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

Working Paper Series
La Follette School Working Paper No. 2013-003
http://www.lafollette.wisc.edu/publications/workingpapers

Advancing the Empirical Study of Performance Management: What We Learned from the Program Assessment Rating Tool

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

dmoynihan@lafollette.wisc.edu

February 20, 2013

forthcoming in *American Review of Public Administration*

Robert M. La Follette
School of Public Affairs
UNIVERSITY OF WISCONSIN–MADISON

1225 Observatory Drive, Madison, Wisconsin 53706
608-262-3581 / www.lafollette.wisc.edu

Advancing the Empirical Study of Performance Management: What we Learned from the Program Assessment Rating Tool

Donald P. Moynihan,

La Follette School of Public Affairs, University of Wisconsin-Madison

Forthcoming at the *American Review of Public Administration*

Abstract

Interest in performance management has never been higher. But what does actual research on this topic tell us about the effects of performance management, reform efforts and governance more generally? Historically, the answer often seemed to be that such reforms did not work very well. This paper focuses on a recent chapter in the history of performance management at the US federal government, the Program Assessment Rating Tool (PART). PART was created by the George W. Bush administration, and ended by the Obama White House. PART, like many management initiatives, came and went. What distinguishes it, however, is the quality and breadth of the research it prompted—research that has increased our knowledge of performance management and reform implementation as well as a whole host of fundamental governance phenomena such as political ideology, administrative burden, performance information use, leadership commitment, and goal ambiguity.

Introduction

In the last decade, the study of performance management has progressed dramatically. In part, this is because of the ongoing popularity of the topic for practitioners. But it is also a function of more rigorous use of research methods to address a wide variety of questions. For example, framing performance information use as a behavioral question of significant importance has encouraged a variety of researchers to model its antecedents and consequences (Moynihan & Pandey, 2010). We are starting to see the use of experimental designs to persuasively model the response to performance data (e.g., James, 2011). The use of mixed-methods and careful modeling have unearthed how performance regimes operate in a contracted-out environment (e.g. Heinrich, 2010; Soss, Fording & Schram, 2011).

In a variety of ways we can therefore make the case that there has been a dramatic improvement in our knowledge of the workings of performance management reforms. One area that is less tractable to some of the improved application of research methods are whole-of-government performance reforms. We usually lack a randomized control or variation in the rates of adoption that might provide inferences about the effects of adoption of such reforms. There are too many potential measures of performance to consider, and attribution problems are too severe – we cannot exclude the possibility that some change in a desired performance goal were not the result of external factors. We might have success stories on some dimension – say money saved - but we also have tradeoff problems. Public organizations pursue multiple goals, and while we might point to an increase in performance in one area, we are likely going to be unaware of whether it came at the cost of a reduction in performance in another (possibly unmeasured) area.

What to do given the difficulties of studying performance management reforms and similar governmentwide reforms? Scholars have still looked for ways to provide insights, and such insights have come. To develop these insights, researchers have historically relied on some variation on the "soak and poke" method associated with Richard Fenno: the researcher mixes observation and interviews of public managers to generate insights about the implementation of the reform. I will refer to this as the "traditional approach", which can be found in the work of giants in the early study of performance systems, such as Aaron Wildavksy and Allen Schick. It

also appears in more recent empirical work (e.g., Frederickson & Frederickson, 2006; Moynihan 2008, Radin, 2006).

This traditional approach has offered us some very real and profound insights. In an earlier era, Wildavksy pointed to the cognitive limitations and political disinterest that made the systematic use of performance data unlikely. More recently, Radin (2006) has argued persuasively that the institutional design of US government is at odds with basic performance management assumptions, and that performance management works less well for some tasks than others. There is a general sense that these processes are burdensome, rarely replacing existing requirements, and complied with begrudgingly. The tone of this work is frequently critical, and a traditional trope is to compare the hopes or explicit goals of reformers with how things work in practice. This has often been useful, casting a skeptical eye on reforms that frequently have uncritical or naive boosters. But these assessments often appear to leave little room for optimism for performance management techniques.

The traditional approach to performance management is frequently mixed with general theorizing about the relationship between these reforms and the existing social processes and behavior in government. It is not always apparent whether the more general theories are a function of the qualitative data, or whether the data is more illustrative. Indeed, in some cases the theoretical argument is explicitly normative.

One very real limitation of the traditional approach is that it is often only as good as the person doing it. It depends on their ability to gain trust, and what they can make of what they observe and hear, and how persuasively they tell that story. Or to put it another way, we are not all Wildavsky. Of course, the quality of more recent work that relies primarily upon quantitative techniques also depends on the researcher, but the terms for testing causal claims and statistical problems can be assessed more transparently, and others can take the same data and evaluate it.

This paper narrates the insights scholars have drawn from studying the Program Assessment Rating Tool, a now-defunct government-wide performance initiative of the George W. Bush administration. One pattern that characterized PART is a greater presence of research that relied on quantitative techniques, and a behavioral approach. PART might have had negligible effects on actual governing (more on that in the paper) but it has undoubtedly

facilitated the empirical study of performance management, providing an array of studies using different techniques, asking different research questions. The resulting studies provide some of the most compelling tests on whether a governmentwide reform achieved its goal. There may have been a parallel level of interest in reforms like the National Performance Review and the Government Performance and Results Act (GPRA) of 1993, but with some notable exceptions (e.g. Thompson, 1999) this interest did not develop reasonable estimates as to the effect of the reform. Studies of PART also offer insights into basic questions of governance and management: the role of political ideology in administrative reform, administrative burden, goal ambiguity, political appointees and the role of leadership. In what follows I do not include all studies of PART, but tend to favor those in peer-reviewed journals, and which were clearly empirical, regardless of whether they were qualitative or quantitative.

It is important to note that the purpose of this paper is not to anoint one methodological technique or theoretical tradition at the expense of another. Indeed, I am in no position to do so even if I were so inclined. My own studies of PART have ranged from the traditional soak-and-poke approach to developing and testing formal models, and I have mixed normative and behavioral theorizing. In the narrative that follows, I weave in work from these contrasting approaches. The central point for the study of performance management, and administrative reforms more generally, is the need to utilize a variety of approaches. We know more about the workings of PART, and PART has helped us to understand other basic questions, precisely because it has been tackled by a variety of approaches that have collectively complemented one another. The key difference relative to earlier reforms is the balancing of the traditional approaches with methodological techniques that sought to test causal theory.

What was PART?

Here I offer a thumbnail sketch of the essential features of PART. Those interested in a more extensive account of the development and history of PART should turn to Dull (2006), or Moynihan (2008). PART was an outgrowth of the 2001 President's Management Agenda, a blueprint prepared by the US Office of Management and Budget (OMB) to represent President

Bush's management goals. Included in this agenda was the goal of integrating performance data and budgeting. PART was a product of this goal.

PART itself was a survey instrument, developed by OMB staff with outside advice. The instrument asked 25-30 questions divided into four categories: program purpose and design, strategic planning, program management, and program results. Based on the responses to those questions, programs were given a numerical score that aligned with a categorical scale of performance ranging from effective, moderately effective, adequate or ineffective. In cases where evaluators felt they could not make a judgment, programs were assigned a "results not demonstrated" judgment, which was generally believed to be a negative assessment on a par with an ineffective grade.

