The Design and Practice of Integrating Evidence: Connecting Performance Management with Program Evaluation

Alex Kroll
Florida International University

Donald Moynihan
Director and Professor, La Follette School of Public Affairs, University of Wisconsin-Madison

August 2017
The Design and Practice of Integrating Evidence:
Connecting Performance Management with Program Evaluation

Abstract

In recent decades, governments have invested in the creation of two different forms of knowledge production about government performance: program evaluations and performance management. Prior research has noted both tensions between these two approaches and potential for complementarities when they are aligned. We offer empirical evidence on how program evaluations connect with performance management in the United States federal government in 2000 and 2013. We show that in the later time period there is an interactive effect between the two approaches, which we argue reflects deliberate efforts by the Bush and Obama administrations to build closer connections between program evaluation and performance management. Drawing on the 2013 data, we also offer evidence that how evaluations are implemented matters, and that evaluations facilitate performance information use by reducing the causal uncertainty that managers face as they try to make sense of what performance data mean.
Practitioner Points

- Performance management and program evaluation produce different knowledge and are more effective when complementary.

- Managers involved in program evaluations are more likely to use performance data because program evaluations make causal inference easier, establishing synergies for performance information use.

- The connections between program evaluation and performance management do not arise naturally, but require a mixture of continuity and design.
Introduction

A fundamental philosophical value of the study of governance is that structural conditions can be designed to improve governmental and therefore societal outcomes. This design value is so widely accepted that it saturates the language of public administration. Consider, for example, the persistent tendency by the advocates of any sort of change to frame their ideas as “reform,” implying the need and ability to engage in a corrective improvement.

The promise and perils of design also suffuse basic tensions in the study of public administration (Levine, Backoff, Cahoon and Siffin 1975). Practitioners who criticize the limited real-world relevance of scholarship are often frustrated by the lack of clear design prescriptions: what should we do? The long-running debate about the role of rationality often centers on barriers that undercut good design, such as the role of politics (Moe 1989), the failure to account for local conditions (Lindblom 1959), or a reliance on simplistic assumptions of human behavior and the functioning of administrative instruments (Andrews, Pritchett and Woolcock 2013; Heinrich and Marschke 2010).

Another, less frequently considered, design issue is the question of how the different components of government fit together. This issue is most often expressed in concerns about how structural silos of resources and expertise constrain the solving of “wicked problems” that demand more collaborative designs both inside and outside government (Emerson, Nabatchi and Balogh 2012; Skelcher, Mathur and Smith 2005). A variation on this problem is how different administrative policies or techniques fit together. Light (1998) narrates how a variety of government changes adopted in the United States over 50 years piled atop one another. Each
could be justified in isolation, but they were not designed to work together as part of a coherent framework.

In this article, we identify a related and common design problem that is informed by these prior critiques: the coordination of different forms of knowledge production about government performance. In particular, we examine how performance management and program evaluation relate to one another, using the U.S. federal government as a case study. To the casual observer, these initiatives may appear so similar as to be interchangeable, but in the U.S. case they are based on different epistemic logics, arose at different times, and are associated with different professional communities. Historically, program evaluations rarely informed performance management efforts, and vice versa, and one approach may be seen as hostile to the other (Blalock 1999). Policymakers have become increasingly aware of this institutionalized separation and in recent years have taken steps to exploit potential complementarities by connecting the two.

The goal of this article is to propose that performance management is more likely to result in evidence-informed decisions if integrated with program evaluations. While the two approaches often compete for attention and resources, we offer evidence that decision-makers can benefit when they are designed to work together. While performance systems provide routine management information at low costs and in a timely manner, they cannot explain variation in performance. Program evaluations, however, are designed to offer such causal knowledge and can therefore help to make sense of performance outcomes.

We use survey data from federal managers in 2000 and 2013, a period marked by an increasing effort to integrate the two approaches. Our primary independent variables are
employee exposure to program evaluations and performance management reforms, and our dependent variable is the reported use of performance data to make decisions, an explicit goal of performance management reforms (Moynihan and Lavertu 2012). We find that employees engaged in performance management are more likely to use performance data if also exposed to program evaluations, but the connections vary between 2000 and 2013. In 2000, there is little evidence of an interaction between program evaluation and performance management reforms, but we see clear evidence of such connections in 2013, implying a closer integration between the two approaches over time.

The article makes three specific contributions to the literature on the adoption and implementation of evidence-based government reforms: First, it offers a rare longitudinal assessment of a reform effort and shows that an initiative that did not “work” at first, worked in the later time period. Second, despite the frequent tendency of performance systems across the world to be dropped and replaced with something new before they take effect (Moynihan and Beazley 2016), the article shows that continuity combined with an element of deliberate design can pay off. Third, we outline and test a specific causal mechanism to explain how the integration of program evaluation can make performance management more valuable.

**Program Evaluation and Performance Management**

Program evaluation and performance management are both structures to generate information on government effectiveness (Nielsen and Eljer 2008), with the goal of learning how to facilitate better governmental outcomes (Newcomer and Brass 2016). Despite such basic
similarities, there are profound differences in their origins, epistemic norms, and, in the U.S. at least, professional communities.

A classic definition of program evaluation is “the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs” (Rossi and Freeman 1993, 5). Evaluations may identify evidence on processes implemented, impact of an intervention, and the cost-benefit ratio of an intervention. The “gold-standard” for evaluations is the randomized-controlled field trial, where an intervention or treatment provided to one population is compared to a control group who is not provided the treatment. Performance management has been defined as “a system that generates performance information through strategic planning and performance measurement routines, and connects this information to decision venues, where, ideally, the information influences a range of possible decisions” (Moynihan 2008, 5). There are basic differences in the frequency of their application and the type of information they offer. Program evaluations tend to be done in a targeted fashion to test if a specific intervention led to a change, while performance management typically features government-wide, routine collection of output and outcome data.

