The Design and Practice of Integrating Evidence: The Connections between
Performance Management and Program Evaluation in the US Federal Government

Alexander Kroll
Florida International University
akroll@fiu.edu

Donald Moynihan
University of Wisconsin-Madison
dmoynihan@Lafollette.wisc.edu

Paper prepared for the Association for Public Policy Analysis and Management (APPAM)
Conference 2015, November 12-14, Miami, FL

Abstract

A fundamental philosophical value of the study of governance is that structural conditions can be
designed to improve governance. 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 these two
approaches, as well as the potential for complementarities if better aligned. We offer empirical
evidence on how program evaluations connect with performance management in the US federal
government in 2000 and 2013. We show that managers exposed to program evaluations are more
likely to use performance data, implying that program evaluations can draw attention to
performance generally. We also 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. Finally, we 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 for the implementation of public
programs.
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 requires no explication or justification, saturating the basic 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 academic study of public administration. Practitioners who criticize the limited real-world relevance of scholarship are often frustrated by the lack of clear design prescriptions. The long-running debate about the role of rationality in the field 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 or the functioning of administrative instruments (Andrews, Prritchett and Woolcock 2013; Heinrich or 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 paper, we identify a related 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 evaluations relate to one another, using the US federal government as a case study. To the casual observer, these initiatives may appear so similar as to be interchangeable, but in the US 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.

Scholars who have examined these two forms of knowledge production have also made the case for a closer integration (e.g. Blalock and Barnow 2001; Hatry 2006; Heinrich 2007; Moynihan 2013; Nielsen and Eljer 2008; Newcomer and Brass 2015), but there is precious little empirical evidence how such integration would work, and the proposed benefits that would result. We address these questions using 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 reported use of performance data to make decisions, an explicit goal of performance management reforms (Moynihan and Lavertu 2012).
We find that employees exposed to program evaluations are more likely use performance data, but the connections vary between 2000 and 2013. In 2000, there is no evidence of an interaction between program evaluation and performance management reforms, but we see evidence of such connections in 2013, implying a closer integration between the two approaches over time. Using structural equation modeling, we provide secondary evidence on how program evaluations matter to performance management reforms, showing that program evaluation facilitates the use of performance data by identifying causal connections between actions and outcomes in a way that performance data itself cannot provide.

The paper offers evidence that program evaluations facilitate the goals of performance management, and that integration between performance management and program evaluations is feasible. In the US case, the connection appears to have emerged from deliberate and ongoing effort first by the Bush administration, and then by the Obama administration. In the discussion section we explore what design lessons the US experience offers for connecting different forms of knowledge production.

**Program Evaluation and Performance Management**

Program evaluation and performance management are both forms of knowledge production that provide information on the effectiveness of government (Nielsen and Eljer 2008), with the goal of learning to facilitate better governmental outcomes (Newcomer and Brass 2015). Despite such basic similarities, there are profound differences in origins, epistemic norms, and, in the US 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 randomized-controlled field trials, where an intervention or treatment provided to one population is compared to a control group who is not provided the treatment.¹ 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 differences in the application of program evaluations and the type of information they offer. Program evaluations tend to be done irregularly, on a few programs, while performance management depends upon routine, frequent collection of 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

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.
2015), 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, with such beliefs reinforced by stories of government or private sector successes, and endorsements by professional groups (Moynihan 2008).

The two approaches reflect 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, as the Johnson administration introduced novel social programs while also requiring formal evaluations of the success of these programs. The late 1960s and early 1970s saw the emergence and growth of new stand-alone research organizations such as Mathematica, and the Manpower Demonstration Research Corporation that specialized in providing program evaluations. In the UK, similar 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 its own organizations and specialist publications, with professional standards for what constituted good program evaluation.

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 2015; 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).

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 US context we examine. 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 2015). 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 bemoans 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, reflecting the standing that a professional status and actual independence of government gives to the evaluation community relative to the bureaucrats who are tasked with implementing performance management techniques to account for their performance.