The tool was implemented by OMB budget examiners, but with extensive consultation with agency staff. Ultimately, the evaluation reflected the judgment of the examiner, but the input of the agency was substantial. If the agency was unhappy with a program score, it could appeal to senior appointees at OMB, or seek a reevaluation at a later period.

Over a five year time period, the OMB assessed just over 1,000 federal programs, representing 98% of the federal budget. PART was hailed as a successful innovation, winning the prestigious Innovations in American Government Award. However, PART ended with the Bush administration. The Obama administration characterized the tool as ineffective at generating true performance information use. Overwhelmed with the economic downturn and dealing with a massive stimulus package, it deployed OMB resources for other purposes (Joyce, 2011).

Did PART Increase Performance Information Use?

To assess whether PART worked, we must define what PART's goals were. To some degree, this is a matter of speculation, but one clear goal was to increase performance information use. The Bush administration criticized GPRA on these grounds, saying ""After eight years of experience [since passage of GPRA], progress toward the use of performance information for program management has been discouraging …Performance measures are insufficiently used to monitor and reward staff , or to hold program managers accountable" (OMB 2001, 27). Did PART do any better than GPRA in this regard? PART certainly placed

more demands on agencies than before (as explained below), but did it increase use? At an aggregate level, the answer to both questions appears to be no.

Managerial Use

Moynihan and Lavertu (2012) seek to estimate the effects of PART using 2007 Government Accountability Office (GAO) survey data. Since PART was implemented across government, there is no group of control and treatment groups to compare, and we do not have cross-time data that allows pre-post comparisons of exposure to PART. In an attempt to solve this problem, Moynihan and Lavertu use self-reported involvement in PART as a treatment to assess its impact, with the assumption that this involvement should increase use. Comparisons of rates of performance information use among those involved with PART and those who were not can be therefore used to infer if PART increased use. If PART is unable to increase performance information use among those exposed to it (about one third of federal managers surveyed), there seems little hope it will increase use among those who are unaware of it. While there is a possible selection effect in play (those who were selected to be involved in performance management might be different from others), it seems to run in the direction of finding a positive result between involvement and use.

Moynihan and Lavertu divide the dependent variables representing performance information use into two categories. *Passive use* represents the use of data to comply with existing performance management requirements, and is represented by survey items tracking the use of data for refining measures and setting goals. *Purposeful use* implies the use of data for decisionmaking and performance improvement, and is represented by items tracking use of data for setting program priorities, changing processes, allocating resources, taking corrective action to solve program problems, setting employee job expectations, and rewarding employees.

Controlling for agency fixed effects and a variety of additional individual and organizational variables, the authors find that involvement in PART is not significantly associated with purposeful performance information use. The one exception is that it is positively associated with the use of data for changing processes, though the size of the effect is not large and it is significant only at marginal levels. In contrast, PART involvement is strongly associated with passive uses of data to refine measures and goals. The authors compare the findings on PART with 2000 survey data tracking the same items with respect to GPRA. The results look

almost identical. Involvement in GPRA is associated with passive use of performance data, but not with purposeful forms of use. In fact, GPRA performs slightly better statistically than PART, associated with using data to change processes at the $p > .05$ for a two-tailed test, and with using data to set program priorities at the .10 level.

The findings are at odds with the goals of PART, but are consistent with claims that critics of PART have made. Overall, the results suggest that despite differences, PART and GPRA generated similar responses among federal managers. They were perceived primarily as a compliance exercise rather than an opportunity to generate purposeful performance information use (though in the section of political ideology that follows, I describe one major exception to this finding). As a formal mechanism of control PART, like GPRA, emphasized creating routines of data collection and dissemination, and gave relatively little attention to routines of performance information use. From this perspective, it is unsurprising that federal managers responded by devoting attention to measurement, and not use.

Moynihan and Lavertu (2012) point to other factors are associated with performance information use. Three are highlighted here, on the basis of relatively strong theoretical and empirical support from prior work on performance information use, as well as findings in the context of PART.

Goal Clarity. A variety of studies on performance information use have emphasized the positive benefits of goal clarity to performance information use (e.g., Moynihan, Wright & Pandey, 2012). This is consistent with basic claim of Locke and Latham's goal-setting theory that more specific goals are more motivational. Moynihan and Lavertu (2012) add an additional twist to the goal-clarity point, showing that managerial perception that it was easy to determine how to use performance information to improve the program was significantly associated with use. The connection between data and action is partly a function of the complexity of the task, but likely also partly a function of those undertaking it. It is likely that greater consideration of causality in discussions of performance may make clarify the connection between measures and actions.

Learning forums. Federal managers who reported having basic discussions with their supervisors and peers about performance information also were more likely to report using performance data for purposeful means. This may seem an obvious point. But discussion of data, and using data is not the same thing. A common design flaw of performance management reforms, including

PART, is the failure to give attention to creating routines of use. Discussion of performance data appears to be a strong social process that must be in place before use occurs.

Leadership Commitment by Leaders. Moynihan and Lavertu (2012) find that perceived leadership commitment to results is positively associated with performance information use in the context of PART, echoing prior results in the context of GPRA (Dull 2009) and elsewhere (e.g., Askim, Johnsen & Christophersen, 2008; Moynihan & Ingraham, 2004). These findings align with some qualitative work that had pointed to leadership commitment as important to PART. For example, in his qualitative study of PART implementation Gilmour (2006) noted that perceived commitment to PART by Secretary of State Colin Powell facilitated lower-level commitment to it.

Budgeting Use

A particular goal of PART was to facilitate the use of performance data for budgeting. The Moynihan and Lavertu (2012) assessment of the impact of PART on managers did not find a relationship between PART involvement and use of performance data for budget execution. Gilmour and Lewis (2006) regress PART scores on changes the President's proposed budget, and find a modest but significant positive correlation. This fits with qualitative accounts of OMB staff taking PART seriously, using it as at least one relevant piece of information in budget decisions (Moynihan 2008; Posner & Fantone, 2007; White, 2012). It also makes a good deal of sense that OMB officials might be at least somewhat likely to pay attention to the results of the tool they developed.

Actual budget appropriations are determined not by the President, but by Congress. Here, qualitative and quantitative work generally coalesces to suggest that PART did not have an effect on appropriations. Qualitative research found that congressional actors paid little attention to PART (Moynihan, 2008; Redburn & Newcomer, 2008), and PART scores did not systematically influence budget decisions. Content analyses of Congressional budget deliberations suggested that PART was not widely used, and appeared to lose influence over time (Frisco & Stalebrink, 2008) and that Congress had plenty of other types of performance information to draw on if it was so inclined (Moynihan 2008). One reason for this resistance is that congressional staffers often doubted the judgments that less-experienced OMB budget examiners drew from PART

(White 2012). Heinrich (2012) offers the most careful attempt to link PART to congressional decisions. Examining 95 Health and Human Service programs, she finds no connection between PART scores and budget changes.

