**John B. Gilmour**
College of William and Mary
**David E. Lewis**
Princeton University

Spotlight on Public Finance

# Does Performance Budgeting Work? An Examination of the Office of Management and Budget's PART Scores

**John B. Gilmour** is a professor of government and public policy at the College of William and Mary. His research focuses on budgetary politics and legislative–executive bargaining. He has published two books, *Reconcilable Differences? Congress, the Budget Process, and the Deficit* (University of California Press, 1990) and *Strategic Disagreement: Stalemate in American Politics* (University of Pittsburgh Press, 1995), as well as articles in the *American Journal of Political Science, Journal of Politics,* and *Legislative Studies Quarterly.*
**E-mail:** jbgilm@wm.edu.

**David E. Lewis** is an assistant professor of politics and public affairs at Princeton University. His research focuses on the presidency, executive branch politics, and public administration. He is the author of *Presidents and the Politics of Agency Design* (Stanford University Press, 2003) and journal articles on American politics and public administration.
**E-mail:** delewis@princeton.edu.

*In this paper, the authors use the Bush administration's management grades from the Program Assessment Rating Tool (PART) to evaluate performance budgeting in the federal government—in particular, the role of merit and political considerations in formulating recommendations for 234 programs in the president's fiscal year 2004 budget. PART scores and political support were found to influence budget choices in expected ways, and the impact of management scores on budget decisions diminished as the political component was taken into account. The Bush administration's management scores were positively correlated with proposed budgets for programs housed in traditionally Democratic departments but not in other departments. The federal government's most ambitious effort to use performance budgeting to date shows both the promise and the problems of this endeavor.*

In the last decade, performance measurement has emerged as the most important public sector management reform in many years, surpassing even management by objectives, total quality management, zero-based budgeting, and program planning and budgeting in the speed and breadth of adoption. Nearly all of the states use some form of performance measurement, and the federal government has also implemented performance measurement in various ways. Closely related to performance measurement is the idea of performance budgeting, or performance-based budgeting, which seeks to link the findings of performance measurement to budget allocations (Joyce 1999). Performance budgeting has been widely adopted abroad (Schick 1990), and, as of 1998, 47 out of 50 states had adopted some form of performance budgeting (Melkers and Willoughby 1998). Both performance measurement and performance budgeting are part of a worldwide effort to transform public management (Kettl 2000).

Starting with the fiscal year (FY) 2004 budget, the Office of Management and Budget (OMB) began to include performance and management assessments of federal programs and to use that performance information in allocating budget resources. This initiative is called PART, short for Program Assessment Rating Tool. In this paper, we explore performance budgeting through an examination of the PART experiment. Specifically, we investigate the role that merit and political considerations played in formulating OMB recommendations for the 234 programs in the president's FY 2004 budget proposal.

This paper has three goals: The first is to assess the extent to which budget allocations in the president's FY 2003 budget were influenced by merit, as measured by PART scores; we found that PART scores and political support influenced budget choices in expected ways. The second goal is to assess the extent to which the observed relationships between performance measures and budgets were a function of political influence on the PART scores themselves. It is possible that the positive relationship between PART scores and the budget was the result of partisan elements of the PART scores. We found that the impact of the PART scores on budget decisions diminished when the political component of the scores was taken into account. A third and final goal is to determine whether performance measures were used in an impartial manner. Given the lack of direct means of translating performance measures into budget decisions, it is possible that favored programs were insulated from negative performance ratings, whereas disfavored programs that could not show results were cut. We found that PART scores were positively associated with Democratic programs but not the rest.

## Performance Budgeting in Practice

Governments adopt performance measurement and performance budgeting for a number of reasons, but probably the most important is the promise these practices hold for determining which government programs produce results and thus deserve budget increases. Unlike private sector enterprises, most government programs are not designed to yield a profit. Without the profit motive, it is difficult to know which programs are generating benefits and which are not. Performance measurement can help with this problem by producing quantitative evidence that shows which programs are accomplishing their purposes. Performance budgeting integrates the results of performance measurement into the budget process, ideally resulting in budget allocations that more closely reflect the relative merit of the programs.

There is little systematic evidence thus far that performance budgeting, as it has been implemented in states and cities, has had a major impact on budgeting decisions. In 1993, the U.S. General Accounting Office reported that "in states regarded as leaders in performance budgeting, performance measures have not attained sufficient credibility to influence resource allocation decisions. . . . [R]esource allocations continue to be driven, for the most part, by traditional budgeting practices" (GAO 1993, 1). A more recent survey of state budget officials by Melkers and Willoughby (2001) indicates that performance budgeting does not have a major impact on how money is allocated. Only 39 percent of those who responded to the survey agreed that "some changes in appropriations were directly attributable" to performance budgeting. But respondents overwhelmingly agreed that performance budgeting had increased their workload. In his essay on performance budgeting, Joyce concludes that "[d]espite the bumper-sticker appeal of these prescriptions . . . the connection between performance and the budget in practice is elusive" (1999, 617). It remains to be seen whether the federal government can be more successful in translating performance measures into budget decisions.

Performance budgeting is a troublesome enterprise because it is difficult to know how to use performance information. If a program performs poorly, does that mean it should be cut because it is wasting money or increased so that it can do better? Few people (apart from some Libertarians) would argue that because the Border Patrol does not succeed in sealing the Mexican border against illegal immigrants, its budget should be slashed. There are many other important programs for which evidence of weak performance could be interpreted as requiring more resources, not fewer, on the grounds that the program's mission is so important that it cannot be permitted to fail. Because of these complications, it is difficult to argue for any kind of mechanistic link between evidence of performance and budget decisions, and the OMB has never claimed any such direct link in its use of PART scores. In performance budgeting, measures still must be interpreted and evaluated in the context of the programs, their mission, and their history.