Tensions and Complementarities

Because program evaluation and performance management occupy the same basic purposes – generating knowledge about government effectiveness – but have existed largely as parallel movements, this has given rise to tensions. One form of knowledge production can take resources and attention from the other. This argument has been largely made on the assumption
that performance management has undercut and justified lower investment in program evaluation (Blalock and Barnow 2001; Johnsen 2013; Nielsen and Eljer 2008). However, the tides may be turning. Of late, the rise of evidence-based policy movement marks renewed attention to evaluation (Haskins and Margolis 2014). Indeed, evaluations are entering the traditional terrain of performance movement. Governments are increasingly investing in knowledge production unites based on behavioral economics, such as “nudge” unit in the UK, the US Science and Behavioral Team, which examine innovations tied to citizen or employee behavior that that are often directly relevant for management decisions on administrative processes.2 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).

Another way to consider the tension between the two approaches is that each has 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 and 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).

---

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.
Proponents of integration have offered 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 offers information about the level and trends in performance, it does not provide causal insights about why performance occurs; program evaluation does (Blalock 1999; Heinrich 2007; Newcomer and Brass 2015). While managers will want to make inferences about the effects of government intervention, they can better do so if informed of 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 the benefits of integration, or how the two might work together. Heinrich (2012) provides an exception, showing that program performance assessments received better scores by the US 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 that “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.”³ Such benefits might occur by providing training to performance management staff, (Blalock and Barnow 2001), encouraging them to adopt evaluation techniques (Newcomer and Brass 2015), or incorporating of program evaluation staff into performance processes (Moynihan 2013). Indeed, Newcomer and Brass (2015) argue that performance management should be actively converted into a sub-field of evaluation.

Given the lack of prior empirical research on this topic, we hypothesize two alternative pathways by which exposure to program evaluation might facilitate performance information use. First we proposed a direct effect on the use of performance data, implying that program evaluation will encourage managers to seek out and use performance data even if there are not efforts to actively integrate the two approaches.

**H₁**: The utilization of program evaluations fosters performance information use.

Our second hypothesis explores the potential for complementarities. Here, we hypothesize the potential for an interactive effect, that is, that when individuals are exposed to both program evaluation and performance management, they are more likely to use performance data.

**H₂**: The positive effect of performance systems on performance information use will be stronger if these systems are accompanied by program evaluations.

---

³ The alternate argument, that program evaluation can benefit from attention to the techniques of performance management, is strikingly absent in comparisons of the two approaches. An exception is Triantafillou (2015) who argues from a democratic theory perspective that the assumed superiority of evaluation knowledge gives rise to anti-democratic tendencies relative to the strong emphasis on accountability in performance management. Perhaps because he starts from a different philosophical perspective, Triantafillou is less sanguine on the potential to integrate the two.
Our final hypothesis seeks to shed light on how program evaluation facilitates performance information use. The most consistent and compelling 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 2015). 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.

H3: Program evaluations foster performance information use through the clarification of causal links.

**Methods and Data**

**Data**

The paper uses two samples of data from a survey conducted by the US 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 in the agencies covered by the Chief Financial Officers Act of 1990. This includes 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, 70 percent of whom responded to the survey (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.

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.

**Measures**

All the measures we use can be found in the appendix. To tap into the concept of performance information use, we employ four indicators which have been utilized in prior work and can be associated with the purposeful dimension of data use (deLancer Julnes and Holzer 2001; Dull 2009; Kroll 2014; Moynihan and Hawes 2012): 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 in 2000 are measured through managers’ involvement in GPRA by combining four dummy variables into an additive index.\textsuperscript{4} Performance management practices in 2013 are captured through managers’ involvement in GPRA by combining four dummy variables into an additive index.\textsuperscript{4} Performance management practices in 2013 are captured through managers’ involvement in GPRAMA. The latter variable is an additive index of three important practices promoted by the Modernization Act that are measured through several items: involvement in cross-agency priority goals, agency priority goals, and data driven reviews (this operationalization was adopted from Moynihan and Kroll 2015). We consider both variables to be formative, not reflective, which is why they are not constructed as latent variables but composite indices. That is, we do not expect that managers provide consistent ratings for all items making up the indices. Instead, it is quite possible that, for example, some managers are highly involved with priority goals, but to a lesser extent with cross-agency goals and maybe not at all in data-driven reviews.