A basic measure of Congressional resistance is that efforts to put PART in statute won little support, partly because of Democratic fears it was a partisan tool, but also because of an unwillingness to delegate institutional prerogatives from the legislature to the White House (Moynihan 2008). Stalebrink and Frisco (2011) developed a quantitative model of legislative support for PART. To measure their dependent variable, they coded public comments made by legislative actors about PART. They found that partisanship did not explain legislative statements of support for PART. Legislators with a business background were more apt to support PART, while those with greater experience or who had received a great deal of campaign contributions from special interests were more resistant to it.

The Costs of PART

If PART did not obviously increase performance information use among managers, or result in performance budgeting, there is good evidence that it generated significant costs in implementation. Reforms do not implement themselves, and implementation costs should be considered as a factor in assessing the worth of these reforms. In the case of PART two particular sets of actors – OMB staff and agency officials – bore the brunt of these costs.

The death of PART was generally not mourned by OMB staff. They had added PART to their existing budgetary responsibilities with few additional resources. OMB staff complained about the burdens of PART (Moynihan, 2008; Redburn & Newcomer, 2008). One GAO report argued that PART was spreading budget examiners attention too thin, a point reinforced by White's interviews of OMB staff (2012). The focus on PART also created opportunity costs, argued White (2012), reducing the ability of OMB to think more analytically about allocative efficiency in judging spending.

Survey assessments of agency staff also found the process burdensome. The GAO survey found that 31% of agency officials surveyed reported they had been involved in PART in some way (Moynihan & Lavertu, 2012). Another survey asked "How much time and effort did your

agency put into the Program Assessment Rating Tool (PART) process?" Answers ranged from "Not much at all" (1) to "Tremendous amount" (5), with an average response of 3.92, well above the midpoint. Qualitative studies have also emphasized the burdens created by PART on agencies (Frederickson & Frederickson, 2006; Gilmour, 2006). Redburn and Newcomer (2008, 3) noted: "Some agency officials are skeptical, or even cynical, about the benefits reaped by program managers and executives from program assessment and evaluation, especially compared to the heavy costs entailed in both and when the resulting data are not used to inform decisions." Some descriptive statistics reported by Gallo and Lewis on the perceived value of PART reiterate this point. They find that more managers believed that PART rarely (22.7%) or never (14.2%) reflected real differences in program performance than believed that that it almost always (2.6%) or generally (14.9%) reflected real differences. An additional 26.6% believed that PART sometimes reflected real differences, and 19% did not know.

Preach to the Choir and Punish the Heretics? Political Ideology and PART

PART represents an opportunity to examine how political ideology may matter to the implementation of management reforms. A research literature on political control of the bureaucracy has emphasized two main tactics: centralization of policymaking authority into the White House and politicization of agencies primarily via the use of political appointees (Lewis, 2008; Rudalevige, 2002). Both parties have used these tools, and seem to use them more aggressively when there is ideological divergence between the President and an agency. These tools are therefore overtly ideological, and given the perceived liberalism of large parts of the federal bureaucracy, tend to be most overt in the case of conservative Presidents (e.g. Golden, 2000; Moynihan & Roberts, 2009). However, there is relatively little attention to how nominally non-partisan "good-government" initiatives are implemented in ways that suggest a partisan bias in implementation on the part of the principal, or response on the part of the agent (for an exception, see Durant's (2008) account of the implementation of the National Performance Review in the conservative Department of Defense).

The study of PART unearthed a series of insights into the role of political ideology in reform implementation. A series of papers show that under a relatively conservative President,

programs in liberal agencies received systematically lower effectiveness scores, were more at risk of budget losses because of these scores, and experienced a greater administrative burden in implementing the reforms. The practical benefits of PART, in terms of greater use of performance data, appear to be reserved for more conservative agencies.

PART was presented as a good-government initiative, deliberately constructed to avoid the criticism that it was a partisan tool. The basic design of the tool incorporated peer-review from outside performance management experts and the National Academy of Public Administration, and was revised to drop questions that might be seen as overtly-political. Mitch Daniels, Director of OMB when PART was created, urged staff to create a non-partisan tool (Moynihan 2008). The assessments were carried out by OMB career staff rather than appointees. Compared to the traditional mechanisms of political control that presidents have used, PART seemed markedly different, argued Matthew Dull (2006), who offers a narrative history of its creation. Dull offered a compelling argument for why the Bush administration invested in making PART a politically neutral tool, asking "Why would this administration, otherwise so strategic in its approach to administration, invest great time and energy building a neutral and transparent instrument like PART?" (2006, 189). The answer is that reforms like PART "seek to orient and revise policy administration by gathering and organizing intelligence and by aligning agency activities with administration priorities. Modern presidents face a dilemma in that policy competence, and the credible information about policy process and outcomes it provides, requires investments of resources and discretion, and with them the acceptance of political risk" (192). There was a political logic to invest in a politically neutral tool, Dull concludes, because the tool could only be beneficial in enabling the administration to pursue its policy goals if it was seen as credible.

My own early assessment of PART tended to align with that of Dull's. Certainly in interviewing members of OMB, they took pains to emphasize that PART was not a partisan tool. Overt partisanship would likely have reduced the willingness of OMB career staff to invest the very great time and effort involved in implementing PART. OMB staff tended to argue that the PART score was not politically influenced, although they noted that the budget decisions that accompanied PART were always a political decision, implying the link between PART scores and budgets would not be equally applied. Some agency staff and Democrats tended to be less sanguine about PART, suggesting that since liberal programs tended to deal with complex social

problems, they would have a harder time demonstrating outcomes. Democratic legislators viewed PART "as a White House tool to cloak ideologically based attacks on traditionally liberal programs under the neutrality of management" (Moynihan, 2008, 134).

It is fair to suggest that while there was some qualitative evidence both for and against the notion that political ideology mattered to PART, there was little evidence that it was an overt strategy. But the weight of evidence that emerges at this point suggests that PART was partisan, if not in thought than in deed. This evidence comes in a variety of forms, but it implies a basic willingness on the part of the reader to accept that certain types of tasks, programs, and agencies can be placed approximately on the liberal-conservative political spectrum. Such a claim has been made, with little contention, within political science (Clinton et al., 2012; Clinton & Lewis, 2008). But systematic use of both the concept and measures of agency ideology is novel within public administration, perhaps reflective of a traditional intellectual preference for keeping matters of partisan politics out of administration. It is worth noting that the bulk of the studies on the ideology of PART come from scholars with a political science disciplinary training.

The potential influence of ideology, or other discretionary factors, is made feasible because PART was not a precise tool applied in exactly the same way by all budget examiners. Despite the very real effort of the OMB to PART a consistent tool, including training and a guidebook on how to implement PART (OMB 2007), qualitative research has pointed out the ambiguity inherent in many of the questions asked in PART, and in defining what constituted an acceptable answer (Posner & Fantone, 2007; Moynihan, 2008). The possibility of an ideological effect is also enhanced because PART reallocated greater power to OMB, which is part of the White House. Prior to the Bush administration, OMB played a secondary role in shaping the performance goals of agencies. PART gave OMB a process by which it could communicate to agencies what their preferred goals should be. Indeed, a content analysis of management recommendations made by OMB to agencies during the PART process found that the majority of these recommendations focused on agency goals and measurement, and not management issues (Moynihan, 2008). Even though most staff there are career officials, part of their job is to represent the goals of the president. Presidential preferences may, consciously or unconsciously, affect the judgment of budget examiners in ambiguous situations that allow for discretionary judgment. And the judgment process was not completely separated from political staff at OMB, since unhappy agency staff could appeal their judgments up to political appointees at the OMB.