There are profound differences between the two on an epistemological level. Heinrich (2007, 255-256) distinguishes between the evidence-based policy movement “in which policies and practices are based on or determined by scientifically rigorous evidence,” while performance management “endeavors to use information on agency or program outcomes to regularly assess government performance and hold managers accountable for results.” Blalock and Barnow (2001) characterize program evaluation as a basic social science technique, while performance management emerges from public and private management concepts rather than the scientific method. While there is a growing social science study of performance management (e.g. Gerrish
2016), this critique remains fair: there is no single performance management model or technique equivalent to the evidentiary rigor of a randomized controlled trial. Governments are moved to adopt performance management based on doctrinal beliefs that measurement holds the key to governmental improvement; beliefs that are reinforced by stories of government or private sector successes and endorsements by professional groups (Moynihan 2008).

The two approaches have different histories, becoming popular at different times. Determining the starting point for each is difficult, but for program evaluation the 1960s served as a seminal period. In the United States, the Johnson administration required formal evaluations of new social programs. The late 1960s and early 1970s saw the emergence of stand-alone research organizations such as Mathematica and the Manpower Demonstration Research Corporation that specialized in providing program evaluations. In the United Kingdom, contemporary expansions in social welfare programs led to the creation of the Social Science Research Council (Heinrich 2007). Program evaluation took on the trappings of a profession, with professional standards for what constituted good practice, its own organizations, and specialist publications.

The starting point for performance management is even more difficult to determine. Governments have measured and categorized social phenomena for centuries, though the idea of using such data as an identifiable management technique does not emerge until the latter half of the 20th century, a period littered with abandoned performance management initiatives. The current era can be given an approximate starting point of 1993, with the passage of the Government Performance and Results Act (GPRA). Despite, at best, mixed evidence of success (Gerrish 2016; Moynihan and Lavertu 2012); the impetus for performance management has not waned. The George W. Bush administration invested a great deal of effort into a performance
assessment of all federal programs called the Program Assessment Rating Tool (PART), and Congress passed the GPRA Modernization Act in 2010 (GPRAMA). The Trump White House has also outlined a new management agenda that calls for a plan to “maximize employee performance” and “performance tracking and accountability” (US OMB 2017).

While it is possible that evaluation and performance functions might be closely connected (Nielsen and Ejler 2008), this has not been the case in the U.S. In a meaningful sense, the communities of evaluation and performance management are different tribes, located in different places, with different languages and beliefs (Newcomer and Brass 2016). The staff that complete performance reporting requirements are rarely charged with considering evaluations, and each group has historically been given different formal guidance from the Office of Management and Budget about how to do their work. The Obama administration bemoaned that “performance measurement and program evaluation are applied in isolation, with agency experts housed in separate units that work independently of each other” (OMB 2014, 66). Indeed, while larger agencies might have a specialized evaluation office, much governmental evaluation expertise is not actually in government, but instead resides in outside organizations. Triantafillou (2015) notes how program evaluation privileges technocratic expertise. The evaluation community enjoys a professional status and independence of government relative to the bureaucrats who are tasked with implementing performance management techniques to account for their performance.
Tensions and Complementarities between the Two Approaches

As parallel movements pursuing the same basic purpose, tensions have arisen between performance management and program evaluation. One form of knowledge production can take resources and attention from the other. Some have argued that performance management has undercut and justified lower investment in program evaluation (Blalock and Barnow 2001; Johnsen 2013; Nielsen and Eljer 2008). However, the tide may be turning. Of late, the rise of the evidence-based policy movement marks renewed attention to evaluation (Haskins and Margolis 2014). Indeed, evaluations are entering the traditional terrain of the performance movement. Governments are increasingly investing in knowledge production units based on behavioral economics, which examine innovations tied to citizen or employee behavior that that are often directly relevant for management decisions on administrative processes. Another example is the emergence of “evidence-based budgeting” (Bornstein 2012), based on meta-analyses of evaluations for different types of programs, making it possible to estimate the comparative return-on-investment for different program options (Pew Charitable Trusts & MacArthur Foundation 2014).

The tension between the two approaches reflects their complementary weaknesses and strengths. The regularity of low-cost performance data compares favorably to the infrequency and cost of program evaluations. But as policymakers discover the limitations of performance data, the value of evaluation-based causal evidence becomes clearer. Put another way, the relative advantages of each form of knowledge production can simultaneously be seen either as a source of tension, as they are kept separate, or as a logic for integration. This perceived complementarity has encouraged those who have enumerated the differences between the two
approaches to make the case for a closer integration (Hatry 2006; Blalock and Barnow 2001; Heinrich 2007; Moynihan 2013; Nielsen and Eljer 2008).

Proponents of integration outline a number of ways the two approaches could connect. First, the outcomes featured in program evaluations and performance management should be consistent, generating evidence on a stable set of goals (Blalock and Barnow, 2001; Hatry 2006; Moynihan 2013). Second, performance data provide a source of usable information to incorporate into program evaluations (Hatry 2006; Nielsen and Eljer 2008). Third, if well-designed performance systems are dependent on logic models, evaluations provide a means to verify links in that process (Blalock and Barnow, 2001). Finally, and most crucially, while performance data offer information about the level and trends in performance, they do not provide causal insights about why performance occurs; program evaluation does (Blalock 1999; Heinrich 2007; Newcomer and Brass 2016). While managers might want to make inferences about the effects of government intervention, they are limited in their ability to do so without evaluation evidence.

