Implementing OMB's Program Assessment Rating Tool (PART): Meeting the Challenges of Integrating Budget and Performance

by

John B. Gilmour*

This article examines the PART initiative of the United States Office of Management and Budget from a practical standpoint: how federal agencies have dealt with the requirements of PART and what strategies they have employed to be successful. The article highlights four challenges that confront both agencies and OMB as they work to complete assessments of all 1 000 programmes and describes approaches that agencies are taking to meet these challenges.

* John B. Gilmour is Professor of Government and Associate Director of the Public Policy Program at the College of William and Mary, Virginia, United States. This article was originally published in 2006 by the IBM Center for The Business of Government, *www.businessofgovernment.org*.

Foreword by Albert Morales, Managing Partner, and Jonathan D. Breul, Executive Director, IBM Center for The Business of Government

On behalf of the IBM Center for The Business of Government, we are pleased to present this report, "Implementing OMB's Program Assessment Rating Tool (PART): Meeting the Challenges of Integrating Budget and Performance" by John B. Gilmour.

In the 1990s, Congress and the executive branch established a statutory and management framework for strengthening government performance and accountability. The Government Performance and Results Act of 1993 was its centrepiece – providing congressional and executive decision makers with objective information on the effectiveness and efficiency of federal programmes and spending. The Bush administration has made integrating performance information into budget deliberations one of five government-wide management priorities under its President's Management Agenda. The Office of Management and Budget (OMB) has been using a Program Assessment Rating Tool, or PART, to assess programme performance and consider the information collected by PART during the annual budget review process.

Professor Gilmour's report examines OMB's PART initiative from a practical standpoint: How have federal agencies dealt with the requirements of PART? What strategies have they employed to be successful? What challenges do they face? His report highlights four challenges that confront both agencies and OMB as they work to complete assessments of all 1 000 programmes and describes approaches that agencies are taking to meet these challenges. The first challenge is for departments and agencies to organise for success. The second challenge of using PART is communicating accomplishments. The third is the challenge of developing suitable measures, and the fourth challenge is linking performance to outcomes.

We hope that this timely and informative report will be useful to public managers who want to understand how federal programmes have responded to PART and learn from the experience of programme-level and bureau-level staff in dealing with the demands of PART.

Executive summary

Performance budgeting and performance measurement, considered together, are the most ambitious public sector management reforms adopted in the last half century. This report is a study of the implementation of a federal performance budgeting management tool called PART, for Program Assessment Rating Tool. The Office of Management and Budget (OMB) initiated it in 2003 and will have completed assessments of all federal programmes by 2007. Performance budgeting addresses persistent difficulties in government budgeting, particularly the problems of knowing where to allocate resources most efficiently and holding programme managers accountable for producing results. Still, there are significant challenges in successfully realising the goals of performance budgeting.

OMB's implementation of performance budgeting is comprehensive and impressive, reflecting careful design and determined follow-through. This report examines the implementation of PART and the four challenges that have faced federal programmes as they have sought to meet its demands.

The first challenge is for departments and agencies to organise an appropriate means of managing the PART assessment process. There are great differences among departments and agencies in the scores given by OMB to their programmes, and it is almost certain that these differences are due in some measure to the nature of the departmental and agency responses.

The second is the challenge of using the PART questionnaire as a means of communicating the accomplishments and shortcomings of a programme to OMB and to other interested stakeholders. Without careful, hard work at mastering the PART instrument, even well-run programmes with good results are not guaranteed a good rating.

The third is the challenge of developing suitable measures. This is a challenge for many programme managers, since they are under pressure from OMB to develop measures of outcomes, and a challenge for OMB as well in that the success of PART as an assessment tool depends crucially on the development of appropriate measures. One of the successes of PART has been how it has encouraged more programmes to adopt outcome measures. Although OMB has stressed the importance of adopting end outcome measures, it has also, in fact, exhibited considerable flexibility and has endorsed output measures in cases where true outcome measures were unavailable or inappropriate.

The fourth challenge is that of interpreting programme performance measures and their associated results in order to understand the extent to which programme managers can be held accountable for their programme's performance. Holding programmes accountable for their observed results is important because it can encourage improvement, but there can be limiting

factors. In some cases outcomes are influenced by factors outside the control of the programme and its managers. In other cases programmes are saddled with statutory designs that impede their ability to produce desired results. It is important for OMB to hold programmes accountable, but especially for that which they can control.

After four years of PART, the greatest accomplishment so far has been in producing useful assessments of 800 programmes. OMB is on track to finish assessments of all federal programmes in 2007. There is evidence that PART assessments have an impact on allocation decisions in the president's budget. Yet, thus far, there is little evidence that PART has caused significant changes in programme management. While it is too soon to expect many programme changes in response to PART, they will be an important achievement of PART in the years to come.

The report includes recommendations for both departments and agencies, as well as OMB.

Recommendations for departments and agencies:

1. Don't give PART to the intern.
2. Get professional help if needed.
3. Work with your OMB examiner.
4. Link PART to your strategic plan.
5. Read OMB's guidance carefully.
6. Provide ample documentation.
7. Measure what you can.
8. Understand OMB's perspective.
9. Renegotiate the definition of the programme.
10. Express measures in non-technical language.

Recommendations for OMB:

1. Formally introduce appropriate flexibility about what constitutes acceptable measures.
2. Provide multiple response categories for answers to PART questions.
3. Distinguish between design and management failures.

1. Introduction

For more than a decade, the federal government has been on a journey to improve performance and accountability by measuring how well its programmes work and to link measures of performance to the allocation of budgetary resources (Joyce, 1999). For example, the goal of the Government Performance and Results Act of 1993 (GPRA) was to refocus efforts of

government agencies on results as opposed to inputs and standard operating procedures. GPRA requires agencies to adopt mission statements, strategic plans, and measures of both programme outputs (such as the number of infants vaccinated) and outcomes (such as the reduction in infant mortality). It is now routine to see a mission statement and strategic plans prominently displayed on the websites of government agencies. After the first decade, the Government Accountability Office (GAO) found that while GPRA created a steady supply of performance information, there was not a strong demand for the use of this information by policy makers or programme leaders.

Shortly after he took office in 2001, President George W. Bush committed to an ambitious agenda of improved government management. A key element of his agenda was to make the government more results oriented by expanding the use of performance budgeting. He directed the Office of Management and Budget (OMB) to work with each agency to recast their budget to include performance information. In addition, in 2003, he expanded this effort by committing to a programme-by-programme assessment of performance. At the time, it was estimated that there were about 1 000 major programmes that might be assessed. President Bush directed OMB to lead this assessment effort, as well. OMB developed an assessment framework, with the assistance of agencies and outside experts, which it named the Program Assessment Rating Tool, or PART. PART has become a vital component of the President's Management Agenda and the administration's performance budgeting initiative.[1]

PART is explicitly designed to build upon the performance information developed by agencies in response to GPRA. The Fiscal Year (FY) 2005 PART guidance to agencies issued by OMB states: "The PART is a vehicle for achieving the goals of GPRA." PART appears to put more "teeth" in GPRA, especially since OMB, which administers PART, develops the president's budget, and its budget decisions are to be influenced to some extent by PART.

In the FY 2004 budget, released in early 2003, the Bush administration numerically rated the effectiveness of 234 major federal programmes, or about 20%. In each of the three succeeding budgets, approximately 200 additional programmes were assessed, for a total of about 800 as of 2006. The grading scheme is relatively straightforward. It was designed by OMB in consultation with an outside advisory panel and the President's Management Council, composed of the deputy secretaries and chief operating officers of the departments and major agencies. It is based on 25 to 30 questions, grouped into four categories, resulting in a total weighted numerical rating ranging from 0 to 100.

Based upon the numerical scores, OMB assigns a management and performance rating to the programmes. These range from the highest rating of

"effective", to "moderately effective", to "adequate", to a lowest score of "ineffective". In addition, the rating of "results not demonstrated" means that the measures the programme's managers developed were not adequate to determine its effectiveness.

The approximately 1 000 programmes are being assessed and reassessed on a five-year schedule. Managers of programmes that have been assessed who are dissatisfied with their score or have instituted improvements can request that they be reassessed sooner.

Box 1. **The PART scoring mechanism**

OMB devised 25 to 30 questions grouped into four categories to assess the performance of agency programmes. Each of the categories contains a series of questions, the answers to which are given a weighted score for relative significance:

1. Programme Purpose and Design (weight = 20%): to assess whether the programme design and purpose are clear and defensible.

Sample questions: Does the programme address a specific and existing problem, interest, or need? Is the programme designed so that it is not redundant or duplicative of any other federal, state, local, or private effort?

2. Strategic Planning (weight = 10%): to assess whether the agency sets valid annual and long-term goals for the programme.

Sample questions: Does the programme have a limited number of specific long-term performance measures that focus on outcomes and meaningfully reflect the purpose of the programme? Does the programme have ambitious targets and timeframes for its long-term measures?

3. Programme Management (weight = 20%): to rate agency management of the programme, including financial oversight and programme improvement efforts.

Sample questions: Does the programme use strong financial management practices? Does the programme collaborate and co-ordinate effectively with related programmes?

4. Programme Results (weight = 50%): to rate programme performance on goals reviewed in the strategic planning section and through other evaluations.

Sample questions: Has the programme demonstrated adequate progress in achieving its long-term performance goals? Does the programme demonstrate improved efficiencies or cost effectiveness in achieving programme goals each year?

Source: Office of Management and Budget.

1.1. *Purpose of this report*

This report examines OMB's PART initiative from a practical standpoint: How have federal agencies dealt with the requirements of PART? What strategies have they employed to be successful? What challenges do they face?

This report highlights four challenges that confront both agencies and OMB as they work to complete assessments of all 1 000 programmes and describes approaches agencies are taking to meet those challenges.

