies accountable for results by tying them to the executive budget formulation process and program funding. Evidence submitted by 95 programs administered by the U.S. Department of Health and Human Services for the PART assessment is analyzed using measures that capture the quality of evidence and methods used by programs and information on characteristics of agencies that might relate to program results and government funding decisions. The study finds that of those programs offering some evidence, most was internal and qualitative, and about half did not assess how their performance compared to other government or private programs with similar objectives. Programs were least likely to provide externally generated evidence of their performance relative to long-term and annual performance goals. Importantly, overall PART and results scores were (statistically) significantly lower for programs that failed to provide quantitative evidence and did not use long-term measures, baseline measures or targets, or independent evaluations. Although the PART program results ratings and overall PART scores had no discernible consequences for program funding over time, the PART assessments appeared to take seriously the evaluation of evidence quality, a positive step forward in recent efforts to base policy decisions on more rigorous evidence.

In the 1990s, "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 (Heinrich 2003; Radin 2000). Although prior administrations 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), transformed these themes and principles 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]

The first major fruit of these efforts was the Government Performance and Results Act (GPRA), enacted 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 passage of the 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 (GAO 2008). Yet some in-depth assessments of the implementation of the GPRA also have been highly critical. Several researchers have suggested that overlaying a results-oriented managerial logic on top of an inherently political process in which 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, the GPRA's performance requirements exacerbated administrative constraints and conflict among program managers and heightened distrust between agencies and legislators.

One of the primary goals of the George W. Bush administration in introducing the Program

Carolyn J. Heinrich is the Sid Richardson Professor of Public Affairs and director of the Center for Health and Social Policy in the Lyndon B. Johnson School of Public Affairs at the University of Texas at Austin. She previously served as director and professor in the La Follette School of Public Affairs at the University of Wisconsin–Madison and associate director of the Institute for Research on Poverty and assistant professor at the University of North Carolina at Chapel Hill. She holds a doctorate in public policy studies from the University of Chicago.
E-mail: cheinrich@austin.utexas.edu

Public Administration Review, Vol. 72, Iss. 1, pp. 123–134. DOI: 10.1111/j.1540-6210.2011.02490.x.

Assessment Rating Tool (PART) in 2002 was to strengthen the process for assessing public program effectiveness and holding agencies accountable for results by making it more rigorous, systematic, and transparent. 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 shift in PART to a focus on assessing the performance of specific *programs,* rather than agencies, reflected the aim for a more "evaluative" approach. The PART questionnaire administered by the Office of Management and Budget (OMB) asked 25 standard questions in each of four topic areas to rate federal programs, with additional questions tailored to particular program types (e.g., competitive grants, block/formula grants, regulatory programs, etc.). The first set of PART questions addressed a program's purpose and design, essentially asking whether the government should be doing this activity (or operating this program) at all, which led to some objections that PART encroached on congressional authority (Radin 2006). The second section on strategic planning expanded on the GPRA in assessing whether the agency set appropriate annual and long-term goals for programs. The third section rated program management, including financial oversight and program improvement efforts. And the fourth set of questions, the hallmark of PART, was 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 was placed explicitly on the programs to justify a positive ("yes") rating with a superior standard of evidence. In addition, the emphasis on results was reinforced by the weighting of the sections, with accountability for results (the fourth section) contributing 50 percent to the calculation of the overall PART score.

In rating program effectiveness, the OMB accepted historical performance data, GPRA strategic plans, annual performance plans and reports, financial statements, and inspectors general's reports, but it also allegedly accorded higher ratings to programs that documented their effectiveness through randomized controlled trials, quasi-experimental methods, and/or longer-term, systematic tracking of outcomes. In fact, recommendations coming from early PART assessments were focused primarily on how program assessment methods could be improved to generate better data on performance rather than improving program performance itself (GAO 2005).

The Bush administration also conveyed (in an executive directive) that a "credible evidence-based rating tool" would ensure that federal programs receive taxpayer dollars *only* when they prove that they achieve results. Unless the use of performance information was linked directly to budgeting activities that drive policy development, argued the OMB, performance management activities would continue to have little impact on policy making and program results (Moynihan 2008).