Calls for closer integration remain at the level of design logic, claiming a potential benefit in connecting two separate but complementary forms of knowledge production about government effectiveness. However, there is precious little empirical evidence on this question. Heinrich (2012) provides a partial exception, showing that program performance assessments received better scores by the U.S. budget office and more funding increases if supported by strong evaluation data.

Our analysis provides evidence on complementarities among federal government managers in how they use performance data. Performance information use has been an explicit
goal of performance management reforms (Moynihan and Lavertu 2012), and we examine
whether and how program evaluation contributes to that goal. This approach is consistent with a
recurring theme in comparisons of the two approaches, which is that the practice of performance
management stands to benefit from incorporating elements of program evaluation. For example,
Nielsen and Eljer (2008) suggest, “evaluation tools may remedy a number of the shortcomings of
performance measurement when applied in performance management and thus contribute to
research-based policy development.” Such benefits might occur by providing training to
performance management staff (Blalock and Barnow 2001; for a study on the variety of potential
training effects, see Kroll and Moynihan 2015), encouraging them to adopt evaluation techniques
(Newcomer and Brass 2016), or incorporating program evaluation staff into performance
processes (Moynihan 2013). Indeed, Newcomer and Brass (2016) argue that performance
management should be converted into a sub-field of evaluation.

We hypothesize that there is a potential interactive effect between the two approaches:
when decision-makers are exposed to both program evaluation and performance management,
they become more likely to use performance data. The claim for connecting program evaluation
and performance management is that program evaluation offers causal evidence on the factors
behind performance. Previous research has pointed to the difficulty of determining the causal
impact of programs as a barrier to performance information use (Moynihan and Kroll 2016).
Without causal knowledge it is difficult to interpret with any degree of confidence what
performance data implies for action. Having a deeper understanding of the causes of
performance should give managers and policymakers more confidence in utilizing performance
data to make decisions.
**Hypothesis:** The positive effect of performance systems on performance information use will be stronger if these systems are accompanied by program evaluations.

**Methods**

This section explains the data and analyses we employ, our measures, and the limitations of the study. We use two surveys of managers in the U.S. federal government and offer several structural equation modeling specifications to examine the proposed hypothesis.

**Data and Analysis Strategy**

The article uses two samples of data from a survey conducted by the U.S. Government Accountability Office (GAO). The survey was stratified by agency and management level and addressed to a random, nationwide sample of mid- and upper-level federal managers from 24 executive branch agencies, which makes the survey representative of about 97 percent of the executive branch full-time workforce, excluding the Postal Service. The first sample we use was collected in 2000. It included 3,816 managers, and a 70 percent response rate (US GAO 2000). The second sample was collected in 2012-13, and it included 4,391 managers with a 69 percent response rate (GAO 2013). The GAO used similar sampling frames and survey questionnaires across years but drew a new random sample of managers for each of the two waves (no employee-level panel data). This makes it possible to run similar models for 2000 and 2013 and compare results across years, but excludes the possibility of a panel analysis.

Having data from two time periods provides an analytical advantage because, as we explain in the discussion section, there were considerable efforts to integrate program evaluation
and performance management between the two points in time. If such efforts were somewhat successful, we would expect to see closer connections between program evaluation and performance management at the later time period. On the other hand, if there is no salient difference between 2000 and 2013, this casts doubt on the potential for governments to connect the two approaches.

To test our hypothesis we assess whether the impact of managers’ involvement in performance management practices on their use of performance data for decision-making is significantly more positive if their programs have also been subject to program evaluations. For this purpose, we use the group comparison option that structural equation modelling (SEM) offers (Acock 2013, 209ff.). In other words, we estimate the effect of performance management on performance information use, while accounting for additional control variables, employing SEM and compare the coefficients for this effect between the groups of managers whose programs have been exposed to program evaluations with those whose programs have not (see figure 1). Using SEM for this comparison has the advantage that it constrains the factor loadings and measurement intercepts to be equal across comparison groups (ibid., 256), so we can be certain that differences in the effect are not due to different understandings by managers of what performance information use means. Furthermore, SEM isolates measurement errors, which yields more accurate estimates, and it offers additional indicators of overall model fit that other techniques do not provide.

[Figure 1]
We also conduct a supplementary test of our hypothesis that offers insights into why program evaluation facilitates performance information use. We expect managers will gain more from performance management practices if these are accompanied by program evaluations because the latter will help establish causal knowledge. This expectation is wholly consistent with the literature described above and, if true, should result in program evaluations having a positive effect on performance data use through the reduction of causality issues. Put another way, managers will be better placed to use performance data if exposed to program evaluations because such evaluations make them better able to make sense of data and link them to action. We will examine such an indirect effect by focusing on those managers from the 2013 sample, who have stated that their programs have been recently evaluated. Managers in 2013 (but not 2000) were surveyed about problems regarding the implementation of, and support for, their program evaluations, which provides us with two reversed measures of evaluation effectiveness. Therefore, we can test whether the quality of program evaluations (i.e., their implementation and support problems) affect data’s missing link to action (a measure of missing causal knowledge) and whether this variable, in turn, influences performance information use. Again, SEM is a useful technique to examine such an indirect effect.

