Playing the Wrong PART: The Program Assessment Rating Tool and the Functions of the President's Budget

The exhaustive literature about the George W. Bush administration's Program Assessment Rating Tool (PART) has not emphasized an issue that appears quite clearly in its terms with Office of Management and Budget (OMB) and congressional staff. Budget analysts can pursue efficiency in two ways. One, the object of traditional analysis, involves assessing programs to understand the relative costs of programs to the agency, and then the effects of worse or less funding. Another approach, termed "budgeting for results," measures program performance so as to eliminate funding programs in a way that increases social welfare. The second approach is much more difficult because it necessitates comparisons of programs of different programs, requiring expertise that often does not exist and is much more easily contained as involving others rather than facts. Both congressional and OMB sources report concerns that PART weakens budgetary analyses by diverging resources from traditional analyses. If one of the goals of PART is to provide analysis that will influence Congress, the focus should be strengthening the capacity to do traditional budget analysis.

What is the role of the president's budget within federal budgeting? Particularly, the role of formal postenrollment budget process, as managed by the Office of Management and Budget (OMB), contribute to the success or failure of the federal budget process?

In the summer of 2008, I decided to revisit this question, which I had studied mostly under four different administrations (1993-2008). In preparation to reviewing the literature produced over that period, I conducted new interviews with 22 mostly senior officials and former officials who had chiefly observed and usually quite extensive experience in the budget process. Among them were two people who had served as director of OMB; two directors of budget review (one from the Clinton and one from the OMB); a Secretary of Commerce who had served as division associate direct (one of the eight civil servants who report directly to the politically appointed program associate director); and two who had served as clerk (staff director) of the House and Senate appropriations committees. Of the seven congressional sources, five had served as the OMB.

These interviews were part of the basis for an overview article (White 2009). But they also provided perspective on a question that has provided substantial controversy in recent public administration and budgeting literature: How can the president use PART to increase the efficiency and the effectiveness of the president's programs in a way that increases social welfare? There are many reasons for why PART was worth the time spent.

There are important reasons for understanding both the role of the president's budget within federal budgeting and the ways in which the budget process may increase the efficiency of federal programs. Presidents and their staffs are also supposed to ensure that budgetary costs only in a way serve the president's interests but would not have little influence separate from the president's veto of the veto and as a party leader if it did not serve some useful function for Congress as well (Caudill 1987; Mosher 1984; Robin 1994; White 1996). The OMB can contribute in a way that is seen as useful even across party and institutional lines by providing a range of analyses that could improve the president's budget or the consequence of decisions. Information about the effects of proposed changes to agency budgets can help decision makers pursue a better scale of outcome to inputs for a given agency. "Scrutinizing the estimates at the agency level is the core function of traditional budget analysis.

As an approach to budgeting, PART in particular and "budgeting by results" in general emphasize comparisons across units rather than allocations within them. The assumption is that programs with better results should—multiple caveats aside—be favored; a portfolio of better-performing results will result in a more efficient government. Yet comparison across programs is much more problematic than evaluation within programs because programs serve different goals. HowPART actually was implemented gave PART users further reason to discount the analyses. At the same time, they complained that they were not getting from the OMB the traditional analysis that they vended and needed. The OMB sources argued that the work needed for PART diversified resources from traditional analysis.

The next part of this essay describes the concerns about the usefulness of PART. The following section explains how these concerns fit with an historical perspective on the purposes of the executive budget and how budget analysis can be used to increase efficiency.

The assumption is that programs with better results should—multiple caveats aside—be favored; a portfolio of better-performing programs will result in a more efficient government. Yet comparison across programs is much more problematic than evaluation within programs because programs serve different goals.