> Performance budgeting is a troublesome enterprise because it is difficult to know how to use performance information. If a program performs poorly, does that mean it should be cut because it is wasting money or increased so that it can do better?

A risk of using performance budgeting is that because its implementation involves subjective judgments, it will be politicized. Certain programs are more appropriate for the use of performance information in determining budget allocations. Many programs provide services that are important but not essential and that compete with or overlap other programs to varying degrees. One could use performance information to shift resources among such programs in order to achieve greater allocative efficiency. Determining which programs are so essential that their failure is unacceptable will never be an impartial process: It is likely that each party will see the programs it likes and supports as essential and unlikely that it will see weak performance as evidence that a program should be cut. Thus, it is possible that the party in power will implement performance budgeting in a politicized way, insulating the programs its favors from negative performance evaluations but cutting the budgets of programs they do not favor that are unable to demonstrate results.

An additional risk in implementing performance budgeting is that the measures employed will reflect political favoritism in addition to merit. It is impossible for performance measures to be perfect assessments of "true merit" in programs, but the measures themselves should not be systematically associated with or determined by the political preferences of the president or governor. When performance measures incorporate a significant political component, they cease to be performance measures and become political measures, and their use in budgeting is not easily distinguishable from standard budgeting practices. In previous work (Gilmour and Lewis 2006), we found that programs established under Democratic presidents received systematically lower PART scores—about 5.5 points lower than programs initiated under Republican presidents. We do not know why this is the case or by what means the disparity was introduced, but this finding suggests that PART scores may measure the political support of programs as well as merit. It could also be that the missions of programs begun under Democratic presidents are inherently less measurable or simply harder to accomplish.

## Performance Measurement in the Bush Administration

In its FY 2004 budget, the Bush administration numerically graded the quality of management in 234 federal programs (20 percent). The grading scheme is relatively straightforward. It was designed by the OMB in consultation with the President's Management Council, an advisory council of lower-level agency political appointees, and includes numerical grades ranging from 0 to 100 in four categories and a total weighted numerical management grade. The four categories are as follows:[1]

1. *Program purpose and design* (20 percent) assesses whether the program design and purpose are clear and defensible.
2. *Strategic planning* (10 percent) assesses whether the agency has set valid annual and long-term goals for the program.
3. *Program management* (20 percent) rates the agency's management of the program, including financial oversight and program improvement efforts.
4. *Program results* (50 percent) rates the program's performance according to the goals reviewed in the strategic planning section and other evaluations.

Grades were determined in each category based on answers to a series of yes/no questions and adjusted for the type of program under consideration (block grant, regulatory, credit, etc.). For example, one question used to assess the quality of strategic planning asks, "Does the program have a limited number of specific, ambitious long-term performance goals that focus on outcomes and meaningfully reflect the purpose of the program?" For this and other questions, the OMB provided background information on the purpose of the question and the elements of an affirmative response. Answers were determined jointly by the agency running the program and an OMB examiner. In cases of disagreement, the answers were resolved through arbitration by the OMB hierarchy, namely, the OMB branch chief and (if necessary) the division director and program associate director. A separate score was calculated and reported for each section; these were summed to produce a total weighted score, which is the PART score used in this paper.

In addition to reporting numerical scores, OMB also assigned management and performance grades to the programs. These range from the highest grade of *effective* to *moderately effective, adequate,* and the lowest score, *ineffective.* In addition, another grade is offered, *results not demonstrated.* Figure 1, a scatterplot of grades by summary PART scores, shows there is a very close relationship between scores and grades, although programs rated *results not demonstrated* have scores ranging from very high to very low. In the figure, we place this rating between *ineffective* and *adequate.*

## Connecting Performance and Budgeting

The OMB claims a significant relationship between PART scores and budget allocations. According to the OMB, "The PART is an accountability tool that attempts to determine the strengths and weaknesses of federal programs with a particular focus on the results individual programs produce. Its overall purpose is to lay the groundwork for evidence-based funding decisions aimed at achieving positive results" (OMB 2003, 9). The Performance Institute, which appears to work closely with the OMB in this endeavor, states that "the president's proposal rewards programs deemed effective with a six percent funding increase, while those not showing results were held to less than a one percent increase" (Performance Institute 2003).

Because the OMB has published its management grades in budget documents and on its Web site, we can examine these claims more closely. It has also published the federal government's FY 2002 appropriations and the Bush administration's proposed FY 2003 and FY 2004 budgets, along with the grades for each program. We focus primarily here on the percent change in the FY 2003 and FY 2004 budgets.[2] This value should reflect the impact of performance assessment on budget allocations. However, one problem

