Assessing Performance Assessment for Budgeting: The Influence of Politics, Performance, and Program Size in Fiscal Year 2005

John B. Gilmour
College of William and Mary

David E. Lewis
Princeton University

INTRODUCTION

Because government agencies do not normally seek to earn a profit, and often operate in environments in which earning a profit is either impossible or undesirable, obtaining objective information by which to evaluate the performance of programs and agencies is very difficult. This leaves legislatures and executive budget agencies in the position of allocating scarce budgetary resources without knowing very well which programs deserve more and which deserve less. V. O. Key (1940) addressed this exact problem when he asserted the "lack of a budgetary theory"—a theory to guide resource allocation. The absence of a budgetary theory or other objective criterion to guide budget decisions no doubt helps explain the explosion of interest in performance measurement and its cousin, performance budgeting (Willoughby and Melkers XX). The promise of performance budgeting is to provide objective information about program outcomes, as opposed to inputs or process measures, that will allow budgeters to make appropriate comparisons among programs and allocate funding in a way that provides a better return on taxpayer money.

The George W. Bush administration, following the practice of most of the fifty states, has instituted its own performance budgeting initiative as part of a larger management agenda that also includes strategic management of human capital, competitive sourcing, improved financial management, and expanded e-government. Following a pilot program in 2002, the administration formally rolled out its performance budgeting initiative in 2003 with the fiscal year (FY) 2004 federal budget. The FY 2004 budget included performance assessments of 234 federal programs in different departments and agencies across the government. Graders assessed federal programs using the Program Assessment Rating Tool (PART), which consisted of a series of yes/no questions answered by Office of Management and Budget (OMB) examiners in consultation with program officials. Answers to the questions on the PART were used to create numerical scores in four

This article was first prepared for presentation at the 2004 annual meeting of the American Political Science Association, Chicago. We thank the Office of Management and Budget for making its performance assessments and process public. We thank Larry Bartels for helpful comments. The errors that remain are the sole responsibility of the authors. Address correspondence to John B. Gilmour at jbgilm@wm.edu.

doi:10.1093/jopart/muj002

management categories—program purpose and design, strategic planning, program management, and program results—and an overall program grade (effective, moderately effective, adequate, ineffective, results not demonstrated).[1]

OMB's purpose in collecting performance information was "making budget decisions based on results" (OMB 2003, 9). The administration argued that it would make budget decisions based in part on program performance, acknowledging that politics would invariably play a role in budget decisions and that performance and budgets have a complicated relationship since low budgets could be a cause of poor performance. The administration's budget and performance integration initiative is the most ambitious effort to implement performance budgeting in the nation's history. It holds tremendous promise for scholars who are trying to better understand this difficult enterprise.

A few works have already begun to assess the success of the administration in using performance information in the FY 2004 budget. These works point to the importance of politics in the PART scores themselves, the importance of having adequate performance measures, and the differential impact of PART scores based on program size in the success of budget and performance integration. Gilmour and Lewis (2005a) found that PART scores and the political content of federal programs influenced budget choices in expected ways. They also found that the impact of management scores on budget decisions appeared to diminish when the political component of the scores was taken into account. Once budget choices were modeled to allow the political content of programs to influence not only budgets but also the PART scores themselves, it became harder to disentangle the unique influence of the scores on budgets. They report that management scores were more important for programs housed in traditionally "Democratic" departments than other programs.

Gilmour and Lewis also report that while PART scores are positively correlated with budget increases or decreases, they are not perfectly correlated with overall program grades. Programs rated "results not demonstrated" have scores ranging from very high to very low. A "results not demonstrated" designation is usually reserved for those programs that have not found and implemented appropriate performance measures. This raises the interesting question of how performance information is used when programs do not have good performance measures.

In its review of the first year of PART, the GAO found that PART scores have a positive and statistically significant on recommended levels in the president's budget, although great deal of variance was left unexplained in these simple regression analyses (GAO 2004, 43). GAO also divided the programs into three groups based on program size, analyzed each separately, and found that for "small" programs the PART had a positive and statistically significant effect on recommended funding levels, but that for "medium" and "large" programs the coefficient was much smaller and not statistically significant. The GAO study is important but limited because it cannot investigate political issues, which prevented it from considering possible political effects on the linkage between PART and budget. Further, the GAO statistical analysis included no controls for type of program or department.

1 The numerical scores are based on the percentage of "yes" answers to questions grouped together by category. For example, if a program had four out of five questions answered with a "yes" in the program purpose and design section of the PART questionnaire, the raw score for that section would be an 80 out of 100.

In the FY 2005 budget OMB included assessments on a new cohort of 176 programs, along with new assessments of the 234 programs assessed in the FY 2004 budget.[2] This new data affords scholars another opportunity to evaluate the extent to which performance information is used in budgeting decisions, paying particular attention to the influence of program political content on grades, the adequacy of performance measures, and the differing utility of performance measurement for small versus large programs. In this article we analyze the impact of performance assessments on the FY 2005 budget and find that PART scores influence budget choices in expected ways, even when controlling for the political content of federal programs. We find evidence that performance has less impact on budget decisions in OMB than the nonperformance sections of PART. We also find some limited evidence that performance matters less for programs without adequate performance measures than for programs that do have adequate performance measures. Finally, we find some evidence that budgets for larger programs are more immune to performance information.

Politics and Program Evaluation

There is little systematic evidence that performance information has had a major impact on budget choices in states or cities (GAO 1993, 1; Joyce 1999; Melkers and Willoughby 2001). Integrating performance information in budget choices can be difficult for a variety of reasons. Poor performance can be caused by a variety of hard-to-identify factors, including budgets themselves. Low budgets can hinder necessary human and capital investments, delay innovation, dampen morale, and increase employee turnover. Cutting the budgets of poor-performing programs does not necessarily lead to program improvements, particularly if the federal program is the only provider of a public service or good.