While both the GPRA and PART aspired to move beyond "counting beans" in using data to make policy and management decisions, research that has compared the GPRA and PART along a number of dimensions points to trade-offs between an approach that engages political actors from both the 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 consequences for resource allocations (Moynihan 2008; Radin 2006). As Breul (2007) conveyed, PART was distinct from the GPRA in that it rendered a *judgment,* including "results not demonstrated."

This research focuses on the implementation and use of information produced by the PART, but with the intent to consider and inform the wider context and ongoing development of federal performance management efforts. More specifically, a central objective of this research is to assess the quality of evidence that agencies provided to the OMB in the PART assessments and to empirically examine relationships between attributes of the evidence and the PART ratings assigned. It is expected, in accord with OMB intentions under the Bush administration, that programs that used more rigorous methods of evaluation and produced better documentation of their results achieved higher overall and program results ratings. 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. Secondarily, this study also explores the relationship between the quality and rigor of evidence provided in PART assessments (and assigned performance ratings) and funding received by programs.

[A] central objective of this research is to assess the quality of evidence that agencies provided to the OMB in the PART assessments and to empirically examine relationships between attributes of the evidence and the PART ratings assigned.

The Barack Obama administration progressively is developing its own initiatives to revamp federal performance management efforts and replace PART. At the request of the OMB director, federal agencies have identified a limited number of high-priority, mission-oriented performance goals for which performance trends will be tracked. In addition, through the new Open Government initiative, the Obama administration intends to make high-quality data available to the public and to promote the use of new methods and technologies in analysis of performance data. It also is expected that the Obama administration will retain the focus on program evaluation that was central to PART, with more of the burden for evaluation likely directed at the agencies and away from OMB budget examiners (Newcomer 2010). These developments suggest that questions about the quality and rigor of the data and evidence supplied by the agencies (and to the public) will continue to be of central importance.

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 policy making. 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 show that some aspects of the quality of evidence submitted for PART reviews were significantly and positively associated with the PART ratings, but not with changes in program funding received. The results suggest limited success of PART but also some promise for the Obama administration's efforts to continue an emphasis on generating evidence of program results and using that evidence to increase support of programs that are "willing to test their mettle" (Orszag 2009).

In general, the study findings show that some aspects of the quality of evidence submitted for PART reviews were significantly and positively associated with the PART ratings, but not with changes in program funding received.

Review of Related Literature

The Development and Implementation 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 reintroduced over time) (Heinrich 2003; Pollitt and Bouckaert 2000; Radin 2000). Active policy and research debates continue along with these waves of reform, Light suggests, 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" (1998, 3). Indeed, the focus on producing "evidence" of government performance has intensified, in conjunction with the expansion of "evidence-based policy making"—that is, policies and practices based on scientifically rigorous evidence—beyond its longtime 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 study by the U.S. Government Accountability Office (GAO), for example, reported friction between the OMB and federal agencies regarding the different purposes and time frames of PART and the 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 (GAO 2005, 7). Some agency officials saw PART's program-by-program focus on performance measures as hampering their GPRA planning and reporting processes, and Radin (2006) noted that appropriations committees in Congress objected to the GPRA's concentration on performance outcomes and its deemphasis of information on processes and outputs.

Some analysts suggest that the development of PART was largely a response to the perceived failure of the GPRA to produce information on "what works" for guiding resource allocations and improving federal program performance (Dull 2006; Gilmour and Lewis 2006). Lynn (1998) argued that an unintended effect of the GPRA was to focus managers' attention on the procedural requirements (or the paperwork) of the reform rather than using the information to improve results. Others see PART as a performance management tool that built on the underpinnings of the 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 questioned this latter provision of PART, asking why the president would expend limited resources on an initiative such as PART, given the dismal record of past initiatives, but also because, if implemented as designed, it would bind the president to "a 'transparent' and 'neutral' instrument that would presumably raise the cost of making political decisions" (2006, 188). In effect, the administration would tie 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 pointed 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.