After a widely criticized first effort for the fiscal year 2003 budget, the OMB in 2002 designed a new instrument, with feedback from employers such as the Council of Accountants for Government and the National Academy of Public Administration, that first was implemented during the budget review in 2002 for the fiscal year 2004 budget submitted to Congress. The OMB's "PARTed" each year, so by 2007 (the fiscal year 2008 budget), 576 programs representing 86 percent of the budget had been assessed (Redburn et al. 2008).
Under PARt, the OMB examined were to rate programs as "effective," "moderately effective," "inadequate," "ineffective," or "could not be rated." OMB's rating scheme, in which the size and complexity of the assessment of whether the program design was logical and purpose clear, 10 percent of whether the agency had a strategic plan with valid performance measures. The OMB's review focused on the programs' management, such as financial oversight, and 50 percent of the ratings of performance based on measures of stated goal achievement. This design could be criticized for evaluating programs in part on standards that were not performance at all. Critics argued that some of the measures, such as criteria of goals, were biased against certain types of programs, such as block grants, for which part of their impact was difficult to precisely measure. (Nunnan 2008; Radin 2008; Sheehan 2009). Even the idea that agencies should be held responsible for performance may be questioned for programs in which the federal government is intended not to have an active role, such as grants in the accountable chain is designed to be weak (Nunnan 2008).

Logically, "effective" had to mean something similar for all programs; the basic categories reflected a view that certain questions are relevant to assessing any program. At the same time, programs differ, and supporters of PAR argued that it recognised this fact better than traditional criteria because it assessed across several categories of the programs (Blanchard 2008). The categories included block/fee-for-service, capital asset and service acquisition, competitive grants, crisis response/relief programs, disaster, and research and development. Starting with a set of 25 questions that were to be asked of direct programs, the instrument provided further questions that could be asked of different kinds of programs. However, one could say that some of the targeted 25 questions were "not applicable" to a program, about 95 percent of the score for a block grant, competitive grants, or credit program would be based on the OMB's review of the score of the program. And 80 percent of the score for a capital program, 83 percent of the score for a regulatory program, and 77 percent of the score for research and development programs would be based on the same measures used for direct programs.

Hence, the instrument's "flexibility" depended greatly on the decision by the OMB's program managers. The examiner would determine the measures used to answer the questions. Thus, in the view of one respondent within the appropriate process, "the problem early on was they basically let individual OMB examiners devise it any way they wanted." Other research has reported the same pattern from the agency perspective. As a result, it was possible for examiners to work with five different examiners, 14 programs, 19 program reviews with re-PARTs, and 60 measures, plus 50 linkages with the OMB. It is amazing to note how they take the guidelines and look at it differently ... it is most clearly dependent on the examiner. That is a serious problem with the PARt." (Page 2009, 32). John Gilmour provided another indication that the PARt advocates believed that having the standards published made the process "transparent," but a negotiation between an OMB examiner and his or her supervisor has the final say. And if the PARt is not, to outsiders, an open program. As OMB examiner who has written in favor of PARt emphasized that "you or anyone else can game the system and make it look as seen seen as OMB examiners. But there is little reason for veteran staff in Congress, or legislators, to credit such broad judgments when made by individual examiners. The GAO report that the analysis of consultation between the OMB examiner and his or her supervisor is necessary because the examiner is dependent on the evidence used for analysis or set priorities as to which programs to analyze was one reason for "most congressional staff we spoke with not using the PARt information" (Green 2008, 19).

Evidence about limitations of PARt PARt contained two aspirations for performance budgeting. One was that it could not only promote accountability but also drive accountability that poor performance could lead agency management to commit to change. PARt did give agency staff strong incentives to assess their program's performance relative to others. As one OMB examinee reported, need to be downgraded on "their management scorecard" within the President's Management Agenda for having no rating. But an incentive to agree on measures is not the same as an incentive to improve performance. The GAO's surveys of federal managers on performance and management issues show that PARt likely constituted an increase in the number of performance measures available within the federal government. The same surveys, however, show no change at all a decade between 1997 and 2008. Federal managers reported large differences in PARt scores: some cut a significant proportion of measures, whereas others did not. Federal managers reported large differences in PARt scores: some did not create a culture of managing for performance. What it did do was give managers at the upper levels of agencies strong incentives to make their programs look good on the measures.

When Gilmour studied agencies that raised their ratings, he found that they "explained their success with almost identical language. 'We learned how to take the test,' they said. They learned 'how to write more acceptable answers' and 'how to develop performance measures that would fit into the PARt' (Green 2008, 23)." The evidence does not support the hope that management of agencies would become more information in the future. Of the distinctions of the difficulties, see Kasid 2010.

