Robert M.
La Follette School of Public Affairs
at the University of Wisconsin-Madison

Working Paper Series
La Follette School Working Paper No. 2005-021
http://www.lafollette.wisc.edu/publications/workingpapers

What Do We Talk About When We Talk About Performance? Dialogue Theory and Performance Budgeting

Donald P. Moynihan
La Follette School of Public Affairs, University of Wisconsin-Madison

dmoynihan@lafollette.wisc.edu

Robert M. La Follette School of Public Affairs
1225 Observatory Drive, Madison, Wisconsin 53706
Phone: 608.262.3581 / Fax: 608.265-3233
info@lafollette.wisc.edu / http://www.lafollette.wisc.edu

What Do We Talk About When We Talk About Performance?
Dialogue Theory and Performance Budgeting

Donald P. Moynihan,
La Follette School of Public Affairs
University of Wisconsin-Madison
E-mail: dmoynihan@lafollette.wisc.edu, Phone: (608) 263 6633

What Do We Talk About When We Talk About Performance? Dialogue Theory and Performance Budgeting

Abstract

This paper examines the Program Assessment Rating Tool (PART) in the federal budgeting process. The early evidence on PART prompts the search for a theory of budgeting that accepts that performance information will influence decisions, but will not be used in the same way from decision to decision, as the espoused theory of performance budgeting suggests. Dialogue theory emphasizes the ambiguity of performance information and related resource allocation choices. An exploratory test of dialogue theory is undertaken through an experiment involving graduate students assessing PART evaluations. The results illustrate a variety of ways in which different individuals can examine the same program and, using logical warrants, come to different conclusions about performance and future funding requirements.

Introduction: The Return of Performance Budgeting

The President's Management Agenda (PMA) of 2001 heralded the most recent effort to introduce performance budgeting to the federal government, based on the seemingly indisputable premise that "everyone agrees that scarce federal resources should be allocated to programs that deliver results" (OMB 2001, p.27). The PMA called for the integration of financial and performance information by increasing the quality and range of data available to decision-makers, assuming that greater technical and allocative efficiency would result (OMB 2001, 21).

These basic propositions are shared with the Government Performance and Results Act (GPRA) of 1993, and other previous efforts to introduce some form of performance budgeting at the federal level (such as Program Planning Budgeting Systems under President Johnson, Management by Objectives under President Nixon, Zero-Based Budgeting under President Carter), and more recently at the state level (Melkers and Willoughby 1998). But GPRA has failed to deliver a strong linkage between performance information and decision-making (Radin 2000, GAO 2004a). The PMA notes that "after eight years of experience, progress toward the use of performance information of program management has been discouraging" (OMB 2001, 27). The failure of GPRA is "because agencies rarely offer convincing accounts of the results their allocations will purchase" (OMB 2001, 27). The cause of this failure, therefore, is laid at the door of agency staff. The remedy, according to the PMA, is for agency staff to integrate performance reviews with budget submissions and allocate costs to outcomes. The goal is to move budgeting to a point where politicians fund more effective programs and reduce or reorganize less effective programs.

The different management proposals outlined in the PMA have had varying degrees of success, but the Office of Management and Budget (OMB) took the call to integrate performance information and the budget process to heart. The OMB piloted the Program Assessment Rating Tool (PART) for the FY03 budget, and has continued to apply it for the FY04 and FY05 budget, and plans to use it in the future. PART is essentially a set of standard questions that OMB analysts, in consultation with agency representatives, use to develop a summary assessment of a fraction of federal programs. For FY03 128 programs were graded, 234 programs were graded in for FY04, and 399 in FY05 (176 of which were new programs). Programs can be rated as ineffective, adequate, moderately effective or effective. If programs fail to provide adequate information they may fall into a separate category, of results not demonstrated. The ratings are to be based on four weighted groupings of questions: program purpose and design (standard weight is 20 percent), strategic planning (10 percent), program management (20 percent) and program results (50 percent).[i] Single-page summary assessments of each program accompanied the President's budget proposal to Congress, and detailed assessments are available on the OMB website for FY04 and FY05.

The espoused theory of PART, represented in the language of the PMA and other reform proposals, is that more performance information will lead to better decisions. This assumes that performance information is objective, standardized, indicative of actual performance, consistently understood, and prompts a consensus about how a program is performing and how it should be funded. Others have argued that the nature of the legislative branch means it is unlikely to expect performance information to have an impact there, but that such information might have an influence in earlier stages of budgetary preparation in the executive branch, and in non-budgetary decisions that might be broadly categorized as management (Joyce and Tompkins, 2002). However, *every* PART summary is accompanied with a numerical funding recommendation to the legislature, and PART is seen as the primary means by which the government is achieving the PMA goal of integrating performance information into the budget. The next section therefore examines the evidence on whether PART scores are influencing budgeting. Following this discussion, I propose an alternative theory of how performance information is used in budgeting. Dialogue theory suggests that performance information is not a neutral way to promote clear consensus and decisions, but is contested, subject to the ambiguity and advocacy of the traditional budget process. The final section of the paper offers a partial illustration of dialogue theory through an experiment where students reexamine the OMB PART assessments, and in some cases offer persuasive disagreements with the OMB decisions.

Examining the Relationship between PART Evaluations and Budget Decisions

Does PART actually have an impact on budget decisions? The creation of PART and accompanying budget recommendations provide some data that can provide at least a partial answer to this question.

A member of the OMB involved with PART suggested that there was no simplistic "reward the winners/punish the losers" relationship between performance and resource allocation. But this formula has been pushed by others. For instance the Performance Institute issued a press release declaring that "Bush's '04 Budget Puts Premium on Transparency and Performance," pointing to a 6 percent increase for effective programs, while those rated ineffective gained less than 1 percent. However, this impression is somewhat misleading when one considers that the OMB's own analysis for FY04 found that programs graded as effective did gain 6.4 percent, but that moderately effective programs gained 6.6 percent, and those rated adequate did even better (8.1 percent). Ineffective programs gained just .07 percent, but programs that failed to demonstrate results gained 4.4 percent. While it might be better not to be deemed ineffective, this is hardly evidence of a systemic preference for higher performing programs. The Government Accountability Office (GAO) (2004a) observes that there are examples of programs whose moderately effective or effective rating was accompanied with recommendation for funding decreases, while some programs rated ineffective enjoyed proposed increases.

