Uncovering the Circumstances of Performance Information Use: Findings from an Experiment

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

dmoynihan@lafollette.wisc.edu

November 2013
Uncovering the Circumstances of Performance Information Use:

Findings from an Experiment

Donald P. Moynihan
La Follette School of Public Affairs
Comments and feedback welcome: dmoynihan@lafollette.wisc.edu

Abstract

One of the most widespread trends in government in past decades has been the diffusion of quantitative measures of performance. This trend is motivated partly by the hope that policymakers will use this data. However, we know little about individual’s basic tendencies to incorporate and use performance data. This paper draws from interactive dialogue theory to argue that the connection between data and decisions will not be automatic, but will depend upon circumstances such as the nature of the data and how it is presented. Because observational studies are limited in their ability to identify such circumstances, this paper argues for an experimental approach that offers greater flexibility in designing theory-based treatments. A vignette experiment methodology approach is used, providing subjects with a variety of budget scenarios, varying the amount and type of information that they receive, and asking them to make budget decisions. The results provide evidence that conditions of goal ambiguity, expectancy disconfirmation, and advocacy alter the potential for performance data to influence resource decisions.
Introduction

One of the most widespread trends in government in past decades has been the diffusion of quantitative measures of performance, premised on the hope that decision-makers will use this data. But while there is a growing literature on self-reported performance information use for certain managerial tasks (e.g. Moynihan & Lavertu, 2012), we know little about people’s basic tendencies to incorporate and use performance data. This paper examines performance information use in the context of specific decisions in a budget environment.

The type of research method that scholars employ enables the pursuit of certain types of questions and precludes others. The vast majority of the literature on performance information use in the public sector relies upon observational studies, and this approach has enabled researchers to examine the effect of environmental, organizational, job and individual factors reported by a survey respondent (Moynihan & Pandey, 2010). In the more limited literature on performance information use in budgeting, observational studies have resulted in researchers assuming that more data, better data, or data that shows variation in performance outcomes will be correlated with aggregated self-reports of data use or actual patterns of budget change. Such studies are important because they address the underlying assumptions that motivate reforms. But they are limited in their ability to specify how the particular characteristics of performance data or the particular context in which they are presented matters. This limits the study and debate about performance budgeting to a relatively narrow set of claims, such as better performance measures result in more resources, or that more measures result in more budget changes. The narrowness of such claims, in turn, gives rise to relatively simplistic debates about whether performance budgeting “works” or not, based on aggregate correlations between the existence of data and some indicator of use. Such claims may be easy to refute, creating the
impression that performance data does not affect budget changes, and that performance budgeting initiatives are doomed to fail.

If we think of performance budgeting as a behavior where individuals use performance data in a decision process, we need better knowledge about how individuals respond to data in a variety of decision situations. This paper draws from a theoretical framework (the interactive dialogue model) that proposes that the relationship between performance data and resource decisions will depend a good deal on the particular characteristics of the data, and the context in which data is presented. This framework allows us to draw on theories in public management (goal ambiguity, expectancy disconfirmation, and advocacy) to identify plausible circumstances in which performance budgeting might occur. An experimental approach, though bringing its own limitations, allows for the design of more nuanced and theoretically-informed treatments of performance information use than possible with observational data. This paper relies upon vignette experiment methodology, providing subjects with a variety of budget scenarios, while varying the amount and type of information that they receive. By examining associations between treatments and variation in resources provided, we can infer whether these treatments alter resource allocation behavior. The results are mixed, but suggest that circumstances of goal ambiguity, advocacy and expectancy disconfirmation matter to whether there will be a connection between performance data and the tendency of individuals to provide resources for a program.

**Performance Information Use in Budgeting**

This symposium reflects a growing and increasingly sophisticated literature on performance information use. The purpose of this paper is to contribute look to one particular aspect of use –
resource allocation – using an experimental approach. A strict definition of performance budgeting is “a budget that explicitly links each increment in resources to an increment in outputs or other results” (Schick, 2003: 101). A simplistic version of performance budgeting reforms promises that performance should correlate with higher resource allocations, as decisionmakers look to continue to fund programs with positive outcomes. But with the exception of situations where budgets are set by a pre-determined performance formula, very few would claim that the strict definition of performance budgeting matches reality (Joyce, 2003). Even so, the ultimate goal of adding performance data to budgets is to change resource allocation behaviors. Observational research that examines the potential for performance data to alter resource allocation behaviors generally searches for broad correlations between the nature of performance data and the use of data to alter budgets. An indicator of performance data is created – such as the number of measures (Ho, 2011), or variation between the data in terms of quality (Heinrich, 2011) or measured performance (Gilmour & Lewis, 2006) – and then correlated with some measure of use (such as self-reported use, or actual change in the budget).

This research has been accumulated through a series of studies drawn from surveys, or observations of patterns of executive budget proposals and legislative appropriations. For example, surveys of state managers relying on self-reported data find that the use of performance information throughout the budget process and in other management decisions was associated with stronger perceived budget effects (Mekers & Willoughby, 2005), and that measurement quality and managerial skills at translating performance measures into decisions were positively related to using performance information for budgeting purposes (Lu, 2007). Similar lessons come from the local level, where Wang (2000) reports that variation in self-reported use of performance data in budgets is predicted by capacity in performance measurement. Ho (2011)
studies the effect of a performance reform in the City of Indianapolis, finding a correlation between actual number of performance measures reported by a department and the number of changes made to the departmental budget.