Beyond these problems, however, budgeting is a political decision influenced by the political content of programs themselves and the political predispositions of key actors in the budgeting process. Indeed, one hindrance in the implementation of the Bush administration's budget and performance integration initiative has been getting appropriators to understand and use this new performance information and believe that it is more than a cover for the Bush administration's own policy preferences.[3] Thus far, PART information has mostly been used on the OMB side of budgetary decisions.

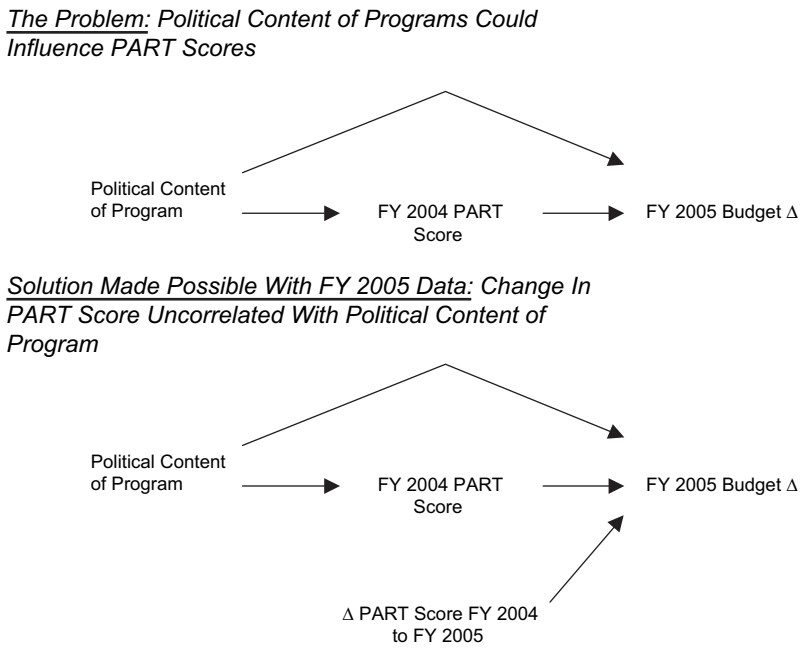
Since subjective judgments are necessarily part of performance evaluation, the PART scores themselves can become politicized. If this is the case, showing a relationship between PART scores and budgets is meaningless since the political content of the program itself ultimately determines its budget. In figure 1 we illustrate this causal process. If we seek to explain budget changes in the federal programs, both politics and merit may play a role. If, however, politics influences our measure of merit, it is extremely difficult to disentangle the influence of merit on budgetary choices.

The FY 2005 PART scores and budget information on the cohort of programs first evaluated in FY 2004 help remedy this problem. We now have two years of data on the set of programs first evaluated by OMB in 2004. Since the political content of the programs

2 Many PART assessments did not change for the 234 programs originally assessed in the FY 2004 budget. In our interviews with OMB officials, they indicated that not all programs were reevaluated and that most reevaluations occurred at the request of the programs themselves.

3 See Gruber 2004c.

Figure 1
Disentangling Political Influence from Program Assessment Rating Tool (PART) Scores



does not change between FY 2004 and FY 2005, but PART scores do change, any change in the budget correlated with changing PART scores should be independent of the political content of the program itself. This will give us one reliable means of determining whether performance information is being used by OMB in budget decisions.

Performance Measures and Goal Displacement

In principle, performance measures can help make budget decisions by focusing attention on accomplishments and away from process. Devising reliable measures of performance presents a tremendous challenge in performance budgeting, and the frequent inadequacy of measures makes implementing performance budgeting difficult, if impossible.

There are several reasons why such measures are hard to produce. One is that some things that government agencies do are inherently hard to measure. For example, the College of William and Mary strives to teach its students to think critically, but that is hard to measure and is not represented at all in the institutional performance measures (though it is in the mission statement). The measures focus on what is measurable—graduation rates, classroom utilization rates, and so on. Another problem is that some agencies have missions that are complex and multifaceted, and in some cases contradictory, which thus require a careful balancing. Balancing among competing objectives is hard to

measure quantitatively. Another difficulty in arriving at good measures is that the agencies often do not like being assessed in this way and find obstacles to measurement. There are strategies for overcoming this aversion, but the more clearly budgetary decisions are tied to measurement, the more agencies will resist measurement. Finally, while it is easy to measure outputs—the things agencies actually do—measuring their impact is harder since there is often a complex, multistage causal chain intervening between actions and intended consequences.

The experience of OMB with PART reflects the difficulty of measuring results since barely more than half of programs assessed so far have been deemed to have adequate measures. OMB assigns programs it has evaluated one of several overall grades—"effective," "moderately effective," "ineffective," and "results not demonstrated." The last category is not intended to reflect well or poorly on a program. Rather, the "results not demonstrated" grade reflects OMB's judgment that the existing measures for a program do not allow an assessment of the program's effectiveness. In the FY 2004 budget, "results not demonstrated" was the most common management grade, accounting for 51 percent of the 234 programs evaluated. In the FY 2005 budget, 41 percent of the 176 programs newly assessed received the "results not demonstrated" grade, which represents some improvement. Still, it was the most common grade given.

The Government Performance and Results Act (GPRA, or the Results Act), enacted in 1993, is another recent effort to make agencies focus on program results and the successor to other related budget reforms, such as PPBS and ZBB, that have sought to overcome the limitations of line-item budgeting (Knott and Hammond 1980; Wildavsky1975). Most previous efforts have had only limited success, and in justifying PART, OMB argued "that while well-intentioned, GPRA did not meet its objectives. Through the President's Budget and Performance Integration initiative, augmented by the PART, the Administration will strive to implement the goals of GPRA" (OMB 2003, 9).