This happened rarely, but the mere threat of it may have compelled budget examiners to marginally favor programs that aligned with appointee ideology rather than risk having their judgment overturned. Finally, political leaders in agencies could affect how PART was implemented. They could make it more or less a priority, affecting how resources were allocated to completing PART, and therefore affecting scores. They could also use PART as a mechanism to closely evaluate programs they may not have trusted were perform well (more on this process below).

How did ideology reveal itself? First, a variety of evidence has established that more liberal agencies received systematically lower scores on PART than more conservative ones. Relying on agency-level preference estimates created by Clinton and Lewis (2008) to categorize programs, Gallo and Lewis (2012) find that programs housed in liberal agencies score systematically lower than moderate agencies, which in turn score lower than programs in conservative agencies. This remains true even when program type and other controls are included. In a narrower context that controls for task, Thomas and Fumia (2011) examine the PART scores for environmental programs. They include a dummy variable for those housed in Environmental Protection Agency programs, since this agency is considered a liberal agency (Clinton & Lewis 2008) and find that environmental programs housed in the EPA receive lower scores relative to environmental program elsewhere in the federal government. A test of the liberal nature of a policy, as opposed to the home agency, is offered by Greitens and Joaquin (2010), who find that programs with redistributive purposes receive lower scores than others.

There is also evidence that ideology mattered to how PART influenced budget. As noted above, there is only strong evidence that PART mattered to budget allocations in the President's proposed budget. Here, Gilmour and Lewis (2006) find that this relationship is explained by programs created under unified Democratic control. Put more simply, liberal programs were exposed to the risks of performance budgeting, while conservative programs were not. They also found that large programs were largely impervious to PART scores for budgeting purposes.

Given these patterns, we might assume that ideological differences might extend to agency behavior in response to PART. There is evidence to support this assumption. Lavertu, Lewis and Moynihan (2012) examine the level of effort reported by managers in responding to PART. They find that respondents in liberal agencies report a greater level of effort than those in

conservative agencies, using equivalent data from different sources. For example, 67% of managers in what Clinton and Lewis (2008) classify as liberal agencies say that to a "moderate," "great," or "very great" extent PART "imposed a significant burden on management resources", compared to 56% of respondents in conservative agencies.

Other basic descriptive statistics support the notion that PART was more burdensome in liberal agencies. A greater percentage of employees report being involved in PART in liberal agencies, and liberal agencies had significantly higher number of programs PARTed. This might be simply because liberal agencies take on smaller programs, but here qualitative work was instructive, suggesting that one mechanism by which ideological differences might have generated greater effort is via the definition of what constituted a program for purposes of PART. Decisions about what a program was were not set in stone, but determined by the OMB in consultation with the political appointees who led agencies (Moynihan, 2008). This created an opportunity for agency leadership who wanted to closely monitor agency activities to use PART to do so by defining programs very narrowly. A case in point is that the liberal Department of Education had more PART programs than the conservative Department of Defense, despite having only one-tenth of the budget (Joyce, 2011). Some PART programs in Education were as small as one million dollars annual budget (Gilmour, 2006). Education also scored very poorly in the PART assessments. The explanation as to why helps to illustrate how political processes might work. "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 programs, and see the PART process as a means of shining a light on those deficiencies" (Gilmour, 2006, 16). The differences between managers in liberal and conservative agencies in terms of their level of effort with PART hold even when controls are included, including individual partisan beliefs, belief that PART had budgetary effects, and the number of PART reviews conducted in an agency. This suggests that we retain an incomplete understanding for how ideology matters to the ways in which reforms generate burdens in agencies.

If effort is one type of agency response that PART created, another is whether managers actually use the data. Here, again, there is evidence that political ideology matters. Lavertu and Moynihan (2012) find that managers in conservative agencies who are involved in PART report

higher level of performance information use relative to peers not involved in PART. By contrast, managers in liberal agencies involved in PART reported no different use of performance information than managers not involved in PART. In short, while PART appeared to increase performance information use among managers in conservative agencies, it had no such effects in liberal agencies. To better discern if this was primarily a function of liberal agencies having different characteristics that made performance management more difficult, the authors compared the responses of those involved and not involved in PART on a series of items that dealt with perceived difficulty in measuring performance, conflict between stakeholders, and difficulty discerning causality. They found that managers in liberal agencies who were involved in PART were likely to regard these items as being significant impediments to performance management compared both to managers in liberal agencies not involved in PART, and all types of managers in other agencies. In other words, managers in liberal and non-liberal agencies tend to rate hindrances to performance management in similar ways, except for managers in liberal agencies exposed to PART. This finding provides more support for the notion that the PART process itself was experienced as more onerous for managers in liberal agencies.

Cumulatively, the findings on PART provide compelling evidence that both the implementation and response to administrative reforms, even ones framed as politically neutral, will be shaped by political ideology. The most positive assessment we can make is that agencies who share the ideological preference of the executive will be more receptive to reforms that he proposes – after all, we see managers in conservative agencies see an increase in performance information use, not a reduction in performance information use as in liberal agencies. Put another way, reforms are more likely to succeed if the political executive is preaching to the ideological choir and more likely to falter among ideological heretics. But the more negative possibility that arises from PART is that reforms are used as a mechanism to investigate and burden the heretics. Managers in liberal agencies exerted greater effort, faced more PART assessments, received lower scores, and found their budgets threatened by PART in a way that peers in conservative agencies did not.

Task Difficulty and Working the Ref: How Did Different Programs Fare Under PART?

A basic claim in the study of performance management is that some programs are inherently more or less suited to performance management techniques. This argument is often derived from James Q. Wilson's (1989) observation that some programs have outputs and outcomes that are more or less easy to observe (Gueorguieva et al. 2009; Radin 2006). Further, programs with multiple goals, and hard-to-measure goals will be less able to justify themselves if performance measures become the coin of the realm. Gueorguieva et al. (2009) illustrate this point by comparing the PART assessments of seven programs, noting that PART makes no distinction between programs where performance is easier or harder to measure.

The design of PART attempted to be flexible enough to deal with different program typologies, and different versions of PART were developed for direct federal, regulatory programs, research and development (R&D) programs, block/formula grants, credit programs, competitive grant programs, capital asset and service acquisition programs. The separation of these program types could be seen as a practical hypothesis that a governmentwide performance management tool could be tweaked to meet the broad differences between programs, treating each fairly, and allowing comparison across type of programs. It is worth noting that the different questionnaires were modeled on the direct federal questionnaire, with some additional questions for the other program types, and so the different forms of PART were much more similar than dissimilar.

The rival hypothesis, best articulated by Radin (2006) is that such tools are inevitably one-size-fits-all, and will be unable to reflect the nuance of different program types. For example, the basic intent of block grants is to allow states to have some measure of discretion in the goals they pursue, even as PART is apt to punish programs that cannot report a single set of goals. R&D programs seek to generate innovation by funding disparate sets of high-risk projects that may take years to generate positive impacts (if they do so at all), while a mechanism such as PART frowns upon failure and demands tangible outcomes in the short-term.