Other quantitative evidence provides stronger support for the idea that PART assessments have a significant relationship in explaining proposed budget increases. The GAO (2004b) finds a significant relationship between PART scores and proposed budgetary increases for FY2004, although PART scores explained only a small portion of the variation. Lewis and Gilmour (2004, 2005) present more ambitious models that also demonstrate a positive relationship between PART scores and proposed budget increases for both FY2004 and FY2005, even when controlling for other possible explanatory factors such as program type, program age and political factors.

The nature of the available data creates some limits on the GAO and Lewis and Gilmour findings. The dependent variable analyzed is the rates of change between previous year estimated appropriation and the President's budget proposal for the current year. The calculation therefore tells us how the executive branch would like to increase the budget, but does not indicate the legislative preferences that actually determines appropriations authority. We know little about whether legislators discuss or dismiss PART, or are even aware of its existence.[ii] While the OMB has sought to educate legislators and their staff, preliminary indications are that Congress is just as unlikely to change its appropriations behavior in light of new performance information as it has with previous performance budgeting efforts (Gruber 2003).

A second limitation is that the delayed passage of budgets by Congress means that the executive branch sometimes does not have an indication of legislative intent for the fiscal year that is about to start, even as it tries to prepare a budget request for the following year. So, the executive branch must use some alternative measure of estimated appropriations. In the FY04 PART summaries, the estimated FY03 figure, which would usually be the amount appropriated by Congress, is actually what the President requested for this program in the FY03 budget request. In creating the FY05 budget the OMB were able to rely on an omnibus appropriation bill before Congress that was expected to pass without major changes, but the PART summaries still did not contain final figures indicative of legislative intent. Therefore, the budget changes calculated are based on the level of increase in the President's proposal over a figure that is an imperfect approximation of previous year allocations.

Even with the caveats listed above, the best research on this topic remains the aforementioned work by the GAO (2004b) and Lewis and Gilmour (2004, 2005), who offer a number of findings. First, PART assessments matter to the President's proposed budget, but they only explain a small portion of the variance of budget changes. PART assessments had a stronger impact on smaller programs in FY04 because, Lewis and Gilmour suggest, such programs have less entrenched support. Perhaps the results also reflect reluctance on the part of the OMB to make major changes in actual dollar totals – cutting 20 percent of a program may be less daunting if the total program is $1 million as opposed to $100 million. However, this result is not replicated with the FY05 data. Of the various PART categories, some matter more than others in influencing budgeting changes. Most important is program purpose and design, positively and significantly related to budget changes in both years. We would expect that program results, worth 50 percent of the PART grade and the rationale for performance reforms, should be positively and significantly related to budget changes. They are in FY04, but are less important than program purpose and design. Program results are not a significant predictor of budget changes in FY05 (Lewis and Gilmour 2004).

Searching for an Alternative: Dialogue Theory

The espoused theory of PART and other forms of performance budgeting – that performance will be rewarded, and failure punished – does not provide a compelling explanation for the results cited above. If PART does influence even the President's proposal, its impact is not very great, and is largely driven by the program purpose and design section, not program results. The anecdotal information about how congressional actor view PART suggests that even the modest influence of PART on the President's proposal are not being replicated when it comes to final appropriations. This is hardly surprising, since it is difficult to find careful observers of the budget process who believe that the addition of performance information will completely transform previous patterns of decision-making. Indeed, this simple rational model has been used most productively as a foil to develop more interesting and more plausible theories about budgeting, primarily the theory of incrementalism (Wildavsky and Caiden 2003).

How does the incrementalist theory of decision-making help to explain the use of performance information in budget decision-making? Incrementalism suggests that performance information does not matter, hence the continuing failure of performance budgeting. Performance information is ignored in favor of previous agreements reflected in last year's base. A comprehensive use of performance information is beyond limited human cognitive capacities, leads to information overload, and is a distraction for policy analysts.

The criticisms of incrementalism remain relevant. Once PART loses its initial gloss of political support, it is questionable as to whether the OMB and agency budgeters can maintain the heroic workload required to analyze even a fraction of the government's programs every year. But PART is somewhat different from previous versions of performance budgeting. PART seeks to reduce rather than create information overload, by having a presumably neutral authority offer a single assessment of multiple measures for a program. The failing of incrementalism as a theory is that it denies that performance information will be used at all, and therefore fails to explore how it might be used. Ultimately, this is because the values of incrementalism are process and consensus, not allocative efficiency or performance.

This paper proposes an alternative theory. Unlike incrementalism, dialogue theory assumes that performance information is used, but not in the simplistic way assumed by the espoused theory of performance budgeting.[iii] Dialogue theory points to the ambiguity inherent in interpreting performance information. There is likely to be no single definitive approach to a) interpreting what performance information means and b) how performance information directs decisions. There are two obvious sources giving rise to rival interpretations. First, the study of organizations and decision-making provides evidence of the ambiguity inherent in organizational life, and the potential for rival interpretations of performance information. Second, we know from the study of the policy process that the design of the budget process creates incentives for particular actors to advance arguments that reflect their institutional role and context, enhancing the potential for disagreement.

Ambiguity in Organizational Life

James March and his collaborators have proposed that organizational life is characterized by ambiguity (March and Olsen, 1976). Ambiguity is "that state of having many ways of thinking about the same circumstances or phenomena" (Feldman 1989, 5). Ambiguity is likely to occur in issues where objectives or issue-definition is unclear, where there is a lack of clarity on causal mechanisms between organizational actions and outcomes, where it is difficult to interpret the past, and where the pattern of individual participation in different decisions is uncertain and changing (March and Olsen 1976).

Feldman (1989) points out that while more information might reduce uncertainty, it will not eliminate ambiguity, since ambiguity is created by different perspectives rather than a lack of information. Ambiguous issues must be interpreted, a process by which actors give meaning to the issue. "Resolution is a matter of agreement rather than proof. To the extent that resolution occurs, it comes from shared understandings, not factual information" (Feldman 1989, 144).

While information helps actors interpret a policy issue, it does not necessarily foster consensus on decisions. “Organizational information processing seems to be driven less by uncertainty about the consequences of specific decision alternatives than by lack of clarity about how we talk about the world – what alternatives there are and how they are related to stories we think we understand, how we describe history and how we interpret it” (March 1987, 160).