Laurence Lynn argues that implementation of GPRA had an impact opposite of what was intended; instead of focusing more attention on results and performance, it encouraged administrators to focus on compliance with the procedural requirements of GPRA. Lynn (1998, 14) writes, "The effects of the U.S. Government's decision to enact and implement GPRA is evident for the most part in the proliferation of products on paper: the output of a seemingly far-reaching technocratic effort, with copious documentation, to create plans, performance standards and targets, the measures by which to assess their attainment . . . . These paper products are generated in response to the process requirements of the reforms. Much of the evaluation of the reforms focuses on the technical adequacy of these paper products and of the processes underlying them." "Ironically," he concludes, "in the light of its focus on performance, the Results Act's principal effect appears to be a heightened emphasis on procedural compliance by agency administrators" (Lynn 1998, 2).

When it is not possible to hold agencies accountable for their results, an alternative is to hold them accountable for their compliance with a set of procedures that purportedly lead to better performance. This would seem to be akin to "goal displacement," whereby the means of accomplishing an important goal takes the place of the goal itself (Merton 19XX). GPRA features procedural requirements, such as devising mission statements, performance measures, and strategic plans. The paper products are an essential element of performance budgeting because agencies must engage in a time-consuming effort to document their mission and short-term and long-term goals. Answering the questions in PART requires agencies to go through much the same process, and in answering the PART

questions, agencies surely rely heavily on work they have done under GPRA. Agencies can produce these documents even when they cannot document their results, which may leave OMB in the position of being forced to rely on procedural compliance rather than results in applying PART in the budget process.

The three sections of PART that do not relate directly to results—purpose, planning, and management—gauge attributes of federal programs that are desirable and might arguably lead to good performance, but they are not measures of performance. Rather, they measure the extent to which agencies have done a good job in producing the "paper products" Lynn discusses. Consider the questions used to assess program purpose and design. In the second year OMB asked five questions:

1. Is the program purpose clear?
2. Does the program address a specific and existing problem, interest, or need?
3. Is the program designed so that it is not redundant or duplicative of any other federal, state, local, or private effort?
4. Is the program design free of major flaws that would limit the program's effectiveness of efficiency?
5. Is the program effectively targeted, so that resources will reach intended beneficiaries and/or otherwise address the program's purpose directly?

Answers to these questions will reflect not just the merits of the programs but also the quality of the paper products produced in response to GPRA. Agencies vary in their responsiveness to the demands of GPRA, but those that take it seriously will be better positioned to offer compelling answers to PART questions about their purpose, the problem(s) they address, and so on. In addition, the questions relating to program purpose are open to political interpretation. One's political orientation might readily influence the answers to questions 2 and 3. None of these factors are closely related to results.

We expect that when good measures are available, OMB will rely on results evidence to determine budget allocations, but that when good measures are unavailable, OMB will rely more heavily other, non-results-oriented criteria.

Program Size and Performance Information

It makes sense that, as GAO (2004) found, the budgets of small programs would be most affected by assessments. Large programs are well established; they have important constituencies and, usually, long histories; they are like large ships with a lot of inertia and are unlikely to be driven off course by a negative (or positive) assessment by an OMB budget examiner. Small programs do not have the same political support as big programs, so with them OMB can apply a PART assessment more aggressively and can even propose to kill a small program that gets a very bad rating in PART. OMB simply cannot kill or significantly change the budget for a big program, no matter how bad, without causing a disturbance.

DATA, VARIABLES, METHODS

Our primary goal in examining the FY 2005 PART scores and budget is to determine whether, and how, PART scores influence budget choices. We examine two cohorts of

federal programs, a cohort of programs first evaluated in FY 2004 and reevaluated in FY 2005 and a cohort of programs being evaluated for the first time in FY 2005. We expect that both of these cohorts will be reevaluated in the FY 2006 budget and that a new cohort of federal programs will be added and evaluated for the first time. The first cohort of programs originally included 234 programs. In FY 2005 it included 223 programs, slightly less than the original 234 programs. The eleven programs that dropped out of the sample were programs that either no longer were organized as distinct programs or were not easily defined as unique programs to start.[4] None were dropped because they had been eliminated between FY 2004 and FY 2005. The second cohort of programs includes 176 programs.

The dependent variable is the percentage change in the budget from FY 2004 to FY 2005. In figure 2 we include histograms of program budget changes from FY 2004 to FY 2005 by cohort. On the left-hand side of the figure is the full sample. One difficulty in dealing with budget data is dealing with outliers. Some programs have budget cuts, and others have budget increases. Including extreme values in the analysis can lead to false conclusions based on a few influential cases. The question becomes where to drawn the line in identifying outliers. The literature on incremental budgeting suggests that budget changes normally are small and that changes in the magnitude of 50 to 100 percent are unusual and outside the confines of the normal budgeting process. If, however, we exclude cases with large budget changes, we bias the results against finding evidence that performance measures matter. In the FY 2005 budget, OMB used performance information to justify large budget changes to a number of programs, partly program cuts of 100 percent.[5] For this analysis we make the somewhat arbitrary choice of excluding cases where budget changes are greater or less than 100 percent.[6] Histograms of truncated samples are included on the right-hand side of figure 2. The distributions look both less skewed and more normal. Another reasonable cutoff would be to exclude all cases with changes greater than 50 percent of a program's budget. We do not do so here both in order to give OMB the benefit of the doubt and to preserve as many cases as possible with small samples. We have estimated all the models at this alternate cutoff; the primary results change very little except that standard errors become larger and estimates become less precise.