The implementation of PART as a tool for mechanically tying budget allocations to program results is also inconsistent with the conception of the budgetary process, described some time ago by Wildavsky (1964), as an expression of the political system. As Wildavsky articulated, a budget simultaneously may be viewed (by diverse stakeholders) as having many different purposes—for example, as a set of goals or aspirations with price tags attached, as a tool for increasing efficiency, or as an instrument for achieving coordination or discipline (among other things). Budgeting is also inevitably constrained by shortages of time and information, by reenactments and long-range commitments made in prior years, and by the sheer impossibility of reviewing the budget as a whole each year. In this context, Wildavsky argued, budgetary reforms, particularly those such as planning, programming, budgeting and zero-based budgeting, that aim to establish a mechanistic link between performance analysis and budget allocations inevitably will fail.

Still, recognizing the incontrovertible role of politics in performance management, the Bush administration entered its first full budget cycle (the fiscal year 2003 budget) with a commitment of significant staff time to developing a credible performance rating tool and an invitation for wide-ranging public scrutiny of the first PART questionnaire draft. According to Dull (2006), early input from the GAO, congressional staff, and other experts and internal advisory committee members led the Bush administration to modify or cut questions viewed as ideologically motivated. The Bush administration also had to confront challenges inherent in assigning ratings to programs in a consistent way, including problems with 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 2004 GAO study of the 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 (GAO 2004, 6). Among its recommendations, the GAO suggested that the 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 counted as an "independent, quality evaluation." It also suggested that the OMB should communicate earlier in the PART process with congressional appropriators about what performance information is most important to them in evaluating programs. The OMB subsequently generated supporting materials to aid agencies and PART examiners in implementing PART, including a document titled "What Constitutes Strong Evidence of a Program's Effectiveness?" that described different methods for producing credible evidence and the hierarchy among them in terms of their rigor.[3]

By January 2009, the OMB and federal agencies had assessed the performance of 1,017 federal government programs, representing 98 percent of the federal budget. In a 2008 survey of senior federal managers, more than 90 percent reported that they were held accountable for their results (OPM 2008). The OMB defined programs as "performing" if they had ratings of "effective," "moderately effective," or "adequate," with the last accounting showing that 80 percent of federal programs were performing.[4] Still, a November 2008 poll of the public indicated that only 27 percent of Americans gave a positive rating (good or excellent) of the performance of federal government departments and agencies.[5] Was PART really making a difference in how the federal government manages performance and, if so, in what ways?

Did PART Work?

In 2005, the OMB was honored with one of the prestigious Innovations in American Government Awards for the development of PART. The award's sponsor described the promising results that the OMB was achieving through PART, in particular, in encouraging more programs to focus on results.[6] The announcement noted that in 2004, 50 percent of federal programs reviewed by PART could not demonstrate whether 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, the OMB reported that 49 percent of programs were rated effective or moderately effective, while the number of programs with "results not demonstrated" had dropped to 17 percent. Only 3 percent of programs were reported to be ineffective, and 2 of the 26 no longer were being funded. In addition, of 127 programs that initially were rated "results not demonstrated," 88 percent improved their scores in a subsequent evaluation (Norcross and Adamson 2008).

These trends in PART ratings appeared to suggest that federal government performance was improving, as was 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 was credible and that the evidence presented by programs during the reviews was of high quality and reflected the achievement of federal program goals. Norcross and Adamson (2008) suggested that programs also could have been getting better at responding to procedural requirements associated with providing information to examiners or that the OMB could have relaxed its criteria. The public's significantly less positive view of federal program performance possibly suggests ignorance on their part of the performance information generated by PART or doubts of its veracity.

The same 2004 GAO report that criticized the early application of PART, however, also lauded PART for introducing greater structure into 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 PART also was 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 the performance management area, to complete the questionnaire. At the same time, contrary to the intent of 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's (2006) analysis suggested that in the absence of acceptable performance information, the 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.