At the federal level, the Government Performance and Results Act of 1993 and the Bush-era Program Assessment Rating Tool (PART) made the integration of budgets and performance data an explicit goal. PART gave rise to a series of well-designed studies that took advantage of the fact that PART generated a single overall judgment on the performance of a program. For example, Gilmour and Lewis (2006) found that programs with higher PART scores tended to enjoy greater increases in the President’s budget proposal, though the size of the relationship was not large. Surveys of managers found that PART did not alter managerial use of performance information for resource allocation purposes on aggregate (Moynihan & Lavertu, 2012) but did increase use in ideologically conservative agencies (Lavertu & Moynihan, 2012). Other research finds that PART scores had little effect on Congressional budget decisions (Heinrich 2012; Frisco & Stalebrink 2008).

An Experimental Approach to Performance Budgeting

Effort to determine if performance budgeting “works” is valuable in responding to prescriptive models, but “performance budgeting is not a theory” (Kelly & Rivernbank, 2003: 6) designed to predict actual behavior. Over time, observational studies have identified some basic claims about how managerial competence with performance systems, or broad contextual factors such as political support or organizational culture, makes performance budgeting more likely. But in terms of understanding the behavioral question at the heart of performance budgeting – how performance data influences a decision – the observational approach relies on relatively
simple assumptions about how the volume of data, data quality, or variation in performance matters. Researchers are constrained in their ability to test and illustrate theories that offer more nuanced accounts of the relationship between performance data, the context in which it is presented, and a particular budget decision.

While experimental designs have shown value in understanding other aspects of the budget process (Bretscheiner, Straussman & Mullins, 1988; Thurmaier, 1992), they have largely not been used to understand the dynamics of performance budgeting. An experimental approach has a number of advantages for studying performance budgeting. Most obvious are the traditional benefits of the experimental technique. There is less potential for some unobserved variable to skew the results or for endogeneity. For example, studies that rely exclusively on self-reported data may be subject to not just data reliability concerns, but also to the risk that managerial perceptions on performance budgeting and variables such as data quality or managerial competence in performance budgeting may be endogenous. This problem may also occur in studies using non-survey based studies. For example, when budget examiners making budget recommendations are also involved in the creating of performance scores (as was the case with PART) it is possible that the budget recommendation for the program might influence the assessment of performance, rather than the other way around. Finally, with an experimental approach, the treatment is relatively unambiguous. This is not always the case with observational studies, when it may be unclear what the independent variables actually reflect. For example, Gilmour and Lewis (2006) note that much of the effect of PART scores on budget change were driven not by the results section of PART, but by program purpose section, which is a measure of the importance of the program’s goal rather than its actual performance. The finding of a correlation between data quality and budget increases (Heinrich, 2012) raises the question of
whether it is better data that are being rewarded, or whether the impact of some programs are simply easier to measure.

The second benefit of an experimental approach is theoretical rather than methodological. An experimental approach allows the researcher to specify tests that flow from theories that assume a more complex relationship between performance information and budgeting. Particular types of performance data or presentational contexts can be created and tested in the context of a particular decision, something not possible in observational studies that aggregate the characteristics of performance data in relatively crude ways (such as the number of measures). A recent paper demonstrates this point. Baekgaard and Nielsen (in press) performed a survey experiment on local elected officials in Denmark, providing information on how well schools in their region were performing. The survey asked officials to complete a hypothetical allocation of resources and about whether their local schools should be reformed. They found that, relative to a control group who received no performance information, elected officials were inclined to increase resources to low-performing schools, and were reluctant to reform high-performing schools. These patterns, argue the authors, align with the logic of blame avoidance theory: politicians feel pressure to do something to help poor performers in a salient policy area, but are happy to leave high performers alone. This logic is notably at odds with the notion that high performance attracts resources (though it may allow for autonomy) and suggests that the risk of political blame is one of the circumstances that affect how performance data relates to resource decisions. In the following sections, I seek to identify other such circumstances, circumstances largely unaddressed by the observational literature on performance budgeting. The next section provides the theoretical motivation for the tests employed in this paper.
The Circumstances of Performance Budgeting

The theoretical motivation of this paper is to understand the ways in which types of performance data and the context in which data are presented alters tendencies to allocate resources. Theory operates at two levels. At a broad level, the paper is motivated by the interactive-dialogue model, which suggests that the link between performance data and decisions will not be automatic, but depend a good deal on the circumstances (Moynihan, 2008). To identify those circumstances, additional theory – on goal ambiguity, expectancy disconfirmation and advocacy – is used to flesh out plausible hypotheses.

Joyce’s (2003) concept of performance-informed budgeting notes that there are many valid factors shape the creation of a budget, of which performance data is just one, making the potential for an automatic correlation between performance data and resources decisions implausible. The interactive dialogue model proposed by Moynihan (2008) builds upon this point, arguing further that there is substantial variation in what is broadly grouped together as performance data, and substantial variation in how it is presented that is likely to affect how it is interpreted and used. This further weakens the potential for an automatic connection between performance data and decisions.