Information selection and use occurs in the context of different beliefs, preferences, and cognitive processes, and will reflect organizational power and politics. Or, as March (1987, 154) succinctly puts it: “[I]nformation in organizations is not innocent.” Information providers will try to shape outcomes by choosing what information will be collected and highlighted. Each measure is representative of values and accompanied by the assumption that the organization should be making efforts that will have an impact on the measure.

Performance assessment tools like PART cannot eliminate subjectivity in how raters understand terms, apply standards and interpret data. The OMB has worked hard to ensure consistency, designing standardized questions, training raters and providing a sixty page guide

(OMB 2003a), and even forming a team to conduct a consistency check on 10 percent of the assessments. However, there will still be differences in interpretation even among OMB staff. As the GAO notes "Any tool that is sophisticated enough to take into account the complexity of the U.S. government will always require some interpretation and judgment. Therefore it is not surprising that OMB staff were not fully consistent in interpreting complex questions about agency goals and results" (GAO, 2004a, 6). For instance, how does a rater evaluate performance the program that is categorized as results not demonstrated, the most common PART assessment for FY04 and FY05? Interestingly, the application of the results not demonstrated categorization was itself contested; it was sometimes applied if there were inadequate performance information, but in other cases it was used if the OMB and agency simply failed to agree on what were appropriate long-term and annual performance measures for the program (GAO 2004b, 25). The understanding of what constituted "independent and quality evaluation" was also a source of disagreement between agencies and the OMB (GAO 2004b, 24). The need for interpretation increased through use of ambiguous terms such "ambitious" "outputs", "outcomes", and "having made progress", and the GAO reported evidence of inconsistency in the use of these terms. Agency officials complained that OMB officials used different standards for defining what measures were outcome-oriented (GAO 2004b, 21). Even efforts to standardize responses through the use of "yes/no" responses created problems, since different OMB staffers had different standards for what constituted a "yes" (GAO 2004a, 7).

Such evidence of subjectivity, even within a single agency that was striving for consistency, is not surprising and should not be viewed as exceptional. Deciphering what constitutes performance is difficult. We can expect such variation in interpretation to increase when the examiners tackle even trickier questions. What does performance mean? What are the next steps for the program?   While performance information tells us something about a program, the data itself does not answer these questions (Blalock and Barnow 2001). Performance data, or simplified assessments of performance data, fail to tell us:

- why performance occurred
- the context of performance
- how implementation occurred
- an understanding of outside influences on performance
- how to choose which program measure is a priority

The absence of this above information makes it difficult to determine what performance actually means. Analysts are usually stuck with interpreting whether performance is satisfactory in light of pervious performance or some target, implicit or explicit.

Garbage-can theory tells us that decision-makers usually draw loose connections between problems and solutions (Cohen, March and Olsen 1972). Given the looseness of such connections, even if two individuals agree on what a piece of performance data represents, they may disagree on what the solution is. A variety of plausible and logical different possible decision options may exist (Toulmin 1958). First of all, the type of action deemed appropriate might not be directly related to funding. The OMB has argued that PART is designed to elicit a dialogue with agency staff, and to generate a variety of recommendations. The GAO (2004a) offers some support to this perspective. In their analysis of the FY04 PART summaries, the GAO found that only 18 percent of PART recommendations were directly related to funding, while another 18 percent were directed toward program management, 15 percent toward program design, and 49 percent focused on program assessment (reflecting the OMB discontent with the high number of programs they judged as failing to demonstrate results). Participants might be talking about performance as a result of the PART assessment, but they might not be talking about performance budgeting.

Even if two individuals agree that the performance information should influence resource allocation, performance data does not tell us how to trade-off between multiple program and agency goals. Performance information does not answer Key's (1940, 1138) question: "On what basis will it be decided to provide x dollars to activity A rather than activity B?" Even if we accurately understood in advance the cost and outcomes of programs, that still does not provide a common basis for comparison since our answer to the question will depend on values. Indeed, the introduction of performance information simply adds related contextual questions: how do we know if more money will improve performance or be wasted? More broadly, how do we understand performance information, and how do we relate it to action?

The Role of Roles

How does the ambiguity of organizational life transfer to the political context? March and Olsen (1976) argues that individual behavior and preferences are endogenous to the organizational context of the individual. Conditions of ambiguity and disagreement will therefore be exacerbated in political institutions that are designed to check one another and represent opposing viewpoints. The political arena is characterized by battles between rival actors to assert hegemony over issue definition (Kingdon 1984). Stone (1997) describes how interpretations of policies are constructed into narrative stories of how the world works. Policy problems are defined and communicated in symbolic terms, and the most important aspect of symbols is their ambiguity. In this context, we can expect that interpretations of performance information will be highly contested in the political arena.

Constituency concerns, ideology and party which will shape how political actors interpret, present and use performance information. Conflict in interpretation is will also be fostered by the norms and incentives associated with particular roles in the budget process. Incrementalist theory described these different roles (Wildavsky and Caiden 2003). Agency staff are likely to be advocates, arguing that performance data calls for more resources. Oversight committees will mix advocacy with a desire to exercise control over the agency. Central agency staff are expected to manage and verify agency claims about performance and the implications for resources. In performing this role, OMB budget examiners consider themselves to be evaluators, and not just bean-counters. The scope of their influence on the agency enforces a norm that they need to consider management issues as well as budget issues (White 1999; Martin, Wholey and Meyers, 1995). PART offers another way to evaluate programs, and to formalize existing practices of performance evaluation that OMB examiners already do. Consistent with Feldman's (1989) observation that information generation frequently appears to be driven by professional and cultural norms, PART is intended to reflect the professional norms of budget examiners, and the underlying values of apolitical and neutral competence.

However, OMB staff are also obliged to take note of the political preferences of the President. An OMB staffer involved in PART commented: "I would note the preferences of the President are considered in the funding recommendations, not the assessments. The preference of the President would likely determine if a poor performing program was fixed or discontinued." This is consistent with another central agency role: to seek to set the policy agenda through the budget process. PART enables the OMB to assert their control over performance measurement as a tool for agenda-setting in a way that GPRA does not. GPRA seeks to incorporate performance measures in the budget, but with little influence on the part of the OMB. Instead, agencies enjoy a high degree of freedom in choosing goals and measures, in consultation with stakeholders and the Congress, leading Rosenbloom (2000) to characterize GPRA as a legislative-centered approach to governing.