Acknowledging the many reasons that poor program performance might not relate to budget outcomes, Gilmour and Lewis (2006) used information on PART performance ratings from 234 programs and the fiscal years 2004 and 2005 budgets to examine the relationship between PART scores and OMB budget decisions. They analyzed changes in PART scores from fiscal year 2004 to 2005, assuming that the political content of the programs would not change, to net out the influence of performance information (which presumably does change). Using the total PART score (and the change in the total score) as the key explanatory variable, Gilmour and Lewis reported positive statistically significant relationships between increases in PART scores and proposed program budget increases. Decomposing the PART score into its four parts, however, they found no link between the performance *results* scores and program budgets. One possible explanation they set forth was that too many programs at this stage had inadequate measures of program performance. They also suggested the importance of looking at the relationship between performance ratings and funding appropriations in future studies.

In a more recent study that included information on 973 programs, Norcross and Adamson (2008) analyzed data from the fifth year of PART to examine whether Congress appeared to use PART scores in making funding decisions. In their analysis, they were able to look at programs that were rated more than once and programs with ratings that improved over time; approximately 88 percent of programs first rated "results not demonstrated" subsequently improved their scores. Their analysis, based on simple cross-tabulations of PART ratings, proposed budget changes (the president's funding request), and congressional appropriations, suggested a tendency for the president to recommend funding increases for effective and moderately effective programs and decreases for ineffective and results not demonstrated programs, and a corresponding (but weaker) inclination of Congress to award higher budget increases to effective and moderately effective programs.

While the Norcross and Adamson study was limited to a simple descriptive analysis, Blanchard (2008), who also used data from a

larger number of programs (over four years, fiscal years 2004–2007), included controls for program type, home department, size, and a few political factors in a multivariate analysis relating PART results scores to budget proposals and appropriations. Looking at simple (increasing) correlations between results scores and changes in budget proposals and appropriations over time, Blanchard suggested that Congress was "easing its way into performance-based funding using the PART performance regime" (2008, 79). Like Gilmour and Lewis (2006), he acknowledged that it is not possible to fully model the political policy process of budgetary changes. Still, unlike Gilmour and Lewis, Blanchard found a positive relationship between PART results scores and congressional appropriations (as well as budget proposals) that was stronger and statistically significant in fiscal year 2006, which confirmed, he concluded, that Congress had "caught on" to how to use the PART results in budget decisions.

A 2008 GAO report alternatively suggested that there had been little progress in getting federal managers to use performance information in decision making. Based on the supposition that congressional "buy-in" to PART would be essential to its sustainability, Stalebrink and Frisco (2009) conducted an 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 hearing report comments, they concluded that PART information rarely was applied in congressional budgetary allocation decisions.

This discussion motivates the central focus of this study: was credible, high-quality information being generated in response to PART that accurately reflected program performance and progress toward program goals? And if so, was it influential in program decision making and in the allocation of budgetary resources? Alternatively, if the quality of information (and supporting documentation) on which program results were judged was weak, it presumably is not in the public interest for there to be direct, tight links between program performance ratings and programmatic or resource allocation decisions. 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.

This discussion motivates the central focus of this study: was credible, high-quality information being generated in response to PART that accurately reflected program performance and progress toward program goals?

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 (DHHS) for the PART process in the years 2002–2007. Although the OMB reported 115 PART reviews for the DHHS at ExpectMore.gov, this number included the 2008 assessment of four programs[7] that came after this study sample was constructed, as well as reassessments of some programs. The DHHS 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 (Heinrich 2007; Radin 2006). Indeed, the 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 not performing (27 percent) was above the average for all programs (20 percent). 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. The intensive nature of the work undertaken to construct new measures for analysis and limited 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 continues to maintain a "Program Performance Archive"[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 measures for this research. The OMB Web site 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 all can 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 whether internal data collection was used, and whether one or more comparisons were made to other programs
- Documentation/evidence of the types of performance measures used by the agencies—long-term, annual, and whether a baseline and/or targets were established
- Whether actual performance outcomes were reported

A basic description of the coding of the PART data to construct the foregoing measures is included in the appendix, as well as descriptive statistics for the variables used in the analysis. The reviews involved reading 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 by multiple researchers for each question asked in each of the four major sections of the PART.