The key independent variable is the total weighted PART score for individual programs. The total weighted score is derived by summing the weighted scores for each section with the weights as follows: program purpose and design (20 percent), strategic planning (10 percent), program management (20 percent), and program results (50 percent). The average score is 63 (up from 60 in FY 2004) for both cohorts, with a minimum of 11 and a maximum of 97.[7] The two highest-rated programs in the first cohort are the National Assessment Program (94, 0 percent FY 2005 increase) in the Department of Education and the Basic Energy Sciences Program (93, 5.2 percent FY 2005 increase) in the Department of Energy. The two highest-rated programs in the second cohort are the

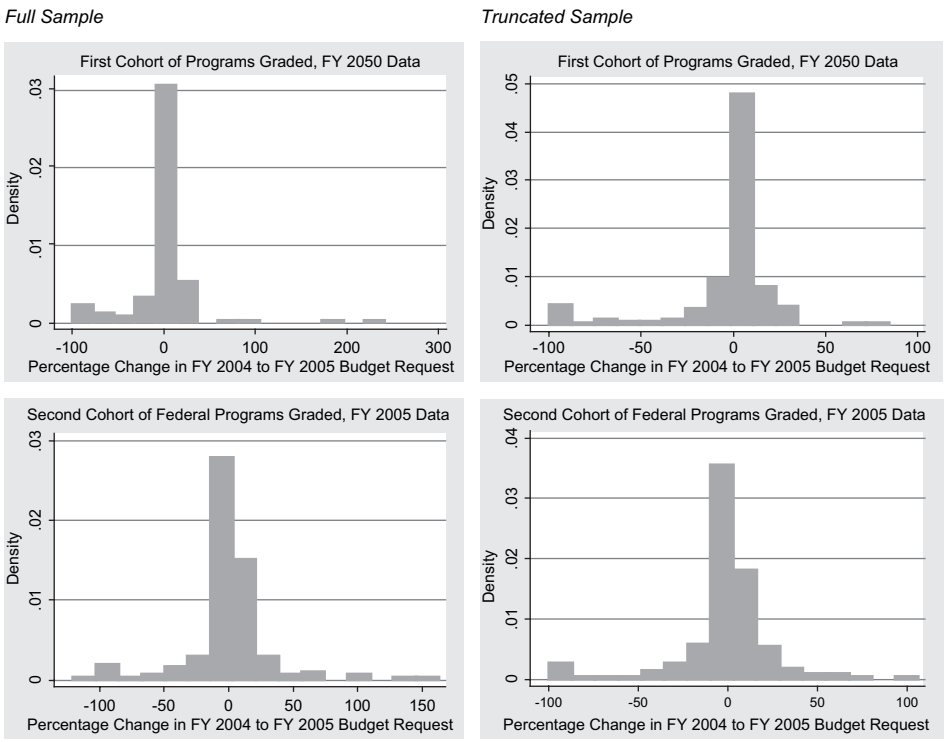
4 The programs dropped were programs in the Energy Department (Environmental Management—R&D, Gas Exploration and Production, Methane Hydrates), the Department of Health and Human Services (Center for Biologics Evaluation and Research, Center for Devices and Radiologic Health, Center for Drug Evaluation and Research, Center for Food Safety and Applied Nutrition, Center for Veterinary Medicine, State and Community-Based Services Programs on Aging), and the National Science Foundation (Geosciences, Tools).

5 See Gruber 2004a, 2004b; Ziegler 2004.

6 In a few cases budgets can be cut by more than 100 percent if programs do not rely solely on appropriations for revenues, as is the case with government corporations.

7 The standard deviation is 17 for the first cohort and 20 for the second cohort.

Figure 2
Histogram of Budget Changes to Graded Federal Programs Fiscal Year 2004 to Fiscal Year 2005



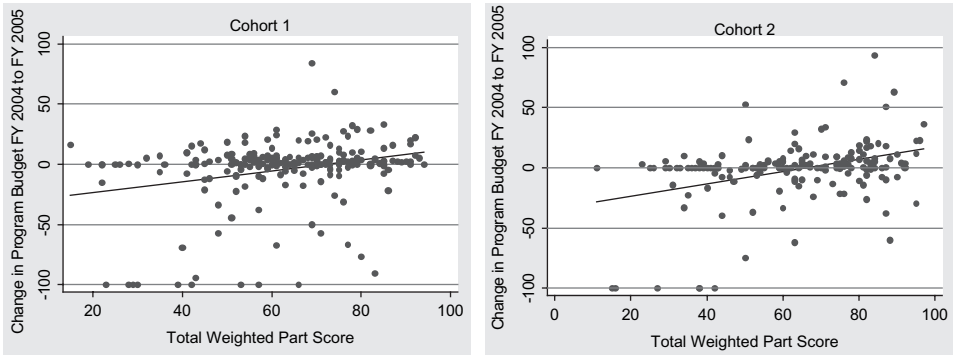
Federal Employees Health Benefits Integrity Program in the Office of Personnel Management (97, 36 percent FY 2005 increase) and the New Currency Manufacturing Program (96, 23 percent FY 2005 percent increase) in the Treasury Department. The lowest-rated programs in the two cohorts are the Veterans Compensation program in the Department of Veterans Affairs (15, 16 percent FY 2005 increase), the State Criminal Alien Assistance Program in the Justice Department (15, −100 percent FY 2005 increase) and the Tribal Courts System in the Department of the Interior (11, 0 percent FY 2005 increase).

In figure 3 we graph the bivariate relationship between total PART score and percentage budget change proposed in the FY 2005 budget. PART scores for both cohorts appear to influence the size of the budget change for a program from FY 2004 to FY 2005. This gives some initial evidence that performance information is being used in program budgeting.

Accounting for Other Factors: Political Content of Programs

A number of other factors could be correlated both with PART scores and budget changes, including the political content of the programs themselves, program age, and program mission. Of particular concern is the influence of a program's political content on both its budget and its PART score. If what a program does influences both its PART score and its budget, the influence of performance on budget choices may be illusory. Measuring the political content of a federal program is extremely difficult, so we take a number of

Figure 3
Impact of Program Assessment Rating Tool (PART) Scores on Fiscal Year 2005 Budget



different tacks. First, we code each program according to what department houses the program. Some departments do work that is more central to the agenda of the Democratic Party than other departments and might provide a reasonable proxy for the political content of the program. We code all programs in the Departments of Housing and Urban Development, Labor, and Health and Human Services and the Environmental Protection Agency with a "1." We also code the Departments of Commerce, Education, and Energy with a "1" since they have been targeted for termination by Republican administrations. Programs housed in all other departments are coded with a "0." Forty-nine percent of programs in the first cohort are housed in what we call "Democratic" departments, and 40 percent of programs in the second cohort are housed in Democratic departments. This is a crude measure because there are some programs in these departments that Republicans like and programs in other departments they do not like, and there are also differences among Republicans in their commitment or hostility to traditionally Democratic departments. President Bush has made an important commitment to education, for example. But to avoid an overall ad hoc approach in constructing this variable, we are relying on our conception of the traditional positions of the parties. We are assuming that collectively the programs coded "1" will be supported more weakly than programs coded "0."