To evaluate the fit of our SEM models, we follow the guidelines by Hu and Bentler (1999) and Kline (2005) and employ the following indicators: This first one is $\chi^2$ which compares the proposed model against a saturated one. Since $\chi^2$ is a significance test, it is highly affected by sample size and almost always significant in large samples (indicating poor fit) and not a reliable fit measure in small samples due to missing statistical power. We, therefore, only report $\chi^2$ as a descriptive measure of fit. A second indicator is the root mean square error of approximation (RMSEA) which reports low fit if a model is not parsimonious (values below
0.08 indicate adequate fit). Additional fit indicators are the comparative fit index (CFI) (values greater than 0.95 suggest good fit) and the standardized root mean square residual (SRMR) for which values below 0.08 indicate a good fit.

**Measures**

All the measures we use can be found in the appendix. To capture the concept of performance information use, we employ four indicators that are consistent with prior work, and which can be associated with the purposeful dimension of data use (deLancer Julnes and Holzer 2001; Dull 2009; Kroll 2014; Moynihan and Hawes 2012): the use of performance data for setting program priorities; allocating resources; adopting new program approaches or changing work processes; and rewarding government employees. In line with previous research, we expect purposeful information use to be a one-dimensional construct for which respondents show a consistent rating pattern across items.

Performance management practices are measured through managers’ involvement in GPRA routines in 2000 and GPRAMA routines in 2013, respectively. Although both types of performance management routines were measured using different scales and numbers of items, we made both variables as comparable as possible by summing up GPRA items and averaging GPRAMA items as well as generating 5-point scales for each variable. That is, we created indices for each variable before running our SEMs and did not model them as latent variables using confirmatory factor analysis for two reasons. First, since the GPRA items are originally dichotomous and the GPRAMA items Likert-scaled, we would have to use two different SEM estimators, making the models for 2000 and 2013 incomparable. Second, both variables tend to
be formative rather than reflective constructs. Conceptually, each of their items is not an equal, transposable measure of performance management, rather it makes sense that managers are involved in some routines but not in others, which can be captured best through composite indices.

The use of program evaluations is measured with a dummy variable. We use factor scores to tap into two types of problems with program evaluations: implementation and support issues. The endogenous variables in our SEMs are performance information use and “missing link to action,” which reflects the difficulty for managers in making causal inferences about actions and outcomes (see the appendix for the full wording). We selected two additional control variables, accountability and leadership commitment, to be able to estimate the effect of performance management and program evaluations on the endogenous variables more accurately. Both controls have been associated with data use in previous work (Dull 2009; Kroll 2015), and we think they are also relevant determinants of data’s missing link to action. Accountability puts pressure on managers to find ways to link data to actions because they are held responsible for performance improvements. Leadership commitment, however, signals that there is top-level support for managers to invest time and resources into exploring and validating their causal assumptions behind indicators. All descriptive statistics and correlations can be found in table 1.

[Table 1]
Limitations

One concern when working with self-reported survey data is the possibility of common source bias. If data on independent and dependent variables are collected through the same method, the systematic error variance shared by these variables may increase which may lead to inflated coefficients and an increased probability for a type-I error to occur. Although we think the use of multiple data sources is generally beneficial, this is not always – as in our case – a feasible research strategy. If we study large samples of managers and we want to measure individual-level variables like performance information use or involvement in performance management and program evaluation practices, self-reports seem to be the most accurate data source and superior to reports by others or secondary data.

Meanwhile, research on the “ins and outs” of single- and multi-method studies draws a more nuanced picture, suggesting that common method variance does not necessarily bias coefficients upwards: “…correlations between variables measured using a common method are simultaneously attenuated due to unreliability and inflated due to common method bias,” resulting in no particular net effect (Conway and Lance 2010, 327). In fact, simulations show that “attenuation from imperfect scale reliabilities offsets inflation due to CMV [common method variance] and that “the odds appear much more in favor of CMV understating relationships” (Fuller et al. 2016, 3197). Further, drawing data from different sources to measure independent and dependent variables may not be as superior to single-source designs as often assumed. Kammeyer-Mueller et al. (2010) conclude that “…data from distinct observers or occasions of measurement can distort estimates of predictor importance at least as much as common source variance” (294). Current research also suggests the need to apply post-hoc
statistical control strategies with caution because “all have significant drawbacks and some have shown poor empirical results” (Conway and Lance 2010, 325).

Selection issues due to unidentified variables can also threaten the design. For example, managers with a more positive predisposition towards evidence-based decision-making may be more inclined to be involved in GPRA or GPRAMA and set up evaluations of their programs. However, whether programs are being evaluated is at least to some extent exogenous to the attitudes of program managers. Rather, it may be agency idiosyncrasies that matter here (resources, size, policy area, organizational culture etc.), which we account for using robust standard errors clustered at the agency level. Although we cannot rule out the possibility that predispositions towards performance management determine selection into GPRA and GPRAMA routines, a reversed effect where involvement in routines shapes attitudes, would just be as plausible, but is not a threat to our design.

Results

Table 2 and 3 show the results for the tests of our hypothesis, separated by the years 2013 and 2000, respectively. Each time, we are interested in examining the effect of involvement in performance management practices on performance information use, contingent on the presence or absence of program evaluations. As illustrated in figure 1, we are mainly concerned with the evaluation of one path in a fairly straightforward structural equation model: GPRAMA \(\rightarrow\) PI Use in 2013 and GPRA \(\rightarrow\) PI Use in 2000. This is why the tables do not show multiple full SEMs but, instead, summarize the findings for these paths in tables 2 and 3 and offer four different specifications of each SEM by iteratively adding the following control variables as additional
robustness tests: accountability, missing link to action, and leadership commitment. (However, all path coefficients for the complete SEMs can be found in the online appendix).