There has been no easy marriage between PART and GPRA. The GAO (2004b, 31-32) noted the tension between PART and GPRA and the lack of congressional involvement in PART. OMB raters often deem existing GPRA measures unsatisfactory and seek the creation of new measures that better reflect their goals. PART is well-suited to the purpose of agenda-setting. The OMB can define issues and shape interpretation by choosing which programs are selected for assessment, which measures are considered relevant and important, which programs are classified as succeeding or failing, and which programs are recommended for greater resources (Feldman 1989).

What Does Dialogue Theory Tell Us?

What does dialogue theory tell us about decision processes and outcomes? Assuming that each actor has enough interest to examine performance information (an assumption questioned by incrementalists) there will be a variety of actors with different interpretations of performance data, and making different arguments about the actions implied. Incrementalism suggested that cognitive limitations eliminated the role of performance information in debate (Wildavksy, 1975). Dialogue theory proposes that while no single actor has a full account of all performance information, such information will be used and placed in the context of a wider argument.

Once we incorporate the assumption of roles and advocacy, we expect actors to selectively present performance data in a context that supports their point of view, and to discount conflicting information. The political nature of decision-making will interact with, rather than be replaced by, performance information. Advocacy and ideology will continue to shape allocation decisions, using performance information as a tool. Rather than present information comprehensively, giving equal balance to all, actors will highlight specific pieces of data, offer plausible explanations for why performance occurs, and how it can be improved. The contribution of performance information to decisions depends on the persuasiveness of the arguments made and the intensity of the interest and preferences of the actors involved. In some cases, what one group of decision-makers conclude is a reasonable interpretation and an appropriate response may be completely at odds with another group. Once these effects are controlled for, decision outcomes are unlikely to have a systematic relationship with performance information in a way that is easily observable to researchers, but may exhibit the conclusions of a reasoned discourse between relevant actors.

The simplest illustration of this point is the basic ambiguity of choice between performance and resources. If a program is consistently performing well, does that indicate that it should receive greater resources, or that it is already amply provided for? Is the poorer performing program a candidate for elimination, of just in need of additional resources? Different actors might take the same set of performance data and offer plausible and logical arguments for either option. In essence, this is what PART does, marshalling performance data, analyzing planning and management efforts, and coming up with a conclusion. But the thoroughness of the process does not mean that another actor could also examine the same performance information and come to a different conclusion. Predictably, agency officials have sometimes criticized PART as being unfair, because it advanced arguments that the agency disagreed with (GAO 2004). The following section of the paper presents an experiment designed to illustrate the ambiguity inherent in the process of interpreting performance data in PART.

An Exploratory Test of Dialogue Theory

The goal of this paper is to conceptualize and develop dialogue theory. A secondary goal is to undertake a preliminary test of the theory, one that is not conclusive enough to be regarded as proof, but which simply seeks to demonstrate whether the patterns of behavior proposed are plausible. I therefore developed an exploratory test of the theory based on an experiment using Masters in Public Service and Administration students at the Bush School of Government and Public Service. The experiment describes variation between program ratings and budget recommendations made by the OMB and students familiar with but not directly involved in the political process. The existence and reasons for such variation support the idea that different actors can employ logical warrants to interpret the same information differently.

In the experiment students were asked to take the position of a staff analyst for a Congressional committee and to analyze 3-5 programs overseen by that committee and evaluated by PART. Students were instructed to both examine the OMB's PART assessments but also to become an expert on the programs in their own right, undertaking research by examining the agency's strategic plan, performance report, website, agency, appropriate think tanks, newspapers and professional magazines and other sources such as GAO reports. The students were asked whether they considered the OMB evaluation to be "fair and accurate, and for each program explain why you agree or disagree with the OMB solution," and to develop funding recommendations to provide to their committee. Students were required to provide a written memo of about 7 single-spaced pages explaining their decisions. They also had to present their findings before the class, who operated as a mock legislative committee, asking pointed questions about the recommendations. After one session, a group of eight students served as a focus group to discuss in greater detail the reasons for the decisions they made. In the analysis that follows I quote student comments made during the focus group or from their memo.

Performing such an experiment is possible only because the PART process exhibits a remarkable degree of transparency in the original evaluation process. Students had access to 1 page summary program-by-program analyses completed by the OMB, as well as more detailed program analyses that provided written responses to a standard battery of questions that contribute directly to the program score, and the performance data used in the analysis. Students could therefore examine in a high level of detail the analytic categories that made up the PART score and rating, the process of reasoning behind the OMB responses, and the relative importance of these factors.[iv]

Twenty-two students evaluated 89 programs, 42 from the FY04 PART summaries, and 47 from the FY05 summaries. The resulting data provides two decision points, a PART program rating (ranging from ineffective to effective), and a budget recommendation. Since 89 programs were evaluated in total there are 178 potential decision points. For three evaluation decisions, and one program budget decision, the student did not express a decision.[v]

Of the 86 valid program rating decision points, the students agreed with the OMB assessment 67 times, and disagreed 19 times. Those 19 disagreements included two instances where the rater had assessed the program two points above the scale set by the OMB, 15 instances of judging the program one point above the OMB scale, one instance of being less generous than the OMB by one point, and one instance of rating the program two points lower than the OMB.

There was much less agreement in terms of actual funding decisions. The raters only agreed with the OMB assessment on 54 of 89 occasions. There was no clear distribution of the disagreements – the raters were almost as likely to cut funding (on 16 occasions) as they were to increase it (on 19 occasions). Comparisons of correlations between PART evaluations and budget recommendations for the sample also suggest that the decision preferences of students were not unusual. Students were about as likely as OMB analysts to match program ratings with proposed budget changes.