Another means of measuring a program's political support is to look at budget change from the year before the program was assessed.[8] If a program's budget was increasing in the year before performance information was available, this reflects a level of political support from the administration for the program. Our expectation is that programs whose budgets were increasing prior to the FY 2005 assessment are more likely to increase after assessment. The average budget change from FY 2002 to FY 2003 was +4.2 percent for the first cohort, and the average budget change from FY 2003 to FY 2004 for the second cohort was +4.1 percent. The effect of budget change prior to assessment should be weaker for the first cohort since the data for budget change comes from two years before the FY 2005 budget was issued.

8 We have also estimated models from the first cohort with their budget change from FY 2002 to FY 2003 before they were subject to PART. In these models the results on the impact of PART score do not change, but the coefficients on the budget variable are consistently insignificant across models. We cannot reject the null hypothesis that percentage budget change from FY 2002 to FY 2003 has no impact on budget change from FY 2004 to FY 2005.

Our final measure of a program's political content is the political configuration at the time that a program was created. We code all programs created under unified Democratic control with a "1" and all programs created under unified Republican control with a "−1." All other programs are coded with a "0." In the first cohort, 60 (18) programs were created under unified Democratic (Republican) Party control. In the second cohort 41 (9) programs were created under unified Democratic (Republican) Party control. Our expectation is that programs created under unified Democratic control will get smaller budgets than other programs.

Controls

In addition to a program's political content, we control for a program's age. Older programs should demonstrate less budget volatility; and because older programs have had to survive multiple authorizations, their age implies a level of political support. The average program is thirty-four to thirty-five years old.

Our final set of controls accounts for what programs do. Program budgets and program performance can be influenced by whether a program regulates, gives grants, or some other function. We include indicators (0, 1) for what a program does. There are seven main program categories: block/formula grant, competitive grant, capital assets and service acquisition, direct federal, credit, regulatory, research and development. Some programs consist of more than one of these activities. The largest category is direct federal programs. We have also estimated models with fixed effects for departments and different cutoffs for the dependent variable and include those in the appendix.

In table 1 we provide regression estimates. We report robust standard errors and indicate significant coefficients in one-tailed tests with asterisks. We employ one-tailed tests since we have directional hypotheses about key independent variables, but we include estimated standard errors for reference. One result that is consistent across the two cohorts is the impact of PART scores on budget increments. The coefficients are all significant. Substantively, the estimates suggest that an increase in PART score by 10 points will lead to a budget increase of 4–5 percent. These results confirm what figure 3 shows, which is that programs that had higher PART scores also got higher budgets.

Surprisingly, the measures of the political content of the programs do not have consistent or statistically significant affects on FY 2005 budget changes. Contrary to what Gilmour and Lewis (2005a, 2005b) report, programs housed in Democratic departments or created under unified Democratic Party control are no more likely than other programs to get budget increases or decreases. Programs that received budget increases in the years before they were evaluated by PART are no more likely to receive increases in FY 2005. The poor performance of the political content variables is surprising, particularly given that budgeting is an admittedly political process. Even the Bush administration and its performance budgeting leaders acknowledge the role of politics in budgetary decision making.

There are several possible explanations for the poor performance of the political content variables as factors explaining budget increases or decreases in FY 2005. First, it is possible that there is substantial collinearity among the measures of political content. We dismiss this possibility since the results are virtually the same in specifications that exclude one or two the measures of political content. Second, it is possible that the budgeting process is different, perhaps less political, in FY 2005 than it was in FY 2004. New

Table 1
Program Assessment Rating Tool (PART) Scores and Budget Changes for Fiscal Year 2004 and Fiscal Year 2005

	Cohort 1		Cohort 2	
Merit				
PART score	0.47** (0.14)	0.47** (0.15)	0.39** (0.13)	0.40** (0.16)
Political Content of Program				
Housed in Democratic department (0, 1)	3.47 (4.19)	4.16 (4.48)	−4.51 (4.96)	−5.94 (5.65)
% Increase in FY 2003 budget	0.04 (0.08)	0.03 (0.08)	0.13* (0.10)	0.24 (0.18)
Unified Democratic/Republican control at creation (−1, 0, 1)	—	4.22 (3.45)	—	−8.11** (3.66)
Other				
Age of program	—	0.07* (0.05)	—	0.08 (0.06)
Block/formula grant (0, 1)	−3.66 (7.49)	−3.55 (8.27)	6.17 (6.75)	7.54 (7.02)
Capital assets and service acquisition	8.33** (4.00)	6.96* (4.79)	−4.88 (16.99)	−8.69 (18.50)
Competitive grant (0, 1)	−1.33 (5.05)	−2.73 (5.62)	1.40 (9.38)	−2.29 (11.91)
Direct federal (0, 1)	5.82 (4.79)	4.18 (5.49)	6.58 (5.40)	4.71 (5.87)
Regulatory (0, 1)	3.04 (4.26)	1.43 (4.95)	18.79** (6.69)	15.00** (6.68)
Research and development (0, 1)	−15.65** (5.95)	−14.83** (6.79)	0.49 (6.63)	0.22 (7.73)
Constant	−35.26** (9.41)	−37.72** (11.08)	−30.62** (9.66)	−30.79** (10.90)
N	204	166	164	131
$F_{9,11}$	2.15**	1.61	2.89**	2.46**
R^2	0.13	0.15	0.21	0.27

Note: Dependent variable is the percentage change in budget from previous year's budget.
*Significant at the .10 level; **Significant at the 0.05 level in one-tailed tests. Robust standard errors reported.

programs have been selected, perhaps on a less political basis than in FY 2004. Though possible, it is unclear how the FY 2005 process is different than FY 2004, unless the transparency of the process has had a disciplining effect on administration budget decision making. The transparent nature of the process may make it harder for the administration to use political justifications in budget proposals. Third, and perhaps most plausibly, it could be that our measures of political content are inadequate. If we could measure the political content of the programs more accurately, it is conceivable that we would get more precise and consistent estimates.