The study of PART helped to provide clearer evidence on how different types of programs fared under a performance management regime. It was relatively easy to compare PART scores across the typologies mentioned above, to see if some programs did systematically better or worse, and this comparison is most systematically captured by Gallo and Lewis (2012). The findings are mixed. On one hand, most types of programs do not score significantly

differently from other types of programs, suggesting that by and large, PART did not penalize most programs because of their distinct characteristics.

But some programs did do better than others. As predicted by Radin, block grant programs did relatively worse compared to other programs (Gallo and Lewis 2012). But contrary to expectations, R&D programs scored relatively higher. Gallo and Lewis (2012) find this effect across different PART programs, while Heinrich (2012) finds it within Health and Human Services programs. Lavertu, Lewis and Moynihan (2012) find that R&D programs are associated with lower effort in completing PART (while regulatory programs are associated with higher effort). What accounts for this surprising finding? The causal mechanism is not clear, but we can speculate based on some circumstantial evidence. This evidence suggests what we might call the "working the ref" hypothesis. Working the referee is a tactic that players, coaches and fans use in sports. By complaining loudly about the unfairness of a particular decision, the player may hope to generate sympathetic treatment for later decisions. In the case of performance regimes, this implies a community of practitioners and stakeholders arguing about the unfairness of methods of evaluation to their particular program or type of program.

With PART, the R&D community articulated their concerns about the unfairness that PART placed on outcomes, and how it did not hold for their type of programs (Gilmour 2006). As early as 2002, the National Academies hosted a daylong workshop where members of R&D agencies expressed concern to OMB officials about the idea of evaluating their programs. The National Academies Committee on Science, Engineering & Public Policy (COSEPUP) was able to draw upon arguments it had developed in critiquing GPRA (COSEPUP 2001) about the need to for OMB to consider broader criteria for science, including relevance and quality that could be best determined by peer review. OMB largely accepted these arguments, and sought to adjust PART accordingly. Indeed, OMB's guidance on PART told budget examiners to apply a different standard to such programs: "agencies should define appropriate output and outcome measures for all R&D programs, but agencies should not expect fundamental basic research to be able to identify outcomes and measure performance in the same way that applied research or development are able to. Highlighting the results of basic research is important, but it should not come at the expense of risk-taking and innovation" (OMB 2007, 76). Despite this, the R&D community remained critical. A report delivered late in the Bush administration continued to

reiterate similar arguments, saying that outcome-based measures are "neither achievable or valid" for R&D programs and **"**have grown out of inappropriate OMB requirements for outcome-based efficiency metrics" (COSEPUP, 2008, 59), and instead suggesting that performance of programs was best reviewed by peers, not by budget examiners. In case studies implementing PART, Frederickson and Frederickson (2006) report that R&D agencies had some success in convincing OMB budget examiners to relax their definition of performance.

The "working the ref" hypothesis fits within the broader notion of performance as an interactive dialogue I have written about elsewhere (Moynihan, 2008), portraying basic notions of performance as something that is negotiated between actors in the policy process. Even so, the relative success of the R&D community is striking, and not emulated by other communities. Any group could lobby for better scores, but why was this group so successful? Another speculative (though testable) claim is that it may have had to do with the high-status of this community, i.e. that the success of working the ref is contingent on who is doing it. Some groups are better able than others to press their claims because of their positive image (Schneider and Ingram, 1993). The R&D community is generally characterized by accomplished scientists who can argue that peer-review offers a functional form of performance assessment superior to tools such as PART. In reviewing the success of the National Institutes of Health (NIH) under GPRA and PART, Frederickson and Frederickson note "it is difficult to underestimate the importance of the status or prestige of NIH third parties" (2006, 112). It is possible that such a group was better able to press its claims than others.

There is other circumstantial evidence for the working the ref hypothesis in other areas. For example if an agency was unhappy with the score allocated to a program, they could request a second review at a later point in time, asking the ref to revisit the prior decision, theoretically on the basis of improved evidence. Gilmour (2006) undertook a qualitative study of a selection programs that were reviewed a second time, and scored higher. What did agencies do to improve their scores? They did not fundamentally change their programs or improve actual performance. Instead, they became more skilled at completing the PART assessment, at understanding what the OMB wanted, and satisfying those desires. Agencies simply got better at presenting their claims via PART. More generally, PART scores improved consistently over the course of the Bush administration (Joyce, 2011, 360), perhaps a function this greater skill of agencies in

completing the assessment, though perhaps also a function of the desire by the White House to show that performance was improving.

PART as a Dependent Variable: Factors Associated with Program Performance

Some studies have used PART as an indicator of program performance. The appeal of using PART for such a purpose is understandable. Studying public sector performance is inherently difficult because of data limitations. Indeed, the dominance of schools-research on schools in this area owes much to the fact that there are widely-agreed upon measures of educational performance (test-scores) in way that is not mirrored by other public functions. Tin general, there are also no definitive measures of performance across functions that allow us to judge the relative effectiveness of programs or organizations pursuing different tasks. Here, researchers often rely on self-reports by agency employees. PART, by developing a comparable single set of performance grades for very different types of programs, provides researchers with a rare opportunity to test theories of effectiveness using a third-party and ostensibly neutral assessment of performance.

Given some of the biases associated with PART scores discussed above, the claim that PART accurately captures performance should not be accepted uncritically. At the same time, even if the scores do include some degree of bias, this does not mean they are uncorrelated with some underlying concept of performance. Studies can seek to control for biases by including controls for the observed sources of biases (e.g. measures of political ideology and program type, in Gallo & Lewis 2012), or by studying PART within a functional area (e.g. Heinrich 2012). In the longer-run, it would be desirable to correlate PART scores to other measures of organizational performance. Its also worth noting that a good deal of published research has used measures similar to PART to summarize performance, most notably the simple five-point measure of UK local government performance, the Comprehensive Performance Assessment, even as Bertelli and John (2010) find evidence of partisan bias in these assessments similar to those unearthed for PART.

It is also possible to use data within PART, rather than the scores themselves, to construct measures of performance. For example, Jung (2012a) uses proportion of targets actually

achieved (reported in the Program Performance Measure section of PART), and aggregates these to develop a measure of agency (rather than program) performance. Jung uses the measure to demonstrate a curvilinear relationship between organizational size and performance – both very large and very small organizations had lower levels of target achievement. Jung's alternate measure of performance appears to reduce the potential for bias observed in the PART scores. On the other hand, OMB evaluators and agency staff play a role in selecting what target measures are included in PART (Moynihan 2008), introducing a potential for bias that has not been examined in the same way as it has for PART scores (e.g., Gallo and Lewis 2012).

Goal Ambiguity

Goal ambiguity has been associated with lower organizational performance (Chun and Rainey 2005). PART data has been used to expand our knowledge of how the relationship between the two variables works by providing new measures of both goal ambiguity and performance.