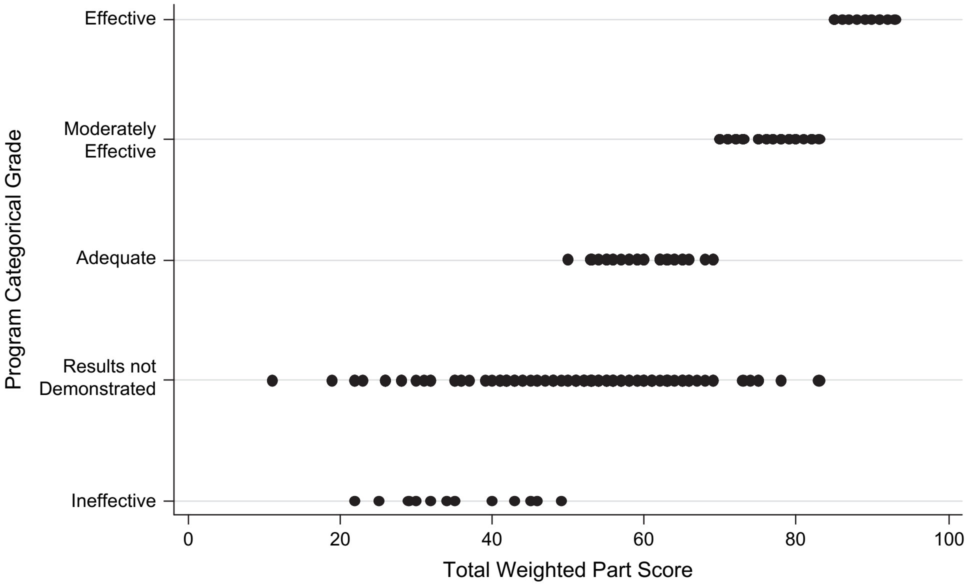

**Figure 1 PART Scores and Performance Grades**

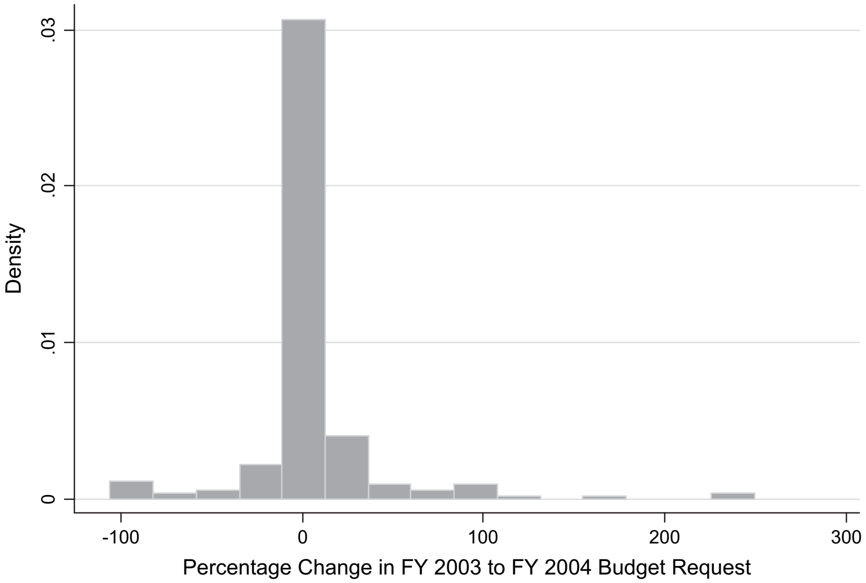

**Figure 2 Histogram of Dependent Variable**

with analyzing the percent change in program budgets is that there are some extreme outliers. Some programs received increases of more than 200 percent, others were cut by 100 percent, and others received incremental increases of varying sizes. We include a histogram of the proposed FY 2004 budget changes in figure 2.

The very large budget changes are a problem for two reasons. First, the process that generates such large changes is different from the process that generates the typical incremental changes in program budgets (Wildavsky 1984). Lumping incremental and nonincremental changes together may be inappropriate if they result from different processes and have different causes.

A second problem is that cases with very large changes are often small programs, and therefore the large percentage increases represent small amounts of money. But in a regression or correlation, such small outlying cases can exert a tremendous and disproportionate influence. Using the raw budget change percentages yields perverse results. For example, there is a negative correlation between budget change in FY 2002–03 and budget change in FY 2003–04. Certainly, some programs will experience a regression to the mean effect following a large increase or decrease, but it is not generally the case that program budget allocations seesaw wildly from positive to negative and back again. Incremental changes are far more common (Wildavsky 1984).

A couple of comparisons between cases with changes greater than 50 percent in FY 2004 and the rest will clarify the difference between programs with large changes and the rest. For cases with an increase or decrease of more than 50 percent, the median budget size was $27 million. For those with changes of less than 50 percent, the median budget was $390 million. For cases with changes greater than 50 percent, the median increase or decrease was 98 percent. For other programs, the median increase or decrease was 4.5 percent.

There is no settled rule for dealing with outliers, but one common way of solving the problem is to log the variable. In this case, however, we cannot log the variable because it includes negative numbers.[3] Another common way of dealing with outliers is to exclude them, using some decision rule to determine which cases are outliers and which cases are not. It is common, then, to perform robustness checks to see whether the decision rule makes a difference. For this study, we excluded cases in which the one-year change was greater than 50 percent. For the FY 2004 budget, this means that 29 cases were excluded. We used another decision rule to exclude all cases that were more than two standard deviations away from the mean.[4] We replicated all analyses in this paper using the two-standard-deviation rule, and the results were actually stronger than the results presented here. It is important to note, however, that the decision to exclude outliers *is* consequential, as it alters some of the regression results in important ways. We will address this further in our discussion of table 1.

Measuring merit is straightforward because we have relied on the OMB's PART scores. The measure of merit used is the PART score. At best, scores of this kind are imperfect measures of results and management, and they may incorporate certain kinds of bias. But it is still reasonable to believe that the scores are significantly correlated with actual merit.

Figure 3 shows a scatterplot of the relationship between the percent increase in budget for programs in the PART and PART summary scores. There is a clear positive relationship, showing that programs with higher PART scores received larger budget increases. This suggests that the administration took performance into account when proposing budgets, provided that the management grades themselves were not politicized.

Measuring political influence in the budget process is more complex. Our expectation was that typically "Democratic" programs would receive less generous budgets and perhaps lower management grades. Measuring the political content of federal programs is difficult because programs usually have supporters on both sides of the aisle and are reauthorized numerous times after their initiation, and because the current administration does not publicize which programs it supports or opposes on ideological grounds.