The first challenge is for departments and agencies to organise an appropriate means of managing the PART assessment process. There are great differences among departments and agencies in the scores given by OMB to their programmes, and it is almost certain that these differences are due in some measure to the nature of the departmental and agency responses.

The second is the challenge of using the PART questionnaire as a means of communicating the accomplishments and shortcomings of a programme to OMB and to other interested stakeholders. Without careful, hard work at mastering the PART instrument, even well-run programmes with good results are not guaranteed a good rating.

The third is the challenge of developing suitable measures. This is a challenge for many programme managers, since they are under pressure from OMB to develop measures of outcomes, and a challenge for OMB as well in that the success of PART as an assessment tool depends crucially on the development of appropriate measures.

The fourth challenge is that of interpreting programme performance measures and their associated results in order to understand the extent to which programme managers can be held accountable for their programme's performance.

Much has been written about the use of performance measures and performance budgeting, but so far little has been written about PART. For example, Harry Hatry (1999, 2001) has written about the different kinds of measures that can be used. Melkers and Willoughby have explored the adoption of performance budgeting requirements at the state level (1998) and how measures are used at the state level (2001, 2004). Philip Joyce (2003) has written about linking performance and budgeting. GAO (2004) has examined the extent to which PART assessments have influenced allocations in the president's budget, as have Gilmour and Lewis (2006). The focus of this report is different, looking instead at how programmes have responded to PART, and the experience of programme-level and bureau-level staff in dealing with the demands of PART.

Box 2. **Methodology**

This report is based on a series of interviews with programme staff and OMB officials. The author interviewed staff associated with 25 programmes in four departments (State, Energy, Labor, and Interior) and the Environmental Protection Agency. The author focused on programmes that increased their ratings, on the premise that this approach would be useful to identify successful strategies for managing the PART process.

1.2. PART *and performance budgeting*

PART is seen as a key element in President Bush's broader push to expand the use of performance budgeting. Progress toward performance budgeting is not just a federal government challenge. Performance budgeting is not easy to do at any level of government, but it holds promise for solving the fundamental challenge of budgeting – knowing where to direct scarce resources in order to gain the maximum public benefit.

A deep frustration in legislatures, governors' offices, and budget offices at all levels of government is that it is difficult – in some cases impossible – to know which programmes are doing good work and which are wasting money. Consequently, ineffective programmes can continue to receive funding year after year, when that money could generate greater public benefit if directed to programmes that produce results. The lack of reliable information about the effectiveness of programmes leads to the adoption of strategies of incrementalism – small increases or reductions at the margin of programme budgets – as a way of dealing with the uncertainty about where to allocate resources for maximum benefit (Wildavsky, 1984).

The aspiration of performance budgeting is immense – to provide decision makers with the information they need to better allocate scarce resources in a way that will yield the greatest benefit. Even modest success in identifying programmes that are effective and those that are ineffective, and facilitating some movement of money away from the ineffective and toward the effective, will be a valuable accomplishment.

A second and perhaps equally important aspiration of performance budgeting is to induce organisational change – to encourage agencies to find better ways of achieving their goals and to improve their results. Allen Schick (2001) points out that behind all performance measurement is "the notion that an organisation can be transformed by measuring its performance." He is pessimistic about this logic, but there are a great many optimists, and the jury is still out on the question of whether and to what extent measurement can induce change. Optimists contend that if agencies cannot document that they are producing results, they will be compelled to change.

Some observers believe that PART will help to induce change by introducing a new level of transparency in government. For example, Clay Johnson, OMB's deputy director for management, stated recently that "transparency leads to accountability, which leads to results. Without transparency, you don't have accountability." In February 2006, OMB unveiled a new website, *www.ExpectMore.gov*, that makes available the assessments of about 800 programmes that have been subjected to PART. With this easily navigated website, the federal government has taken a giant and unprecedented step to make available to its citizens assessments of individual government activities. ExpectMore.gov divides programmes into two groups, those that are "performing" and those that are "not performing". By exposing programmes that are not performing, OMB is surely hoping to compel them to improve, and to give their constituents and stakeholders arguments to demand improvements. These efforts have been recognised by the broader government improvement community. In 2005, PART was awarded a Ford Foundation Innovations in American Government award.

This recognition is remarkable, especially given that the states, not the federal government, have led the way in adopting performance budgeting in the United States. Performance budgeting has been widely adopted abroad (Schick, 2001), and as of a 1998 report, 47 out of 50 states had adopted some form of performance budgeting (Melkers and Willoughby, 1998). Because performance budgeting is not defined uniformly, it is hard to know what that means. States may have aspired to link programme outcomes and budgetary decision making, and adopted systems of performance measurement, but few have in place a system that has any kind of explicit link between performance and budgets. Evidence suggests that the impact of state performance budgeting systems on legislative priority setting in the budget process is modest (Melkers and Willoughby, 1998).

Although the federal government has been slow in adopting performance budgeting, its current approach is particularly comprehensive and impressive. For example, the care taken in devising PART to be objective, and the background analysis and documents that support it, reflect careful thinking about the challenges of assessing performance across a wide spectrum of government programmes.

1.3. *The PART process*

The PART process begins with the release each year of the list of programmes to be assessed that year. The definition of what constitutes a "programme" is developed jointly between an agency and OMB (see Box 3). The programme officials then begin their task of formulating suggested answers to the questions, along with explanations and evidence. The PART document is now completed on line. The budget examiner for the programme

Box 3. **What constitutes a "programme"?**

OMB is using the PART to assess programmes across the entire federal government over a period of five years, conducting about 200 programme assessments each year. That means the federal government will be sliced into approximately 1 000 programmes. The problem is that programmes are not always readily demarcated from the rest of an agency or bureau. The OMB guidance recommends that agencies begin with the budget when they try to identify programmes for PART, and try as much as possible to equate programmes with budget accounts. This has the advantage of making it easier to translate PART findings back into the budget document. But as OMB acknowledges, "programme activities in the budget are not always the activities that are managed as a programme in practice."

An example of a budget category that was inappropriately identified as a programme was the Department of Interior's (DOI) "Land and Water Conservation Fund – Land Acquisition". It straddles three separate bureaus in Interior. In its PART assessment, it received answers of "no" to question after question because it simply was not a programme. As the examiner reported, "During the PART process, it became apparent that land acquisition is not a true program but rather an activity or tool that serves a variety of disparate programs across multiple DOI bureaus. In the future, land acquisition would be better evaluated in relation to each of these programs."

Defining or identifying programmes is of crucial importance. Agencies and OMB need to be careful not to define programmes in terms of budget categories when that would yield a set of programmes that is not managed as a coherent, free-standing entity with a coherent set of goals. It is probably easier at first to go along with defining programmes in terms of the budget, but that may lead to weak assessments.

Programmes as identified under PART vary in size. The Bureau of Labor Statistics (BLS) is treated as a single programme with a budget of over USD 500 million. BLS conducts numerous surveys annually, and each of these could be considered a separate programme. That would have made too many small programmes, so it was decided instead to evaluate the entire organisation as a whole. On the other hand, the Department of Education has taken the approach of slicing programmes very small. An example is the B.J. Stupak Olympic Scholarship Program, a very small college scholarship programme with a yearly budget of about USD 1 million. The Department of Education has many other small scholarship programmes, each assessed separately. The Individuals with Disabilities Education Act (IDEA) has been broken up into at least seven separate programmes that require a PART evaluation. Such PART assessments are time-consuming, and dividing departments very finely means spending a lot of resources assessing programmes that are limited in scope. BLS is probably assessed at too high an aggregation, and the Department of Education programmes are too disaggregated. But no common approach across the federal government has been adopted.

reviews materials submitted by the programme and decides which answers to give for each of the questions. Needless to say, programme officials give themselves more yeses than the examiners do. Programme officials who do not agree with the assessment can appeal up the chain of command in OMB. There are appeals each year, and a few are successful.

Programmes are given scores based on the proportion of "yes" answers a programme is awarded in each of the four sections. Although OMB does not report an overall score for programmes, one can easily calculate summary scores using the official weights for each section and a spreadsheet programme. Based on the overall scores, OMB assigns grades to programmes: ineffective, adequate, moderately effective, or effective. Table 1 reports the range of overall scores corresponding to each grade.

Table 1. **Converting scores to grades**

Numerical score	Grade
85-100	Effective
70-84	Moderately effective
50-69	Adequate
0-49	Ineffective

If a programme lacks measures deemed adequate by OMB, it is rated "results not demonstrated" (RND). Programmes rated in 2005 that were given grades of RND had overall scores ranging from a low of 11 to a high of 83. The lowest scoring programme, with an 11, was the Tribal Courts programme in the Interior Department. The highest score, a 97, was earned by the Inspector General Oversight of Federal Employees Health Benefits Program (FEHBP) in the Office of Personnel Management, a small programme with an USD 11 million budget.

The scores and grades are not just for show: an important goal of PART is to link budget decisions with assessments of outcomes and overall programme quality, although OMB is also clear that these assessments are not the only factor in budget decisions. A high rating will not necessarily be rewarded with a budget increase, and low-rated programmes may receive increases because they may have been too under-funded to be effective.

An important feature of PART is its emphasis on measuring outcomes rather than outputs. GPRA also requires outcome measures, but PART takes this to a new level. Scholarship on performance budgeting has discussed the different kinds of measures at length, distinguishing between outcomes and outputs (Hatry, 1999, 2001). The PART guidance[2] is clear: "Measures should reflect desired outcomes… Outcome measures are most informative, because these are the ultimate results of a program that benefit the public. Programs must try to translate existing measures that focus on outputs into outcome measures by focusing on the

ultimate goal of the program…" The OMB examiners who do the PART evaluations insist that programmes find true outcome measures whenever possible, pushing sometimes reluctant programme managers to look for creative ways of assessing their results. An important exception to the insistence on outcome measures is research and development programmes, for which the OMB guidance acknowledges outcome measures may be inappropriate because results cannot be predicted in advance of the research.