Each row in tables 2 and 3 represents a different model specification. The left-hand side of the tables compares the effect of GPRAMA/GPRA on PI use by program evaluation yes-no groups, and it reports the unstandardized and standardized regression coefficients as well as p-values for the effect in each group. The right-hand side displays the model r-squared by group and observations for each model as well as a Wald test. The Wald test examines the null hypothesis that the effect of GPRAMA/GPRA on PI Use across program evaluation yes-no groups is equal. P-values below 0.05, therefore, indicate a rejection of the null hypothesis because the GPRAMA/GPRA effect is significantly different across groups. The note below each table contains information on the overall model fit of the SEM. In each table, the “poorest” fit is reported, suggesting that all models have a better fit than the one documented in the note.

[Table 2 & 3]

Table 2 focuses on the 2013 sample. It shows that involvement in GPRAMA has a significant positive effect on performance information use across all specifications and program evaluation yes-no groups. We also see that the magnitude of the regression coefficients is consistently greater in the group where GPRAMA involvement is accompanied by program evaluations compared to the reference group. The Wald tests show that these effect differences between groups are significant in all models, suggesting there is robust evidence for an interaction effect between performance management and program evaluation.
The r-squared for the model without any controls indicates that just the GPRAMA effect accounts for 16.4% of the variation in performance data use in the presence of program evaluations. Even in the model that includes all control variables and has an explained variation of 38.2%, the standardized coefficients show that a one standard deviation increase in GPRAMA involvement for the program evaluation group accounts for a 26.3% standard deviation change in data use. Cumulatively, this is evidence in favor of an interaction effect and an impressive effect size of managers’ GPRAMA involvement, particularly if combined with program evaluations.

Table 3 paints a different picture for the 2000 sample. Although mostly positive, not all GPRA effects on data use are significantly different from zero – that is, some p-values are below, others above, the 0.05 threshold. Further, the GPRA effects between groups look fairly similar, except for the effect in the model without any control variables. This is confirmed by the Wald tests, where only one out of four is significant, suggesting no reliable differences between program evaluation yes-no groups. While the control variables contribute as much as they do in table 2 to explaining performance information use, it is the GPRA effect, which explains no more than 3.3%, even with no controls present. Despite the fact that the overall model fit of the SEM is very good, we see little or no evidence that performance management benefitted from program evaluations in 2000, and we find a GPRA effect which is smaller than that of GPRAMA in 2013.

Figure 2 offers more insights into how program evaluations can foster performance information use. We constrain our sample to responses from managers who stated their programs were subject to evaluations in 2013. These managers were surveyed about the effectiveness of the program evaluations they have been involved in, and their responses fall into two factors associated with implementation (e.g., concerns about the credibility, timeliness, usability, and
validity) and support (e.g., lack of top executive and political commitment, lack of resources) problems (see also the principal component analysis in the appendix). The lower the quality of program evaluations, the harder it becomes for managers to link performance data to actions (Betas are 0.27 and 0.32). Put another way, only effective program evaluations can help generate causal knowledge to help managers make sense of performance data. If such knowledge is absent and links between data and action missing, we will see less performance information use (Beta=-0.17).

The figure also shows that the effects of program evaluation-related problems are only indirect and mediated through the “missing link to action” variable. The indirect path for implementation problems is significant at p=0.066 and the indirect path for support problems at p=0.001, respectively. The results therefore provide evidence of the indirect effect of program evaluation on performance information use via providing a better understanding of causality. They also suggest that as problems with program evaluations mount, this weakens the ability to make causal assessments, thereby reducing performance information use. The fit of the SEM is, again, very good, and the range of all factor loadings for implementation problems is 0.76-0.86, for support problems is 0.73-0.81, and for performance information use is 0.55-0.88.

[Figure 2]
Discussion

Our study makes several theoretical and practical contributions. It offers the longitudinal assessment of a reform effort and shows that the initiative under examination has become more successful over time. Theories of cross-time maturation of management systems are relatively rare in our field – frequently, time is not really theorized in public administration scholarship (Oberfield 2014) and, at best, treated as a lagged variable. Efforts to theorize and test designs have also been observed to be relatively rare (Barzelay and Thompson 2010).

The article provides evidence for the possibility that good design can coordinate different forms of knowledge production about government performance. The connection between program evaluation and performance management makes sense on paper, but rarely works out in practice. We show that the two can be connected: the results are consistent with the hypothesis that program evaluations can help managers to get more out of performance management via improved causal knowledge.

However, we only see this connection in 2013, not 2000. What accounts for the difference between the two time periods? A seemingly-obvious possibility is the maturation of these different systems. Reforms are more likely to succeed when they are given some measure of continuity across administrations. Although theoretically appealing, there has been very little empirical evidence for such continuity effects, particularly in the area of evidence-based accountability systems. And as obvious as the value of allowing a reform to incrementally develop might seem, a frequent tendency of performance management efforts across the world is that they are dropped and replaced with something new before they take effect (Moynihan and Beazley 2016). Our study has shown that continuity enables different types of reforms to connect
with one another. However, our analysis also shows that it is not just continuity, but also an element of deliberate design, that helps improve the system over time. For example, Moynihan and Lavertu (2012) show that the 2000-2007 time period does not show that different iterations of a system generate more performance information use.