Jung (2012b) used PART data to construct measures of different types of goal ambiguity, which presumably could be replicated from similar type of performance systems. Target specification ambiguity, defined as "lack of clarity in deciding on the quantity and/or quality of work toward the achievement of a program's performance goals" (Jung 2012b, 681), was measured as the proportion of explicit goals in PART not accompanied with specific targets. Time-specification goal ambiguity (also referred to as timeline ambiguity, Jung 2012a) is "the lack of clarity in deciding on the distinction between annual and long-term goals in each program" (Jung 2012b, 684), and is measured as the proportion of PART goals that are not clearly labeled as annual or long-term. Finally, evaluative goal ambiguity – the "interpretive leeway that a statement of organizational goals allows in evaluating the progress toward the achievement of the mission" (Chun and Rainey 2005, 533) – is measured as the ratio of outcomes measures to output measures for PART.

Target specification ambiguity is negatively related to PART scores based on a wide sample of PART programs (Rainey and Jung 2009) and a study of environmental programs

(Thomas and Fumia 2011). Jung (2012a) finds that target-specification ambiguity and time-specification ambiguity are negatively associated with agency performance using the proportion-of-targets-achieved indicator of performance. Heinrich (2012) offers some overlapping findings. Looking at the quality of evidence of Health and Human Service programs, she finds that the use of long-term measures, baseline measures or targets and independent evaluations, are generally associated with higher PART scores. She also notes that while program assessments show a preference for presenting qualitative data, programs that relied upon quantitative data tended to enjoy higher PART scores.

Jung (2012a) also finds that evaluative goal ambiguity is associated with lower proportion of PART targets achieved. Heinrich's (2012) more limited sample of Health and Human Service programs does not reveal a significant correlation between evaluative goal ambiguity and PART scores, but does find that this type of ambiguity is associated with lower funding by Congressional appropriators, a relatively rare instance of a systematic association between performance data and legislative decisions. Meanwhile, Thomas and Fumia (2011) find that a related type of goal ambiguity, accuracy in labeling measures as outcomes, is associated with higher PART scores.

Clinton et al. (2012) use PART to develop a measure of policy certainty that is conceptually similar to goal ambiguity: the higher proportion of programs with valid PART performance measures, the higher is policy certainty. They find that this measure is negatively related to agency autonomy (as measured by length of agency-relevant public law). When Congress has more valid performance measures at its disposal, it appears to delegate less to agencies.

Political Appointees

PART has also shed light on ways that political leadership mattered to performance. Using early PART scores as measures of performance Lewis (2008) finds that programs run by political appointees generally scored lower than career managers. The plausible interpretation, offered by Lewis, is that career managers were more skilled, and better able to generate performance. It may also be the case that career managers cared more about protecting their

program and were willing to contribute effort to scoring more highly on the PART assessment. Gallo and Lewis (2012) retest the role of political appointees on PART performance using all PART assessments, though limiting their analysis only to programs where survey data suggests managers believe that PART scores were meaningful in discerning real performance differences.

With this-better validated measure of performance, and controlling for agency type and ideology, Gallo and Lewis again find that political appointees score more poorly on PART relative to career managers. Furthermore, they find that political appointees who were selected on the basis of campaign experience score even worse than other appointees. The results offer perhaps the most persuasive evidence of the performance costs associated with the use of political appointees. In both studies, length of term leading the bureau is associated with higher PART scores, pointing to the particular advantage of leadership stability for performance.

Discussion: What Have We Learned From PART?

Table 1 summarizes some of the key findings from the study of PART, organized by a series of dependent variables. Rather than simply restate these findings, let us consider what research questions the study PART raises. If PART never took place, would the trajectory of future public management research be much different?

Perhaps the most importance substantive lesson from the study of PART is the need to explicitly model political variables in studying administrative outcomes. More persuasively than research on any comparable reform or presidential administration, studies relying on PART pointed to the affect of political appointees on performance, and the role of political ideology in shaping how performance is defined, how a reform is implemented, and how it is received. These studies show that that political appointees are associated with lower performance outcomes, and that agency political ideology matters a good deal. Programs in liberal agencies received systematically lower PART scores, and were more subject to budget risk because of those scores. Managers working in liberal agencies were more likely to regard PART as burdensome, and less likely to be spurred to use performance data as a result. The prominence of these political variables can be at least partly attributed to marked presence of political scientists working on PART, more fluent in political concepts, and more willing to assume that political ideology

Table 1: Selected Summary of Findings from Quantitative Research on PART		
Concept (measure)	Key Predictors	Explanation
Performance (PART scores)	Agency political ideology	Programs in conservative agencies received higher PART scores than those in liberal agencies (Gallo & Lewis 2012; Thomas & Fumia 2011)
	Political appointees	Political appointees, especially campaign staff, lower performance (Lewis 2008; Gallo & Lewis 2012)
	Program type	R&D programs received systematically higher scores; block grants and redistributive programs received lower scores (Gallo & Lewis 2012; Greitens and Joaquin 2010)
	Goal ambiguity	Target-specification goal ambiguity associated with lower performance (Jung & Rainey 2009; Thomas & Fumia 2011); lower PART scores if programs failed to provide quantitative evidence, long-term or baseline measures, or evaluations (Heinrich 2012)
Performance (proportion of PART targets achieved)	Goal ambiguity	Target-specification, timeline and evaluation goal ambiguity (Jung 2012a) associated with lower performance
	Organizational size	Curvilinear effect; small and large organizations perform less well (Jung 2012a)
Reform implementation (performance information use by managers)	Involvement with PART * agency political ideology	Involvement with PART does not increase purposeful performance information use on aggregate, but does increase passive use (Moynihan & Lavertu 2012) and purposeful performance information use for managers in conservative agencies (Lavertu & Moynihan, 2012)
Performance budgeting (President's budget proposal)	PART program scores	Modest but significant correlation between PART scores and program budget changes (Gilmour & Lewis 2006)
	Agency political ideology	Programs created under Democratic control more subject to risk of budget losses due to PART (Gilmour & Lewis 2006); large programs protected from budget risks because of PART
Performance budgeting (legislative budget)	PART program scores	No connection between PART scores and changes in appropriations for Health and Human Service programs (Heinrich 2012)
Performance budgeting (legislative budget)	Goal ambiguity	Evaluation goal ambiguity reduces appropriations for Health and Human Service programs (Heinrich 2012)
Performance budgeting (managerial budget execution)	Involvement with PART * agency political ideology	Involvement with PART does not increase performance information use for managerial resource allocation on aggregate, but does increase use for those in ideologically moderate and conservative agencies (Moynihan & Lavertu 2012; Lavertu & Moynihan 2012)
Administrative burden (time and effort devoted to completing PART)	Agency political ideology	Respondents in liberal agencies report higher level of burden in completing PART than those in conservative agencies (Lavertu, Lewis & Moynihan 2012)
	Program type	Less burdensome to complete PART for R&D programs; regulatory programs more burdensome (Lavertu, Lewis & Moynihan 2012)
Agency autonomy (length of statues)	Policy certainty	Higher policy certainty (proportion of PART programs with valid performance measures) results in lower discretion (Clinton et al. 2012)

matters even to good-government reforms. The findings challenge those of us who come from a traditionally public administration background to take politics seriously enough to understand how it shapes the workings of administrative life.