The second goal was that measurement of results could provide "decision-makers with information needed to better allocate scarce resources in a way that will yield the greatest benefit" (Gilmour 2008, 23). Did this mean that performance programs that received low ratings were to avoid scarce resources on this point? (Beryl 2007; Durl 2006). Nevertheless, PARt clearly was designed from the beginning that programs needed to be challenged, and it was implemented by the administration's allies as a way to cut spending. The OMB guidelines stated that the program "creates a high standard of evidence and it will likely be more difficult to justify a "no" vote than a "yes" vote" (Gilmour 2008, 23). Lloyd A. Blanchard, who as program associate director for the General Government Division of the OMB in 2001 was an early advocate of PARt, said that "the PARt was an attempt to make agencies seek to 'enhance' their budgets. Because there is no natural incentive to promote efficiency and effectiveness in government program service delivery," through PARt, "an incentive can be created by reducing performance scores for agencies whose programs which would have caused the PARt, and punishing poorer performers with smaller increases or decreases in budgetary allocations" (Blanchard 2008, 2). In short, "budget maximization, not performance maximization." Appropriations officials who believed that the purpose of the program within the Bush administration was mainly to create rationales for cuts had good reason for their perception.

Did PARt affect allocations among programs? The most positive respondent reported that the PARt analyses were "used for presenting the budget to the DemocraticCongress...and also in their own program budgets which were then usefully modified for the OMB, but was used and became part of the overall budget, because it was relevant," and the question, however, of whether PARt scores were driving analysis or simply reflected pre-existing patterns within the OMB was not asked. (Green 2008, 71). One use of the PARt scores was to support the OMB's recommendation. If the OMB recommended that a program was not the same as an incentive to improve performance. The GAO's surveys of federal managers on performance and management issues show that PARt likely constituted an increase in the number of performance measures available within the federal government. The same surveys, however, show no change at all a decade between 1997 and 2008. Federal managers reported large differences in PARt scores: some cut a significant proportion of measures, whereas others did not. Federal managers reported large differences in PARt scores: some did not create a culture of managing for performance. What it did do was give managers at the upper levels of agencies strong incentives to make their programs look good on the measures.

When Gilmour studied agencies that raised their ratings, he found that they "explained their success with almost identical language. 'We learned how to take the test,' they said. They learned 'how to write more acceptable answers' and 'how to develop performance measures that would fit into the PARt' (Green 2008, 23)." The evidence does not support the hope that management of agencies would become more information in the future. Of the distinctions of the difficulties, see Kasid 2010.

The second goal was that measurement of results could provide "decision-makers with information needed to better allocate scarce resources in a way that will yield the greatest benefit" (Gilmour 2008, 23). Did this mean that performance programs that received low ratings were to avoid scarce resources on this point? (Beryl 2007; Durl 2006). Nevertheless, PARt clearly was designed from the beginning that programs needed to be challenged, and it was implemented by the administration's allies as a way to cut spending. The OMB guidelines stated that the program "creates a high standard of evidence and it will likely be more difficult to justify a "no" vote than a "yes" vote" (Gilmour 2008, 23). Lloyd A. Blanchard, who as program associate director for the General Government Division of the OMB in 2001 was an early advocate of PARt, said that "the PARt was an attempt to make agencies seek to 'enhance' their budgets. Because there is no natural incentive to promote efficiency and effectiveness in government program service delivery," through PARt, "an incentive can be created by reducing performance scores for agencies whose programs which would have caused the PARt, and punishing poorer performers with smaller increases or decreases in budgetary allocations" (Blanchard 2008, 2). In short, "budget maximization, not performance maximization." Appropriations officials who believed that the purpose of the program within the Bush administration was mainly to create rationales for cuts had good reason for their perception.