What conclusions, albeit tentative, can be reached? First, raters were generally reluctant to disagree with the ratings of the OMB. While the next section will discuss the reasons for disagreement, it is first worth considering the reasons for agreement. Students pointed to the information advantage enjoyed by the OMB, and the persuasiveness of the PART format, which allows the OMB to lay out detailed information to supports its recommendations. One student pointed to what she saw as the tight logical connections between the information presented and recommendations: "everything in the PART analysis was clearly laid out; they presented information and came to conclusions as a result of that."   To be sure, the PART summaries are persuasive. PART offers to go through mounds of performance data, speak with an authoritative voice on the implications for management and resources, underpinned by values of efficiency and effectiveness. Unless someone is deeply interested in the program, and has the basis upon which to form a contrary opinion, they are likely to accept the OMB version.

Students were more likely to challenge the PART assessments if they had some prior knowledge of the program, or if they could find an alternative source to support their arguments, such as the GAO or a think-tank. In the focus group, a number of students commented that if they had access to agency officials they would have had a better basis for challenging the PART outcomes. One student, who had experience working in the federal government, commented: “there was insufficient evidence for me to disprove their rating. We were looking at public documents, not talking to officials. If you talk to program staff you can find direction to get supplemental information that can help you advocate a position.” The experiment therefore underestimates the likelihood of disagreement between parties because agency officials and legislative committee staff will not suffer from the same measure of information asymmetry as the students, and will have more strongly-held preferences. We can expect that agency actors, stakeholders, and committee staff will be in a better position to launch plausible alternative interpretations to the PART summaries.

Students were more comfortable in disagreeing with the level of funding. In part, this is because the limited PART evaluation scale encourages raters to move from the initial evaluation score only if they experience a high level of disagreement, whereas students who did not feel deep disagreement might have kept the same rating measure, but added or subtracted an incremental sum from the President’s recommended sum (this happened 20 times). Students willing to disagree with the OMB on the program evaluation were also more likely to disagree on funding allocation. Disagreement on the evaluation and disagreement on the budget recommendation is positively correlated at .247, significant at .022 (two tailed).[vi]

The Logic of Disagreement

The point of the experiment is not so much the rate of agreement, but the reasons why disagreement occurred. This helps inform how performance information is interpreted, and in turn, used. In general, students used the following rationales when disagreeing with the grades and the funding decisions:

- PART ratings were unreliable, usually based on a disagreement about the adequacy of the underlying performance data or an alternative interpretation of what the data meant.
- Programs with results not demonstrated, poor financial controls, or which have achieved their original goals should not receive increases in funding
- Programs with strong positive assessments should receive higher funding
- Programs should receive increases in line with programs with similar PART ratings
- Program problems, while real, are being eliminated through improved planning and management; program cuts are not appropriate
- Clear relationship between resources, need and program delivery trumps performance information
- Stakeholder views trumps performance information

Students disagreeing with the PART ratings questioned how performance was assessed. One complaint was that functions that were inherently difficult to measure were forced to fit into the PART framework, and as a result, more likely to be considered ineffective, or categorized as results not demonstrated. In the focus group discussion, one student said: “I disagreed when they [the OMB] took something that could not be made more specific, and wanted it to be more specific” Another noted. “A lot of the programs that I evaluated, it was so difficult to quantify the results, so the program took a hit because it was something that was almost immeasurable.” One student argued that a program dedicated to the development of hydrogen technology should receive a higher evaluation, based on alternative interpretation of the context of the program. The student argued that the program had contributed real progress in moving from a petroleumbased economy, and was meeting its difficult-to-measure benchmarks on the way to a very ambitious and important long-term goal that required a certain amount of patience.

The issue of appropriate measurements appeared to be especially contentious with education-support programs. The Montgomery GI Bill is designed to facilitate veterans' return to civilian life by providing education benefits. Because the program lacked a direct measure of this goal, it was classified as results not demonstrated. The student argued that the rating failed to take into account clear benefits that the program was providing, and pointed to the high number of participants and strong management ratings as adequate reflection that the program was succeeding at a demonstrable level. The rater argued that the results expectations of the OMB were extremely difficult to satisfy. Indeed, given the impact of factors beyond the control of the program, assessing the contribution of an education program to how veterans readjust to civilian life would seem to call for a well-designed program evaluation rather than a simple performance measure (Blalock and Barnow 2001).

One student disagreed with both the PART evaluation and budget recommendation for the TRIO Upward Bound program. The program was originally envisioned to help prepare disadvantaged students entering college, but a 1999 study found that it was most effective at helping high-school students likely to drop out. The OMB rated the program as "ineffective" and recommended that the program be reorganized to more actively serve this high-risk population. The student, relying partly on the perceptions of professional education groups, argued that this was a mistake: "It appears the OMB is more concerned with changing the mission to reflect the program's current performance rather than finding ways to improve performance in pursuit of the original mission…While the most at-risk students would arguably receive some benefit from being targeted by the program, the students who are already performing relatively well would suffer the most. What the PART evaluation fails to consider is whether or not the benefit to at-risk students would be great enough to off-set the harm caused to the other students in the program. There is no useful cost-benefit analysis included in the program's PART evaluation." The student argued that without more evidence, the program should be rated as "results not demonstrated", and provided with greater resources as it strives to better fulfill what the PART analysis acknowledges is a worthwhile mission. The program was also an example of how PART recommendations, by seeking to shift programmatic resources to different client groups, clearly fall within the boundary of policy changes. A similar example comes from a student's disagreement with the OMB's position on the Hope Revitalization Program, intended to improve the quality of low-income housing. The student argued that the recommendation to eliminate the program was based on an excessively narrow reading of the program policy mandate, which continued to be relevant. She pointed to the original enabling legislation to argue that the program was "also intended to create viable mixed income communities. The number of deteriorated units is greater now than the number identified when the program began. As HOPE VI projects are being completed, the benefits of redeveloped mixed income communities are becoming clear. In addition, housing authorities have developed the experience and expertise required to complete these projects in a timely manner."

For another education program, TRIO Support Services, the student also recommended increased funding above the President's recommendation, arguing that a study that showed a 9 percent higher rate of participant bachelor degree completion relative to a control group suggested that the "results not demonstrated" rating underestimated the program performance. Another student argued against a proposed 12.5 percent cut to the Even Start literacy program, arguing that poor performance reporting mechanisms between grantees and the Department of

Education, rather than actual program failure, was responsible for the “ineffective” rating, and that more data was needed.