The interactive dialogue model argues that while performance information is portrayed as something that is objective, consistently understood, and a basis for building consensus, this is often not the case. Instead, performance data is rarely comprehensive – for any moderately complex program, there are additional ways to measure performance. Performance data is also ambiguous – two individuals can examine the same performance data and come to different understandings of what it means. Performance data itself will fail to provide much contextual information (for example, why performance did or did not occur, the context of performance,
how implementation occurred, an understanding of outside influences on performance, how to choose which program measure is a priority, whether the outcome represents value for money) that individuals need to interpret its meaning, and may not even provide a clear pattern of improvement or decline. This increases the importance of how the presentation of data offers cues to the audience about the meaning of performance information. Aggregating different types of performance data together as a single unit of analysis misses these significant variations between the nature of data and the context of its presentation.

The central implication of the interactive dialogue model for research on performance budgeting is that we need to uncover the circumstances of performance information and the context of its presentation to learn how it will affect resource decisions. But the model falls short in specifying what those circumstances are (with the exception of the strong emphasis on the importance of advocacy, explained in greater detail below). The remainder of this section outlines three such circumstances, which are then tested in experiments: advocacy, goal ambiguity and expectancy disconfirmation.ii

Given the variety of experiments, it is helpful to summarize theoretical expectations in a single table.

*Insert Table 1 here*

*Advocacy*

The interactive dialogue model argues that advocates play a crucial role in not just selecting and disseminating performance information, but also in offering an interpretation of the meaning of that data (Moynihan, 2008). This argument is consistent with the central role that advocates, both bureaucrats and external interest groups, are given in broadly differing accounts of politics
The interactive dialogue model predicts that performance data will be selected, disseminated, and framed by advocates seeking to serve the interests of different governance institutions, political parties, and policy perspectives. They will offer arguments that convey their interpretation of information. But for such a strategic approach to work, it assumes that individuals are actually amenable to different policy narratives that will alter their interpretation of performance data.

Hood (2006) makes a similar point, noting that bureaucrats may “spin” performance information to make it appear more palatable. Spinning may take the form of claiming credit when performance scores look positive, or seeking to place blame on external factors when performance scores are declining. For an example of the latter behavior, Charbonneau and Bellavance (2012) demonstrate that Canadian municipalities with lower performance scores are more likely to report justifications for their performance than those that perform better. While the term “spin” evokes negative connotations, advocates may play a valuable informational role in offering an interpretation of the data, explaining what the data means, how relevant it is, why performance did or did not occur, and how it should affect the budget. In the experiments below, we examine whether the comments of advocates about performance data affect the resource allocation decision that subjects make.

Goal Ambiguity

Goal ambiguity “refers to the extent to which an organizational goal or set of goals allows leeway for interpretation, when the organizational goal represents the desired future state of the organization” (Chun & Rainey, 2005: 531). Goal ambiguity has become an overarching concept that includes a number of sub-concepts (Chun & Rainey, 2005; Jung, 2012), which have been
generally associated with lower organizational performance (e.g. Jung, 2013). The way in which such ambiguity affects budgets have not been widely examined. Here, we examine two ways in which goal ambiguity in performance measures may influence allocation decisions.

The first type of goal ambiguity is evaluative goal ambiguity – the “interpretive leeway that a statement of organizational goals allows in evaluating the progress toward the achievement of the mission” (Chun & Rainey 2005: 533). This has been measured in terms of the relationship between outcomes and outputs, with more outcomes associated with lower evaluative goal ambiguity (Jung, 2012).

Performance reforms, such as the Government Performance and Results Act and PART, share a similar preference for outcomes over outputs (Moynihan, 2008). Outcomes may also be seen as being inherently higher quality data, and more clearly communicating the inherent value of the program. This is one variable where observational studies provide insight on the relationship between this type of goal ambiguity and budget outcomes: Heinrich (2012) finds programs from the federal Department of Health and Human Services that relied more on outcomes received higher budgets. Relative to output measures, the use of outcome measures is therefore expected to have a greater impact on resource allocation.

The second aspect of ambiguity examined reflects priority goal ambiguity, defined as “the level of interpretive leeway in deciding on priorities among multiple goals” (Chun & Rainey, 2005: 535). This type of ambiguity captures the common tendency of public programs to feature multiple and potentially competing goals. For a performance system, this creates a very real potential that multiple measures will offer contradictory accounts of organizational performance. For example, studies of substance abuse programs and educational outcomes have observed that
the perceived performance of these programs often depends a good deal on which measures are selected (Heinrich & Fournier, 2004; Nicholson-Crotty, Theobold, & Nicholson-Crotty, 2006). At a psychological level, contradictory performance data created by priority goal ambiguity is expected to create a sense of dissonance, which subjects may resolve by ignoring the information that contradicts their views, or by altering their beliefs about the program. Chun and Rainey (2005) find some support for their hypothesis that priority goal ambiguity is associated with perceptions of lower performance. In an experiment below, we test if the addition of performance data showing conflicting results on a competing measure of performance affects budget support for a program.

*Expectation Disconfirmation*

Another potential circumstance that will affect the perception and response to performance data is the expectations that individuals have about what performance could or should be. This claim draws from the expectation disconfirmation model, which proposes that an individual’s sense of satisfaction with services is shaped by their expectations (Morgeson, 2013). While primarily based on private sector research on customer satisfaction, the expectation disconfirmation model has had limited application in the public sector, primarily to understand citizen satisfaction with public services.