Did PARt affect allocations among programs? The most positive respondent reported that the PARt analyses were "used for presenting the budget to the DemocraticCongress...and also in their own program budgets which were then usefully modified for the OMB, but was used and became part of the overall budget, because it was relevant," and the question, however, of whether PARt scores were driving analysis or simply reflected pre-existing patterns within the OMB was not asked. (Green 2008, 71). One use of the PARt scores was to support the OMB's recommendation. If the OMB recommended that a program was not
For years, the Committee has directed departments and agencies to improve the budget justification document quality and presentation. The most recent OMB recommendation was to improve the analysis and information. While the Committee has seen some improvement, much more justification is needed to be done properly. The President’s Budget in a Nutshell Program (PART), introduced in 1992, and still used for budget justification, is too limited. The Committee finds little use for a budget justification which does not review specific details of the measurable indicators and success used to evaluate a program’s performance, relevance, or adherence to underlying authorization statutes. Further, the Committee has less patience for a program’s funding level, rather than the "PART score," getting to green; or "this is what OMB provided." (U.S. House 2008)

Menzenbaum’s respondents made much the same points. As one put it, "No one has a problem with the agency showing the overall picture, but we cannot find information that tells us this is how they spend the money and how they are spending the budget in the current year if appointed. Without that we’re not convinced. We have to respect the critical role of what the budget is, and how you would propose to spend it, which would be part of a conversation about how you would use the money. We want to know how you plan to control your budget and what performance data you have on your projects.

The OMB veterans shared the concern about the opportunity costs of PART. The basic question may be phrased as, does PART improve budget analysis? The overall answer from my OMB respondents is that it could make a contribution, but not if it displaces traditional analysis, and not particularly if the examiners are knowledgeable to begin with. The earlier quote about PART being useful for less experienced examiners expresses part of this perception. Another source called PART "a structured approach to doing something good examiners were doing all along." Yet the more common response about PART was that, in one senior examiner’s official words, it “really a lot of a cost.” This was viewed as a significant cost.

"One thing I used to do,” one division associate director recalled, "was sit down with examiners and identify two or three things we would need to learn more about from March to September … we could not do the same thing year after year. We would check up every month between February and September.” In the late 1980s, IRC’s use of funds and amount of time spent dealing with Congress was viewed as making it difficult for examiners to learn about the actual operations of agencies (White 1991). By 2008, the OMB was much less directly engaged with Congress (White 2009), but, in one career veteran’s words, "there has been an expansion of legislative tasks, and the results is that examiners do not do what they are supposed to do -- even at the last minute!" The trouble with PART for appraisers is that it does not answer these questions. As the House Appropriations Financial Services and General Government Subcommittee complained in the 2008 report on oversight, "We are still suffering minimal progress in the area of oversight.

PART and other management initiatives, therefore, diminished the capacity to do traditional budget work. The problem was made worse when, in response, examiners were told to spend a lot more time on PAR and less time on their traditional duties. But when this happened, examiners found that they had a lot more time to do other things. By 2008, when the executive budget process was over, it served not only to advertise for the President’s budget but also to help Congress do its own analysis.

Appropriators might be able to convince their colleagues in Congress that Agency X does not need an increase, or can live with a 2 percent increase or cuts that can be absorbed by a field program, and that they will need to do some hard work to think about the operational needs of Agency Y. Traditional analysis asks for a micro-level perspective; in essence, what would it mean to have any of the programs cut by 2 percent? As such, traditional analysis would be the key to meeting the requirements of a budget that is passed by an agency? (Taniu, O’Neill, and Hohen 2005). It is quite another matter to give two programs PART scores and say that one is more deserving than another because it earned a higher rating, especially when the scores are adjusted by the performance of the program. The scores are assigned by individual budget examiners whose authority of expertise (never miss) is not recognized by outsiders.

The mismatch between executive analysis and congressional needs was evident to agency budget officers. "What is an agency to do when the OMB insists that it emphasize performance information, including budgetary information that the appropriators want traditional justification?" (Kamensky 2007). The answer, as the director of the Budget Service for the Department of Education presented it, was to give the appropriators the information they requested. Performance information could be included in narratives, as it normally has been when agencies had some, but with descriptions of the contextual and policy issues that accompanied the "performance budget" information be submitted in a separate tab (Shelly 2006). Appropriators will always take more information, and it is credible and helps them meet the political challenge of passing their bills. But even my OMB respondents, who recommended PART, believed that PART should be a supplement to traditional budget analysis, and that means that, if PART significantly reduces the time needed for traditional analysis, it is not a good investment.