Finally, it is possible that the political content of programs influences budgets through the PART scores rather than directly. In table 2 we address this problem by including the change in the PART score from FY 2004 to FY 2005, in addition to the FY 2005 PART score, for the first cohort of programs. The political content of programs should not change between FY 2004 and FY 2005, but the PART score can and does change. If a change in the PART score is positively correlated with a change in the budget, this is evidence that PART scores influence budgets rather than the political content of programs themselves. The estimates in table 2 show that changing PART scores are significantly related to changing budgets. An increase in PART score from FY 2004 to FY 2005 of 10 points is estimated to increase a program's budget by 4–5 percent, holding constant the influence of other factors, including its actual PART score.[9]

The first two tables have shown that PART scores and changes in PART scores are correlated with FY 2005 budget changes. Measures accounting for the political content of the programs and measures accounting for program age or program function also have no consistent effect on budget changes. Gilmour and Lewis (2005a) found that PART scores mattered more for programs housed in Democratic departments. When we regressed budget changes on PART score by samples defined by Democratic or non-Democratic department, however, we could not reject the null hypothesis that PART scores mattered equally for programs in all departments. In fact, the estimates suggested that PART scores mattered more for programs in non-Democratic departments.

When Do PART Scores Matter for Budgets?

Several analyses of the FY 2004 PART implementation suggest that PART scores may have different impacts on budget choices depending on the adequacy of program performance measures and program size. To examine the impact of PART scores depending on the adequacy of performance measures, we analyze the components of the PART scores separately. PART scores are the weighted sum of four components—program purpose, program planning, program management, and results. Only the last of these reflects results. To ascertain how important results and performance are in making budget decisions, we

9 Of course, if program political content is measured poorly or program political content was used to determine increases or decreases in PART score, there is omitted error in the regression correlated with the dependent variable. To account for this possibility, we estimated a two equation model via two-stage least squares where PART score is the dependent variable in one equation and budget change is the dependent variable in the other equation. We include three exogenous variables that predict PART score that should be uncorrelated with the budget change variable: Senate-confirmed appointee, commission structure, fixed term for appointee. These variables should influence management because of their impact on manager tenure and planning but have no direct effect on budget change (Gilmour and Lewis 2005b). They may have an influence on the budget level (McCarty 2004).

Table 2
Impact of Change in Program Assessment Rating Tool (PART) Score on Fiscal Year 2005 Budget

Merit		
PART score FY 2004	0.47** (0.14)	0.47** (0.16)
Change in PART score	0.44* (0.29)	0.49** (0.20)
Political Content of Program		
Housed in Democratic department (0, 1)	3.39 (4.20)	4.22 (4.61)
% Increase in FY 2003 budget	0.04 (0.08)	0.03 (0.08)
Unified Democratic/Republican control at creation (−1, 0, 1)	—	4.20 (3.49)
Other		
Age of program	—	0.07* (0.05)
Block/formula grant (0, 1)	−3.33 (7.77)	−3.82 (8.50)
Capital assets and service acquisition (0, 1)	8.62** (4.49)	6.74* (4.85)
Competitive grant (0, 1)	−0.99 (5.91)	−2.97 (6.14)
Direct federal (0, 1)	6.20 (4.70)	3.92 (5.88)
Regulatory (0, 1)	3.32 (4.68)	1.23 (4.93)
Research and development (0, 1)	−15.31** (6.57)	−15.06** (6.98)
Constant	−36.68** (9.49)	−37.40** (11.50)
N	204	166
$F_{9,11}$	1.99**	1.47*
R^2	0.13	0.15

Note: Dependent variable is the percentage change in budget from previous year's budget.
*Significant at the .10 level; **Significant at the 0.05 level in one-tailed tests. Robust standard errors reported.

substituted each of the four components of the PART—purpose, planning, management, and results—in place of the total PART score in the equations. This analysis is reported in table 3, where we examine cohort 1 and 2 separately. We find for both cohorts that purpose has a larger coefficient than results, although for cohort 1 the difference is small. For cohort 1, the purpose variable has the largest coefficient, closely followed by results. For cohort 2, the purpose variable overwhelms the other components of PART scores. This is surprising, because in creating the overall PART score, the results score is weighted more than the others. Summarizing table 3, it appears that performance had virtually no impact on budget decisions for the second cohort, and a small impact in the first cohort.

To consider the possibility that inadequate measures of performance limit the impact of the results measure on budget decisions, we re-estimated the models in table 3 separately for programs with adequate measures and for those without. We include each program's purpose and design score as an independent variable rather than the whole PART score. (We could not include other components of the PART score because adequacy of performance measures helped determine scores in these areas.) If, in the absence of good performance information, OMB makes decisions on the basis of what it can measure, we should see a larger coefficient for the purpose variable among programs without good measures. The results of this analysis, reported in table 4, are mixed. For cohort 1, we see the opposite of what we expected, with the programs with good measures having a larger purpose coefficient. This is very surprising and raises the question of why OMB did not make more use of the performance information it had. For cohort 2, programs without

Table 3
Impact of Program Assessment Rating Tool (PART) Components on Fiscal Year 2005 Budgets