In some cases students agreed with the PART assessments, but felt that the OMB had not followed through on the logic of PART because of a failure to cut, or at least limit the growth of poor performers or programs that have completed their missions. An extreme example comes from the area of law enforcement, where the student agreed with the OMB’s results not demonstrated rating for the Community Oriented Policing program, and accepted the OMB’s argument that the program had succeeded in its mission of providing funding to encourage the adoption of the community policing philosophy: “The federal government intervention was a boost to get the manpower issue caught up and that has been accomplished.” The student disagreed with the OMB’s recommendation that the program should be dramatically cut (from $748.4 million to $97.1 million). Instead, he argued instead that the program should be simply eliminated, with any remaining useful aspects moved to another part of the Department of Justice.

A similar example comes from the Project-Based Rental Assistance program. The OMB rated it as “ineffective” and proposed FY04 funding above the FY02 actuals (although less than FY03 proposed figure). The student agreed with the PART summary, but recommended zeroing out the program and moving the money to another Housing Voucher program that was deemed moderately effective. The alternative programs were described in the PART assessment as having advantages in terms of greater individual choice, cost-effectiveness, less federal liability, and more local regulation. The student, therefore, was not contradicting the PART analysis, but instead pushing the implications of the analysis to its logical conclusion, i.e. take funds from a program that is not effective, and put it towards a better alternative. OMB examiners used the same logic elsewhere. The OMB proposed moving resources for USAID Climate Change to another department, while the student disagreed, noting that the program had clear goals but not measures: “In order to know whether funding should be altered, effective performance measures must be in place. A better solution would be to encourage the development of such measures over the next year and then consider a funding change.”

Other students argued that programs should be able to demonstrate results before gaining large increases. One student who looked at different education programs argued for the principle that programs categorized as results not demonstrated—four of the five programs she assessed—should receive very limited, if any, increases, and consistently cut the President’s proposed increases to these programs. Another student examining medical care and disability compensation programs for veterans adopted the same principle, arguing that the proposed increases should be limited as long as the program failed to demonstrate results.[vii] One student spelled out the logic of this approach: “[I]s there really an incentive to create effective performance measures if the program will receive increased funding regardless? If the priority is to institutionalize the PART as an effective evaluation tool, the Appropriations Committee should react to poor PART evaluations with budgetary restraint.” Similar arguments were made by a student who suggested that a proposed 75 percent increase in the Wildlife Habitat Incentives Program should be resisted until the program had been more thoroughly evaluated. A related argument was that adequate financial controls needed to be in place before funding increases were made. One student objected to a proposed increase in funding for Student Aid Administration, citing a GAO report that found ongoing problems with fraud and error.

The logic of this approach – a consistency between performance, evaluation and funding – was extended to situations where students saw programs that received positive evaluations receiving less than programs with lower ratings. An example is where one student argued for limiting an increase in the Home Investment Partnership Program below what the President recommended, on the grounds that this “moderately effective” program was gaining a greater increase than other programs that were deemed effective. The overall pattern of PART funding recommendations may therefore become part of the logic of comparative “fair play” that agencies use to push for maintaining parity in increases (Wildavsky and Caiden 2003). For example, one student explicitly cited the PART budgeting patterns themselves in arguing for a greater increase. She agreed with the OMB rating for the Soil Survey Program, but pointed out that the OMB provided other moderately effective programs with an average eight percent increase, while this particular program was only receiving a one percent increase.

The need for evidence of a direct connection between resources and performance could be deflected if agencies could make a plausible case that they were making changes to make programs more effective. For example in the Farmland Protection Program of the Natural Resource Conservation Service (NRCS), the OMB proposed, and the student agreed, that increases were justified despite a results not demonstrated rating because the agency was in the process of creating a plan to remedy the program shortcomings. The student noted that “According to the PART report, the NRCS understands and recognizes this problem and has already contracted with different universities and outside research groups to formulate an outcome-oriented performance measure that can be used to evaluate the long-term effects of this program. The results of these studies were not complete at the time of the PART analysis. Though the agency is having the matter reviewed, it is still difficult to quantify the environmental benefits of an agricultural easement. The studies did, however, buy NRCS some time." Accepting that program problems exist, and having a management plan to remedy those plans therefore becomes a credible tactic to delay or prevent budget cuts.

An alternative rationale was to call for higher funding and discount the PART evaluations where students saw a clear relationship between resources, need and delivery of appropriate services. Such an interpretation views program performance measures as secondary to factors that drive program demand. The program needs the resources to do its job. The most extreme version of this logic is for entitlements, where program performance is irrelevant for budgeting decisions, although it may have management implications. Other programs might be considered quasi-entitlements, in that they enjoy strong political support and represent a payment or service to however many individuals qualify. An example is the Burial Benefits program in Veteran's Administration, whose costs are driven by the mortality of eligible veterans.

The logic is also applied by both the students and the OMB to non-entitlements where program resources are perceived as central to achieving a vital function. This reflects the finding of Lewis and Gilmour (2004) that the program purpose and design scores, which represent the perceived importance of the function, were the only element of the PART analysis that had a significant influence on funding recommendations for both FY04 and FY05. Examples that students identified include the GI Bill previously mentioned, a program designed to deal with cybercrime, and a safety and security program for nuclear materials. In each case the student and OMB disagreed on the evaluation, but both agreed to continue increases in program allocations on the grounds of program importance. A student rating Wildland Fire Management agreed with the PART assessment of "results not demonstrated," but disagreed with the decision to cut funding, pointing to the saliency of the program in light of recent wildfire disasters: "This year more than ever we have witnessed the devastation of wildfires and how important it is to have the resources to prevent, fight and recover from wildfires." To some degree the PART assessment acknowledged this basic point by providing a 100 percent score for the program's purpose. Another student called for raising funding for the space shuttle mission above the President's proposed increase, citing the Columbia disaster and the need for a greater investment in safety.