The use of program evaluations is measured employing a dummy variable. We use factor scores to tap into two types of problems with program evaluations: implementation and support issues. Another important variable is “missing link to action” which captures difficulties in linking performance data to action because of causal complexities. The other variables are controls that have been previously identified as important predictors of performance information use (Kroll 2015). All descriptive statistics and correlations can be found in table 1.

\textsuperscript{4} Moynihan and Lavertu (2012) have used the same items to measure involvement in GPRA but created a dummy variable instead of an additive index. We employ an index to better use the variation in responses, and to make the 2000 and 2013 measures more comparable. However, the findings we present in this paper are robust to either measurement approach.
Estimations and Model Fit

We use two types of analyses in this paper. First, we use OLS regression analysis to examine whether performance management has a stronger effect on information use for decision-making if such systems are accompanied by program evaluations. In particular, we test interaction effects and estimate separate models for 2000 and 2013 to account for the differences in context in understanding the hypothesized effects. Second, we use structural equation modeling to further examine the indirect pathways by which program evaluations may foster performance information use. Such models allow us to simultaneously estimate multiple regression equations (indirect and direct effects) and use varying sets of control variables, while offering intuitive results. Path diagrams with standardized regression coefficients, based on maximum likelihood estimations, illustrate the proposed models and offer several indicators for assessing the fit of the overall model, rather than just individual equations.

To evaluate model fit, we follow the guidelines by Hu and Bentler (1999) and Kline (2005) and assess our measurement and structural models using several 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 $\chi^2$ and its degrees of freedom 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). $\text{P}_{\text{close}}$ is the $p$ value associated with the RMSEA, and it tests the null hypothesis that there is a close fit, which is why
values above 0.05 provide evidence in favor of the hypothesized model. Additional fit indicators are the comparative fit index (CFI), the Tucker-Lewis index (TLI; values greater than 0.95 for both indices suggest good fit), and the standardized root mean square residual (SRMR) for which values below 0.08 indicate a good fit.

Limitations

One concern when working with self-reported survey data is the possibility of common source bias. Statistical findings can be spurious if all variables stem from the same data source and are likely to share error variance. However, prior work suggests that performance information use, which serves as our dependent variable, is less vulnerable to common source bias than variables that are considered more socially desirable (Meier and O’Toole 2013). The interaction effect we propose is not prone to common source bias. In a recent review article, Jakobsen and Jensen (2014, 24) conclude that common source bias, if present, is likely to attenuate interaction effects, not inflate them.

It is difficult to examine common source bias statistically. One technique is Harman’s single factor test that is based on an unrotated factor analysis of all items, suggesting the existence of bias if the majority of variation can be explained by a single factor, while no other significant factors emerge. In their widely cited article, Podsakoff et al. (2003) argue against the use of this method due to sensitivity issues, but they offer a few more promising statistical remedies, most of which can only be used in combination with structural equation modeling.

We adopt one of these approaches and control for the effects of a single unmeasured latent method factor. This method is an advancement of Harman’s single factor test, allowing not
just for the identification of a common method factor but also to account for it statistically. To do so, we let all the items in our hypothesized structural equation model load on a common method factor and examine whether the statistical relationships still hold once this factor has been added to the model. This approach is not without limitations, and a common method factor does not only pick up on variance due to shared measurement error, but also due to relationships among variables other than the hypothesized ones. While this makes it impossible to differentiate among specific causes of bias, it also makes the common-method factor approach very conservative, as the overestimation of bias will reduce the precision of our estimates in ways that disfavor our hypotheses (Podsakoff, MacKenzie, and Podsakoff 2012). As we explain in the results section below, our hypothesized indirect effect remains significant, even after the inclusion of a latent method factor.