The OMB examiners who do the PART assessments have been successful in prodding programme managers to adopt better measures. Adopting measures is not new, since they have been required since 1993 by the Government Performance and Results Act, but under GPRA less was at stake. With PART, there is far more emphasis on adopting end outcome measures, with a link between assessments and budget decisions. Further, the threat that a programme will be labeled "results not demonstrated" is an important incentive to have programmes adopt measures acceptable to OMB examiners. Departments with too high a proportion of programmes rated RND cannot get a green on the performance section of the President's Management Agenda (PMA) scorecard.[3] In one department, OMB examiners informally told officials that if they did not reduce the number of programmes rated RND, OMB would consider reducing the department's administrative budget. GAO officials who have examined PART contend that in some departments a rating of "ineffective" is preferred to RND, because of the impact on the PMA scorecard.

Individual programmes have made significant progress. Initially a very large proportion of programmes were graded "results not demonstrated", meaning that OMB examiners did not approve the programmes' measures. In subsequent years, many programmes have been able to replace their RND grade with a "real" grade by adopting suitable measures. Figure 1 compares grades assigned to the initial cohort of programmes assessed in the FY 2004 budget with their grades as reported in the FY 2007 budget. These programmes have had several years to respond to critiques and suggestions from OMB. As Figure 1 makes clear, most of the programmes initially graded RND have been able to get a real grade. Programmes that were initially assigned a real grade of either ineffective, adequate, moderately effective, or effective have also had some success in getting higher grades. Of 103 programmes with real grades, 15 were able to have their grade raised; five had their grade lowered.

In addition to assessing outcomes, the greatest strength of PART is the inherent reasonableness of the questions that form the basis of the process. It is hard to argue with an assessment that asks programmes to have a clear statement of programme purpose, good planning, strong financial management practices, and so on. These are questions, participants acknowledge, that are good for a programme to consider and answer.

Figure 1. **Comparison of FY 2004 and FY 2007 programme grades**
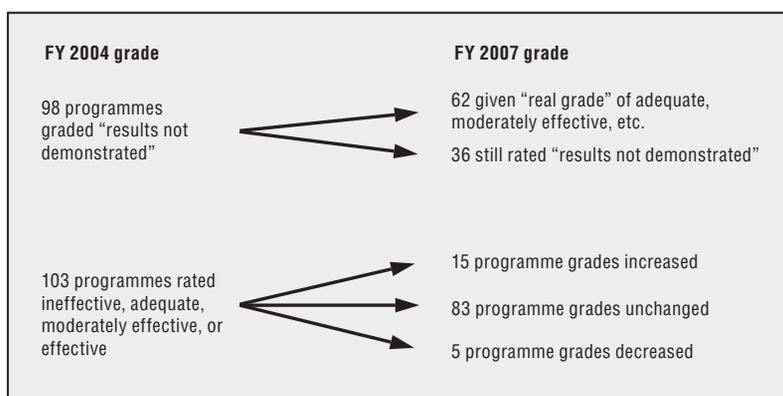
Note: Thirty-four of the FY 2004 cohort of programmes could not be compared with FY 2007 grades because they were either merged into other programmes or otherwise changed such that they could not be paired with a programme in the FY 2007 budget. These have not been included in this figure.
Source: OMB documents.

Based on this author's research and interviews, PART is taken very seriously at the programme and bureau level. Management systems imposed from above always meet a certain amount of skepticism and resistance, and that is true with PART. But attitudes have changed as programme managers have seen the determination and persistence of OMB in implementing PART. Agency officials who might have thought PART was a passing fancy that could be waited out realise now that it is here to stay for at least the duration of the Bush administration, and that the quest to better link performance to budget will continue. PART will be even less likely to go away if a bill sponsored by Representative Todd Platts (R-PA) becomes law. His bill, entitled the Program Assessment and Results Act, would amend GPRA by adding a provision requiring OMB to conduct programme assessments every five years. It has been reported out of committee but has not yet seen floor action.

2. Challenge 1: Organising for success

There are large, important differences between departments in their programme assessment ratings under the PART process. Is this due to the inherent nature of the programmes these departments operate, or is it related to something else?

The Departments of State, Treasury, and Defense have been unusually successful in managing the PART process, receiving a large proportion of high ratings while avoiding low and RND ratings. There is every reason to believe that the seriousness with which a department takes PART has an effect on its scores and ratings.

The State Department has been particularly successful in managing its PART initiative. It has a high proportion of effective and moderately effective ratings, no ineffective ratings, and few RND. Only 16% of all programmes assessed across the government had earned a rating of effective by the release of the FY 2007 budget, but 34% of programmes at State that had been assessed had effective ratings. Treasury and Defense also had large proportions of programmes earning effective ratings. The Department of Labor had only one programme with an effective rating, while Veterans Affairs and the Environmental Protection Agency had none. Fifty-five per cent of programmes assessed in the Department of Education were rated RND.

Table 2 ranks departments by their relative success with PART; it shows the percentages of programmes that have been rated effective and those rated RND. A useful summary of a department's success is to subtract the percentage of programmes with a rating of RND from the percentage of those rated effective. Departments with a positive score are doing reasonably well. Those in negative territory are having problems. According to this calculation, State comes out on top because it has many programmes in the effective category and few in RND. Education comes in last because it has few rated effective and most rated RND. To some extent, differences between departments can result from the greater difficulty some departments have in assessing outcomes, and the Department of Education probably has a harder time than most. But the international programmes administered by the State Department have outcomes that are no less difficult to assess than education programmes. The stark differences observed in Table 2 must be due to more than the inherent differences in departmental missions.

The State Department also stands out in the extent to which it has succeeded in getting initial ratings raised. Programme officials dissatisfied with the rating first assigned can (with the consent of department-level officials) ask to be reassessed before the normal five-year period ends. Of the 234 programmes first assessed in the FY 2004 budget, 62 had raised their score by the release of the FY 2006 budget. Of those, all but 11 replaced a rating of RND with a real rating, meaning that they had gotten approval for their measures of results. Of the 11 instances in which a programme was initially assigned a real rating and subsequently received a higher rating, five were in the State Department. Outside the State Department, improving a rating is unusual. It makes a lot of sense, then, to focus on what the State Department has done to manage PART so effectively.

Two important characteristics mark the State Department's approach to PART. First, top leaders in the department made it absolutely clear that they took PART seriously. Second, the State Department's Bureau of Resource Management plays a central role in organising the bureau-level responses to PART. Success at the State Department is a product of two non-exclusive factors. Part of the

Table 2. **Percentage of programmes rated effective and results not demonstrated, by department, FY 2007 budget**

	Percentage rated effective	Percentage rated "results not demonstrated" (RND)	Effective minus RND
State	38	7	31
Treasury	44	16	28
Defense	34	13	21
Transportation	20	0	20
Energy	22	8	14
Commerce	18	18	0
Labor	4	11	−7
Justice	11	19	−8
Environmental Protection Agency	0	13	−13
Health and Human Services	11	27	−16
Agriculture	6	27	−21
Homeland Security	16	38	−22
Housing and Urban Development	4	32	−28
Interior	8	37	−29
Veterans Affairs	0	33	−33
Education	3	55	−52

Source: Author's calculations from OMB-provided data.

success stems from the programmes being generally well run. In addition, staff at State have somehow figured out how to work the system with maximum effectiveness. They work hard at complying with the PART process, filling out the questionnaires carefully, and developing measures acceptable to OMB.

While a range of officials at the State Department give a number of explanations as to why they have done so well, they uniformly stress one: that former Secretary of State Colin Powell took the PART process seriously from the start. Powell and former Deputy Secretary Richard Armitage let it be known through memos and other communications to the bureaus that they considered PART an important priority and wanted State to do well. Interviews with officials at the State Department made it clear that they understood PART was considered important at the very highest levels and, consequently, at lower levels as well.

Powell stands out among recent Secretaries of State in his commitment to department management. Some of his recent predecessors had a reputation for not investing much time or energy in matters of internal management, but Powell was deeply concerned with management of the department. He was reportedly belittled in some parts of the Bush administration for being a mere manager, but the opinion at Foggy Bottom is that Powell's emphasis on management issues has had a real impact on the capacity of State to effectively address its mission. In addition to emphasising PART, Powell also made a practice of appointing individuals who were good managers as his assistant secretaries.

The second factor in State's success is that personnel at the bureau level are not left to their own devices to figure out PART. Instead, the Bureau of Resource Management (RM) takes the lead in organising bureau-level responses. Work begins in January when bureau staff have a list of the programmes slated for PART evaluation that year. By February, working groups for each of the programmes are up and running. The experience of personnel at RM is important because they have worked with numerous programmes and have a sense of the kinds of measures that OMB likes and wants, and they can help devise acceptable measures for a programme. Finally, when the PART evaluations come out, RM convenes "after action" reviews to see what went right and wrong.

No doubt other departments have adopted similar approaches. In one bureau at the Department of Energy that has done very well on PART, the bureau director has each of the programmes self-assess with the PART instrument each year. He holds them accountable for the extent to which their self-assigned answers match those handed down by OMB when the programme is formally PARTed.

The State Department is also notable for the extent to which it seeks to have low-rated programmes reassessed. If a programme gets a rating below effective, typically the department seeks to have the programme reassessed the following year. State appears to be relentless in seeking new assessments; some programmes have been assessed three times. At State, it seems that anything less than an effective rating is unacceptable, so they keep trying until they get an effective rating. This stands in marked contrast to other departments which, by and large, have not sought to change ratings once assigned, except to get rid of the dreaded RND. Officials in other departments were surprised to learn that so many State Department programmes had been reassessed, because they were under the impression that they could go through the process again only after making dramatic changes or under exceptional circumstances.