A frequently evoked implication of this theory for public managers is the need to limit citizen expectations to avoid disappointment (van Ryzin, 2006; James, 2009; Poister & Thomas, 2011). How can governments shape expectations about performance? One of the primary tools at their disposal are performance targets. Targets communicate the expectations that governments set for themselves and invite others to hold them accountable for. Targets might alter expectations about performance about constitutes program success and failure. The same actual
underlying performance may be perceived as more or less impressive if accompanied by performance targets that are respectively modest or ambitious. If targets anchor expectations to a level that is not achieved in practice, this is likely to reduce support for the program, reflected in lower resource allocation. For example, two budget officials may examine the same record of performance for a program. However, if one of the officials also sees that the program had set targets above what it actually achieved, this is assumed to create a sense of disappointment, and weaken support for the program. The final experiment tests this hypothesis by providing the same performance scores to both the control and treatment group, but by also providing target data for the treatment group.

**Data and Method**

The analysis employs a basic version of vignette experiment methodology (Finch, 1987). Vignettes provide short descriptions of a context and choice for a respondent to deal with, controlling for the variety of potential influences, and isolating the effect of hypothesized causal factors in a systematic way. One hundred and forty graduate students at a large state university were recruited to complete the vignettes at the end of one of their class periods. The instructor of the class introduced the researcher who explained the survey. The researcher emphasized that participation was voluntary and anonymous, and that participation was not related to their class. Students were offered a $20 financial incentive to participate. Participants were provided a paper survey that they completed in the classroom, taking an average of 20 minutes to do so (though no time constraints were in place). The response rate among those offered the survey in class was 72%. Sixteen percent of students were from social work graduate program, 28 percent from business, 50 percent from public affairs, and 6 percent from other programs.
Each subject received the same basic introduction, a short scenario with the following text:

This survey provides you with a series of short descriptions about public sector programs. The examples are hypothetical, although the programs are based on real services. In this scenario, you are an employee in the County Budget Office. You are responsible for making a series of annual budget recommendations to the County Executive. You will be provided a variety of information, including a summary of the purpose of the program, previous year budgets in constant dollars, and performance data.

Using the information given to you in each description, please make a decision about how much the budget of each program you examine should change for the upcoming yearly budget. There is no “right” or “wrong” answer, and your responses do not have to add up to a specific total. However, the County Executive has declared that she expects the upcoming budget to increase by approximately 3% overall, although some programs might see larger budget increases, and some programs budgets might be reduced.

The survey then introduced a series of vignettes that presented a scenario for different public programs. Each vignette offered a one-paragraph description of the program, describing its function (provided below). All vignettes contained the program budget allocations for the previous three years in order to provide the subjects with a base budget that serves as a reference point for future allocations. The vignettes were based on real county government programs, but no actual government names were mentioned. A range of basic local functions were represented in the vignettes: water, parks, health, social services, job training, and policing. Subjects were asked to make a budget recommendation (in dollars) for the coming year, based on the information provided to them. The order of vignettes was randomized, in order to prevent responses to earlier vignettes systematically shape responses to later vignettes. For each vignette there was a roughly equal chance that the respondent could be assigned to the treatment or control (in practice, the assignment of treatment for the vignettes ranged from a minimum of 45% to a maximum of 52% of the responses).
All vignettes used in this analysis are provided in the results section. Each one includes a control and treatment. The treatment group generally received the same information as the control group, but also received some additional form of performance data, or was provided a characterization of the data (for example, the comments of an advocate). The goal is to observe whether those who received the treatment allocate resources in a significantly different way from the control group.

**Dependent Variable**

The dependent variable is the percentage budgetary change from the previous budget year that the subject makes for each vignette. While the majority of responses provided by subjects reflected incremental changes, they also included responses where the subject may have proposed cutting the program completely, or dramatically increasing it. The average response for the dependent variable was .009 (i.e., about a one percent increase in the proposed budget) and the standard deviation was .214. Responses that included significant budget changes violate the assumption of a normal distribution necessary for an independent samples t-test. The solution of converting the dependent variable to a log to generate a normal distribution is not desirable because zero, a frequent response among subjects, cannot be transformed and would be dropped from the sample. This paper therefore employs another statistical solution, the Wilcoxon Mann-Whitney rank-sum test, a non-parametric analogue to the independent samples t-test that does not assume a normal distribution. For the purposes of interpretation, it is important to note that the negative sign for z-scores for the rank-sum test indicates that the treatment group receives higher budget allocation, and vice versa, so that a hypnotized positive increase will be supported when the rank-sum score is negative and significant.⁴
An experimental approach provides an ideal setting for focusing on the individual, and offers high internal validity. The downside of an experimental approach is external validity. While the majority of respondents come from programs in public affairs and social work, and have some basic familiarity with the public setting, we cannot determine if actual budget officials would respond in the way the subjects do. This is an inherent limitation of the experimental approach that must be weighed against the benefits described above.

Some aspects of the experiment are intended to mirror actual budget scenarios, i.e. the provision of previous allocations, and the recommendation to stay within a three percent budget constraint. The actual one percent aggregate increase that subjects recommended was lower than this, and consistent with actual budget increases in the local county government at the time (which had a .1% increase in the 2009-10 time period).

**Experiments and Results**

The results of the model are provided in Table 2, and each of the individual experiments is explained below. All include the verbatim information provided both to the treatment and control groups, with the treatment labeled and put in italics to ease comprehension (though no such labeling occurred in the actual vignettes).

*Insert table 2 here*

*Experiments 1 & 2: Advocacy*

The first two vignettes test the role of advocates. Both the treatment and control group receive identical performance information about the number of arrests for a police program,
which is increasing over time. The treatment group receives a comment from a member of the organization that argues that the crime increases are due to recession and that more resources are needed.