In some respects the contribution of PART is methodological, offering new techniques to address existing questions. This is true of questions of performance, which allowed for tests of established variables such as size and goal ambiguity. For goal ambiguity, studies using PART data not only connected it to lower performance, but also to greater autonomy and lower resource allocation. The ability to pursue these questions and the examples of how to do so has an effect that goes beyond methodology, shifting the winds that direct new research in one area or another, and inviting similar studies for other well-established variables, such as employee motivation.

The study of PART is also likely to have an impact in studying how performance management reforms fare. Table 1 summarizes a relatively weak case for the notion that PART achieved its goals of fostering performance budgeting or performance information use. While there is relationship between PART scores and the President's proposed budget, there is not evidence of such an effect on legislative decisions or budget execution by managers. In these studies, PART research offered performance information use – either self-reported or inferred from decision outcomes – as a benchmark for understanding the impact of performance management reforms. This variable was already the subject of growing attention (Moynihan & Pandey 2010), but studies involving PART looked for means to isolate the effects of a particular reform on this form of behavior. Though not always easy to achieve, the study of PART evokes a logic of quasi-experimental design as a standard for studying performance management, a significant difference from the traditional approach.

The study of PART provides a sense of the possibility of balancing qualitative and quantitative work in performance management. Table 1 is incomplete because it focuses on quantitative work. This is partly because of the difficulty of summarizing the more complex concepts that feature in qualitative work. But in a wide variety of areas, we see overlap between qualitative and quantitative assessments of PART, e.g. on the burdens imposed by PART on agency staff, or on the importance of leadership commitment. Quantitative analysis has relied upon qualitatively-based insights on the workings of PART to develop reasonable hypotheses, establish the most plausible direction of causality with cross-sectional data, and explain results,

and to avoid the problem of armchair theorizing. Qualitative work has also provided a better sense of the mechanics of PART and the sometimes untidy process of social construction that gave rise to the contents of PART scores, reminding us of the adage popularized by the former head of the Bank of England, Sir Josiah Stamp (1929, p.258-259): "The government are very keen on amassing statistics. They collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the *chowky dar* (village watchman in India), who just puts down what he damn pleases." Qualitative work gave us a glimpse and understanding of *how* PART emerged, who the village watchmen were and why they acted. But qualitative work did not always offer a consistent set of claims, and in some cases, quantitative research resolved issues raised by qualitative work, such as how well different program types fared under PART. Quantitative findings also sometimes go beyond mere tests of hypotheses generated from qualitative work. The body of work on the role of political ideology does more than resolve a debate on this topic, but provides a well-developed theoretical basis for understanding the effects of ideology on reform more generally.

At the same time, there are areas where quantitative work offers little. These are primarily to do with normative arguments about the purpose of reforms. A good example is a set of work that have argued that performance regimes serve to displace attention to democratic values– such as equity (Radin 2006), transparency (Piotrowski and Rosenbloom 2002), constitutional values (Rosenbloom 2007), and citizenship (Wichowsky and Moynihan 2008) – or other important tasks, such as policy-making (White 2012). Empirical work can help to illustrate this point, or test the degree of goal-displacement. But the key argument here is a normative one about how management reforms should reflect democratic values. Such arguments deserve a domain in public administration research, though should not be our sole focus.

The point is not that one approach is better than the other. Rather, the article notes prior approaches to assessing the impact of reforms such as PART have relied almost exclusively on a combination of qualitative work and normative theorizing, and that balancing such with the greater generalizbaility of large-N work allows for more persuasive tests of causal claims. Such work will be strongest if it can apply a mixed methods approach. Failing that, a basic willingness

to draw insights from multiple methodological approaches helps to avoid a blind-spot of any one approach.

Conclusion

As we conclude, one might ponder why PART was so well-studied relative to similar reforms. There are a number of plausible reasons. First, general improvements in methodological training meant that there were simply more well-trained people able to study the topic (with a major debt to political scientists who worked on the topic). A second reason is the transparency of the tool itself. To the everlasting credit of the Bush administration the application of the PART tool was made unusually transparent. The questions asked, the answers given, the judgments based to those answers were all recorded and made publicly available on the internet. The PART tool provided desirable measures to study with its ineffective-effective scale and accompanying budget recommendations, and academics would also code the contents of the PART assessments for additional measures. A third factor was the development of data that could be related to PART. The GAO collected invaluable survey data that it shared with academics. Academics developed supplementary data, including surveys that asked directly about PART, and categorizations of agency ideology, and public statements about PART.

A final point raised by the study of PART has to do with the relationship between academia and the world of practice. There is a trade-off between practitioner needs and the ability of academics to provide timely insights on reform. Practitioners want insights on the new reform they are currently implementing. But the first wave of research on PART did not appear until 2006-2008, and comprehensive analyses are just now emerging. Academics had little to report directly on PART until PART was fully in place and nearing its conclusion. One may blame this partly on the relaxed schedules of academic production, but it also takes time to develop meaningful data on a phenomena. The process of evaluating the effects of a reform is inherently retrospective. This may disappoint practitioners who have moved on to something else, and creates pressures on scholars to equate relevancy with studying contemporary reforms. But the scholarly community provides a greater service to practice by using its skills to evaluate reforms thoroughly. Commentary on the present reform will be most insightful only if it is

informed by a deep understanding of the past, which, as the Book of Ecclesiastes notes, has a tendency to repeat itself: "What has been will be again, what has been done will be done again; there is nothing new under the sun." Our ability to identify underlying variables in reform and governance and how they mattered in the past will aid us best if we are to influence the future.

References

Askim, J, Johnsen A., & Christophersen K.A. (2008). Factors Behind Organizational Learning from Benchmarking: Experiences from Norwegian Municipal Benchmarking Networks. *Journal of Public Administration Research and Theory 18*, 297-320.

Bertelli, A.M. & John, P. (2010). Government Checking Government: How Performance Measures Expand Distributive Politics. *The Journal of Politics 72*, 545-558.

Chun, Y.H., & H.G. Rainey. (2005). Goal Ambiguity and Organizational Performance in U.S. Federal Agencies. *Journal of Public Administration Research and Theory 15*, 529-557.

Clinton, J.D. & Lewis, D.E. (2008). Expert Opinion, Agency Characteristics, and Agency Preferences *Political Analysis 16*, 3-20.

Clinton, J.D. Bertelli, A., Grosse, C. & Lewis, D.E. (2012). Separated Powers in the United States: The Ideology of Agencies, Presidents and Congress. *American Journal of Political Science 56*, 341-354.

Committee on Science, Engineering, and Public Policy (COSEPUP). (2001). *Implementing the Government Performance and Results Act for Research.* http://books.nap.edu/catalog.php?record_id=10106#toc

Committee on Science, Engineering, and Public Policy (COSEPUP). (2008). *Evaluating Research Efficiency in the U.S. Environmental Protection Agency* http://books.nap.edu/catalog.php?record_id=12150#toc

Dull, M. (2006). Why PART? The Institutional Politics of Presidential Budget Reform. *Journal of Public Administration Research and Theory 16*, 187–215.

Durant, R. (2008). Sharpening a Knife Cleverly: Organizational Change, Policy Paradox, and the "Weaponizing" of Administrative Reforms. *Public Administration Review 68*, 282-294.

Frederickson, D.G., & Frederickson, H.G. (2006). *Measuring the Performance of the Hollow State*. Washington D.C.: Georgetown University Press.