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 data set assembled by Lee, Rainey, and 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. For example, Chun and Rainey (2005) discussed how ambiguity in organizational goals and directives and limitations in professional and managerial capacity contribute to challenges in specifying goals and measuring performance, making it more likely that public officials and managers will rely on measures of inputs, processes, and outputs rather than attempting to evaluate agency or program outcomes and impacts. In their study, they also found a relationship between financial publicness and goal ambiguity and described tensions between political needs for goal ambiguity and goal clarification that is critical to measuring performance outcomes. Their data were extracted from sources including the GPRA strategic plans, the Budget of the United States Government, Congressional Quarterly's Federal Regulatory Directory, and the Central Personnel Data File of the Office of Personnel Management for a sample of 115 agencies, although only 15 of those agencies are represented among the programs included in this study.

Descriptive statistics of variables from this data set that were used as control variables in the analysis are also presented in the appendix, albeit with the limitation that they are measured at the agency rather than the program level. 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 also was 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 attributable 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, as well as between the rigor of evidence (and results scores) and the 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) argued that the first three sections of PART, which are concerned with purpose, planning, and management, measure the extent to which federal programs produce the required paperwork under the GPRA. They also noted that some of the questions regarding program purpose are open to politicization (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 accordance with OMB guidelines that defined what constitutes strong evidence (i.e., information produced through random assignment experiments and quasi-experimental or nonexperimental methods with matched comparison groups),[10] I expect higher ratings for programs that provide quantitative evidence, that are independently evaluated, and that report longer-term measures of outcomes and establish explicit performance targets. These standards for strong evidence are applied in other contexts as well, such as by the U.S. Department of Education in its What Works Clearinghouse that was established in 2002 "to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education" (see http://www.whatworks.ed.gov/). The clearinghouse employs strict standards for evaluating program effectiveness, requiring a randomized trial, regression discontinuity methods, or a quasi-experiment (with equating of pretest differences) to meet evidence standards. This characterization is not intended to imply that qualitative evidence is always of inferior quality, but rather, it reflects the guidelines established for judging performance in the examiner program reviews.

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

H_1: 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.

C_{1a}: Results scores will be *positively* related to the number of question responses backed by quantitative evidence and to the number of question responses backed by external (or independent) evaluations.

C_{1b}: Results scores will be *negatively* related to the number of question responses backed by qualitative evidence *only,* to the number of question responses with no supporting evidence, and to the number of question responses backed by internal evaluations *only.*

H_2: 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 two corollary hypotheses apply in relation to overall PART scores, in addition to the following:

C_2: Overall PART scores will be *higher* for programs that use long-term measures of program outcomes, for programs that annually measure progress toward long-term goals, and for programs that establish baseline measures and targets for assessing performance.

Finally, taking into consideration prior research on PART and statements by former president Bush suggesting that better documentation of results was as important as improved performance, I test whether measures reflecting higher-quality evidence provided to demonstrate program results are positively related to funding received by the programs.

H_3: 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 the appendix show what types of evidence programs are more and less likely to offer in response to the PART questions on 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 percent and 7 percent, respectively). Only in response to the question asking whether independent evaluations of sufficient scope and quality indicate that 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 objectives. On the positive side, more than 90 percent of programs report regularly collecting timely performance information, and more than 80 percent 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 subgrantees, contractors, cost-sharing partners, and other government partners was coded as internal. The findings that nearly half of the programs did not make comparisons to other programs with similar purposes and that the independent evaluations conducted generated mostly qualitative evidence on performance are 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 frequently are not initiated by the programs themselves, and thus, some programs may have had an advantage in these performance reviews that does not reflect organizational efforts to improve performance evaluation (and even may 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 = .001), and for each question for which only qualitative evidence is offered in response, the results score is reduced by 0.047 (p = .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 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 an alternative specification, controls for the year in which the PART assessment was performed were added to account for the fact that the OMB clarified its definitions and expectations for evidence quality over time. More than half of the programs in this study were assessed in 2002–2004 (prior to the GAO review of PART). However, these indicators are not statistically significant (in this or subsequent models) and coefficient estimates of the key explanatory variables differ by less than one-hundredth (0.01), suggesting that the relationship between the rigor of evidence and results scores did not change substantially over the study years.[11]

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 because of 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 and capital asset programs receive significantly higher PART results scores.[12] 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 readily might construct plausible arguments to explain why research and development programs such as the National

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 $\alpha < 0.10$; **coefficient statistically significant at $\alpha < 0.05$.