1. The County Sheriffs Office is responsible for a wide variety of law enforcement activities. The Narcotics Task Force focuses on illegal drug use. A number of media reports in the last year have focused on an increase in drug-related crime, and have been critical of local law enforcement. Treatment: The County Sheriff acknowledges that drug use and related crime may be up, but says “We usually see this sort of thing spike when the economy is in recession. This means that the Narcotics Task Force needs more money to tackle this problem.”

<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narcotics Task Force funding</td>
<td>641,000</td>
<td>712,000</td>
<td>732,000</td>
</tr>
<tr>
<td>Number of arrests</td>
<td>78</td>
<td>72</td>
<td>83</td>
</tr>
</tbody>
</table>

For the next vignette subjects receive performance information about a parks program.

Again, both control and treatment group receive the same performance data. In addition, both groups also receive a comment explaining the data from an advocate of the organization. The treatment group also receives a critical comment characterizing the data as consistent with interest groups politics.

2. The Division of Parks and Recreation offers access to a variety of parks in the county. Most parks feature at least one full-time employee-staffed welcome center as well as a number of part-time-staffed recreation programs, such as Nature for Kids, which introduces children to the outdoors and physical activities. The Director of the Division points to the positive assessments that users of the park service provide. “People who come to the parks really enjoy it, and that shows in the data,” she says. Treatment: But the head of the local taxpayers association suggests that parks system is not a good use of public resources. “You see a relatively small number of people using the park system again and again. Of course they give the system good ratings – they are getting a free service that the rest of us are paying for.”
<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Park funding</td>
<td>1,225,000</td>
<td>1,301,000</td>
<td>1,342,000</td>
</tr>
<tr>
<td>Number of visits to county parks per year</td>
<td>122,489</td>
<td>119,987</td>
<td>126,629</td>
</tr>
<tr>
<td>Customer satisfaction ratings with park programs (1 = very dissatisfied, 5 = very satisfied)</td>
<td>4.23</td>
<td>4.36</td>
<td>4.30</td>
</tr>
</tbody>
</table>

With experiment 1, the hypothesis is that the positive comments of the advocate will serve to provide contextual information that will make subjects more sympathetic to the program and result in higher funding. The results are consistent with this hypothesis (significant at a .05 level if a one-tailed test is used). For experiment 2, the expectation is that the comments of the critic of the program will reduce support for the program. This does not prove to be the case, and indeed the direction of the sign for the treatment group is positive rather than negative, perhaps suggesting that a perceived attack on the program generated sympathy for it, though the results are not significant.

**Experiments 3 & 4: Goal Ambiguity**

The next experiments examine the role of goal ambiguity. Experiment 3 six tests the effect of evaluative goal ambiguity. Here, the subjects are presented with detail about a preventative health care program. The control group receives information about an output – the number of clients who have received the funding. The treatment group does not receive the output data, but instead receives outcome data: estimated savings due to the program. Note that the trend in terms of change is similar for both the treatment and control – scores improve for the second year, and decline the third year. There is therefore no clear pattern of improvement or decline.

3. The Department of Health Services offers a program called Health Check, which is a preventive health check-up program made available for anyone under the age of 21 who is currently enrolled in Medicaid. Health Check provides a head-to-toe medical exam, immunizations, eye exam, lab tests, growth and development check, hearing check,
nutrition check, and teen pregnancy services. The goal of the program is to prevent the incidence of more serious and more expensive health situations.

<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health Check funding</td>
<td>232,000</td>
<td>244,000</td>
<td>269,000</td>
</tr>
<tr>
<td>Control: Clients treated</td>
<td>1232</td>
<td>1401</td>
<td>1325</td>
</tr>
<tr>
<td>Treatment: Estimated savings due to preventive care</td>
<td>383,000</td>
<td>402,000</td>
<td>389,000</td>
</tr>
</tbody>
</table>

The outcome measure here is particularly powerful, implying that additional dollars invested in the programs generates net savings (although the vignette does not specify whether these savings accrue to the government or are more general societal savings), and the treatment also suggests that the outcome is a direct function of the program. Subjects provided significantly higher amounts to the program when they received outcome data rather than when they received the output information.

Experiment 4 tests the effects of priority goal ambiguity to understand if the addition of conflicting results on another goal alters a resource allocation decision. A classic example of priority goal ambiguity is in job training programs, which have the basic goals of placing employees into some sort of work and finding quality work opportunities for trainees. These goals have been found to conflict, with programs often prioritizing placement at the expense of a high-paying position (Heinrich & Marschke, 2010). In the following vignette, subjects are introduced to such a program. Control groups receive information about the placement of employees, which generally shows a positive upward trajectory alongside a gradual increase in funding. The treatment group also receives information that clouds this picture of success. Even though placement rates are increasing, the treatment group can observe that the level of pay that graduates of the program receive is declining.

4. The Department of Social Services runs a job training program for citizens facing long-term unemployment. The training involves taking a six-week course designed to improve various skill-sets such as basic computer skills, and also offers and career counseling.
The goal of the program is to help to place individuals in secure employment that pays a living wage.