As a first cut, we tried to loosely group programs as Democratic or Republican according to the department in which they were housed. Because certain departments within the executive branch work in areas that are more central to the agenda of a particular party, we believed that departmental affiliation might provide a reasonable proxy for political favor. We created a *Democratic department* variable to distinguish between programs in disfavored departments and those situated elsewhere. The Republican Party has been somewhat hostile to a number of cabinet-level departments and independent agencies. For example, it has proposed eliminating the Departments of Commerce, Education, and Energy. In addition, the Departments of Housing and Urban Development (HUD), Labor, and Health and Human Services (HHS), as well as the Environmental Protection Agency (EPA), all have agendas that are central to the Democratic Party but not to the Republicans. All programs in the PART housed in one of these departments were coded 1, and the rest were coded 0. This is a crude measure, as there are some programs in these departments that Republicans support and programs in other departments they do not support, 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. But to avoid an overall ad hoc approach to constructing this variable, we relied on our conception of the traditional positions of the parties. We assumed that, collectively, the programs coded 1 would be supported more weakly than programs coded 0. It might have been better to have a panel of experts evaluate all 234 programs and make individual determinations of whether each appeared to be favored or disfavored by the administration, but such codings are highly subjective. Furthermore, many of the programs were sufficiently small and obscure that few coders would have had knowledge of all of them, and their decisions would have been based largely on guesswork.

One can imagine other ways of assessing political support for programs. The seven departments included in the Democratic department variable were opposed by Republicans on varied grounds. Four (HUD, HHS, Labor, and the EPA) have missions that generally match the Democratic Party's agenda. Another three (Education, Energy, and Commerce) have missions that have been opposed by the Republican Party on the grounds that their missions are inconsistent with markets or federalism. Furthermore, some reviewers of this article contended that because the Bush administration is not hostile to education programs and has not sought the elimination of the Commerce or Energy Departments, we should not

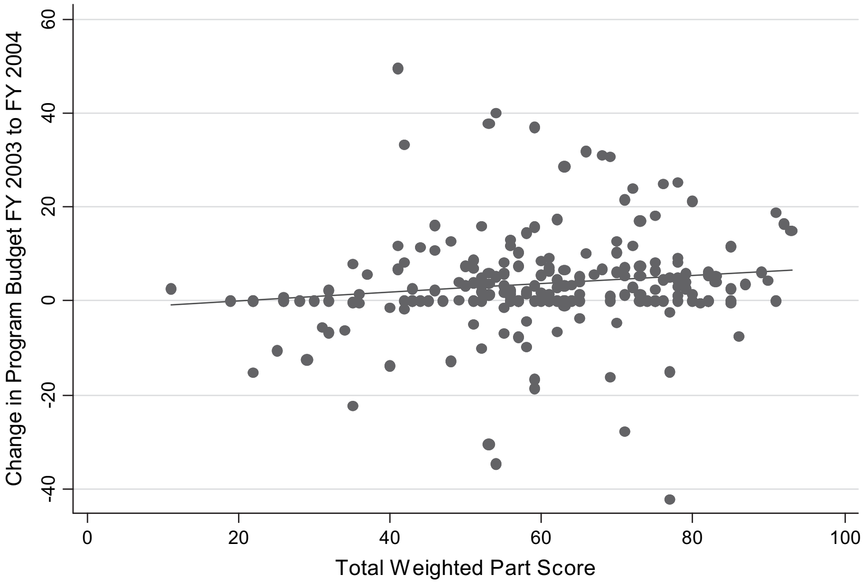

**Figure 3 Impact of PART Score of FY 2004 Budget**

lump them in with the core Democratic departments. Thus, in some models, we divided the Democratic department variable in two separate variables: one consisting of "core" Democratic departments and the other consisting of departments that the Republican Party has proposed eliminating.

With these caveats, we graphed the PART scores and proposed budget changes according to whether the programs were housed in departments typically associated with the Democratic Party's political agenda (figure 4). The PART scores appear to be more highly correlated with budget increases or decreases for more Democratic programs. This suggests that merit evaluations may be more important for traditionally Democratic programs, whereas other program budgets are insulated from the influence of merit evaluations.

Another way to measure the political content of a federal program is to analyze the political environment at the time the program was established. Because programs begun under Democratic Congresses or Democratic presidents might exhibit characteristics that endear them to Democrats and not to Republicans, we created dummy variables for *Democratic president* (0,1), *Democratic Congress* (0,1), and *unified government* (0,1), as well as an interaction of these three variables (0,1). One shortcoming of these coarse measures of program content is that they do not capture bipartisanship in program support, subsequent program authorizations, or variation in ideology among politicians from the same party.

Prior budget support can be at least partly a measure of political favor. Programs that received larger increases from FY 2002 to FY 2003 are likely to be more favored by the administration than programs that received smaller raises or even cuts. Therefore, we devised a second budget variable to measure the percent budget change between the amount appropriated in FY 2002 and the amount requested by the president in FY 2003.[5]

We shall proceed in three stages: The first stage is a regression analysis that investigates the role of PART scores and other political variables on budget allocations. Second, because PART scores may have been partially determined by political factors, such as party control at the time of a program's creation, it is possible that the observed influence of PART scores on the budget was actually a function of political considerations. We estimated a model of FY 2004 budget change with two-stage least squares. The third stage examines whether PART scores were used in an impartial manner. To accomplish this, we estimated the regression models separately for programs in traditionally Democratic departments and for all other programs.