Center for Health Statistics and those administered by the Agency for Healthcare Research and Quality and the National Institutes of Health 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 Centers for Disease Control and the National Institutes of Health, 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. 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 show 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 earlier, the first three sections of the PART questionnaire are concerned less directly with program results and more focused on 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 no evidence is provided, the overall PART score is reduced by

0.567 (or about 30 percent 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.[13] In addition, there are negative, 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 were probably based more on what programs reported they did 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) characterized as a measure of managerial capacity—is negatively related to overall PART scores, although only weakly. Indicators for the year in which the PART program was assessed (not shown) again did not add any explanatory power.

As discussed earlier, the "teeth" of PART (at least rhetorically) were supposed to be the budgetary consequences attached to the performance ratings, and the Bush administration emphasized that better documentation of results would be as important as high performance ratings themselves for program funding. This motivated a third hypothesis that changes in program funding following the PART assessments (fiscal year 2008) would 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—also were included the model predicting the change in program funding (in millions of dollars) from the year prior to the program's PART assessment to fiscal year 2008. The results (presented in column 1 of table 2) show that no 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

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
External evidence-results (#)	272	752	296
	(436)	(690)	(414)
No evidence-results (#)	614	248	422
	(853)	(604)	(680)
Qualitative evidence only-results (#)	–815	–896	–1060
	(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)
% management/analysis positions	–133	–143	–151
	(229)	(228)	(225)
Log # full-time employees	–3710*	–3244*	–3860*
	(2009)	(1620)	(2079)
Age of agency	74	61	74
	(69)	(62)	(69)
Congressional salience	15769**	17086**	16799**
	(4479)	(4861)	(5169)
No long-term measures	–2867	–3992	–3366
	(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 $\alpha < 0.10$; **coefficient statistically significant at $\alpha < 0.05$.

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 the more basic relationship that has been investigated in prior studies of PART: are changes in funding related to program performance, as measured by the PART ratings (holding other agency and program characteristics and congressional salience constant)? As discussed earlier, previous studies generated mixed findings on this question, with Gilmour and Lewis (2006) reporting no link between the performance *results* score and program budgets, and Blanchard (2008) finding a positive relationship between PART results scores and congressional appropriations that grew stronger over time. The model in column 2 of table 2 adds the program's score on section 4 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. The results of these regressions show no relationship between either the PART results (section 4) score and funding changes, or between the overall PART score and funding changes. In each of these models, 56 percent to 58 percent of the total variation in program funding changes is explained, apparently primarily by agency characteristics (agency size, evaluative ambiguity, and congressional salience). In addition, indicator variables for the year the program was rated were not statistically significant when added to the model (not shown in this table), offering no support for Blanchard's suggestion that Congress became more effective at using PART results in its budget decisions over time.

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. In all, 14 percent of the DHHS programs saw funding declines (by up to 100 percent, or a loss of all funding), and the others saw increases ranging from 0.5 percent to 44 percent. Approximately one-third of the programs realized funding changes (measured in millions of dollars) of ±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.

As Moynihan (2008) noted, while OMB staff maintained that partisan preferences did not affect PART assessments, they did acknowledge their influence on resource allocations. This is expected, added Moynihan, given that the PART "explicitly feeds into the highly political budget process" (2008, 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. Indeed, even if the PART was objective in reflecting the nature and quality of evidence on program performance in its ratings, one would be naive to expect the program ratings to have a direct or mechanical link to program funding, given the political nature of budget decisions and the many constraints imposed by limited information, long-range budget commitments, and reenactments on budgetary allocations.