	Cohort 1	Cohort 2
Merit		
Program purpose and design	0.33** (0.15)	0.58** (0.17)
Strategic planning	0.01 (0.13)	0.19* (0.12)
Program management	0.19* (0.12)	−0.07 (0.17)
Program results	0.24** (0.11)	−0.03 (0.10)
Political Content of Program		
Housed in Democratic department (0, 1)	−1.57 (4.62)	−6.29 (5.43)
% Increase in FY 2003 budget	−0.07 (0.09)	0.21* (0.16)
Unified Democratic/Republican control at creation (−1, 0, 1)	0.99 (3.53)	−7.17** (3.47)
Other		
Age of program	−0.09 (0.07)	0.08* (0.06)
Block/formula grant (0, 1)	−0.51 (9.41)	8.62* (6.44)
Capital assets and service acquisition (0, 1)	14.46** (5.48)	−2.62 (15.57)
Competitive grant (0, 1)	−0.38 (5.15)	0.86 (11.23)
Direct federal (0, 1)	13.94** (7.41)	3.07 (5.40)
Regulatory (0, 1)	3.24 (5.84)	14.69** (6.61)
Research and development (0, 1)	−6.71* (4.97)	−2.95 (7.13)
Constant	−50.50** (15.89)	−59.97** (18.72)
N	159	131
F (14 *df*)	1.93**	2.50**
R^2	0.22	0.38

Note: Dependent variable is the percentage change in budget from previous year's budget.
*Significant at the .10 level; **Significant at the 0.05 level in one-tailed tests. Robust standard errors reported.

good measures have a far larger purpose coefficient than those with good measures.[10] This evidence suggests that with cohort 2, at least, OMB was rewarding procedural compliance rather than actual performance, particularly for programs that lacked good measures.

In its 2004 report on the PART system, the GAO noted that PART scores matter most for small programs. Small program budgets were more amenable to large percentage increases or decreases consistent with performance information. PART scores had a smaller impact on the budgets of larger programs, even when entitlement programs were excluded. In table 5 we replicate GAO's analysis using FY 2005 data. We include indicator variables for small (<$75 million) and medium-sized ($75 million < program budget < $500 million) programs and interact these indicators with the PART score variable. The base category is large federal programs (>$500 million). If PART scores matter most for small and/or medium-sized programs, the coefficients on the interaction terms should be positive and significant. In table 5 this is what we see. The positive coefficient on the interaction terms suggests that the impact of PART scores for budgets is larger for small and medium-sized programs than for large programs. While small and

10 We note, however, that we cannot reject the null hypothesis in statistical tests that the coefficients are the same size. We can only point to different point estimates. More data will be required to determine whether goal displacement is occurring.

Table 4
Importance of Program Purpose Scores for Fiscal Year 2005 Budgets by Adequacy of Performance Measures

	Cohort 1		Cohort 2	
Merit				
Program purpose score	0.33* (0.21)	0.50** (0.25)	0.75** (0.26)	0.57** (0.22)
Change in program purpose score	0.12 (0.42)	−0.16 (0.25)	—	—
Political Content of Program				
Housed in Democratic department (0, 1)	−0.52 (7.76)	3.32 (5.94)	−9.60* (7.20)	−5.00 (8.07)
% Increase in FY 2003/4 budget	−0.08 (0.17)	0.06 (0.11)	0.12 (0.18)	0.27* (0.20)
Unified Democratic/Republican control at creation (–1, 0, 1)	2.39 (4.47)	9.61* (6.10)	−1.10 (6.66)	−8.28* (5.07)
Other				
Age of program	0.12 (0.11)	0.17** (0.06)	0.03 (0.14)	0.09* (0.06)
Block/formula grant (0, 1)	5.50 (15.05)	7.47 (12.91)	7.33 (8.69)	13.28 (10.64)
Capital assets and service acquisition (0, 1)	6.28 (10.19)	28.18** (9.68)	17.25** (8.44)	−7.95 (20.76)
Competitive grant (0, 1)	8.16 (12.74)	9.64 (9.26)	43.54** (11.64)	−4.99 (12.17)
Direct federal (0, 1)	9.74 (10.16)	9.02 (10.02)	7.84 (8.82)	1.66 (7.34)
Regulatory (0, 1)	9.88 (9.32)	17.47** (8.92)	11.81* (7.68)	15.80* (10.39)
Research and development (0,1)	−11.34 (14.27)	—	2.98 (10.68)	−2.74 (9.92)
Constant	−42.58** (21.58)	−67.51** (26.89)	−74.00** (26.37)	−52.58** (23.92)
Good Performance Measures?	No	Yes	No	Yes
N	85	78	57	74
F (11 *df*)	0.59	1.82**		2.26**
R^2	0.10	0.27	0.41	0.40

Note: Dependent variable is the percentage change in budget from previous year's budget.
*Significant at the .10 level; **Significant at the 0.05 level in two-tailed tests. Robust standard errors reported.

Table 5
Impact of Performance on Fiscal Year 2005 Budget by Program Size

	Cohort 1	Cohort 2
Merit		
PART score	0.06 (0.14)	0.23* (0.16)
PART score × Small program	1.28** (0.43)	0.13 (0.30)
PART score × Medium-sized program	0.38* (0.25)	0.37 (0.31)
Political Content of Program		
Housed in Democratic department (0, 1)	4.50 (3.87)	−4.78 (5.87)
% Increase in FY 2003/4 budget	0.01 (0.07)	0.26* (0.18)
Unified Democratic/Republican control at creation (−1, 0, 1)	3.54 (3.55)	−6.63** (3.79)
Other		
Small program (<$75 million)	−84.87** (30.47)	−10.73 (21.19)
Medium-sized program ($75 million < program budget <$500 million)	−27.88* (18.34)	−29.22* (21.84)
Age of program	0.08** (0.05)	0.07 (0.06)
Block/formula grant (0, 1)	−1.36 (7.41)	4.88 (7.81)
Capital assets and service acquisition (0, 1)	7.42 (5.09)	−8.55 (16.03)
Competitive grant (0, 1)	3.09 (6.06)	−4.50 (11.66)
Direct federal (0, 1)	6.61 (5.33)	3.41 (6.17)
Regulatory (0, 1)	3.10 (5.18)	13.21** (7.46)
Research and development (0, 1)	−13.70** (6.47)	−0.92 (8.39)
Constant	−12.85 (11.79)	−16.33 (10.66)
N	166	131
F (15 *df*)	1.70**	1.97**
R^2	0.24	0.29

Note: Dependent variable is the percentage change in budget from previous year's budget.
*Significant at the .10 level; **Significant at the 0.05 level in two-tailed tests. Robust standard errors reported.

medium-sized programs are more likely to get cuts or smaller increases overall, PART scores do have a significant impact on budget changes from FY 2004 to FY 2005. For small programs, an increase of 10 points is estimated to increase a program's budget from 1.3 to 13.5 percent. For a medium-sized program the increase is closer to 4 percent.