Selection issues due to unidentified variables can also be a threat to the design. For example, managers with a more positive predisposition towards performance management may be more likely to be engaged in GPRA or GPRAMA and, at the same time, more likely to be regular performance data users. Though we cannot rule out this possibility, there are reasons to believe that such a mechanism is unlikely to drive the results we find in this paper. First of all, the engagement in performance management routines is not fully voluntary and, instead, dependent on a managers’ rank – a variable which we identify and control for. The higher up managers are, the more likely it is, that they are involved in all three GPRAMA routines, rather than just one or two. Secondly, the logic behind potential selection effects is often not as clear as we might think. For example, it could also be that experiences with GPRA and GPRAMA determine manager’s attitudes towards performance management which, in turn, affect their data use. Such a mediation is just as plausible as the selection effect mentioned before, but it would
not be a threat to our design. We have few reasons to believe that there is systematic selection into program evaluations (as table 1 shows, the correlations between this variable and all others are fairly low). One potential selection factor could be the policy area managers work in, which we control using agency-clustered standard errors.

Results

Table two shows the results of the regression analyses relevant for hypotheses one and two. Our main focus is on how the utilization of program evaluation and its interactive impact with performance management routines (involvement in GPRA and GPRAMA) influence managers’ performance information use. The control variables are mostly significant and their effects are consistent with prior research (Dull 2009; Kroll 2015; Moynihan and Lavertu 2012). The table separates the findings for 2000 and 2013 and presents three models for each of the two samples. All models control for relevant confounding factors while we add the primary variables. The first model examines the individual effect of program evaluation on managers’ data use, the second one includes performance management as a control variable, and the third model adds the interaction term.5

Hypothesis one proposed that the utilization of program evaluations fosters performance information use. Across years, we find support for this hypothesis. Model one shows that

---

5 The data for both samples were 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. It is also noteworthy that that sample size drops in 2012 after the GPRAMA variable has been added but remains the same after the addition of the GPRA variable in 2000. This is because the GPRA index had to be constructed based on four dummy variables, and we conserved observations from managers with no involvement in GPRA whatsoever as 0. This is to rule out the possibility that null effects in 2000 are simply due to a smaller sample size (<800). However, the results presented in table 2 are robust to either coding choice for the GPRA variable.
managers involved in program evaluation report higher use of performance information for decision making than those who are not. Model two, however, shows differential results for both samples, implying that the relationship between program evaluation and performance information use may be contingent on the performance system in place at different times. After accounting for managers’ involvement in performance management (GPRA and GPRAMA), the impact of the program evaluation variable remains significant in 2000, but this effect vanishes in 2013, which points towards the importance of reform context we will turn to in our discussion below. Put another way, the positive effects of program evaluation on performance information use occurred relatively independent of GPRA in 2000, but are mediated by GPRAMA in 2013. This may also reflect the point that GPRA itself was poorly configured to encourage performance information use (Moynihan and Lavertu 2012) while GPRAMA appeared to be positively associated with performance information use (Moynihan and Kroll 2015). We also need to keep in mind that the effects of both variables, performance management and program evaluation, seem to overlap, meaning they share a good amount of variance in 2012 ($r=0.27$, see table 1). A majority (58%) of all managers who report on average at least “some involvement” in GPRAMA are also involved in program evaluations.