What the State Department has done probably does not seem unusual or extraordinary, but some other departments appear to follow a far more relaxed procedure. Bureau-level staff in other federal departments said that their department served mostly as a conduit, passing along documents from OMB. The bureaus were largely on their own in responding to PART. Even in departments with a reputation for taking the PMA seriously, the bureaus work largely on their own.

The Department of Education also stands out, but for having very low assessments. More than half of its programmes are rated RND, and only two out of 74 programmes are rated effective. The State Department has sought to get the highest possible ratings for its programmes, but the Education Department has taken a different approach, not using PART to validate its successes, but to create motivation for transformation. According to Robert

Shea, the OMB manager of the PART initiative, Education's low ratings do not reflect a lack of interest in performance management. Rather, the leaders at the Education Department believe the department is burdened with many ill-conceived, poorly designed programmes, and see the PART process as a means of shining a light on those deficiencies. They accept a low baseline if that is what it takes to get programmes redesigned.

3. Challenge 2: Communicating accomplishments

The biggest question on the minds of agency staff charged with working on a PART evaluation is always, "What can I do to get a better rating?" The most important prerequisite to a good score is having a strong, well-managed programme. PART is not a perfect yardstick for assessing programmes, but it is implausible that a weak programme will be able to get one of the higher ratings. Good programmes have, however, gotten bad ratings. It takes hard, careful work for a programme to get a score that reflects its true merit. As an Energy Department budget official explained, success on PART is a matter of both "achievement and articulation". For programme managers, a key challenge imposed by PART is to learn how to use the PART questionnaire as a tool to communicate what they do, why it is important, and what they have accomplished.

A useful strategy for learning about agencies' success with PART and the impact of PART is to look at programmes that have managed to raise their ratings. Of the 234 programmes assessed in the initial PART ratings released in 2003, 62 were able to convert a "results not demonstrated" score to a real score by 2006. Of the programmes that got a real grade, 15 were able to raise their grade by 2006. Were the raised grades evidence of improved results? Better programme management? That only 15 programmes got higher grades suggests that PART is not yet causing large-scale management innovation or change. The key factor mentioned by officials of programmes that increased their grades was that they learned how to use the PART instrument better.

Officials from several of the programmes that raised their ratings explained their success with almost identical language. "We learned how to take the test," they said. By this they meant that they learned better how to use the PART instrument as a device to communicate their accomplishments to their examiner. Questions are generally clear, but can be interpreted in different ways. There can be disagreements about the appropriate evidence to document claims. Over time, programme staff learned better how OMB is viewing the questions, and learned to write better – or more acceptable – answers. Equally important, they learned how to devise outcome measures that passed muster with OMB.

Officials did not attribute their success to changing the programme. None of the officials interviewed for this project claimed that they had introduced

significant management improvements or changed programme design in order to raise their rating. In one sense, this is disappointing, as an important goal of performance measurement is to induce programme improvement. Yet it should not be particularly surprising that making programmes better was not an element in raising scores. PART has been in place for only a few years, not enough time to turn around a weak programme. In the short run at least, the best strategy for getting a better score is to become better at "taking the test" – in other words, to better communicate.

Over time, however, if programme managers find that efforts to communicate more effectively do not result in higher ratings, they may feel compelled to make changes in programme management. In one case where the PART rating for a Labor Department programme was increased because major changes had been made in its management, the change was not in response to PART but in response to GPRA (see Box 4).

This study examined only a fraction of all programmes that have been assessed with PART, and it is possible that other programmes demonstrate more evidence of management change in response to PART. Improving management of federal programmes is an extremely important goal, one that will be achieved only as the result of perseverance over a longer period of time than has yet elapsed under PART. An overall assessment of PART and its contribution to the quality of federal programme management will have to wait.

4. Challenge 3: Overcoming measurement challenges

For both OMB and the programmes being assessed by PART, an important and enduring challenge is developing good measures of performance and outcomes. OMB has appropriately taken a strong stance in favor of end outcome measures. But because OMB has maintained a high standard for approving measures, a large number of programmes have been lumped in the "results not demonstrated" category. The problem for OMB is that, without measures of outputs, it is difficult to engage in performance budgeting.

Given the tremendous diversity of federal programmes, simple prescriptions for measuring effectiveness are likely to fail to assess the real merits of some. The PART guidance recognises important differences among federal programmes by dividing them into multiple types – direct federal, research and development (R&D), block grant, and so on – and providing questions specifically designed to assess each type. However, in one important way, PART seeks to impose a near uniform standard on all programmes: an insistence on finding end outcome measures. The guidance makes one special exception: it does not require R&D programmes to devise outcome measures, acknowledging that this would be infeasible. Because federal programmes are so varied in design and aspiration, devising good outcome measures requires

---

Box 4. **How measuring performance led to organisational change in a Department of Labor programme**

The process of devising performance measures requires programme managers to engage in a certain amount of introspection, since they must think carefully about what it is their programme does and why. On occasion, careful thinking can lead to a moment of clarity and a fundamental rethinking of how a programme does what it does. Such was the case for a little-known programme in the Department of Labor, the Federal Employees Compensation Act (FECA). It provides compensation for non-military federal employees who become injured or ill on the job. In short, it is a disability insurance programme for federal workers. Disabled workers receive a benefit equaling 75% of their previous pay, which is paid for by the disabled worker's agency.

For years the programme focused on processing claims and paying out benefits. This became a problem since the beneficiaries tended to stay on the rolls a long time; the total bill for their disability payments was growing and placing an increasing burden on agencies.

A change came when the programme devised measures under the auspices of GPRA. According to Shelby Hallmark, head of the Office of Workers Compensation Programs, which administers FECA, this produced a complete shift in their thinking about the programme and led to the development of a measure of "lost production days" (LPD). The programme thereafter had a new goal: getting injured workers back into employment and minimising LPDs, not just processing claims. Prior to adopting the LPD measure, the programme did not track return to work at all.

The shift in orientation from a process to an outcome orientation has not been without difficulties. Claims examiners were used to doing things the old way and resisted taking on the new tasks associated with tracking LPD. In addition, they have not been able to meet their targets, although the programme is not far off, earning a "moderately effective" rating from OMB.

An important message in the experience of the FECA programme is that organisational change comes slowly, sometimes very slowly. GPRA was passed in 1993, and 13 years later FECA is still working to implement the changes induced by complying with it. PART may also someday have an important impact on programme management, but it's far too early to tell.

---

ingenuity and flexibility from the programme officials as well as some flexibility from OMB. In practice, OMB recognises the difficulties of assessing different kinds of programmes and has been more flexible than the guidance suggests in allowing programmes to adopt measures that are not true outcome measures.

Since PART is intended to assess results, its success depends vitally on the ability of programme managers to identify suitable measures of outcomes or results. This has been a weakness of PART, as well as the broader efforts of promoting performance budgeting, because good outcome measures can be very hard to come by. In the first several years of PART, about half of the programmes assessed received an initial rating of "results not demonstrated" (or RND), meaning that they did not have what OMB regarded as adequate measures. Paradoxically, the large number of programmes labeled RND is evidence that OMB is holding agencies to a high standard and not accepting whatever measures the programmes propose. In subsequent years, some of those first labeled RND have had their measures approved and accordingly received a real rating. By the fourth year of PART assessments, released with the FY 2007 budget, the proportion of RND ratings had dropped to about a quarter.

OMB has come down squarely in favor of "end outcome" measures, rather than output measures, and offers important guidance. Measures should as much as possible be of outcomes rather than outputs. Outputs are generally seen as easier to measure, but outcomes are the preferred type of measure because good outcome measures should encourage the agency to expend its efforts in solving problems and having a real impact on the world. In addition, an efficiency measure is also required.

When programme managers are asked to develop outcome measures, there is predictable pushback or resistance. A typical response is, "What we do can't be measured." There is some truth to this claim, as there really are accomplishments of programmes that defy measurement. But programmes that are accomplishing results must leave some kind of mark on the world, and the challenge for programme administrators is to think creatively about ways of measuring it (see Box 5). After initial resistance and significant prodding from OMB, many programme managers have devised inventive and useful measures of their results.

Reading the PART guidance issued by OMB, one would get the impression that OMB is unyielding in its insistence on end outcome measures. But an examination of programmes that have gotten high ratings indicates that OMB examiners are actually flexible and open to persuasion in the case of particular programme outcome measures that are either impossible or inappropriate. This section surveys the difficulties that some programmes have had in assessing outcomes, the solutions they found, and OMB's response.

4.1. *Measuring results of programmes with "lumpy" outcomes*

Some programme outcomes have an either/or quality that makes it difficult to use outcome measures as a means of tracking progress. The Secure Transportation Asset Program in the Department of Energy has as a goal "to

---

Box 5. **What to measure?**

What is an **outcome** to OMB examiners? The literature on performance budgeting has discussed the concepts of outcomes and outputs at length, and has developed distinctions within each of these categories. Harry Hatry has produced useful discussions about the different conceptions of outcome measures. He distinguishes between "intermediate" and "end" outcomes. An intermediate outcome is something such as fire truck response time, which bears an obvious relationship to the true goal of extinguishing fires. An end outcome is the goal that is sought by the programme, which might be fire damage prevented or dollar value of fire damage. The difference is that the end outcome is something actually enjoyed or experienced by customers or clients of a programme. According to the OMB guidance, "outcomes describe the intended result of carrying out a program or activity. They define an event or condition that is external to the program or activity and that is of direct importance to the intended beneficiaries and/or the public. For a tornado warning system, outcomes could be the number of lives saved and property damage averted." OMB has come down clearly in favour of end outcomes.

The major alternative to outcomes as the object of measurement is **outputs**. According to OMB, "outputs describe the level of activity that will be provided over a period of time, including a description of the characteristics (*e.g.* timeliness) established as standards for the activity. Outputs refer to the internal activities of a program (*i.e.* the products and services delivered)."

OMB prefers outcome measures: "Outcome measures are most informative, because these are the ultimate results of a program that benefit the public. Programs must try to translate existing measures that focus on outputs into outcome measures by focusing on the ultimate goal of the program."