The maturation of reforms might be necessary, but perhaps not sufficient. Design and persistence also matter. The efforts to connect performance management and program evaluation benefited necessarily from the deliberate efforts to connect the evaluation and performance management efforts. While discussion sections normally seek to draw out implications of the findings, in this case, we can best do so via a deeper understanding of these design efforts that took place between 2000 and 2013.

By 2000, GPRA was just beginning to be fully implemented after a long pilot period. GPRA legislation mentioned program evaluation and called on agencies to include it in performance reports, but in practice evaluations were rarely mentioned in these reports. From an agency perspective, GPRA reporting requirements took resources and attention away from program evaluations (Blalock and Barnow 2001; GAO 1998). Bush officials saw GPRA as ineffective and created an alternative approach to performance management called the Program Assessment Rating Tool to assess and rank almost all federal programs (Moynihan 2013). The involvement of the U.S. Office of Management and Budget (OMB) made agencies markedly more responsive to government-wide expectations about what constituted good evidence than was the case under GPRA.

The Bush administration sought to explicitly connect performance management and program evaluation in a number of ways. PART directed attention to program level outcomes
rather than the agency level focus of GPRA. In the first governmentwide effort of its kind, it compelled agencies to include “independent and quality program evaluations” along with performance measures as evidence to support the value of their programs. OMB enforced this requirement to the point that the quality of evaluations affected PART scores (Heinrich 2012). The attention to evaluations were only one aspect of PART, and it would be inaccurate to characterize PART itself as a program evaluation. Observers of these efforts also noted uneven implementation. The changes did not always lead performance management staff to fully or enthusiastically embrace evaluations. Agency staff complained that OMB expectations for evaluations were unrealistic, only accepting randomized controlled trials (Metzenbaum 2009; Moynihan 2008). Yet, PART was important in that it compelled performance management staff to learn about program evaluations, and it was accompanied with guidelines about “What Constitutes Strong Evidence of a Program’s Effectiveness?” (OMB 2004).

Despite disbanding PART, the Obama White House remained committed to supporting both performance management and especially program evaluation. The OMB increased attention and funding for program evaluation (Joyce 2011). The Obama administration was even more explicit than the Bush administration in calling for a closer integration of the two approaches. Such an integration has been a constant theme of the Analytical Perspective section of President Obama’s annual budget proposal, which served as the blueprint for the White House management agenda. Within this section, performance management and program evaluation are given equal billing as dual strategies in the development of better outcomes. They are portrayed as complementary mechanisms to facilitate learning and the creation of “a culture that sees program evaluation, statistical series, data analytics, and performance measurement as valuable, complementary tools, since each has different strengths” (OMB 2014, 65).
While GPRAMA did not explicitly deal with the potential for integration between performance management and evaluation, it put in place some mechanisms to make such integration more likely. Historically, performance reporting requirements such as GPRA were seen as extensions of the budget process and often left to budget staff. GPRAMA formalized a Bush-era innovation, which was to appoint specialist performance staff, Chief Operating Officers and Performance Information Officers, in each agency. For such staff, performance was not a distraction from the primary job of preparing the budget, but a key task unto itself. Such staff were more likely to be attentive to directives from the OMB on the benefit of evaluations and more likely to build bridges with the evaluation community. OMB (2015) guidance for implementing the requirements of the Modernization Act make repeated references to how the Chief Operating Officers and Performance Information Officers were to connect evaluation staff and evaluation evidence with broader performance management process.

OMB staff held workshops on program evaluations, and such seminars have occurred in individual agencies, with the goal of educating managers on what program evaluations can do (Moynihan 2013). According to a former OMB official who oversaw performance efforts, Performance Information Officers have collectively: “built a robust learning and support network that has helped agencies improve their data-driven reviews, strategic planning, and, with the evaluation community, integration of evidence-based methods and agency performance management practices” (Metzenbaum 2014, 11). The OMB also tried to build a “demand for reliable data, its analysis, and complementary evaluations” (OMB 2014, 70) through the reporting mechanisms established in GPRAMA, such as quarterly reviews in each agency and strategic annual reviews between the agency and OMB.
Overall, there was therefore clear evidence of a sustained effort, driven largely by the OMB, to integrate program evaluation and performance management. This effort drew on a variety of tactics, including rhetorical encouragement on the merits of such an integration; educational efforts mostly directed toward performance management staff on the nature and benefits of program evaluation; the specialization of the performance function beyond a budget activity; efforts to bring together members of the program evaluation and performance communities within agencies through venues such as quarterly performance reviews; and formal requirements for the provision of evaluation information in performance management reporting processes. While we cannot empirically determine the relative importance of any individual tactic, the formal demands by the OMB for evaluation evidence during the Bush and Obama years have made it impossible for agencies to shrug off the other efforts as window-dressing.

One lesson with regard to the design of the reform is that it took a powerful actor to drive such connections on a government-wide basis. The OMB, through a mixture of craft and formal authority, has enveloped evaluation evidence into performance management tools, pushing the performance management community in government to accept evaluation as part of a broader “evidence infrastructure” to guide their work. The OMB is a uniquely powerful actor in directing administrative attention to management issues, and it is doubtful that much progress would have been made without its efforts. It is important not to overstate the degree of convergence that has occurred. Many of the observations on the limited connections between the two approaches continue to hold (Newcomer and Brass 2016). At the same time, real progress has been made.
Conclusion

In examining the relationship between program evaluation and performance management, this article has offered empirical evidence for previously advanced claims of an interactive effect between the two approaches. The practical implication is that there are benefits in trying to connect the two approaches. We also offer evidence that the way by which evaluations facilitate performance information use is by reducing the causal uncertainty that managers face as they try to make sense of what performance data actually mean to the implementation of public programs.