DISCUSSION AND CONCLUSION

The PART system has been in operation for two complete years now. One cannot escape the conclusion that the scores are correlated with proposed budget increases. This was true for FY 2004, and it is true for FY 2005. In this sense, the administration's performance budgeting initiative is a success. Several outstanding questions about the effort remain, however. First, how is performance information influencing appropriations? The administration has been criticized by Congress, the GAO, and public interest groups for not selling their performance initiative to Congress.[11] Indeed, in one hearing held on Capitol Hill, a longtime appropriations staff member frankly acknowledged that he did not know

11

what PART was.[12] If performance information influences administration proposals but does not influence appropriations, federal program managers will have less of an incentive to comply with the PART system and use it as a guide for management improvement. In budget-performance integration the budget is the tool by which agencies and programs are punished or rewarded for meeting performance goals, but at this point it is still unclear how performance information is being used in appropriations. If it is working its way into appropriations through administration proposals only, how much so? The FY 2006 budget will include actual appropriations figures for the FY 2004 budget cycle and the first cohort of programs graded. This will allow researchers the first opportunity to see how effective the PART system is in influencing actual program dollars.

Second, how much influence does the political content of the program have on both its PART score and on program budget increases or decreases? In FY 2004 programs housed in Democratic departments got systematically lower budgets than other programs. PART scores also mattered more for programs housed in these departments. In FY 2005, however, political content measured in this way had no systematic effect on either budgets or the way that PART scores influenced budgets. More data and better measures of program political content will help resolve the issue of how much politics influences both performance measurement and budgets.

Third, what impact does the lack of good performance measures have on performance budgeting? Our analysis found that performance and results do not play as prominent a role as might be expected from an effort at performance budgeting. For both cohorts, the "results" portion of PART has less influence on budget decisions than other, less outcome-related sections of PART. It seems plausible that the underreliance on performance measures and overreliance on other measures is due to a shortage of good performance measures. The greatest weakness of PART, an important impediment to the implementation of real performance budgeting, is the shortage of good performance measures. In the first year, half of the programs lacked adequate measures; in the second year, more programs had measures deemed adequate, but 40 percent still lacked adequate measures. The results are not conclusive, but it appears likely that results influence budget allocations more for programs with good measures than for programs that lack good measures. To make PART performance budgeting in fact as well as name, better measures are urgently needed.

Because too few programs have good measures of results, OMB cannot rely much on the results section of PART and instead appears to use that which they can measure in making budget choices. What they can measure, unfortunately, has little relationship with performance and is much closer to the kind of proceduralism that performance budgeting is intended to displace.

REFERENCES

General Accounting Office (GAO). 1993. *Performance budgeting: State experiences and implications for the federal government*. Washington, DC: Government Printing Office.

———. 2004. *Performance budgeting: Observations on the use of OMB's program assessment rating tool for the fiscal year 2004 budget*. Washington, DC: General Accounting Office. Available at http://www.gao.gov/new.items/d04174.pdf (accessed November 7, 2005).

12

Gilmour, John, and David E. Lewis. 2005a. Does performance budgeting work? An examination of OMB's PART scores. *Public Administration Review*, forthcoming.

480 ———. 2005b. Political appointees and the competence of federal program management. *American Politics Research*, forthcoming.

Gruber, Amelia. 2004a. Bush seeks $1 billion in cuts for subpar programs. *Government Executive Magazine*, 30 January.

———. 2004b. Program cuts figure into Bush plan to control deficit. *Government Executive Magazine*,

485 2 February.

———. 2004c. President and CEO: George W. Bush's ambitious management reform agenda is only beginning to show results. *Government Executive Magazine*, 15 July:36–46.

Joyce, Philip G. 1999. Performance-based budgeting. In *Handbook of Government Budgeting*, ed. Roy T. Meyers. San Francisco: Jossey-Bass.

490 Key, V. O., Jr. 1940. The lack of a budgetary theory. *American Political Science Review* 34: 1137–44.

Knott, Jack, and Thomas Hammond. 1980. *A zero-based look at zero-based budgeting*. New Brunswick, NJ: Transactions Books.

Lynn, Laurence E., Jr. 1998. Requiring bureaucracies to perform: What have we learned from the

495 U.S. Government Performance and Results Act (GPRA)? Working Paper, University of Chicago. Available at http://harrisschool.uchicago.edu/About/publications/working-papers/pdf/wp_lynn.pdf (accessed November 7, 2005).

McCarty, Nolan. 2004. Bargaining over authority: The case of the appointment power. *American Journal of Political Science* XX.

500 Melkers, Julia E., and Katherine G. Willoughby. 2001. Budgeters' view of state performance-budgeting systems: Distinctions across branches. *Public Administration Review* 61:54–64.

Merton, Robert K. 1968. Bureaucratic structure and personality. In *Social Theory and Social Structure*, by R. K. Merton. New York: Free Press.

Office of Management and Budget (OMB). 2003. *Performance and management assessments, budget*

505 *of the United States government, fiscal year 2004*. Washington, DC: Government Printing Office.

Wildavsky, Aaron B. 1975. *Budgeting: A comparative theory of budgetary processes*. Boston: Little, Brown.

Willoughby, Katherine G., and Julia E. Melkers. 2001. Assessing the impact of performance budgeting: A survey of American states. *Government Finance Review* 17:25–.

510 Ziegler, Mollie. 2004. Budget axe weeds out poor performers: 15 programs get the ax. *Federal Times*, 2 February.