PART requires agency programme managers to adopt long-term measures, annual measures, and efficiency measures.

---

safely and securely transport nuclear weapons, weapon components, and special nuclear materials." Consequently, it might choose for an outcome measure "thefts of nuclear bombs or fuel". Though seemingly a good measure, this would in fact be misleading, since there has never been a theft of nuclear material (in the United States at least). There might be no thefts in a given year even if security is poor. This programme requires a measurement of "security" which is an output, not an outcome. The measures chosen include:

- annual average scheduled overtime hours per federal agent;
- annual percentage of mission hours reduced by shipment optimisation;
- cumulative number of safeguard transporters in operation;
- cumulative number of federal agents at the end of each year.

These are all output measures, but they may be the best way to assess "security". These measures have been approved and the programme is rated "moderately effective".

Another programme in the Department of Energy, the Elimination of Weapons-Grade Plutonium Production Program, has a similar problem. Its mission is building new fossil-fuel power plants in Russia to replace the energy supplied to two cities by plutonium-processing plants. The outcome desired is the removal of these sources of fissionable plutonium. It will take about ten years to complete construction of the new plants, at which point the plutonium-processing plants can be shut down and decommissioned. The desired outcome is closing the plants, but using that outcome as a measure would produce weird and misleading reporting: nine straight years of not achieving the outcome and then, in the tenth year, 100% attainment of the outcome. Instead, the programme measures progress toward the goal of completing the new plants and measures the tons of plutonium produced in the former Soviet Union:

- percentage of construction completed on fossil-fuel plant in Seversk that will facilitate the shutdown of two weapons-grade plutonium-producing reactors;
- percentage of construction completed on fossil-fuel plant in Zheleznogorsk that will facilitate the shutdown of one weapons-grade plutonium-producing reactor;
- metric tons of weapons-grade plutonium produced per year in the Russian Federation.

The first two are process or output measures, but the third would seem to be an outcome measure. OMB has not approved the measures.

4.2. *Measuring results of enforcement programmes*

Enforcement programmes have a problem with outcome measures because there is often a dynamic relationship between effectiveness of enforcement and incidents that the enforcers are trying to stop. The State Department's Border Security Program or BSP (also known as the Visa and Consular Services Program) has this kind of problem. Its goal is to maintain secure borders and keep potential terrorists from entering the United States. At one point, programme managers had considered using as its outcome measure the number of prohibited persons stopped from entering the country. The problem with this measure, they realised, was that if they did a great job of increasing security at the nation's borders, prohibited persons would themselves choose not to seek entry, which would mean that there would be few apprehensions at the border. To score well on this measure, the bureau might be forced to encourage known terrorists to enter the country, just so

they could be stopped. Of course, the BSP would never do such a thing, but a measure that suggests an agency is not doing its job when it is actually doing a great job is seriously flawed. BSP solved this problem by carefully crafting its statement of programme purpose. As stated in BSP's PART response, "the purpose of the Border Security Program is to protect American citizens both here and abroad and safeguard US borders **through improvements in consular programs, processes, and systems**." By saying that the purpose is to introduce improvements, the statement of purpose invites assessment of outputs – the improvements introduced. Their measures are clearly process and output oriented:

- development of a biometric visa programme for the United States;
- number of Consular Management Assessment Team (CMAT) assessments;
- number of days between receipt of routine passport application by Passport Services and issuance of a passport;
- percentage of passport applications processed to issuance within a certain number of days' receipt.

OMB approved the measures and gave the programme an "effective" rating.

The Federal Air Marshal Service in the Department of Homeland Security has a mission similar to BSP, providing security on commercial aircraft. In this case, staff selected a combination of outcome and output measures. Counting terrorist attacks on airplanes is a true outcome measure. The other two listed are output measures:

- number of successful terrorist and other criminal attacks initiated from commercial passenger aircraft cabins with FAM (Federal Air Marshal) coverage;
- level of operational FAMs verified as meeting recurrent training requirements;
- level of FAM coverage on flights with identified threats. (Targets and actual data are classified for security reasons.)

Despite the inclusion of an end outcome measure, OMB was unimpressed and gave the programme an RND rating.

4.3. *Measuring results of data collection organisations*

Programmes that have a mission of generating data have trouble measuring their outcomes. The Bureau of Labor Statistics (BLS), like other agencies whose mission it is to produce statistics, has a hard time with outcome measures. As BLS staff began the PART process, they asked themselves, "What is the intended goal of our statistics?" The answer they came to was "better informed business decisions". That is a difficult concept to measure. So, instead of trying to measure the quality of business decisions, they decided to measure their outputs, in particular the timeliness with which they release the

24 separate studies they conduct each year. They measure the extent to which they get their studies out on schedule, which is a nontrivial accomplishment but far from a true outcome measure. As a surrogate for an outcome measure, they assess customer satisfaction with their products:

● percentage of scheduled releases issued on time;

● customer satisfaction with BLS data and assistance;

● number of months elapsing between collection and publication of detailed employee-benefits data, with no increase in production spending.

OMB has accepted these measures, and the bureau has earned an "effective" rating.

The US Geological Survey (USGS) is a scientific agency that produces a great deal of data. The Geologic Hazards Assessments Program provides earth science data to reduce loss of life and property from volcanoes, landslides, and other geological hazards. If the Geologic Hazards Program is doing its job well, the end outcome should be a safer country, since people will better understand the risks of earthquakes and other dangers and be able to stay out of their way or prepare for surviving them. Measuring the actual impact of this data will be nearly impossible, however, putting the Geologic Hazards Program in much the same position as BLS. But because the programme is classified as R&D, it need not employ outcome measures (BLS is classified as direct federal). Rather than attempt to measure actual outcomes, the USGS measures progress toward data collection goals, the usability of its data, and customer satisfaction. These are measurable steps that are likely to lead to the desired but unmeasurable outcome:

● percentage of potentially active volcanoes monitored;

● number of urban areas for which detailed seismic hazard maps are completed;

● number of counties, or comparable jurisdictions, that have adopted improved building codes, land-use plans, emergency response plans, or other hazard mitigation measures based on USGS geologic hazard information.

The third of these is the most interesting: the programme attempts to assess the actual use that customers have made of its data. OMB has given the programme a "moderately effective" rating.

Producers of statistics have a difficult time documenting their actual results, and so do agencies that make grants to support scientific research. The Energy Department's Office of Science administers a programme called Basic Energy Sciences. It provides grants to support energy research, the goal of which is to "expand the scientific foundations for new and improved energy technologies". Documenting that the research they fund is actually accomplishing its goal is difficult, since the impacts of research projects undertaken today may not be seen for years, and may not be predicted in

advance. OMB's "Research and Development Investment Criteria" explicitly acknowledge the difficulties. The criteria require programs to demonstrate relevance, quality, and performance, but not outcomes.

The Basic Energy Sciences examiner for OMB urged the programme to adopt "Committees of Visitors" (COVs), groups of distinguished scientists knowledgeable in the field who come to the agency and review the funding decisions made by the Office of Science. This idea was borrowed from the National Science Foundation. The COVs review the proposals and assess whether proper procedures have been followed in awarding grants; they ensure that the proposals being funded meet recognised standards of good research. While they cannot assess the future outcome of the research that is funded, they can ensure that the research meets high standards. The presumption is that if the correct procedure is followed in distributing grants, good science will result and the aims of the programme will be advanced. The COVs are universally considered to be a good tool at the Office of Science. Associate directors have found COV feedback very helpful. COVs have made significant criticisms that have been taken seriously and recommended changes that have been adopted.

4.4. Measures that are outputs for one programme and outcomes for another

With some programmes, there is no outcome that can be distinguished from an output. In the Overseas Buildings Operations Bureau (OBO) of the State Department, the Capital Security Program is systematically rebuilding more than 100 American embassies. Many of the older US embassies are located in urban areas where there is insufficient space around the embassy to provide for adequate security. Thus the embassies are being relocated and rebuilt outside of city centres. In a sense, the desired outcome is enhanced security, which might be measured in terms of attacks against embassies, injuries to personnel, or damage to facilities. Instead, the programme has chosen to measure its success in the timeliness and cost-effectiveness of building new embassies. These are output measures, but important ones that are closely associated with the achievement of the desired outcome of enhanced security:

- percentage of capital security construction projects completed within the approved construction budget;
- percentage of capital security construction projects completed within the schedule authorised in the construction contracts;
- number of new capital security construction projects awarded.

One can even argue that, in a case like this, simply building new embassies that satisfy security requirements is the outcome desired. OMB has approved these measures and the programme is rated "effective".

4.5. *Programmes with statutory limitations on measures*

In nearly everything they do, public programmes are limited by their authorising statutes. Some programmes are set up in a way that deliberately prevents them from focusing on end outcomes, requiring them instead to emphasise compliance with a set of procedures. In a sense, such limitations can constitute serious design flaws, because they prevent a programme from accomplishing as much as it might or accomplishing goals in efficient ways. But Congress often has reasons for designing programmes as it does, and programmes may make good political sense, even if from a rational policy standpoint they are less than optimal. Members of Congress can be aware of these weaknesses but be uninterested in revising them, because the perceived flawed programme design may address a particular political need. Should programmes be held accountable to an end outcome measure if Congress has withheld the powers needed to accomplish the outcome? This is a difficult question. The position OMB has taken is that all programmes must be able to produce results. Design flaws are no excuse. This is a good position to take, since low ratings for programmes can highlight design flaws and stimulate Congress to take corrective action. But it can be discouraging when programme managers are held accountable for performance failures beyond their control.