The most critical reviews of PART have argued that it involves an approach to budgeting that ignores politics. PART advocates, Beryl J. Forman, Palmore, and business advocate II, are attracted to the dimensions of budgeting that flow from the immediate management approach, including a reliance on the technical skills of the budget staff, a belief in a national resource allocation model, a belief that information is neutral and can be determined to be true or false, a search for optimiz- ation or self-limiting approaches, an emphasis on efficiency norms, a tendency to use the private or business sector as a model, and a strong reliance on the executive branch of government. (2008, 116–17)

These criticisms are all significant, but the argument that Radin iden- tifies should not be interpreted in the meaning that the executive cares about efficiency norms and the legislature does not.

The rationale for PART tells the federal government as a portfolio of programs and refines the budgeting question as, in "which of these programs or initiatives should I fund?" As a result, the analysis switches from how resources would be used within an individual program to comparisons among programs, however, it seems to be value neutral. Because almost all programs serve
different values, efficiency comparisons across programs are use, in practice, value choices. Not only are such comparisons easily criticized, but also the execution of such comparisons is significantly more difficult than evaluation of an agency's work plan. Congress, naturally, is not interested in analysis that creates value judgments based on a framework that looks unrealistic but quite desirable. If one looks at the details, it is based on individual budget examiners’ excessive discretion.

The Budget Bureau's Traditional Contribution to a More Rational Budget Process

I am not suggesting that individual members of Congress always are pleased to be told that their pet programs cannot use more money. Thus, Congress has pushed the presidential budget organization and the legislative budgetary—particularly the budget bureau and the appropriations—because the budget process is one battle in a contest for control of federal agencies. Yet the appropriate, if not the necessary, budget bureau is developed particularly over its first 50 years served this congressional interest in countering agency and agency advocates’ claims about how much their programs “need.”

This is not the place for a full review of the history of the federal government's executive budget (good monographs include Bertha 1977; Berthoud 1984; Teichman 1978.) For a congressional perspective on the 1921 act, see U.S. Senate 2008; for other sources, see White 2009. As Colvin (1987) argued, the core issues of the executive bureau and the congressional bureau—is a fiscal policy or a policy design, yet the procedure actually assessed better fit that design. The president received new and significant “fiscal maneuver” powers, but Congress retained its ability to “disuse” of the presidential proposals.

The president's budget helped Congress by deflecting some blame. If the president proposed cost-cutting that Congress was willing to accept, legislators could accept the proposal and blame him. Even if it only reduced their share of the blame, that was better than having the blame all on themselves. If Congress rejected a presidential proposal, and so had to replace it with another in order to keep spending levels, legislatures at least could get some credit from the interests they protected.

The second contribution made by the presidential budget process was more technical: it provided oversight of agency proposals and thus a way to “scrutinize” them to reduce “fat and increase efficiency within programs. As Simon (1997) and Valkovsky 1966 both explained, “efficiency” was not the same as “economy,” the goal of simply reducing spending. They defined efficiency as the ratio of outputs to inputs, so that it could be increased either by raising the output for a level of inputs or by lowering the input for a level of output. Efficiency would increase even more than inputs, making the organization less “efficiency.” In this sense it is not a politically biased concept, and so it can be useful to political actors who otherwise might be in conflict with each other.

As Schick explained, “expedient for running operations and services are...the accusation that budget makers look to short-term savings to avert reversals and expenditures” (2004, 96–97). In the United States, this has been true of both the budget bureau and the appropriations committees, so if the budget bureau can help identify such efficiencies, that is often a service to Congress. Different politicians would put programmatic efficiencies to different use, but this only shows that increasing programmatic efficiency is the closest thing to a neutral competence. Thus, in my interview with two decades ago, a senior former official said that “the ideas I gave up were a central competence budget analysis. As I told one staff, that meant if it was a Republican administration trying to minimize cost, it was a Democratic administration, how to maximize value for the money we had.”

Even now, appropriators will say that having a presidential review is necessary. The best example of why is the National Cancer Institute’s “Bayes budget.” The former designated to subject various program’s priorities directly to Congress. When asked directly whether that is useful, an appropriator replied, “No, it’s just too expensive, nobody can use it.”