## Results

The first set of models, shown in table 1, uses ordinary least squares regression to assess the influence of various factors on budget allocations in simple models without controls. In all models, the dependent variable is the change in the OMB's recommended levels from FY 2003 to FY 2004. The mean budget change was 3.6 percent and the standard deviation was 11.3. The largest changes in the sample, after excluding outliers, were a decrease of 42 percent and an increase of 49 percent. We report robust standard errors and indicate significance at standard levels in one-tailed tests because we have directional hypotheses about the impact of both merit and political factors on budgets.[6]

One key finding is that the PART score variable has a positive coefficient and is statistically significant in all models. This suggests, at least preliminarily, that merit

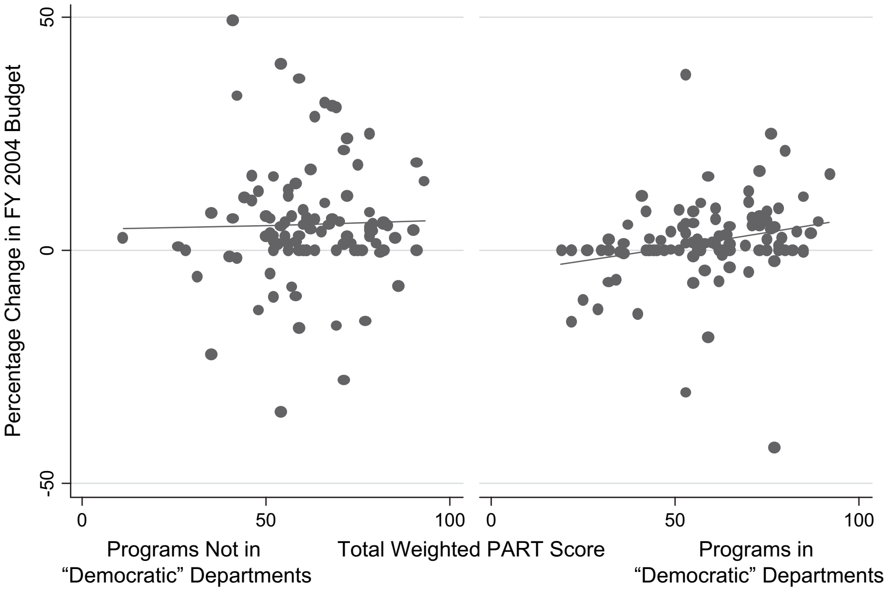

**Figure 4 PART Scores and the FY 2004 Budget by Democratic Departments**

**Table 1** Models of Fiscal Year 2004 Program Budget Increases or Decreases

| | | | | | | |
|---|---|---|---|---|---|---|
| **Merit** | | | | | | |
| PART score | 0.08** (0.04) | 0.08 (0.04) | 0.08** (0.04) | 0.12** (0.04) | 0.11** (0.04) | 0.11** (0.04) |
| **Political content of program** | | | | | | |
| Housed in Democratic department (0,1) | −3.46** (1.62) | — | — | — | −4.39** (1.69) | — |
| Housed in core Democratic department (0,1) | — | −1.62 (1.71) | — | — | — | −3.87** (1.87) |
| Housed in department proposed for closing by Republicans (0,1) | — | −5.27** (1.92) | — | — | — | −4.98** (1.80) |
| Percent increase in FY 2003 budget | — | — | 0.11* (0.08) | — | 0.10* (0.08) | 0.10 (0.08) |
| Democratic president (0,1) | — | — | — | −3.26 (3.33) | 0.31 (2.89) | 0.30 (2.90) |
| Democratic Congress (0,1) | — | — | — | −5.42** (2.84) | −0.99 (2.07) | −0.96 (2.05) |
| Unified government (0,1) | — | — | — | −5.89** (3.19) | −3.52* (2.45) | −3.36 (2.48) |
| Interaction (0,1) | — | — | — | 7.73 (6.15) | 1.75 (4.73) | 1.70 (4.71) |
| Constant | 0.64 (3.13) | 0.39 (3.10) | −1.03 (2.37) | 1.82 (3.51) | 1.38 (2.66) | 1.39 (2.64) |
| Number of observations | 205 | 205 | 189 | 174 | 161 | 161 |
| $R^2$ | 0.04 | 0.05 | 0.04 | 0.06 | 0.12 | 0.13 |

Note: ** significant at the .05 level in one-tailed test; * significant at the .10 level in one-tailed test. Robust standard errors reported.

did play a role in the determination of program budgets. Not surprisingly, the political content of the programs appears to have influenced the proposed budgets. The Democratic department variable was negative and statistically significant in the models. Breaking the Democratic department variable in two produced modest changes. In one model, the variable for departments proposed for elimination had a larger coefficient than the variable for core Democratic departments, and in another, they were nearly identical. The variable measuring budget change from FY 2002 to FY 2003 also had a positive sign and was marginally significant. Using the political configuration at the time a program was created to assess its content produced more ambiguous results. The estimates themselves suggest that programs established under unified Democratic control received systematically lower budgets. In divided government (defined as anything other than unified governance), the presence of a Democratic president or a unified Democratic Congress decreased a program's budget. Surprisingly, however, programs created under unified Republican control fared as poorly as those created under unified Democratic control and worse than those created under divided government. A closer examination of these programs revealed that 18 programs were established under unified Republican control. Of these 18 programs, five dated to the Civil War or Reconstruction periods. Interestingly, three programs in the model were initiated in 2001, and of those, two received either no increase or a budget cut.

This analysis, which does not consider possible political influences on PART scores, indicates that PART scores have a real impact on budget allocations, as do other political factors, such as Democratic department and budget change. Measures of program political content based on party control of the branches of government provided more ambiguous results.

Had we performed this analysis with the excluded outlier cases, the results would have been different. The coefficient for the PART score variable would have been larger, and the Democratic department variable would still have been negative but not statistically significant. The variable for budget change in FY 2003 would have had a negative sign. Thus, the finding that programs housed in Democratic departments received less funding is contingent on excluding outliers.

Using the coefficients in model 1, we can estimate the impact of changes in some of the independent variables on budget allocations. The Democratic department variable had a coefficient of –3.5 or –4.4, meaning that all else being equal, a program in one of the Democratic departments would have received between 3.5 percent and 4.4 percent less than a program in another department. The PART score variable had a coefficient of 0.08 to 0.12, depending on the model. An increase from one standard deviation below the mean to one standard deviation above the mean (an increase of 33.6 points on the PART scale) would correspond to an increase in program budget of 2.7 percent to 4.0 percent.