The article also shows that the implementation of, and support for, program evaluations matter for determining their effectiveness. It first required the development of a consensus that designing such connections was possible and desirable as well as a period of sustained effort. But we draw our conclusions cautiously. In particular, we encourage additional qualitative research on the issue. Case studies focusing on selected agencies could shed more light on causal mechanisms, offering the potential to test different design approaches of marrying performance management and program evaluation.

For other countries that have adopted performance management and program evaluation tools as part of a comprehensive reform package, such divides may never arise in the first place. Indeed, other countries seeking to develop a strong evidentiary basis to understand public sector performance would do well to avoid the institutionalized separation that arose in the U.S. over the course of decades. But if such a separation exists, the case offers some insights into how to bridge it.
There is also a lesson for the design and coordination of government reforms more broadly: simply because there is the potential to connect two logically aligned governmental forms of knowledge production, there is no guarantee that such a connection will occur by itself. While government reformers are often in a rush to replace the systems in place with new ones, they should recognize that patience and persistence are also virtues. The willingness and ability to consistently tinker with, improve, and connect government systems over a sustained period of time can pay dividends.
Endnotes

1 Heinrich (2007) points out that there are some respects where randomized controlled trials are not feasible to mirror actual processes, and observational studies sometimes have advantages in that they can better reflect actual selection processes into programs.

2 Though different from large-scale evaluations of social programs that have historically defined program evaluation, the application of behavioral economics in government consistently relies on randomized controlled trials and offers claims of greater casual insight than performance data.

3 Violations of these assumptions will result in poorly fitted models. Using SEM’s group comparison option to test for interaction effects is only possible if the moderating variable – like in our case the existence of program evaluations – is dichotomous.

4 The data for the 2000 and 2013 samples were both drawn from the same agencies. However, the GAO lumped some of the agencies together into an “other” category when disseminating the 2000 data, which is why the number of usable agency clusters is lower in 2000 than in 2013.

5 In 2009, the OMB Director issued Memo M-10-01 “Increased Emphasis on Program Evaluations,” in 2012 “Use and Evidence and Evaluation in the 2014 Budget,” and in 2013 “Next Steps in the Evidence and Innovation Agenda.”
References


Table 1: Descriptives and Correlations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Year</th>
<th>Mean (SD)</th>
<th>Range</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Performance Information Use</td>
<td>Pooled</td>
<td>3.53 (0.92)</td>
<td>1-5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. GPRAMA</td>
<td>2013</td>
<td>2.81 (0.90)</td>
<td>1-5</td>
<td>0.36</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. GPRA</td>
<td>2000</td>
<td>2.49 (1.63)</td>
<td>1-5</td>
<td>0.17</td>
<td>NA</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Program Evaluation</td>
<td>Pooled</td>
<td>0.52 (0.49)</td>
<td>0-1</td>
<td>0.11</td>
<td>0.20</td>
<td>0.26</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. PE Implementation Problems</td>
<td>2013</td>
<td>2.34 (0.95)</td>
<td>1-5</td>
<td>-0.12</td>
<td>0.07</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. PE Support Problems</td>
<td>2013</td>
<td>2.62 (1.11)</td>
<td>1-5</td>
<td>-0.18</td>
<td>-0.04</td>
<td>NA</td>
<td>NA</td>
<td>0.53</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Accountability to Results</td>
<td>Pooled</td>
<td>3.77 (0.98)</td>
<td>1-5</td>
<td>0.36</td>
<td>0.25</td>
<td>0.07</td>
<td>0.05</td>
<td>-0.10</td>
<td>-0.12</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8. Missing Link to Action</td>
<td>Pooled</td>
<td>2.48 (1.12)</td>
<td>1-5</td>
<td>-0.31</td>
<td>-0.10</td>
<td>-0.13</td>
<td>-0.06</td>
<td>0.42</td>
<td>0.45</td>
<td>-0.17</td>
<td>1</td>
</tr>
<tr>
<td>9. Leadership Commitment</td>
<td>Pooled</td>
<td>3.63 (1.10)</td>
<td>1-5</td>
<td>0.38</td>
<td>0.29</td>
<td>0.20</td>
<td>0.07</td>
<td>-0.18</td>
<td>-0.32</td>
<td>0.41</td>
<td>-0.25</td>
</tr>
</tbody>
</table>

Note: All latent variables are treated as item-averaged index variables in this table.
Table 2: SEM Results for the Effect of GPRAMA on PI Use in 2013

<table>
<thead>
<tr>
<th>GPRAMA → PI Use,</th>
<th>Controls</th>
<th>Wald Test&lt;sup&gt;a&lt;/sup&gt;</th>
<th>R&lt;sup&gt;2&lt;/sup&gt; by Group</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>by presence of program evaluations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>St. Beta</td>
<td>b</td>
<td>St. Beta</td>
<td>χ&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>no</td>
<td>.328</td>
<td>.321</td>
<td>.366</td>
<td>.406</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>yes</td>
<td>.217</td>
<td>.215</td>
<td>.313</td>
<td>.344</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td></td>
<td>.199</td>
<td>.198</td>
<td>.291</td>
<td>.325</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td></td>
<td>.157</td>
<td>.156</td>
<td>.237</td>
<td>.263</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
</tbody>
</table>

Notes: All models fit with χ<sup>2</sup> < 90.73, RMSEA < 0.056, CFI > 0.98, SRMR < 0.037; <sup>a</sup>the Wald test examines the null hypothesis that a constraint, setting the effect of GPRAMA on PI Use across program evaluation yes-no groups to be equal, would be valid; standard errors adjusted for 24 agency clusters; p values of SEM coefficients in parentheses.