The federal Perkins Loan Program is an example of a programme that received a weak PART assessment (ineffective) because of design flaws. Administered through the Department of Education, the Perkins Loan Program provides loans to college students based on need. The flaw in the programme is that the money is distributed through the intermediary of the school, and the aid formula gives more money to schools that have been in the programme longer. Those colleges and universities obviously like this arrangement, but it does not allocate money to the students who need it most. The critique in the PART report is serious:

> The program's institutional allocation formula (*i.e.* how much program funding is given to each school to offer Perkins aid) is designed to heavily benefit postsecondary institutions that have participated in Campus-Based programs for a long time, at the expense of more recent entrants or new applicants. Since these long-standing institutions do not have a higher proportion of needy students, this allocation formula tends to limit the program's ability to target resources to the neediest beneficiaries.

The obvious preference of OMB and the Department of Education is to eliminate the programme altogether and shift its resources to other, better-targeted student loan programmes. From the standpoint of efficient allocation of resources, it would make sense to dissolve this programme and fold it into

other loan programmes. Congress, however, still supports the programme and continues to fund it. The Bush administration has asked for legislative changes to the programme, but Congress has so far declined.

OMB has assigned the Perkins Loan Program a grade of "ineffective". And while this failing grade is based on problems apart from design issues, it is hard to see how the programme will ever get a good grade until the design issues are resolved.

Medicare is another instance of a programme that was graded down for having "design flaws". Medicare may have flaws, depending on how one sees its mission. Is its goal to provide quality healthcare to beneficiaries, or to efficiently and fairly administer a law passed by Congress? If the former, Medicare has significant flaws; if the latter, it does not. The PART evaluation for Medicare gives it a "no" on question 1.4, which asks if a programme is free of major design flaws, and describes the flaw that OMB sees in the programme:

> Several features of the Medicare program reflect its outdated statutory design. For example, unlike most private health insurance, Medicare does not protect beneficiaries against high out-of-pocket costs, *i.e.* it does not provide catastrophic protection... Updating the statutory design will allow Medicare to better serve beneficiaries.

The design flaw, if one wishes to call it that, stems from the original creation of Medicare as an entitlement programme providing "social insurance". In the social insurance concept, beneficiaries "earn" their benefits by paying for them with "contributions"; consequently, there is no means test. Because of this design feature, Medicare provides health insurance for many individuals who need no help buying insurance. And because all beneficiaries receive an equal benefit without any means testing, Medicare provides a fairly limited benefit, conspicuously omitting coverage for catastrophic illnesses, which in turn limits its ability to affect outcomes.

These are legitimate criticisms of Medicare, but the lack of catastrophic coverage is not an oversight. In the 1980s, Congress created a programme to pay the cost of catastrophic illnesses and then quickly killed it in the face of determined opposition by wealthier retirees who paid for the programme. Good arguments can be made for changing the structure of Medicare. For example, an alternative programme design that allocates more resources to the neediest beneficiaries could do more to improve health outcomes within existing budgetary constraints. But at present there is little interest in either Congress or the administration to initiate a major redesign of Medicare. Given that Medicare is stuck with its "flawed" design for the foreseeable future, it probably makes more sense to assess how well the centres for Medicare and Medicaid services implement the programme as devised by Congress.

In fact, the programme performance measures for Medicare reflect the statutory limitations and consequently assess intermediate rather than end outcomes:

● percentage of Medicare beneficiaries who receive influenza vaccination and pneumococcal vaccination;

● percentage of women who receive a biennial mammogram;

● percentage of diabetic beneficiaries who receive diabetic eye exams.

In addition, the programme has measures to assess efficiency and some outputs, but none to assess end outcomes such as access to health insurance or overall health of beneficiaries. OMB has approved the measures and given the Medicare programme a rating of "moderately effective". If not for two no's on questions that were attributed to statutory problems, Medicare would have earned a grade of "effective".

Some programmes at the Environmental Protection Agency (EPA) have similar problems with design limitations. Ideally, environmental programmes should have goals that can be expressed as end outcomes, such as cleaner air or water. One can easily see the advantage of holding EPA accountable for end outcomes; for a programme that has the goal of cleaning the nation's waters, it makes sense to assess how well they are doing by measuring water quality. Yet officials at EPA point out the problems with holding them accountable for end outcomes. Many EPA programmes have statutory designs that limit their authority to achieve their mission. Consequently, there are important sources of pollution they cannot regulate. For example, EPA has no authority over local zoning and land-use decisions, even though local decisions have an important impact on water pollution. These are limitations created deliberately by Congress, which Congress is unlikely to change.

Nonetheless, EPA programmes "were held across the board to end outcome measures", according to an EPA official involved with the PART process. Their programmes did not do well in the assessments, at least not at first. In the first year of PART assessments (the FY 2004 budget), ten EPA programmes were assessed and all but one were rated RND. OMB did not like the measures employed by EPA, which were not adequately outcome oriented. Subsequently, EPA has worked hard at devising outcome measures and has succeeded in getting many of the programmes out of RND status. According to the most recent assessments, only six of the 46 EPA programmes that have been assessed have ratings of RND, indicating that EPA has done a remarkable job of identifying outcome measures. Still, their measures overall are still low, with only 17% of their programmes earning a rating of either "effective" or "moderately effective".

End outcome measures are very useful, but they are not possible to devise or appropriate to use in all circumstances. In many cases, OMB has exhibited more flexibility in approving measures than the PART guidance would suggest. While

there can be certain advantages in having flexibility in adopting measures, this comes at the cost of sacrificing the considerable advantages of outcome measures. OMB's Shea contends that this flexibility is actually a weakness in the implementation of PART. In the future he would like to see those programmes that still have output measures push harder to find suitable outcome measures. This tension remains a considerable challenge for OMB and PART.

5. Challenge 4: Linking performance to outcomes

Once suitable outcome measures have been adopted, a challenge remains in knowing how to use them. One cannot attribute all blame and responsibility for changes in outcome measures to a programme, as a recent story from the corporate world makes clear. On 30 January 2006, ExxonMobil announced that its annual profit for the prior year had been USD 36 billion, a record for American companies, beating its own previous record by more than USD 10 billion. Judging by its end outcome measure of profitability, ExxonMobil was doing a wonderful job. However, informed observers of the oil business recognised that the immense profits did not necessarily reflect excellent performance by the company and its executives, since the profit was due mostly to causes beyond the control of company executives: high international oil prices and supply interruptions from Hurricane Katrina.

Outcome goals are considered better than output goals because they reward a programme for producing results, not for going though the motions of bureaucratic routines. But using outcome measures as an accountability tool may not be appropriate, because we cannot always attribute changes in outcome measures to the actions of a programme. Outcomes are often a step or two removed from the outputs of the programme, and have causes other than the programme. Thus we do not always know whether movement in a measure, good or bad, is due to actions of the programme or to external causes. This is the "attribution problem". It cuts to the heart of performance measurement and performance budgeting. Stated simply, the attribution problem is knowing how much of a change in an outcome indicator can be attributed to actions of the programme and knowing how much control over an outcome it is reasonable to expect of a programme. An important challenge in using outcome measures is being able to **hold programmes accountable to the right degree** for their attainment of outcome measures.

A programme in the State Department that chose pure outcome measures highlights both the promise and limits of using end outcome measures. Programmes in the international affairs field often have extremely lofty goals that are difficult to influence or measure directly. Nonetheless, it still makes sense to try to assess impacts on outcomes, although the measures must be treated with caution. The Support for Eastern European Democracy (SEED) Program in the Bureau of European and Eurasian Affairs in the State Department makes

hundreds of grants to countries of the former Soviet Union and Eastern bloc with the purpose of promoting democracy and increasing market orientation. Developing measures of the success of the programme was difficult because it makes literally hundreds of grants – grants that are very dissimilar in nature and that cannot be assessed by any common benchmark. Staff explained that the first time they completed the PART survey, they received a low rating. After that they went back and completely redid their measures, adopting new ones that were audacious in the extent to which they assessed true end outcomes.

To measure "democratisation", they adopted the independent Freedom House's "freedom scores" for countries to which they gave grants. To measure market orientation, they adopted other measures that were similarly ambitious: actual measures of market orientation in the economy. In adopting these measures, they took real risks, because these are variables over which they have some, but far from complete, control. Yet they did exactly what they should have and selected measures that assess the outcome that the programme is supposed to influence.

At the same time, the outcome measures are coupled with measures of outputs, organisational structure, effective co-ordination, and efficiency. The overall set of measures adopted seeks to assess a balanced array of features of the programme, from some over which the programme has a great deal of control to others over which it has only loose control. This seems to be a sensible approach to assessing a programme with such an immense aspiration.

With the outcome measures there is a problem of attribution: knowing how much of the change in the selected measures can actually be attributed to the work done by recipients of the programme's grants. If the countries that receive grants become more democratic, is that because the grants are doing what they should, or is it because other forces were driving democratisation in Eastern Europe? There is no way of telling, because there are no direct measures of the actual impact of the individual grants and no measures of other causes of democratisation. The same is true of the measure of economic change.

The OMB examiner who approved these measures understood fully that there would be problems of attribution. Not only could the programme be given credit for improvements that it was not responsible for, but also the programme could be blamed for reverses of democratisation that it could not have prevented. She asked the staff members of the programme whether they were willing to be held accountable to those measures; they said they were. What is good about having such outcome-oriented measures is that they give the programme all the right incentives. Staff have every reason to make grants that will really have an impact on the desired outcomes, and they will have an incentive to remove funding from programmes that are not accomplishing anything useful.

But what if there are military coups that remove democratically elected governments from power or if governments reverse market-oriented reforms? The examiner for the programme indicated that if forces clearly beyond the control of the programme had caused a decline in measures, it would not be reasonable to hold the programme responsible. These measures must be interpreted within the context of other available information. However, this is an imperfect process, and it is possible that the programme will be credited with successes it did not cause or blamed for failures it could not prevent. Still, the use of end outcome measures keeps attention focused on the purposes the programme was created to serve.