The models in table 1 did not control for other factors whose omission from the regressions might have biased the estimates of those variables we care about. In table 2, we reestimated the models with appropriate controls. In particular, we included three types of controls. First, we controlled for program age; older programs should demonstrate less budget volatility and have had to survive multiple authorizations, implying a level of political support. The average program was 35 years old (SD = 32); the youngest was established in 2001 and the oldest in 1802 (patent and trademark programs).

**Table 2** Models of Fiscal Year 2004 Budget Increases or Decreases

| | | | | |
|---|---|---|---|---|
| **Merit** | | | | |
| PART score | 0.11** (0.05) | 0.11** (0.04) | 0.12** (0.05) | 0.12** (0.04) |
| **Political content of program** | | | | |
| Housed in Democratic department (0,1) | −12.89** (5.15) | — | — | −12.27** (4.37) |
| Percent increase in FY 2003 budget | — | 0.11* (0.08) | — | 0.10 (0.08) |
| Democratic president (0,1) | — | — | −3.78 (3.25) | −1.20 (2.62) |
| Democratic Congress (0,1) | — | — | −5.99** (2.89) | −2.38 (2.10) |
| Unified government (0,1) | — | — | −7.42** (4.00) | −5.51** (3.29) |
| Interaction (0,1) | — | — | 9.86* (6.26) | 6.16 (5.10) |
| **Other** | | | | |
| Age of program | −0.05* (0.03) | −0.04* (0.03) | −0.04 (0.05) | −0.02 (0.04) |
| Constant | 11.67 (8.74) | −2.57 (4.15) | 5.29 (6.09) | 9.48** (5.42) |
| Include program fixed effects | Yes | Yes | Yes | Yes |
| Include department fixed effects | Yes | Yes | Yes | Yes |
| Number of observations | 176 | 163 | 174 | 161 |
| $R^2$ | 0.20 | 0.26 | 0.23 | 0.27 |

Note: ** significant at the .05 level in one-tailed test; * significant at the .10 level in one-tailed test. Robust standard errors reported. Coefficients for program and department fixed effects were excluded to make the table manageable. These estimates are available from the authors upon request.

We also included indicator variables for the type of program. Patterns of budgeting vary according to what programs do. That is, block grant programs may be evaluated in systematically different ways from regulatory or direct federal programs. To account for these differences in program type, we included dummy variables for each type—competitive grant, block/formula grant, regulatory, capital assets and service acquisition, credit, direct federal, and research and development programs.[7]

Similarly, we included fixed effects for the departments housing the programs. Each department has a unique history, context, and relationship with congressional committees, as well as a political bent that can affect the budgeting for specific programs. In particular, different departments may have systematically larger or smaller budget constraints affecting program budgeting decisions that have nothing to do with the programs themselves.

The models in table 2 generally confirm the results reported in table 1. Importantly, merit evaluations continued to have a significant impact on the FY 2004 budget proposal, about the same magnitude as the larger effects reported in table 1. The coefficient indicating that a program was housed in a traditionally democratic department was larger, indicating that such programs received 12 percent to 13 percent smaller budget changes than programs in other departments.[8] The coefficient for budget change in FY 2003 was still positive and large, but the standard error was a bit larger than it was in table 1. The variables accounting for partisan composition at the time of a program's inception followed the same pattern as in table 1.

In the next step, a two-stage analysis was used to ascertain the extent to which (if at all) the political influences on PART scores undermined the finding that merit influences budget allocations. The finding that merit, in the form of PART scores, has an effect on budget choices might be the result of political influences on the PART scores. Elsewhere, we found that PART scores were influenced by the party of the president at the time a program was first created (Gilmour and Lewis 2006), with programs created under Republican presidents getting scores about 5.5 points higher than programs created under Democratic presidents. It follows that the increase in budgets we observed for programs with higher PART scores might actually be the result of political influences on the PART scores. We used a two-stage regression analysis to solve this puzzle.

We reestimated the models from table 2 with two-stage least squares. The system of equations estimated both the PART score and the size of the proposed budget change. To estimate such a system of equations, we needed to include appropriate instruments or variables that would influence the PART score but not the proposed budget change directly. We included indicators for programs administered by political appointees, programs administered by commissions, and programs whose managers serve for fixed terms, all variables used to estimate models of PART scores in Gilmour and Lewis (2006).[9] We report only the second-stage estimates in table 3, but we note that, apart from the instruments, none of the variables were successful in estimating the PART scores. The signs on the political factors were correctly signed and occasionally close to significance but not systematically so.

The most easily identifiable difference between the estimates in table 2 and those in table 3 is the sign and significance of the coefficients for PART scores. In the two-stage least squares, the coefficients were negative and insignificant, whereas they were positive and

**Table 3** Two-Stage Least Squares Models of Fiscal Year 2004 Budget Increases or Decreases

| | | | | |
|---|---|---|---|---|
| **Merit** | | | | |
| PART score | −0.07 (0.17) | −0.06 (0.12) | −0.09 (0.15) | −0.08 (0.13) |
| **Political content of program** | | | | |
| Housed in Democratic department (0,1) | −13.10** (7.33) | — | — | −17.18** (6.03) |
| Percent increase in FY 2003 budget | — | 0.10 (0.09) | — | 0.09 (0.09) |
| Democratic president (0,1) | — | — | −5.27* (3.87) | −2.60 (3.04) |
| Democratic Congress (0,1) | — | — | −7.27** (3.26) | −3.63* (2.30) |
| Unified government (0,1) | — | — | −9.62** (5.11) | −7.83** (3.76) |
| Interaction (0,1) | — | — | 12.68** (7.60) | 8.94* (5.80) |
| **Other** | | | | |
| Age of program | −0.040* (0.03) | −0.03 (0.03) | −0.00 (0.05) | 0.01 (0.04) |
| Constant | 22.49** (12.85) | 17.46 (9.92) | 31.20 (13.07) | 22.24 (10.37) |
| Include program fixed effects | Yes | Yes | Yes | Yes |
| Include department fixed effects | Yes | Yes | Yes | Yes |
| Number of observations | 165 | 156 | 163 | 154 |
| Adjusted $R^2$ | 0.16 | 0.22 | 0.19 | 0.22 |