Table 3: SEM Results for the Effect of GPRA on PI Use in 2000

<table>
<thead>
<tr>
<th>GPRA → PI Use,</th>
<th>Controls</th>
<th>Wald Test&lt;sup&gt;a&lt;/sup&gt;</th>
<th>R&lt;sup&gt;2&lt;/sup&gt; by Group</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>by presence of program evaluations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>St. Beta</td>
<td>b</td>
<td>St. Beta</td>
<td>χ&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>no</td>
<td>.045</td>
<td>.074</td>
<td>.106</td>
<td>.182</td>
</tr>
<tr>
<td></td>
<td>(0.093)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>yes</td>
<td>.077</td>
<td>.127</td>
<td>.077</td>
<td>.133</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td></td>
<td>.065</td>
<td>.106</td>
<td>.066</td>
<td>.114</td>
</tr>
<tr>
<td></td>
<td>(0.080)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td></td>
<td>.045</td>
<td>.075</td>
<td>.046</td>
<td>.080</td>
</tr>
<tr>
<td></td>
<td>(0.095)</td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.003)</td>
</tr>
</tbody>
</table>

Notes: All models fit with χ<sup>2</sup> < 94.12, RMSEA < 0.059, CFI > 0.97, SRMR < 0.037; <sup>a</sup>the Wald test examines the null hypothesis that a constraint, setting the effect of GPRA on PI Use across program evaluation yes-no groups to be equal, would be valid; standard errors adjusted for 17 agency clusters; p values of SEM coefficients in parentheses.
## Appendix: Measures

<table>
<thead>
<tr>
<th>Variable</th>
<th>Operationalization</th>
</tr>
</thead>
</table>
| Performance Information Use, α=0.868 | For those program(s)/operation(s)/project(s) that you are involved with, to what extent, if at all, do you use the information obtained from performance measurement when participating in the following activities?  
  • PI Use 1: Setting program priorities  
  • PI Use 2: Allocating resources  
  • PI Use 3: Adopting new program approaches or changing work processes  
  • PI Use 4: Rewarding government employees I manage or supervise |
| GPRA (2000), α=0.868 | Additive index based on the following four dummy variables. During the past 3 years, have you been directly involved in any of the following GPRA-related activities?  
  • Developing ways to measure whether program performance goals are being achieved  
  • Gathering and analyzing data to measure whether programs are meeting their specific performance goals  
  • Using measures for program performance goals to determine if the agency’s strategic goals are being achieved.  
  • Assessing the quality of data used in measuring performance.  
  (1= yes; 0=no) |
| GPRAMA (2013), α=0.814 | Item-averaged index of involvement in three GPRAMA routines: cross-agency priority goals, agency priority goals, and data-driven reviews.  
(After listing of existing cross-agency goals): To what extent, if at all, do you agree with the following statements as they relate to one or more of the cross-agency priority goals listed above?  
  • I have been involved in creating the cross-agency goals  
  • The program(s)/operation(s)/project(s) I have been involved in contribute to the achievement of one or more cross-agency priority goals  
  • I have collaborated outside of my program(s)/operation(s)/project(s) to help achieve the cross-agency priority goals.  
(After listing of agency priority goals) To what extent, if at all, do you agree with the following statements as they relate to [agency name] priority goals?  
  • I have been involved in creating my agency's priority goals |
The program(s)/operation(s)/project(s) I am involved with contribute to the achievement of one or more of my agency's priority goals.
I have collaborated outside of my program(s)/operation(s)/project(s) to help achieve one or more of my agency's priority goals.

To what extent, if at all, do you agree with the following statements as they relate to [agency name] quarterly performance reviews?

Overall, the program(s)/operation(s)/project(s) that I am involved with has been the subject of these reviews.

## Program Evaluation

Have there been any program evaluations, either currently underway or completed within the last 5 years, for the program(s)/operation(s)/project(s) that you are or were involved with?

(1 = yes; 0 = no)

## Program Evaluation Implementation (IP) and Support Problems (SP)

Based on your experience with the program(s)/operation(s)/project(s) you were involved with over the past 5 years that were evaluated, to what extent, if at all, have the following factors hindered the agency's use of the program evaluation(s)?

| IP 1: Concerns about the credibility (validity or reliability) of study results | 0.89 |
| IP 2: Difficulty obtaining study results in time to be useful | 0.85 |
| IP 3: Concerns that the evaluation did not address issues that are important to decision-makers | 0.82 |
| IP 4: Difficulty distinguishing between the results produced by the program and the results caused by other factors | 0.75 |
| SP 1: Lack of ongoing top executive commitment or support for using program evaluation to make program or funding decisions | 0.05 |
| SP 2: Lack of ongoing Congressional commitment or support for using program evaluation to make program or funding decisions | -0.05 |
| SP 3: Lack of resources to implement the evaluation findings | 0.05 |

## Accountability to Results

Agency managers/supervisors at my level are held accountable for the results of the program(s)/operation(s)/project(s).

## Missing Link to Action

Difficulty determining how to use performance information to improve the program.

## Leadership Commitment

My agency's top leadership demonstrates a strong commitment to achieving results.

Note: Unless otherwise stated, agreement with the items is measured using a 5-point Likert scale.