The Migration and Refugee Assistance-Protection Program in the State Department's Bureau of Population, Refugees, and Migration (PRM) has chosen as an outcome measure reductions in the number of refugees displaced across national borders in the world: "per cent reduction of long-standing global refugee population due to achievement of durable solutions." PRM's goal is to reduce the number of refugees by 25% by 2009. That is an end outcome measure in the purest sense, coupled with an ambitious target. This outcome measure is combined with a variety of measures of outputs and intermediate outcome measures. As outcome measures should, this will tend to keep the programme focused on achieving desired results.

The activities of PRM programmes tend to reduce the number of refugees, but many other causes contribute to the number of refugees. It would be unreasonable to conclude that PRM was failing in its mission if another war were to break out in the Balkans, increasing the number of refugees. But if the bureau cannot be held accountable for such increases in the number of refugees, it can be similarly difficult to know when it should be given credit for declines in the number of refugees. In the case of this programme, as with the SEED Program, examiners must evaluate the evidence of results in the context of other factors that might have an impact on the outcome indicator.

5.1. *Interpretation challenges of attributing programme performance to outcomes*

An important challenge for OMB examiners is knowing how to interpret and use outcome data. Performance budgeting has an appeal in that it appears to provide the information needed to make more rational budget decisions. But on closer examination, it is clear that outcome data, while important, are only part of the picture. Because outcomes are often several steps removed from the actual actions of a programme, there is always some question about the extent to which changes in outcome indicators are due to the programme or to other causes. Following are some examples of how attributing outcomes to programme performance can be problematic.

Sorting out multiple causes of outcomes. Many of the outcomes sought by government programmes are subject to many separate causes. A programme to reduce premature births might be doing a good job, but a separate cause beyond its control, such as an epidemic of crack cocaine, might erase all of the programme's accomplishments. If the programme were not doing its work, the outcome indicators might look even worse. Programme activities are but one of a number of factors moving the outcome indicator. Alternatively, causes apart from programme activities might be improving the outcome, which might improperly credit the programme with accomplishments it did not generate.

Accounting for time lags between performance and results. Some programmes have measurable outcome goals, but even when the programme is doing what it should, there may be a long time lag between programme activities and observable impacts. Research activities may produce benefits that are not seen for many years. The Superfund Program of EPA cleans up toxic waste sites, and an important goal of this remediation process is to reduce groundwater pollution. But it may take years for groundwater quality to improve.

Having limited authority to act. In some cases programmes have, by law, too little authority to have a major impact on outcomes or they labour under statutory design flaws. EPA, which is charged with cleaning up the nation's water bodies, has authority over certain kinds of contaminants and pollutants but lacks the power to control others. For example, residential development has a major impact on water quality of the Chesapeake Bay, but EPA has no control over local zoning decisions. While EPA actions stop some pollutants from entering the bay, overall water quality may still decline (although not as fast as it would without EPA regulations in place).

5.2. *The need for flexibility*

When developing measures to satisfy PART and completing the PART questionnaire, programme officials need to think hard and exercise creativity in devising the most outcome-oriented measures they can. At the same time, they need to be alert to the possibility that they can persuade their examiner that other kinds of measures are more appropriate.

In some cases it may be more straightforward to hold programmes accountable for intermediate outcomes than for end outcomes. Intermediate outcomes are helpful because: *i)* they have a close relationship to the actual goal of a programme; and *ii)* they are more readily under the control of a programme. This is not to say that a quest for end outcome measures should be abandoned. They are important, since they help to show whether a programme is contributing to the solution of the problem it is intended to mitigate.

Interpretation of outcome measures will always require some judgment. OMB's Shea says "there is no on-off switch" whereby a programme either is or

is not responsible for observed outcomes. "You can never make decisions solely on the basis of measures," he explains. "This is a framework for having a discussion about finding ways to make programs perform better." His is a sensible approach to take in dealing with end outcome measures.

6. Conclusions and recommendations

6.1. *Conclusions*

OMB has succeeded in implementing an intelligent and effective system of performance budgeting. It has produced 800 assessments in four years, and will complete assessments of all federal programmes in 2007.

The pre-eminent strength of PART is the reasonableness of the questions. OMB has worked diligently to produce a set of questions that directs public managers to think carefully about important issues. Numerous observers have said they found the questions to be good and sensible, focusing attention on important aspects of public management. Answering the questions causes programme managers to think in useful ways about how their programme is designed and run, and how it produces results.

Further, there is reason to think that the programme ratings are objective and able to distinguish between effective and ineffective programmes. A danger in any system of programme assessment is that differences in the scores will not manifest true differences in programme management and programme quality, but instead reflect differences in the skill of individuals in filling out the form or other similar irrelevant factors. It is difficult to assess the validity of PART scores – that is, whether they are measuring something real – because there is no other available external gauge that is known to be a reliable measure of programme quality. In this sense, PART is a pioneering effort. However, this author's research and interviews indicate that PART ratings, even if not a perfect measure of programme quality, measure real differences in programmes.

Senior career officials who had participated on the programme side in multiple assessments generally believed that the programmes that received the higher scores were in fact the better programmes. Within bureaus at least, assessments are ranking programmes correctly, but that does not tell us if comparisons across bureaus and across departments are also valid. Further, the analysts and programme managers interviewed by the author – virtually all careerists – almost uniformly believed that the exercise of completing the PART questionnaire was good for programmes.

PART is notable for the emphasis it places on results and outcomes rather than processes and outputs. However, it appears that some of the programmes that have received effective scores do not have true outcome measures; instead, budget examiners for these programmes have approved measures of outputs or processes. The decisions of the examiners in these cases seem reasonable,

given the difficulty or inappropriateness of measuring outcomes in those particular cases. Just as people in state and local government often criticise the federal government's insistence on "one size fits all" solutions, a rigid insistence on the sole use of outcome measures fails to acknowledge the tremendous diversity of federal programmes. Whenever possible, budget examiners should insist that programmes develop outcome measures, but they should also recognise that, for certain programmes, other kinds of measures are most appropriate. It appears likely, however, that OMB will continue prodding programmes that do not yet have outcome measures to identify some.

Programmes face an important choice in adopting measures. OMB clearly has a strong preference for end outcome measures. Insofar as a programme can identify measures that are acceptable to OMB, that can lead to a better score. If OMB signs off on the measures, the immediate reward is that the programme will escape the dreaded "results not demonstrated" category. Initially, OMB has required only that the measures have been identified and adopted, not that there necessarily be any data to plug into them. That comes later. The downside to adopting true outcome measures is that the programme will be held accountable based on those measures, and if it cannot meet targets or demonstrate improvements in the measures, OMB may deem the programme ineffective.

The strategic dilemma for a programme is this: by adopting a true outcome measure, a programme may achieve a better score right now; but it must then be held accountable to that measure down the road. Adopting a true outcome measure is a good idea if events are moving in the right direction, because then the programme may be credited with successes it did not cause. If it is clear that failure to achieve an outcome goal is due to circumstances the programme cannot possibly control, it may escape being held accountable. Thus it is possible that a programme can take credit for improvements it does not cause while avoiding blame for problems it cannot control. This is an ideal situation for the programme, but it does not promote transparency and accountability.

Like GPRA, PART focuses attention on outcomes and results, but it has not yet succeeded in bringing about significant changes in programme management. Improvements in ratings have come predominantly from learning better how to use the PART instrument, not from introducing important changes in programme design or management. Performance budgeting is intended to improve resource allocation by devoting budget increases to programmes that perform well. But by emphasising results rather than process, performance budgeting is also intended to encourage programmes to improve management and change the means by which they accomplish their goals. If programme officials find that existing processes do not allow them to achieve the desired results, they may be compelled to rethink how the programme operates. The officials interviewed for this report did not report that the improved ratings for their programmes came from management changes.

The absence of major management changes should not be surprising at this point. PART has been in place for only a few years, and the kinds of management innovations that PART should ideally bring about will take years to emerge. Reform will occur when programmes find they cannot get good scores without real change. Further, the author spoke, for the most part, with individuals in programmes that received good scores, programmes with the least need to change. Assessing the success of PART as an incubator of management reform will take time and patience. This is difficult for OMB officials because they are understandably in a hurry to demonstrate that their efforts have generated successes. But producing useful assessments of 1 000 federal programmes in five years will in itself be a tremendous accomplishment.

6.2. Recommendations

The insights gleaned from this initial assessment of the OMB PART initiative can be summarised as a set of recommendations for agency personnel who prepare the materials used by OMB in conducting the assessments and ratings, as well as recommendations for OMB itself.

6.2.1. Recommendations for departments and agencies

The experience of programmes that have successfully navigated the challenges of PART assessments provides ten useful insights that may help other agencies improve future assessments.

Recommendation 1: Don't give PART to the intern. Answering the questions well and persuasively requires extensive knowledge of the programme. Experienced staff need to be involved in preparing the PART. There are cases of programmes delegating the PART questionnaire to a low-ranking or new employee, or in one case giving it to a consultant who was not deeply familiar with the programme. Less certain is the question of who specifically in the department should prepare the PART responses. In most cases, it appears that the primary responsibility is given to an individual in the planning or budget office at the bureau level. In many programmes that have been successful with PART, a fairly senior official has responsibility for it. Typically this person works closely with a more junior person. But whoever runs the PART needs to be in close touch with people who know the programme very well. In science-based programmes, the scientists themselves are typically involved in answering the questions. In addition, make sure the people handling the PART are good writers. PART is not a writing test, but a good writer can answer questions artfully and make a programme look its best; a poor writer can fail to communicate the performance of a good programme.

Recommendation 2: Get professional help if needed. Staff at one programme that received a weak rating on PART recognised that they faced a major challenge in devising measures of results. They were convinced that the

programme was well run and designed, but that documenting the accomplishments would be difficult. To solve this problem, the programme rehired a former staffer with extensive programme evaluation experience, providing them with an individual with a unique and valuable combination of skills. Experts in programme evaluation may be able to help a programme explain what it does and to devise measures that OMB will approve.