<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job training funding</td>
<td>344,000</td>
<td>363,000</td>
<td>366,000</td>
</tr>
<tr>
<td>Clients placed in jobs for 6 month period or longer</td>
<td>72%</td>
<td>74%</td>
<td>75%</td>
</tr>
<tr>
<td>Treatment: Average hourly pay of job placements</td>
<td>$8.15</td>
<td>$8.03</td>
<td>$7.25</td>
</tr>
</tbody>
</table>

The expectation is that the information showing declining pay rates will have a negative significant effect on the willingness of subjects to fund the program. This proves to be the case (p>0.05 for a one tailed test). The results of the experiment suggest that when multiple pieces of performance information create ambiguity rather than clarity about the meaning of performance, this will reduce the potential for performance data to influence budget allocations. If goal priority ambiguity is indeed a frequent characteristic of public programs, it provides a prominent reason why performance data has a limited influence on budget decisions.

*Experiment 5: Expectancy Disconfirmation*

The final vignette tests the potential for expectancy disconfirmation to affect resource allocation. Here, subjects were again given information on job placements. The pattern shows a gradual increase in number of clients placed at a rate consistent with a gradual increase in funding. The treatment group also receives information about targets that the program had. While there is an increase of actual placements over time, these placements never reach the target rate, and the gap between actual placements and targets grows each year. Both the treatment and control group receive the same information on performance, but the treatment group is also shown that performance has not kept up with targets.
5. The Mental Health section of the Department of Health Services funds a work placement program. The goal of the program is to help those with mental health problems find and maintain voluntary or paid employment. The program partners with local non-profits and a select number of for-profit providers that offer employment for the program’s clients. Program administrators seek to match available positions with client skills and interests, and offer regular consultation with employees and employers to assess progress.

<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work placement funding</td>
<td>79,000</td>
<td>81,000</td>
<td>84,000</td>
</tr>
<tr>
<td>Number of clients placed in positions for six months or more</td>
<td>Treatment: Target</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Actual</td>
<td>31</td>
<td>33</td>
</tr>
</tbody>
</table>

The addition of performance targets is expected to anchor expectations about what constitutes good performance above what was actually achieved, thereby reducing support for a program. In effect, the respondent is induced to experience heightened expectations and a sense of disappointment via the use of performance targets set above actual performance achieved. The results are consistent with this hypothesis: the treatment group who are provided the target information provide lower allocations to the program than those that just received the actual performance scores (see table 1).

**Discussion**

This section discusses the findings. First, the results are summarized. Then the aggregate theoretical insights of the experiments for performance budgeting are considered, before finally offering some more specific insights based on particular vignettes.

The findings offer mixed insights about how advocacy might matter. The results of experiment 1 show that the positive comments of a public employee about performance alters the budget allocation for a program, but the negative comments of an external critic of a program
(experiment 2) are not significant. The results of experiment 3 and 4 show that performance data with both goal priority ambiguity and evaluative goal ambiguity are associated with lower resource allocations. The results of experiment 5 suggest that disconfirmation of expectations created by performance targets reduces resource allocation among the test subjects.

At an aggregate level, what do these experiments tell us about performance information use for resource allocation? The results are consistent with the interactive dialogue model characterization of performance information use as a form of behavior that is circumstantial, with individuals using performance data to alter allocations depending on the nature of the data and how it is presented. This does not mean that each circumstance is so idiosyncratic that it is impossible to study systematically. Instead, it implies the need to use theory to identify and test plausible circumstances in which performance data will be used. This paper makes a start on this task, providing evidence that the presence of advocates, conditions of goal ambiguity, and the experience of expectancy disconfirmation alter the potential for performance data to be used for resource allocations. The treatments tested here are not esoteric or unusual conditions, but frequently occur in the budget process – advocates will argue about performance data, complex programs will experience goal ambiguity, and targets are used to set expectations.

In interpreting the results it is important to note that in only one of the experiments with significant findings (experiment 3) did the treatment include an actual change to the performance data presented. In the other experiments, both the control group and treatment group saw the same basic performance data. The treatments provided additional information that were hypothesized to alter the respondent’s perception of the value and meaning of the performance data, and thereby affect the respondents willingness to provide resources for the program, i.e. the context in which they understood the initial data was altered. This additional information was
another type of performance data (experiment 4), the comments of an advocate (experiment 1), and performance targets (experiment 5). The results therefore provide strong support for the assumption that the response to a piece of performance data is framed by the context in which it is presented. Contextual information does not change performance data itself, but can help the decisionmaker form a belief about the meaning of data, and what it implies for resource allocation.

The results also provides a series of relatively specific insights about the practice of performance budgeting, offering novel findings that advocacy, goal ambiguity and expectancy disconfirmation matter to how individuals relate performance data to resource allocation decisions.

The interactive dialogue model portrays policy actors in the policy world doing their best to select, present and interpret data consistent with their goals and beliefs (Moynihan 2008). Performance data is part of a narrative that policy actors tell to create impressions and beliefs. The findings on advocacy provide some evidence that this form of strategic behavior matters to successful persuasion of the subjects tested. Investing time and effort in shaping the interpretation of performance data is a rational act, at least in the experimental context presented here. The results further suggest that the hopes for performance budgeting may rest as much upon the communication skills of those selecting and disseminating the information as it does the availability and nature of the data itself.