Note: ** significant at the .05 level in one-tailed test; * significant at the .10 level in one-tailed test. Robust standard errors reported. Instrumented variable: PART Score. Instruments: political appointee manager, commission, fixed term for appointee. Coefficients for program and department fixed effects were excluded to make the table manageable. These estimates are available from the authors upon request.

significant in the previous models. Two explanations are that (1) the PART scores were politicized and the merit component of the scores had little effect on the budget, or (2) our estimates are flawed, perhaps because of the difficulty of measuring the political components of programs. We have no way of distinguishing between these possibilities. It is also possible that merit matters for some programs but not for others and that lumping them all together muted the true impact of merit. It is to this third possibility that we now turn.

In a politicized usage, PART scores might be used in evaluating programs that the administration does not view favorably, whereas other programs could be insulated from performance information. To test the hypothesis that PART scores might be used to evaluate Democratic but not Republican programs, we estimated the models discussed previously but separated programs in Democratic departments from the rest. The results were strikingly different for the two models: For the programs in Democratic departments, the budget change variable and the PART score were both positive and statistically significant. In the model for programs in non-Democratic departments, however, the PART score variable was negative. The FY 2003 budget change variable was close to 0 and insignificant. This indicates that evaluations of management quality matter for programs traditionally supported by Democrats but less so for Republican programs.[10]

Not surprisingly, political considerations and merit influence budget proposals for federal programs, although in a nuanced way. Neither the administration nor anyone else has argued otherwise. The administration has claimed all along that it would use the PART scores to determine budget increases and decreases but that some programs that are well managed may be cut and some that are poorly managed may receive increases. Interestingly, however, merit evaluations appear to matter more for programs in traditionally Democratic departments.

## Conclusion

Despite spreading enthusiasm for performance budgeting at the federal, state, and local levels of government in the United States, significant problems limit its implementation. The most important of these is the impossibility of devising an automatic or impartial means of translating performance information directly into budgeting allocations. A program that is performing poorly might perform better if given additional resources, whereas another very successful program may need no more than its current allocation. A number of factors, among them political preferences, could easily interfere with the translation of measures into budget recommendations. An additional difficulty is that if the measurement process itself is not neutral, political considerations may warp the assessments, as well as their application. In practice, performance budgeting may reflect merit no more than traditional budgeting.

In a limited yet still important way, PART scores influence the OMB's budgetary allocations. Given the overwhelming importance of politics in making budgets, it is significant that PART scores have some impact. Despite this success, it is discouraging that the impact of PART is limited to Democratic programs. Advocates of Democratic Party budgetary goals can take some solace in these findings. They should expect that a Republican administration will reduce funding for programs that Democrats care about. Predictably, programs housed in Democratic departments received, on average, increases of 1.8 percent, compared

**Table 4** Two-Stage Least Squares Models of Fiscal Year 2004 Budget Increases or Decreases

| | Democratic Departments | | Other Departments | |
|---|---|---|---|---|
| **Merit** | | | | |
| PART score | 0.23** (0.11) | 0.18* (0.11) | −0.30* (0.21) | −0.24 (0.28) |
| **Political content of program** | | | | |
| Percent increase in FY 2003 budget | 0.10* (0.07) | 0.11* (0.07) | 0.03 (0.15) | −0.03 (0.15) |
| Democratic president (0,1) | −0.79 (2.49) | −1.40 (2.23) | 1.52 (9.29) | −4.05 (9.41) |
| Democratic Congress (0,1) | −0.98 (2.25) | −1.66 (1.88) | −1.99 (4.28) | −8.72 (6.68) |
| Unified government (0,1) | −6.18** (2.80) | −7.98** (2.67) | −0.58 (6.03) | −5.84 (7.31) |
| Interaction (0,1) | 4.58 (4.80) | 6.57* (4.40) | −1.23* (12.32) | 14.08 (16.78) |
| **Other** | | | | |
| Age of program | −0.01 (0.03) | 0.01 (0.03) | 0.00 (0.07) | −0.05 (0.09) |
| Constant | 9.34* (6.29) | −8.06 (7.84) | 26.14** (13.37) | 35.01 (26.70) |
| Include program fixed effects | No | Yes | No | Yes |
| Include department fixed effects | No | Yes | No | Yes |
| Number of observations | 87 | 87 | 67 | 67 |
| Adjusted $R^2$ | 0.20 | 0.39 | — | 0.20 |

Note: ** significant at the .05 level in one-tailed test; * significant at the .10 level in one-tailed test. Robust standard errors reported. Instrumented variable: PART Score. Instruments: political appointee manager, commission, fixed term for appointee. Coefficients for program and department fixed effects were excluded to make the table manageable. These estimates are available from the authors upon request.

with 5.6 percent for other programs. The differential use of PART scores suggests that reduced funding for Democratic programs is at least being allocated in an efficient manner that will generate the most benefit for the money.

Although this paper has reported only a very modest connection between measured performance and budget decisions by the OMB, the impact on appropriations may be smaller still. This paper assessed the impact of PART scores only on OMB recommendations, not actual appropriations. It is likely that the impact of PART scores will be further attenuated as the president's budget is considered in Congress. As of July 2003, indications were that staff members of the appropriations committees in Congress had little understanding or awareness of PART scores and little interest in them (Gruber 2003). The OMB may be able to persuade congressional committees to take performance evaluation seriously, but the committees may also choose to disregard this kind of performance information and rely on other criteria in formulating appropriations bills.