Frisco, V. & Stalebrink, O.J. (2008). Congressional Use of the Program Assessment Rating Tool. *Public Budgeting and Finance 28*, 1-19.

Gallo, N. & Lewis, D.E. (2012). The Consequences of Presidential Patronage for Federal Agency Performance *Journal of Public Administration Research and Theory, 22,* 195-217.

Gilmour, J.B. (2006). *Implementing OMB's Program Assessment Rating Tool (PART): Meeting the challenges of integrating budget and performance*. Washington, DC: IBM Center for the Business of Government.

Gilmour, J.B., & Lewis, D.E. (2006). Assessing performance budgeting at OMB: The influence of politics, performance, and program size. *Journal of Public Administration Research and Theory* 16:169-86.

Golden, M. (2000). *What motivates bureaucrats?* New York: Columbia University Press.

Greitens, T.J. & Joaquin, M.E. (2010). Policy Typology and Performance Measurement: Results from the Program Assessment Rating Tool (PART). *Public Performance & Management Review 33*, 555–70

Gueorguieva, V., Accius, J., Apaza, C., Bennett, L., Brownley, C., Cronin, S, & Preechyanud, P. (2009). The Program Assessment Rating Tool and the Government Performance and Results Act: Evaluating Conflicts and Disconnections. *American Review of Public Administration 39*, 225–245.

Heinrich, C.J. (2010). Third-Party Governance under No Child Left Behind: Accountability and Performance Management Challenges, *Journal of Public Administration Research and Theory 20*, i59-i80.

Heinrich, C.J (2012). How Credible is the Evidence, and Does It Matter? An Analysis of the Program Assessment Rating Tool. *Public Administration Review 72*, 123-134.

James, O. (2011). Performance Measures and Democracy: Information Effects on Citizens in Field and Laboratory Experiments. *Journal of Public Administration Research and Theory 21*, 399-418.

Joyce, P.G. (2011). The Obama administration and PBB: Building on the legacy of federal performance-informed budgeting. *Public Administration Review 71*, 356–67.

Jung, C.S. (2012a). Navigating a Rough Terrain of Public Management: Examining the Relationship between Organizational Size and Effectiveness. *Journal of Public Administration Research and Theory* doi:10.1093/jopart/mus040

Jung, C.S. (2012b). Developing and Validating New Concepts and Measures of Program Goal Ambiguity in the U.S. Federal Government. *Administration & Society 44*, 675-701.

Lavertu, S., Lewis, D.E., & Moynihan, D.P. 2012. Administrative Reform, Ideology, and Bureaucratic Effort: Performance Management in the Bush Era. Paper presented at Association of Public Policy Analysis and Management annual meeting, November 8-11, Baltimore, Maryland.

Lavertu, S. & Moynihan, D.P. (2012). Agency Political Ideology and Reform Implementation: Performance Management in the Bush Administration. *Journal of Public Administration Research and Theory* doi:10.1093/jopart/mus026

Lewis, D.E. (2008). *The Politics of Presidential Appointments*. New York: Cambridge University Press.

Moynihan, D.P. (2008). *The Dynamics of Performance Management: Constructing Information and Reform.* Washington D.C.: Georgetown University Press.

Moynihan, D.P. & Ingraham, P.W. (2004). Integrative leadership in the public sector: A model of performance information use. *Administration & Society 36*, 427-453.

Moynihan, D.P., and Pandey, S.K. (2010). "The Big Question for Performance Management: Why do Managers Use Performance Information?" *Journal of Public Administration Research and Theory* 20, 849-866.

Moynihan, D.P. & Lavertu, S. (2012). Does Involvement in Performance Reforms Encourage Performance Information Use? Evaluating GPRA and PART. *Public Administration Review 7,* 592-602.

Moynihan, D.P., Wright, B.E., & S.K. Pandey. (2012). Setting the Table: How Transformational Leadership Fosters Performance Information Use. *Journal of Public Administration Research and Theory 22,* 143-164.

Moynihan, D.P. & Roberts, A.S. 2010. The triumph of loyalty over competence: The Bush administration and the exhaustion of the politicized presidency. *Public Administration Review 70,* 572-581.

Piotrowski, Suzanne J., and David Rosenbloom. 2002. Nonmission-based values in results oriented public management. *Public Administration Review* 62: 643-57.

Posner, P.L. & Fantone, D.L. (2007). Assessing Federal Program Performance: Observations on the U.S. Office of Management and Budget's Program Assessment Rating Tool and Its Use in the Budget Process. *Public Performance & Management Review 30,* 351-368.

Radin, B. (2006). *Challenging the Performance Movement: Accountability, Complexity and Democratic Values*. Washington, DC: Georgetown University Press.

Stamp, Josiah (1929). *Some Economic Factors in Modern Life*. London: P. S. King & Son

Rainey, H. & Jung, C.S. (2009). Program Types, Goal Ambiguity, and Performance in U.S. Federal Programs. Paper presented at the 2009 Public Management Research Conference, Columbus, Ohio, October 1-3.

Redburn, F.S. & Newcomer, K. (2008). Achieving Real Improvement in Program performance and Policy Outcomes: The Next Frontier. Washington D.C.: National Academy of Public Administration.

Rosenbloom, D.H. (2007). Reinventing administrative prescriptions: The case for democratic-constitutional impact statements and scorecards. *Public Administration Review 67,* 28-39.

Rudalevige, A. (2002). *Managing the President's Program: Presidential Leadership and Presidential Policy Formulation.* Princeton, NJ: Princeton University Press.

Soss, J., Fording, R., & Schram, S. (2011). The Organization of Discipline: From Performance Management to Perversity and Punishment. *Journal of Public Administration Research and Theory 21*, i203-232.

Schneider, A. & H. Ingram. (1993). Social Construction of Target Populations: Implications for Politics and Policy. *American Political Science Review 87*, 334-347.

Stalebrink, O.J. & Frisco, V. (2011). PART in Retrospect: An Examination of Legislator's Attitudes toward PART. *Public Budgeting and Finance 31*, 1-21.

Thomas, C. & Fumia, D. (2011). The Effect of Program Size and Goal Ambiguity on Performance: An Analysis of PART Assessments for 165 Environmental Programs. Paper presented at the National Public Management Research Conference, Syracuse University, June 2-4, 2011.

Thompson, J. (1999). Devising Administrative Reform that Works. *Public Administration Review 59*, 283-293.

U.S. Government Accountability Office (GAO). (2008). *Government Performance: Lessons Learned for the Next Administration on Using Performance Information to Improve Results. Washington, DC: U.S. Government Printing Office. GAO-08-1026T.*

US Office of Management and Budget (US OMB). 2001. *The president's management agenda*. Washington, DC: Government Printing Office.

U.S. Office of Management and Budget (OMB). (2007). Program Assessment Rating Tool Guidance 2007-02. http://stinet.dtic.mil/cgi-bin/GetTRDoc?AD=ADA471562&Location=U2&doc=GetTRDoc.pdf

Wichowsky, A., & Moynihan, D.P. (2008). Measuring How Administration Shapes Citizenship: A Policy Feedback Perspective on Performance Management. *Public Administration Review* 68: 908-20.

Wilson, J.Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It.* New York: Basic Books.