The evidence concerning hypothesis one suggests that program evaluations can work independent of performance management to increase performance information use (as was the case in 2000), but that is not necessarily the case (as in 2013). Hypothesis two further examines
the interconnections between program evaluation and performance information by considering interaction effects. The expectation was that the positive influence of performance systems on performance information use will be stronger if these systems are accompanied by program evaluations. Model three examines this hypothesis. Again, we get different results for our two samples. The interaction effect is not significant in 2000, while it is positive and significant in 2013. Since the program evaluation variable is dichotomous, we can also find its marginal effects for 2013 in model three. Managers involved in GPRAMA report high levels of data use, even if they do not utilize program evaluations (b=0.42*). However, if performance management is accompanied by program evaluations, managers’ data use significantly increases above and beyond the pure GPRAMA effect (b=0.42+0.33=0.75*).\(^6\)

Hypothesis three stated that program evaluations will foster performance information use through the clarification of causal links. To test this hypothesis, we turn to the 2013 data because a) our prior analysis in table 2 indicated that program evaluation under the Obama administration helped managers getting more out of their performance data, and b) only this data set offers additional variables we will use to examine whether the quality of program evaluations matters as well. Using structural equation modeling, figure one shows in bold our hypothesized model as well as all significant regressions coefficients after the controls used in table two have been included in the model to avoid omitted-variable bias. The arrows for the control variables were

\(^{6}\) The model-fit statistics indicate that the interaction effect, despite its significance in 2013, only explains a small portion of the variance in performance information use. Our theory does not postulate that program evaluations (through direct, indirect, or interactive links) are among the strongest predictors of information use. Instead, we argue that their effects are often neglected, despite the fact that they are significantly different from zero. It is also noteworthy that interaction effects, in general, have rather marginal magnitudes. We estimated that the significant interaction term in model 3 (2013) has an effect size of \(f^2=0.004\), which is twice as large as the median effect size reported in 106 articles published in psychology journals, determined by Aguinis et al. (2005) in a 30-year review. “\(f^2\)” is the ratio of systematic variance accounted for by the moderator relative to unexplained variance in the dependent variable.
added based on what we know about these variables from prior research and based on the statistical relationships we found among variables when fitting the model. For example, senior executive status is positively related to more discretion which fosters data use. At the same time, senior managers are more likely to be involved in GPRAMA. As reported below the figure, the model fit for the hypothesized model is very good.

We find support for hypothesis three. Program evaluations foster performance information use because they reduce the data’s missing link to action and help establishing casualties. This indirect effect is significant at a level of p=0.02. At the same time, program evaluations have no direct effect on data use, once we have accounted for their indirect impact. The opposite is true for performance systems like GPRAMA – they have a direct effect on information use, but lack the potential to establish causalities (reduce the “missing link”).\(^7\) The hypothesized indirect effect of program evaluation remains significant (p<0.05), even if we account for the existence of a latent common method factor.

[Figure 1]

Figure three offers a supplementary analysis to better understand the indirect effect of program evaluation proposed in hypothesis three. So far we have compared managers who have been involved in evaluations with those who have not. We now constrain the 2013 sample to

\(^7\) In our hypothesized model, we expect program evaluation and GPRAMA to be correlated, without giving this relationship a causal direction. However, in order to be able to treat GPRAMA as an endogenous variable when adding the controls, we had to replace the correlation with a causal arrow. This technical modification does not affect our results.
only those involved in program evaluations and examine whether differences among evaluations matter. In particular, we look at two types of issues associated with evaluations: implementation (e.g., concerns about the credibility, timeliness, usability, and validity) versus support (e.g., lack of top executive and political commitment, lack of resources) problems. As reported below, the model fit is good, and we see that both types of issues are distinguishable constructs (the appendix includes the results of another type of factor analysis which confirms this finding). As before, the effect of program evaluation on performance information use is indirect (p<0.01), not direct. The novel insight that this figure provides is that program evaluation problems increase existing issues related to data’s missing link to action which, in turn, reduces information use. Put another way, the result provides additional evidence of the indirect effect of program evaluation on performance information use via providing a better understanding of causality, but also suggests that as problems with program evaluations mount, this weakens the ability to make causal assessments, thereby weakening performance information use.

Discussion: Connecting Performance Management and Program Evaluation

The results provide a mixture of insights. First, there is a positive effect of program evaluation on performance information use in one time period, but an interactive effect at a later period. The findings therefore suggest that program evaluation can facilitate a key goal of performance management, and also that it may be possible to build closer links between the two approaches. The difference in the results from 2000 and 2013 demand a deeper understanding of
contextual differences, and in particular what the Bush and Obama administration did to integrate program evaluation and performance management.