The National Forest Improvement and Maintenance program was rated as "adequate" by the OMB, and failed to get a higher score because of a low score in the results/accountability sub-section because "the program has a significant deferred maintenance backlog (estimated at $13 billion) and the Forest Service has been unable to demonstrate that it can maintain its current infrastructure needs" (US OMB 2003b, 25). OMB recommended cutting the program. The student agreed with the PART evaluation, but argued that the backlog of deferred maintenance was due to lack of adequate resources, and recommended a funding increase. She reasoned that "The agency may be more successful in achieving its annual goals and in reducing the backlog of projects if it had more resources. I encourage increasing the agency's budget until the backlog is significantly reduced. This should be achieved in an incremental manner, the agency should be given the opportunity to prove that it can effectively utilize additional resources in a manner that is effective in reducing the project backlog." To a large degree this argument was based an interpretation of the information contained in the program's PART analysis, to the effect that reducing the backlog is central to agency goals, is directly related to resources, and is currently underfunded. Given these warrants, the disagreement the student made with funding recommendations appears reasonable.[viii]

In rare cases, students discounted the role of PART because of politically pragmatic stakeholder concerns. While agreeing with the PART assessment, one student argued for higher increases in funding for the Veteran's Administration Medical Program for largely political reasons, noting that the Secretary had lobbied for a higher level of funding, and that pursuing such an increase would raise stakeholder trust.

Overall, the above examples suggest that the students employed reasonable warrants upon which to disagree with OMB conclusions about PART ratings and funding recommendations. The plausible nature of the student arguments is demonstrated by the fact that OMB examiners employed many of these same rationales – on what constituted a reasonable basis for assessing performance, on how tightly evaluations should connect to funding, on the importance of demonstrating results, and on the relative importance of program need and the connection of resources to service delivery.

Conclusion

This paper has proposed an alternative approach to considering the use of performance information in budgeting decisions. The theory, and the test of the theory, are clearly exploratory, and some caveats are therefore in order.

Dialogue theory suggests that we can expect to see variation in interpretation of performance information due ambiguity of performance information and the influence of roles in the political process. The experiment helps us understand how variation in interpretation between analysts employing reasonable and logical warrants can still lead to different conclusions. The experiment does not directly speak to the influence of roles in the political process, beyond examining the key role of the OMB. This is a reasonable starting point for an analysis of PART– after all, if it does not have an impact among the executive branch, it is not likely to change legislative decision-making. But future research must take Congress into account. A useful approach would be to simply examine whether and how legislators consider performance information, either through analysis of committee hearings or via interviews with legislators and their staff. This research option has the virtue of offering highly valid understanding of how the key actors interpret performance information and the different rationales for using it, well beyond the modest beginning developed in the experiment presented in this paper.

One limit to the experiment is that it does not include funding constraints that budgeters face, although it turned out that the subjects did not display a marked tendency to increase funding in the absence of such constraints. As a theory of budgeting, dialogue theory also has weaknesses similar to that of incrementalism, in that it tells us only about how decisions are made for a fraction of federal spending, since most of the budget is tied up in mandatory spending, and what discretionary spending remains appears to be at least partly determined by last-minute decisions included in large omnibus appropriations package, although even here most of the details tend to finalized previously in appropriations subcommittees.

Caveats aside, what are the specific contributions of dialogue theory? In summary, the theory rejects the incrementalist argument that performance information is not used at all. In contrast to the rational espoused theory of performance budgeting, dialogue theory includes the role of politics, rejects the idea that performance information is objective enough to be uniformly understood in the same way and to prompt similar responses among different actors in the budgetary process. Rather, performance information is used, but the meanings assigned to such data are subjective, and will be interpreted and debated among different actors consistent with their values, training, motivations, partisan positions and cognitive characteristics. Funding judgments that decision-makers believe have been based on a reasoned debate on performance may differ significantly from one decision to another, and cumulatively not demonstrate a strong influence of performance information, and may even appear to be random.

It is not the intent of this paper to suggest that PART is an exercise without use. Performance information aids interpretation. It sometimes can be used to help explore ambiguity and produce alternative conceptions of an issue by fostering a dialogue on what the information means and how it should shape action. The addition of information will reduce uncertainty, and may render some interpretations and arguments more credible than others. While advocacy and ambiguity will persist, it will at least be better informed. Dialogue theory suggests that it may be too much to expect that incremental funding choices will be strongly influenced by PART ratings. Such lofty expectations would set PART up to be another performance budgeting failure. Instead, a more limited goal for PART is to simply encourage a dialogue where decision-makers consider funding outcomes *and* management choices in the context of performance evaluations.

References

Blalock, Ann B. and Burt Barnow. 2001. Is the New Obsession with Performance Management Making the Truth About Social Programs. In Dall Forsythe, ed. *Quicker, Better, Cheaper?: Managing Performance in American Government,* 485-517. Albany, NY: The Rockefeller Institute.

Cohen, Michael, James March and Johan P. Olsen. 1972. A Garbage Can Model of Organizational Choice. *Administrative Science Quarterly* 17:1-25.

Gilmour, John, and David E. Lewis. 2005. Does Performance Budgeting Work? An Examination of OMB's PART Scores. *Public Administration Review,* forthcoming.

Gilmour, John B. and David E. Lewis. 2004. *Assessing Performance Assessment for Budgeting: The Influence of Politics, Performance, and Program Size in FY 2005.* Paper presented at the 2004 annual meeting of the American Political Science Association, Chicago, IL, September 2-5.

Gruber, Amelia. 2003. OMB Ratings Have Little Impact on Hill Budget Decisions. *Government Executive,* June 13, Daily Briefing. Available at: http://www.govexec.com/dailyfed/0603/061303a1.htm

Joyce, Philip G., and Tompkins, Susan. 2002. “Using Performance Information for Budgeting: Clarifying the Framework and Investigating Recent State Experience.” In Kathryn Newcomer, Ed Jennings, Cheryl Broom, and Allen Lomax, eds. *Meeting the Challenges of Performance-Oriented Government,* . Washington D.C.: American Society for Public Administration.

Kingdon, John. 1984. *Agendas, Alternatives and Public Policy*. Boston: Little & Brown.

Key, V.O. 1940. The Lack of a Budgetary Theory. *American Political Science Review* 34(6): 1137-1140.

March, James G. 1987. Ambiguity and Accounting: The Elusive Link Between Information and Decision Making. *Accounting, Organizations and Society* 12(2): 153-168.

March, James G. and Johan P. Olsen. 1976. *Ambiguity and Choice in Organizations.* Bergen: Universitetsforlaget.

Martin, Bernard H., Joseph T. Wholey, and Roy T. Meyers. 1995. The New Equation at the OMB: M+B=RMO. *Public Budgeting and Finance* 15(4):86-96.