Conclusion and Implications

The Bush administration's Program Assessment Rating Tool 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 present a mixed review of PART's effectiveness. The empirical analysis using data on 95 Department of Health and Human Services programs with newly constructed measures to evaluate the quality and rigor of evidence supplied by the programs showed some consistent, 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. Although the ratings of program results and the overall PART scores had no discernible consequences for program funding over time, the apparent seriousness with which evidence provided by the programs was evaluated is a positive step forward, particularly given the current administration's stated focus on "building rigorous evidence to drive policy" (Orszag 2009).

The Obama administration has made clear its intent to continue efforts to strengthen program evaluative capacity, and it has signaled its commitment with an expansion of the Institute for Education Sciences and increases in evaluation budgets of federal agencies such as the Department of Labor and others (Orszag 2009). Like the Bush administration, it also is calling for more experimental evaluations and other rigorous methods of performance analysis, including initiatives at the Department of Health and Human Services that are promoting more rigorous and systematic use of performance data. And very similar to the language in the PART initiative, former OMB director Peter Orszag (2009) explicitly stated that they are "providing more money to programs that generate results backed up by strong evidence."

> In the context of increasing public demands for accountability that include high-stakes pressures to demonstrate performance improvements, policy makers frequently have little choice but to consider and use a mix of different types of information and methods in producing annual performance reports.

Thus, while PART clearly is being replaced, key elements of the PART initiative are still intact in the new administration's performance management efforts. Correspondingly, the Obama administration likewise will have to confront some of the same challenges in producing more credible evidence and reliable knowledge of program outcomes, such as the not infrequent incompatibil-

ity of these efforts with the requirements of other performance management initiatives such as the 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, policy makers frequently have little choice but to consider and use a mix of different types of information and methods in producing annual performance reports.

This, of course, begs another question: what other objective information and factors, besides program performance, should influence funding allocation decisions? For example, are there substantial, measurable factors that contribute to congressional salience that could be made more explicit or transparent in funding allocations? For purposes of accountability, the public should, at a minimum, understand how much influence program performance really has on funding decisions and what the trade-offs 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 and accessible way) and allowing for full disclosure, as President Obama has promised, would be another step forward. As Light's (1998) analysis suggests, while PART is now history, similar policy tools or reforms, even beyond the Obama administration's efforts, surely will follow, as the public interest in seeing evidence of government performance is unlikely to abate anytime soon.

Appendix: Newly Constructed Measures and Variable Descriptives

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

Acknowledgments

I thank the University of Wisconsin–Madison for support of this research through the Regina Cawley Loughlin Scholar funds and Maureen Quinn and Samuel Hall for their dedicated research assistance. I also thank the manuscript referees and participants of the 2009 Public Management Research Conference for their very helpful comments.

Notes

1. For a history of the National Partnership for Reinventing Government, see http://govinfo.library.unt.edu/npr/whoweare/history2.html (accessed September 15, 2011).
2. See http://www.gpoaccess.gov/usbudget/fy04/pdf/budget/performance.pdf (accessed September 15, 2011).
3. See http://www.whitehouse.gov/omb/assets/omb/performance/2004_program_eval.pdf (accessed September 15, 2011).
4. See http://www.whitehouse.gov/omb/expectmore/rating/perform.html (accessed September 15, 2011).
5. "In the Public We Trust," Partnership for Public Service and Gallup Consulting, November 2008.
6. See http://www.innovations.harvard.edu/awards.html?id=7496 (accessed September 15, 2011).
7. This information is from August 2009; the programs newly rated in 2008 included Centers for Disease Control Division of Global Migration and Quarantine, Health Information Technology Research, Office of Medicare Hearings and Appeals, and Substance Abuse Drug Courts.
8. See http://www.whitehouse.gov/omb/performance_past/ (accessed September 15, 2011).
9. The other two researchers, besides the author, who coded the data were Maureen Quinn, formerly an analyst at the Government Accountability Office and a legislative analyst at the Wisconsin Legislative Audit Bureau and now at the Allegheny County Economic Development Office in Pennsylvania, and Sam Hall, formerly a researcher at the Urban Institute and now a law student at the University of Michigan.
10. See http://www.whitehouse.gov/omb/assets/omb/performance/2004_program_eval.pdf (accessed September 15, 2011).
11. The results with PART assessment year indicators for all models presented in the tables of results are available from the author upon request.
12. The number of observations in the models with agency characteristics drops to 89 from 95 because of 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.
13. Again, the measure of the number of questions for which internal evidence was provided was excluded because of multicollinearity problems that emerged after adding additional variables to the model.