The budget bureau across its history sometimes has pursued another way so improve efficiency within agencies: improving agency management. The budget bureau itself has turned the bureau into a neutral budgetary competence, however, is particularly controversial (Berman 1977; Dickman and Rudelik 2007; Monbaron 1984; Posen 2007; Radia 2008; Stone 1950; Tanaka, O'Neill, and Holom 1983; Tanok 1998). The problem arises, how to define a plausible management function would require another article. Because my focus was not on budget formulation, it was not a topic in any interview, it is therefore difficult to believe, though it appears that DPC did much to improve federal management}

The best that can be said is that it increased the number of measures (GAO 2008; Gilsa 2008; Mezene 2009).

The process of proposing and then appropriating budgets for programs that are now called “discretionary” programs is not, of course, the only part of federal budgeting that matters. It has become a smaller part, in both dollar and procedural terms, of budgeting for the past few decades (Schick 1980; White 2009; White and Wilder 1991).

Nevertheless, throughout the past decades of conflict and procedural change, a large part of the Bureau of the Budget and its allocations of discretionary spending to individual agencies. It remains by far the largest part of the OMB’s activity. Throughout a quarter century of interviewing the appropriators committees, appropriations usually have expressed to me the interests in having a president’s budget that provides information that they can use to figure out how to get the most value for money out of individual agencies. Part did not help.

Conclusion

So what, now, for PART in particular, and for “budgeting for results” in general?

PART now has been added to what Blinder, Kraen and Reifer point to the “long, irregular history” of performance budgeting in the United States, a history, thus, as of 2003, had “failed to make a substantial impact on the budget process” (2003, 31). I have argue that the fundamental problem of PART, as of all such measures, is it seeks to analyze program’s overall efficiency to as to compare programs. Such comparisons are not useful and therefore are not used by Congress—because they inevitably involve value judgments that legislators will not accept from the OMB, and because the kinds of measures and expertise required are extremely easy to question. In the case of PART, the outcome measures were (and, I believe, in some cases) that increasing programmatic efficiency was not effective approach in programmatic metrics varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obviously appropriate to programs metric varied greatly on functional factors such as whether they actually measured the impact of the various other questions were applied even if not obvious

Acknowledgments

This article uses unattributed quotations from open-end interviews with participants in federal budgeting, all of which were conducted on condition that I would not quote nor attribute views to any individual without direct written permission. In addition to the interviews conducted in 2008, my understanding of the role of the program budget has been informed by interviews conducted by interviews conducted in 1989 and 1990 with 22 OMB staff and 11 departmental or agency budget officers. Since 1983, I have conducted interviews with 300 or more managers or staff in the budget bureaus and appropriations processes. About a quarter of these have been with members of Congress the rest were with a mix of execu

References


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 in the Office of Management and Budget to the application of the Program Assessment Rating Tool (PART), introduced in 2002 to score rigorously, systematically, and transparently many public programs effectiveness and hold agencies accountable for results by tying them to the executive budget formulation process and program funding. Evidence submitted by 95 programs administrated 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 less programs offering some evidence, most was internal and qualitative, and fewer half of not some lose 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 in long-term and annual performance goals. Importantly, almost 80% of 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 assessment 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; Rabin 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 reinvigorated teams (internal and government-wide), reinvented laboratories within agencies, town hall meetings and reinvigoration summits, and new legislation to mandate performance management at the federal level (Kamensky 1999).

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 disputes have that there is a definitive transformation in federal government capacity and infrastructure for managing for results, that is, in use of outcome-oriented strategic plans, performance measures, and reporting of results (GAO 2008). Yet same in-depth assessment of the implementation of the GPRA also have been highly critical. Several researchers have suggested that overlapping results-oriented management 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; Rabin 2000, 2006). Rabin (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 Assessment Rating Tool (PART) was to address some of the criticisms of the GPRA. First, the PART was designed to measure the quality of evidence provided by agencies and to hold agencies accountable for results by tying them to the executive budget formulation process and program funding. The PART assessors are to use a rigorous, systematic, and transparent approach to evaluate programs’ performance, providing a more comprehensive and objective assessment than the GPRA. The PART aims to improve government performance by focusing on results, rather than just compliance with regulations. By doing so, the PART is intended to promote a culture of continuous improvement and performance excellence within government agencies.