Melkers, Julia E., and Katherine Willoughby. 1998. The State of the States: Performance-Based Budgeting Requirements in 47 out of 50. *Public Administration Review* 58(1): 66-73.

Roberts, Nancy. 2002. Keeping Public Officials Accountable through Dialogue: Resolving the Accountability Paradox. *Public Administration Review* 62(6): 685-669.

Radin, Beryl. 2000. The Government Performance and Results Act and the tradition of federal management reform: square pegs in round holes. *Journal of Public Administration and Research Theory*, 10(1): 111–35.

Rosenbloom, David H. 2000. *Building a Legislative-Centered Public Administration: Congress and the Administrative State, 1946-1999.* Tuscaloosa: University of Alabama Press.

Stone, Deborah. 1997. *Policy Paradox: The Art of Political Decisionmaking.* New York: W.W. Norton and Company.

Toulmin, Stephen. 1958. *The Uses of Argument.* London: Cambridge University Press.

US General Accounting Office (GAO). 2004a. *Performance Budgeting: OMB's Program Assessment Rating Tool Presents Opportunities and Challenges for Budget and Performance Integration* GAO-04-439T. Washington DC: GAO.

US General Accounting Office (GAO). 2004b. *Observations on the Use of OMB's Program Assessment Rating Tool for the Fiscal Year 2004 Budget.* Washington DC: GAO.

US OMB. 2001. *The President's Management Agenda.* Washington DC: US Government Printing Office.

US OMB. 2003a. Budget Procedures Memorandum No. 861. Available at: http://www.whitehouse.gov/omb/budget/fy2005/pdf/bpm861.pdf.

US OMB. 2003b. *Budget of the United States Government: Performance and Management Assessments*. Washington DC: US Government Printing Office.

US OMB. 2004a. Department of Agriculture PART Assessments. Available at http://www.whitehouse.gov/omb/budget/fy2005/pma/agriculture.pdf. Accessed June 8, 2004.

US OMB 2004b. Department of Interior PART Assessments. Available at http://www.whitehouse.gov/omb/budget/fy2005/pma/interior.pdf. Accessed June 8, 2004.

White, Barry. 1999. "Examining Budgets for Chief Executives." In Roy Meyers, *Handbook of Government Budgeting,* 462-484. San Francisco: Jossey-Bass.

Wildavsky, Aaron A. and Naomi Caiden. 2003. *The New Politics of the Budgetary Process,* 5th ed. New York: Longman.

Wildavsky, Aaron A. 1975. *Budgeting*: *A Comparative Theory of Budgeting Processes.* Boston: Little Brown and Company.

Notes

[i] OMB examiners may adopt an alternative weighting of questions if they consider appropriate (OMB 2003a, 15).

[ii] Why not supplement the PART summaries with additional information from either the White House budgetary appendixes or appropriations bills? The short answer is that program definitions and levels of aggregation sometimes vary between the different branches, which would make some program comparisons misleading. Initial efforts to develop a more complete dataset using these sources suggested marked differences in listing and calculation of program figures. This impression was confirmed in an interview of an OMB employee involved with PART who advised against trying to make such comparisons. In the appendixes of the President's proposal some programs are not listed in the same way as they are listed in the PART summaries. They are aggregated into a larger category or disaggregated into a smaller category. Loans and grants in particular tend to have markedly different numbers in the appendixes than in the PART summaries. What about appropriations bills? Again problems of calculation make comparison problematic. For instance in FY2004 the OMB and Congress disagreed with how to allocate retirement costs, with the effect that agency and program figures in the President's budget proposal are inflated relative to appropriations bill.

[iii] There is a large literature devoted to issues of discourse and dialogue e.g., Roberts (2002). The dialogue theory presented in this paper is not a direct product of that literature, although it does share assumptions the exchange of ideas between professionals as part of the deliberative process. A key difference is my acknowledgement of power is this exchange, since the design of political institutions and the budget process means that some political actors consistently have greater influence than others. This theory in this paper owes more to incrementalism, the study of ambiguity in organizational life, and argumentation theory, which proposes standards for how actors construct reasonable arguments and practical uses of working logic in normal and often irrational decision processes (Toulmin 1958).

[iv] Detailed PART summaries are provided at http://www.whitehouse.gov/omb/budget/fy2004/pma.html, and http://www.whitehouse.gov/omb/budget/fy2005/part.html

[v] However, for the budget recommendation decision, the rater asked for greater information, and expressed doubt that the committee should commit to a greater than 100 percent program increase based on the evidence presented by the White House. On this basis, I consider it reasonable to assume that this budget recommendation indicates disagreement with the OMB recommendation, and essentially calls for lower funding.

[vi] In order to control for the role of outliers I converted the variables into a simplified disagreement scale, where students evaluating the program higher than the OMB is scored at -1, evaluating the program lower is 1, no disagreement is 0; students recommending higher budgets than the OMB is -1, recommending a lower score is 1, no disagreement is zero.

[vii] It should be noted that this is an entitlement program so that funding outcomes are determined by legislated criteria.

[viii] The specific comments that gave rise to this judgment include: "There is a clear need for improving and maintaining the safety and economy of roads, trails and facilities on NFS [National Forest Service] lands. Increasing or decreasing funds has a direct impact on the conditions of NFS roads, trails and facilities" (OMB 2004a, 199); "The program is very delinquent with capital maintenance and improvements" (OMB 2004a, 200); and "To the extent that funds are available, this program has made great strides in achieving its outcome goals. However, due to the lack of funding, the long term goals still remain unattainable and may need to be revised or more clearly defined to more accurately reflect what can reasonably be accomplished given limited resources. As it stands, the Forest Service is only able to demonstrate that it is reducing the growth rate in the deferred maintenance backlog. It is clear that additional efforts are needed to align infrastructure with available resources and to leverage resources from other sources" (US OMB 2004a, 205). Elsewhere, the OMB applied the same logic of that of the subject, recommending increases for a similar program (the National Park Service Facility Management, also rated adequate) pointing out that "Increasing or decreasing funding for these programs has a direct impact on the condition of that infrastructure" (US OMB 2004b, 198) even though "NPS [National Park Service] has not yet shown how budget requests link to particular performance targets. Nor has it documented how different funding levels would achieve different results" (US OMB 2004b, 200).