Recommendation 3: Work with your OMB examiner. Individuals whose programmes have raised their PART rating repeatedly emphasised the importance of working with their OMB examiner. In some cases, the examiner came to the agency and spent the day with programme personnel working on their performance measures. Involving examiners in the process of devising measures may increase their buy-in of the measures, and may also increase their understanding of why certain measures were adopted. Make friends with your budget examiner. Programme officials with good ratings often praise their OMB examiner. They spend a lot of time educating their examiner, and the examiner can in turn help programme staff with the task of articulating accomplishments.

Recommendation 4: Link PART to your strategic plan. PART does not take place in a vacuum, detached from planning. Programmes with strong strategic planning already in place have a better experience with PART. Strategic planning under GPRA prepares programmes to answer the PART questions, and encourages programme staff to think clearly about means-ends relationships. Individuals who have worked on successful PARTs stress the seamless connection between their strategic planning and PART efforts. PART is not a replacement for GPRA; PART is a natural extension of the planning process put in place by GPRA.

Box 6. **Recommendations for departments and agencies**

1. Don't give PART to the intern.
2. Get professional help if needed.
3. Work with your OMB examiner.
4. Link PART to your strategic plan.
5. Read OMB's guidance carefully.
6. Provide ample documentation.
7. Measure what you can.
8. Understand OMB's perspective.
9. Renegotiate the definition of the programme.
10. Express measures in non-technical language.

Recommendation 5: Read OMB's guidance carefully. OMB supplies detailed guidance for examiners that lays out the criteria a programme must meet to get

a "yes" answer to each question. Programme officials need to pay very close attention to those criteria and address them exactly as they answer the questions. Programme officials should read the explanation supplied for each "no" answer and make sure that the examiner has justified it precisely in terms of the guidance. Some programme officials have noticed that examiners have strayed from the guidance in justifying "no" answers, holding programmes to higher standards. Make sure that the examiner is holding a programme to the exact criteria outlined in the guidance, not another standard he or she feels is appropriate.

Recommendation 6: Provide ample documentation. OMB wants evidence to document answers. Programmes that have successfully negotiated the PART process emphasise the importance of providing complete, voluminous documentation for all claims. That means thousands of pages of evidence to back up every claim made on the PART questionnaire. The point is not to inundate or intimidate an examiner with an ominously large stack of paper, but to anticipate questions the examiner might have. Examiners may not have time to read all of the documentation provided, but it is important that, if they look for something, they can find it. Thus this also means that the documentation should be carefully organised and easy to navigate.

Recommendation 7: Measure what you can. There is a lot of pressure for programmes to adopt outcome measures, but it has been possible in some circumstances to gain approval for measures of outputs. The PART guidance makes an exception for research and development programmes. But other kinds of programmes have been able to persuade an examiner that outcome measurement is impossible or inappropriate in their particular case. To see if this is possible, look at programmes that have had output measures approved to see if their circumstance applies to you.

Recommendation 8: Understand OMB's perspective. Individuals at the programme level who are answering the PART questions should make sure they are interpreting the questions the same way that OMB and their examiner are. Staff at one programme that managed to raise their rating said they found that the examiner looked very differently at some of the questions. The second time around they were better able to anticipate what the examiner wanted.

If the OMB examiner does not like the measures a programme proposes, it can be useful to look at other similar programmes elsewhere in the federal government that have had their measures approved. Perhaps it is possible to emulate or devise analogues to successful measures. In the State Department, the Bureau of Resource Management serves as a central clearinghouse and helps programme staff identify the kinds of measures that have been approved elsewhere at State. Emulate successful measures. Get to know your

examiner. Spend time with your examiner. Get your examiner to discuss measures in advance: get buy-in.

Recommendation 9: Renegotiate the definition of the programme. Many of the programme definitions or demarcations are idiosyncratic at best. Defining programmes in terms of budget accounts does not work in all cases. The Bureau of Population, Refugees, and Migration initially defined programmes in terms of international organisations to which they made contributions for refugee assistance: the United Nations High Commissioner for Refugees or the International Committee of the Red Cross. The bureau has since decided to redefine the programmes in terms of functions performed, such as "protection". In the Interior Department, the "Land and Water Conservation Fund-Land Acquisition" programme got a poor score, at least in part because it is not a programme but an "activity". In the first year of PART, a large number of small programmes in the Department of Health and Human Services were evaluated; by the second year, many had disappeared, apparently merged with other programmes that were assessed.

Recommendation 10: Express measures in non-technical language. An important goal of PART is transparency, but technically worded measures are opaque. Some examiners have technical backgrounds, but most do not. Nor do most members of Congress. Measures that ordinary people can understand are likely to be better received by OMB and Congress. It is important for government programmes to document their accomplishments in ways that the people who support and pay for them – members of Congress and the taxpayers – can readily understand and appreciate.

6.2.2. Recommendations for OMB

Likewise, the following insights may help OMB better navigate some of the challenges raised in this report.

Recommendation 1: Formally introduce appropriate flexibility about what constitutes acceptable measures. It is laudable that OMB has pressed diligently for outcome measures whenever possible. It is equally laudable that examiners have exercised discretion and allowed some programmes to substitute other kinds of measures when appropriate. However, the PART guidance is not as clear as it should be about the circumstances under which something other than outcome measures is acceptable. The guidance indicates: "Programs that cannot define a quantifiable outcome measure – such as programs that focus on process-oriented activities (*e.g.* data collection, administrative duties, or survey work) – may adopt a 'proxy' outcome measure." But, in fact, other kinds of programmes have had output or intermediate outcome measures approved. It would be helpful to programme-level staff and OMB examiners alike if they had clearer indications about when alternatives to end outcome measures are acceptable.

Recommendation 2: Provide multiple response categories for answers to PART questions. While the questions are sensible, the problem is that they must be answered with either a "yes" or "no". Since the phenomena being assessed with the questions tend to be continuous variables, forcing the answers into two categories necessarily introduces error. The real answers to the question about clarity of programme purpose must reside along a continuum stretching from, say, 1 to 10. Suppose OMB decides that scores of 8 and above get a yes, and those below get a no. Such a system gives the same score to programmes with a 1 and a 7, even though they are very different, and gives very different scores to programmes with a 7 and an 8, even though they are very close. Such scores will not reflect reality as well as they might. Examiners report that they often have difficulty deciding how to answer a question, which means there are a lot of close calls, and thus a lot of error. The solution is a simple one of allowing more response categories. Some of the questions in Section 4 permit a response of a "large extent" rather than just "yes" or "no". Permitting intermediate responses to all questions would yield more accurate assessments.

Recommendation 3: Distinguish between design and management failures. There is no shortage of programmes that fail because Congress saddled them with a design that makes political sense, but which inhibits the ability of managers to produce good results. With PART, OMB is standing up to Congress, pointing out design problems, and insisting that all programmes produce results. As OMB sees it, congressionally mandated bad design is no excuse. But it can be discouraging to agency programme managers if their programmes are designated ineffective or results not demonstrated because of a programme design that Congress foisted on them and they cannot control. It would be useful if PART ratings made a distinction between: *i)* failures that are caused by a congressionally mandated programme design; and *ii)* failures caused by poor programme management. One could also add a third category: failures caused by unpredictable circumstances beyond the control of programme managers, such as natural disasters, wars, or inadequate resources. The solutions to these problems are completely different. Managers who do a good job of running a flawed programme need recognition for their achievements, just as Congress needs to be continually reminded of the importance of eliminating statutory impediments to programme effectiveness.

Box 7. **Recommendations for OMB**

1. Formally introduce appropriate flexibility about what constitutes acceptable measures.

2. Provide multiple response categories for answers to PART questions.

3. Distinguish between design and management failures.

Notes

1. See *www.whitehouse.gov/omb/budintegration/pma_index.html*.

2. The guidance is available at *www.whitehouse.gov/omb/part/index.html#guidance*.

3. The scorecards can be viewed at *www.whitehouse.gov/results/agenda/scorecard.html*.

References

Gilmour, John B. and David E. Lewis (2006), "Assessing Performance Budgeting at OMB: The Influence of Politics, Performance, and Program Size", *Journal of Public Administration Research and Theory*, 16, pp. 169-186.

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

Hatry, Harry (1999), *Performance Measurement: Getting Results*, Urban Institute Press, Washington DC.

Hatry, Harry (2001), "What Types of Performance Information Should be Tracked?", in Dall Forsythe (ed.), *Quicker Better Cheaper*, Rockefeller Institute Press, Albany, New York, United States.

Joyce, Philip G. (1999), "Performance-Based Budgeting", in Roy T. Meyers (ed.), *Handbook of Government Budgeting*, Jossey-Bass, San Francisco, California, United States.

Joyce, Philip G. (2003), "Linking Performance and Government: Opportunities in the Federal Budget Process", IBM Center for The Business of Government, Washington DC.

Kettl, Donald (2005), *The Global Public Management Revolution*, 2nd edition, Brookings, Washington DC.

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, pp. 63-73.

Melkers, Julia E. and Katherine G. Willoughby (2001), "Budgeters' View of State Performance-Budgeting Systems: Distinctions Across Branches", *Public Administration Review*, 61, pp. 54-64.

Melkers, Julia E. and Katherine G. Willoughby (2004), "Staying the Course: The Use of Performance Measurement in State Governments", IBM Center for The Business of Government, Washington DC.

Romero, Simon and Edmund L. Andrews (2006), "At ExxonMobil, A Record Profit but No Fanfare", *New York Times*, 31 January.

Schick, Allen (2001), "Getting Performance Measures to Measure Up", in Dall Forsythe (ed.), *Quicker Better Cheaper*, Rockefeller Institute Press, Albany, New York, United States.

Wildavsky, Aaron B. (1984), *The Politics of the Budgetary Process*, 4th edition, Little Brown, Boston, Massachusetts, United States.