While it might be tempting to argue that the lesson of experiments 3 and 4 are to reduce goal ambiguity, such changes should be made with caution. Some programs are inherently complex, with multiple and competing goals. If this complexity is reduced it could have perverse results. The vignette on priority goal ambiguity reflects this point. The control group who only received
information about job placements for the hypothetical job-training program were at a disadvantage relative to the treatment group who were also given information on job quality. Focusing on one goal at the expense of another may create a sense of greater clarity, but at the expense of excluding information that is relevant to decision outcomes, and possibly incentivizing displacement of the unmeasured goal (Heinrich & Marschke, 2010). The lesson is therefore not to artificially exclude measures that represent fundamental program goals. If this lesson seems straightforward, in practice it is not. Practitioners struggle with priority goal ambiguity, and may see a virtue in simplicity. For example, the 2010 GPRA Modernization Act instructs US federal agencies to focus attention on 2-8 high-priority goals. There is a logic for doing so, but it may result in the exclusion of relevant data in budget decisions.

The findings on evaluative goal ambiguity also have practical implications, and some risks. The results suggest that decision-makers are more apt care about outcomes than outputs in making resource decisions. One reason that outcomes are not more widely used is because of uncertainty about the connection between program effort and the outcome data. In such cases, a rush to outcomes might backfire. But if program managers can develop outcome data that is clearly linked to program efforts, the findings suggest it will increase attention to performance data in resource decisions. This may involve the creation of non-traditional type of performance data. In particular, the findings suggest the benefit of using estimates from evaluations as performance data. The outcome measure employed in the vignette (dollars saved due to health prevention activities) require evaluations to generate estimates of savings per visit that could be extrapolated to generate performance measures for other populations. In the vignette, the outcome is clearly labeled as an estimate. While evaluations are sometimes used to buttress performance measures, estimated impacts from evaluations are generally not converted to
performance indicators. The reluctance to establish estimated performance outcomes is understandable, because of the potential for manipulation, and especially in cases where the program target population may differ from the population upon which the original estimates were based. But there is some evidence from practice that the creation of such evaluation-based performance estimates is feasible and attractive. The Washington State Institute for Public Policy provides state legislators with return-on-investment estimates for a variety of public programs based on meta-analyses of evaluation studies. This approach is being promoted by influential think-tanks in public policy as a best practice to replicate (Pew Center for the States, 2012). The findings here show that such evaluation-based estimates matter to resource allocation in experimental conditions.

Previous studies of expectancy disconfirmation have cautioned governments that raising performance expectations will result in greater dissatisfaction among citizens if those expectations are not met (James, 2009; Poister & Thomas, 2011). The results here give rise to a parallel lesson for performance information use. The use of targets creates a risk for programs, for when those targets are not met – even if performance scores are improving – this negatively alters a willingness to provide funds for a program. This, in turn, creates an incentive for agencies to not publicly report targets. This strategy may not be possible. Instead agencies may seek to set modest targets they know they can achieve or limit effort so as to curb higher expectations (Hood, 2006).

Conclusion: A Research Agenda for an Experimental Approach to Performance Information Use
This paper uses an experimental approach to study how performance data is used. A number of treatments examined how different types of performance information and the manner in which it was presented can alter the perceptions of its audience. The variety of treatments provides some understanding of why discussions of performance information as a unitary phenomenon is a simplification that fails to appreciate the substantial differences between performance data and the different ways they might be persuasive. Much depends on the nature of the data and how it is presented to represent a coherent policy narrative.

The experimental approach employed offers an alternative to observational studies of performance information use, providing greater flexibility to create treatments that test conditions under which performance data matters. Given the lack of prior work employing an experimental approach, the findings presented here should be considered preliminary. Research following this approach can be pursued in three distinct ways.

First, additional work can further test the conditions identified in this paper (advocacy, goal ambiguity and expectancy disconfirmation). The vignettes (as with real life) provide specific contexts, and could be tested with different specifications to verify their robustness. Further testing could also explore what makes these conditions relevant. For example, the experiments using advocate comments suggest that advocates can influence how others interpret the data for resource allocation purposes, but more is needed to understand when they will be persuasive. We might speculate that the relative success of the advocacy treatment in experiment 1 (the sheriff who critiqued the performance data) was enabled because the position of the advocate or the content of his comments conveyed greater expertise than the advocate in experiment 2 (the representative of a taxpayer group). Additional experiments could test how the particular characteristics of advocacy – such the position of the advocate, the substantive content of their
comments, the tone in which they are expressed – matter to its ability to alter how performance data is understood. Similar variations in vignette design could further knowledge about the influence that goal ambiguity and expectancy disconfirmation have on budget decisions.

The second route for future research is to uncover other conditions that shape how performance data influences budget allocations. Given the wide range of potential theories that could be employed there is little to add, but to note that concept of performance information use implies the study of a particular behavior – the use of data to make a decision – implying that behavioral theories based on plausible psychological processes have the best opportunity to succeed.

Finally, future research could deal more persuasively with the external validity issue that limits the generalizability of the type of experiments pursued here. Experimental designs that survey actual budget officials (ranging from agency officials and central budget officials who prepare budgets, legislative staff and elected officials involved in passing budgets, and managers involved in budget execution) would offer a greater claim on external validity (e.g., Baekgaard & Nielsen, in press). Given the frequency of surveys of public managers, it would be relatively easy to introduce vignettes in these surveys. Studies that provide the same experiments to different groups of officials in the budget process would be deeply insightful, examining how individuals who occupy different roles respond to the same treatments. By extension, individual characteristics such as experience or beliefs could be tested as moderators to the treatments.