References

Blanchard, Lloyd A. 2008. PART and Performance Budgeting Effectiveness. In *Performance Management and Budgeting: How Governments Can Learn from Experience,* edited by F. Stevens Redburn, Robert J. Shea, and Terry F. Buss, 67–91. Armonk, NY: M. E. Sharpe.

Breul, Jonathan D. 2007. Three Bush Administration Management Reform Initiatives The President's Management Agenda, Freedom to Manage Legislative Proposals, and the Program Assessment Rating Tool. *Public Administration Review* 67(1): 21–26.

Chun, Young Han, and Hal G. Rainey. 2005. Goal Ambiguity and Organizational Performance in U.S. Federal Agencies. *Journal of Public Administration Research and Theory* 15(4): 529–57.

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

Frederickson, David G., and George H. Frederickson. 2006. *Measuring the Performance of the Hollow State.* Washington, DC: Georgetown University Press.

Gilmour, John B., and David E. Lewis. 2006. Does Performance Measurement Work? An Examination of the Office of Management and Budget's PART Scores. *Public Administration Review* 66(5): 742–52.

Heinrich, Carolyn J. 2003. Measuring Public Sector Performance and Effectiveness. In *Handbook of Public Administration,* edited by B. Guy Peters and Jon Pierre, 25–37. Thousand Oaks, CA: Sage Publications.

———. 2007. Evidence-Based Policy and Performance Management: Challenges and Prospects in Two Parallel Movements. *American Review of Public Administration* 37(3): 255–77.

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

Lee, Jung Wook, Hal G. Rainey, and Young Han Chun. 2009. Of Politics and Purpose: Political Salience and Goal Ambiguity of U.S. Federal Agencies. *Public Administration* 87(3): 457–84.

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): 231–37.

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

Newcomer, Kathryn. 2010. Putting Performance First—A New Performance Improvement and Analysis Framework. In *Framing a Public Management Research Agenda,* edited by Jonathan D. Breul, 7–16. Washington, DC: IBM Center for the Business of Government.

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. Working paper, Government Accountability Project, George Mason University.

Orszag, Peter R. 2009. Building Rigorous Evidence to Drive Policy. *OMBlog,* June 8. http://www.whitehouse.gov/omb/blog/09/06/08/BuildingRigorousEvidencetoDrivePolicy/ [accessed September 15, 2011].

Osborne, David, and Ted Gaebler. 1992. *Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector.* Reading, MA: Addison-Wesley.

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

———. 2006. *Challenging the Performance Movement: Accountability, Complexity, and Democratic Values.* Washington, DC: Georgetown University Press.

Sanderson, Ian. 2003. Is It "What Works" That Matters? Evaluation and Evidence-Based Policy-Making. *Research Papers in Education* 18(4): 331–45.

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

U.S. Government Accountability Office (GAO). 2004. *Performance Budgeting: Observations on the Use of OMB's Program Assessment Rating Tool for the Fiscal Year 2004 Budget.* Washington, DC: U.S. Government Printing Office. GAO-04-174.

———. 2005. *Performance Budgeting: PART Focuses Attention on Program Performance, but More Can Be Done to Engage Congress.* GAO-06-28.

———. 2008. *Government Performance: Lessons Learned for the Next Administration on Using Performance Information to Improve Results.* Washington, DC: U.S. Government Printing Office. GAO-08-1026T.

U.S. Office of Personnel Management (OPM). 2008. Senior Executive Services Survey Results. http://www.opm.gov/ses/SES_survey_results_complete.pdf [accessed September 15, 2011].

Wildavsky, Aaron. 1964. *Politics of the Budgetary Process.* Boston: Little, Brown.