The results of this research bear out the difficulties of introducing performance-based budgeting. The ordinary least squares regression analysis reported in tables 1 and 2 shows that PART scores had an impact on budget choices, but the two-stage analysis in table 3 shows that, after controlling for political influences on PART scores, they had no discernable impact on the budget. But political factors did have a significant impact in both the one-stage and two-stage analyses. The disparity between the findings in tables 1 and 3 is at least partly resolved by table 4, which shows that PART scores influenced budget allocations for programs housed in Democratic departments but not other programs. This last finding underscores the difficulty of using performance information in an impartial way. It appears to be easier to implement performance budgeting with programs that one does not support.

## Acknowledgments

The authors thank Larry Bartels, Nolan McCarty, and Donald Kettl for help on this project and the OMB and Performance Institute for publishing the administration's management grades. Any errors that remain are our own.

## Notes

1. See OMB (2002, 2003).
2. Budget change is calculated as [(FY 2004–FY 2003)/FY 2003] * 100.
3. In an analysis in which we used a logged budget variable, including outliers but excluding cases with negatives increases (or cuts), the findings were not different from those reported here.
4. This amounts to excluding all cases in which the one-year change was greater than 80 percent. We estimated all of the models in which the one-year change was greater than 70 percent, 60 percent, and 40 percent. In general, these models confirmed (with some variation) what is reported here and available from the authors.
5. Because President Clinton proposed the FY 2002 budget in January 2001, the FY 2003 budget was the first put together by the Bush administration.
6. We report robust standard errors because a Breusch-Pagan test indicated that we could reject the null of constant variance ($p < 0.00$).
7. See OMB (2002).
8. The two alternate measures of the Democratic department variable produced nearly identical results in the models of table 2 and therefore are not reported here. The same is true for the results in table 3.

9. Although an argument could be made that the factors included here may have influenced budget change directly, we think they are appropriate instruments. First, if modeling change in the proposed budget for its own sake, we would be unlikely to think of any of these variables as being important determinants of budget decisions (see, however, McCarty 2004). Second, any effect these variables might have on the budget is likely to be channeled through their impact on management. If commissions historically receive smaller budgets, for example, it is likely because of frequent criticisms about management and planning (Arnold 1998). Third, we note that none of the variables had a coefficient distinguishable from zero in bivariate regressions on the dependent variable. Finally, the dependent variable we modeled is a difference, and the impact of the factors that we included should more proximately affect the budget level rather than the change from year to year. McCarty (2004) argues that programs that are insulated from presidential control (by devices such as fixed terms) are likely to receive higher budgets than uninsulated programs. However, in his model, these factors influenced the budget decisions of Congress rather than the president. It is not clear how the degree of independence from the president affects presidential budget requests. One could speculate that programs that are insulated from presidential control receive lower presidential requests than other programs because the president has less influence. However, congressional response could mitigate this effect. Therefore, it is unclear whether organizational structure has any direct and systematic influence on program budgets.
10. This also appeared to be true when we defined Democratic departments differently. When we reestimated the models in table 4 with different definitions of Democratic department, the results were close to those reported in table 4. One difficulty with reestimating these models, however, is that the new definitions decreased the sample size for the Democratic department regressions to 47 and 40 programs, respectively. In all cases, the coefficient on PART scores was positive (as expected) and in some cases larger than the coefficients reported in table 4. Not surprisingly, however, reducing the sample size from 87 to 47 increased the size of the standard errors. When Democratic department was defined as departments targeted for termination by Republicans (Education, Energy, and Commerce), the PART score coefficient was significant at the .05 level or .10 level and larger than those in table 4. When Democratic department was defined as the EPA, HHS, HUD, and Labor, the coefficient was positive but smaller and insignificant.

## References

Arnold, Peri E. 1998. *Making the Managerial Presidency: Comprehensive Reorganization Planning, 1905–1996*. 2nd ed. Lawrence: University Press of Kansas.

Gilmour, John B., and David E. Lewis. 2006. Political Appointees and the Competence of Federal Program Management. *American Politics Research* 34(1): 3–21.

Gruber, Amelia. 2003. OMB to Brief Appropriators about Program Ratings. *Government Executive*, July 2. www.govexec.com/dailyfed/0703/070203a2.htm [accessed June 7, 2006].

Joyce, Phillip G. 1999. Performance-Based Budgeting. In *Handbook of Government Budgeting*, edited by Roy T. Meyers, 597–619. San Francisco: Jossey-Bass.

Kettl, Donald F. 2000. *The Global Public Management Revolution*. Washington, DC: Brookings Institution Press.

McCarty, Nolan. 2003. The Appointments Dilemma. *American Journal of Political Science* 48(3): 413–28.

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

———. 2001. Budgeters' View of State Performance-Budgeting Systems: Distinctions across Branches. *Public Administration Review* 61(1): 54–64.

Performance Institute. 2003. Bush's '04 Budget Puts Premium on Transparency and Performance. News release, February 3. www.performanceweb.org/press/2003/ [accessed June 7, 2006].

Schick, Allen. 1990. Budgeting for Results: Recent Developments in Five Industrialized Countries. *Public Administration Review* 50(1): 26–34.

U.S. General Accounting Office (GAO). 1993. Performance Budgeting: State Experiences and Implications for the Federal Government. Washington, DC: Government Printing Office. GAO/AFMD-93-41. http://archive.gao.gov/d36t11/148507.pdf [accessed June 7, 2006]

U.S. Office of Management and Budget (OMB). 2002. Instructions for the Program Assessment Ratings Tool. www.whitehouse.gov/omb/mgmt-gpra/bpm852_add1_att-b.pdf [accessed June 7, 2006].

———. 2003. Budget of the United States Government, Fiscal Year 2004: Performance Management and Assessments. Washington, DC: U.S. Government Printing Office.

Wildavsky, Aaron B. 1984. *The Politics of the Budgetary Process*. 4th ed. Boston: Little, Brown.