The three routes for future research sketched here are not mutually exclusive, but elements of each could be fruitfully combined. At a minimum, those interested in the study of performance
information use could look to an experimental approach as a mechanism to theorize and test when and how this form of behavior occurs.
<table>
<thead>
<tr>
<th>Theory</th>
<th>Key claim tested</th>
<th>Treatment (experiment #)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advocacy</td>
<td>Arguments of advocates will shape understanding of data and how it is weighed in decision process, altering support for a program</td>
<td>Arguments of advocate about performance data to defend program (#1) or criticize program (#3)</td>
</tr>
<tr>
<td>Goal ambiguity</td>
<td>Programs with performance data that reflect goal ambiguity receive lower support</td>
<td>Evaluative goal ambiguity: Using outcomes rather than outputs (#3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Priority goal ambiguity: Addition of conflicting performance data reflecting alternative goal (#4)</td>
</tr>
<tr>
<td>Expectancy disconfirmation</td>
<td>Performance targets anchor expectations about what constitutes good or bad performance, altering support for a program</td>
<td>Addition of missed targets information (#5)</td>
</tr>
<tr>
<td>Experiment #, treatment</td>
<td>Z score</td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td><strong>Advocacy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Arguments of advocate about performance data to defend program</td>
<td>-1.737*</td>
<td></td>
</tr>
<tr>
<td>2. Arguments of about performance data to criticize program</td>
<td>-.153</td>
<td></td>
</tr>
<tr>
<td><strong>Goal Ambiguity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Evaluative goal ambiguity: using outcomes rather than output</td>
<td>-3.674***</td>
<td></td>
</tr>
<tr>
<td>4. Priority goal ambiguity: Addition of conflicting performance reflecting alternative goal</td>
<td>1.863*</td>
<td></td>
</tr>
<tr>
<td><strong>Expectancy Disconfirmation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Addition of missed targets information</td>
<td>1.966*</td>
<td></td>
</tr>
</tbody>
</table>

Wilcoxon rank-sum (Mann-Whitney) test, negative sign for z score indicates treatment group receives higher budget allocation, and vice versa. ***p<.001; ** p<.01; * z<.05; + p<1, two-tailed. Using a two-sided standard, experiments 4 and 6 would be significant at .05 level. N varies between 137 to 139.
References


Endnotes

¹ Here I refer to peer-reviewed studies that systematically track some pattern of budgetary behavior or outcomes as a dependent variable, and test how performance data correlates with this variable. This excludes a good deal of work
that describe the nature of a performance budgeting reform, or take a case-based approach to offer hypotheses about what makes reforms more or less successful, but whose intent is not to test theory.

ii A deep tradition of experimental research, largely from psychology, demonstrates that how data is presented alters decision outcomes. This work draws from prospect theory (Kahneman and Tversky 1979), and more generally the logic of “framing effects” (Frisch 1993). For example, Davis and Bobko (1986) found strong support for a job training program when the same performance statistics were presented in terms of levels of success rather than failure rate (see also Levin, Schneider, and Gaeth 1998 for a summary of similar work). But it is important to note that the actual manipulations pursued in this paper follow only loosely fit with how Kahneman and Tversky (1981, 453) define decision frames: “the decisionmaker’s conception of the acts, outcomes and contingencies associated with a particular choice.” Rather than reframe what is essentially the same piece of information (equivalency framing), our approach is closer to studies of communication in political science, which add or change salient cues. This approach is sometimes described as “emphasis frames” which “focus on qualitatively different yet potentially relevant consideration” (Chong and Druckman 2007, 115). Another important difference with framing research is that we give less attention to explaining underlying psychological processes (though this logic is present in the literature we draw on to develop our hypotheses), in favor of developing a fuller understanding of the central question posed by the interactive dialogue model, i.e., how policy actors alter their behavior in the contemporary context of greater performance data.

iii The response rate is based on the total number of usable responses divided by the total number of students enrolled in the course when the survey was administered. Factors that reduced the response rate were student absences on the day of the class and the need for students to attend other classes when the survey was administered. Of the 140, 11 were recruited by a general email seeking participation among students in the public affairs programs who had not previously been able to participate, though these are not included the response rate. The proportion of males and females among the subjects (53%) was similar to the sample pool (which was 50%).

iv As a robustness test alternate versions of the model reported in table two were created. These robustness tests adjusted outliers to a more plausible level of change that still reflected the general intent of the respondent to make a non-incremental change to the program. This was done in two ways, using different thresholds for non-incremental changes. First, following Kemp’s (1982) proposal of 10% as an appropriate level of non-incremental change, recommended changes above 10% to were converted to a 10% change. We also adopted Jones, Baumgartner and True’s (1998) suggested threshold of “large increase” at 20% and “large decrease” at 15% to modify outliers. The results of the experiments were robust to these alternate specifications.

v While not reported here due to space constraints, two additional experiments were designed to test the assumption implied in observational studies of performance budgeting: that performance data, by itself, alters decisions. In both experiments, the treatment group received performance data, and the control group did not, and in neither case was this treatment associated with a change in performance data use. In one experiment, the data had no clear relationship with performance, illustrating one very good reason why data may not inform budget decisions, since the data may not offer a pattern that informs policymakers how to change resources. In the second experiment, there was a clear correlation between performance and previous funding, but this still did not exert a change in resource allocations. The point is not that performance information never matters to allocation decisions – other experiments here show that it does – but that such a relationship is not automatic. The results also imply that if broad correlations between performance data and changes in resource allocations are not found in observational data, it may not be because such a relationship does not exist, but that a crude aggregation of performance data excludes more nuanced but systematic relationships between data and budget decisions.