In 2000, GPRA had been passed for seven years, but was only 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 these reports gave little attention to evaluations. From an agency perspective the demanding reporting requirements created by GPRA took resources and attention away from program evaluations (Blalock and Barnow 2001). In 1998, the GAO reported that: “some evaluation officials hoped the Results Act would increase the use of evaluation results in decisionmaking, while others feared that the large investments required to produce valid and reliable outcome data across whole agencies would compete for funds currently used for more in-depth evaluations of individual programs’ effectiveness” (GAO 1998, 5).

Bush officials largely saw GPRA as ineffective, and instead created an alternative approach to performance management called the Program Assessment Rating Tool (Moynihan 2013). PART was used to assess and rank almost all federal programs through the life of the Bush Presidency. Agencies completed the assessments, but the US Office of Management and Budget (OMB) in turn assessed their information and determined the final scores. Agencies therefore had to be markedly more responsive to OMB expectations about what constituted good evidence than was the case under GPRA.

OMB used PART to formally connect performance management and project evaluation in a number of ways. First, PART directed attention to program level outcomes rather than the agency level focus of GPRA (Heinrich 2012). Second, PART compelled agencies to include
“independent and quality program evaluations” along with performance measures as evidence to support the value of their programs. OMB took these standards seriously enough that the quality of evaluations affected PART scores (Heinrich 2012). These changes may not have not led performance management staff to fully or enthusiastically embrace evaluations – indeed, agency staff completing the assessment frequently complained that the expectations for evaluations were unrealistic, or that OMB would not credit evaluations that were not randomized controlled trials (Moynihan 2008). Despite these limitations, PART was important in that it compelled performance management staff to learn about program evaluations, OMB developed guidelines about “What Constitutes Strong Evidence of a Program’s Effectiveness?” (OMB 2004) for agency staff completing PART. The PART process also gave performance management staff reason to seek out existing evaluations for their programs, or even to commission such evaluations in the future.

Despite disbanding PART, the Obama White House remained committed to supporting both program evaluation and performance management. The OMB released a series of memos that sought to increase attention and funding to program evaluation.\footnote{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.”} In 2010 the Obama administration supported the passage of the GPRA Modernization Act, which put in statute many of the practices initiated by the OMB. The White House has also been even more explicit than the Bush administration in calling for a closer integration of the two approaches.

The idea of better integrating program evaluations with performance data has been a mainstay of the Analytical Perspective section of President Obama’s annual budget proposal, which serves 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. As the administration wore on, greater overlaps appeared, in particular the idea of evaluation and performance data as complementary mechanisms to facilitate organizational learning: “The best government programs use a broad range of analytical and management tools, which collectively comprise an “evidence infrastructure,” to learn what works (and what doesn’t) and improve results. In doing so they support a culture of continuous feedback and improvement...It is 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 the GPRA Modernization Act did not explicitly deal with the potential for integration between performance management and integration, it put in place some changes that made 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 guidance for implementing the requirements of the Modernization Act make repeated references to how the Chief Operating Officers and Performance Information Officers are to connect evaluation staff and evaluation evidence with broader performance management process, suggesting that they “should make sure that the agency gathers and analyzes performance and other evidence, including evaluations and other research as needed, to better understand the problems they are
trying to tackle, the effectiveness of past efforts to address problems, factors affecting change, and the costs of delivery” (OMB 2015).

OMB staff have 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’s has also tried to build a “demand for reliable data, its analysis, and complementary evaluations” (US OMB 2014, 70) through the performance reporting mechanisms established in GPAMA. GPRAMA required agency priority goals, where senior decisionmakers promise progress on a handful of goals, which will be discussed quarterly in each agency, and has advised agencies to include evaluation staff as part of their quarterly reviews (OMB 2015). GPRAMA also requires a strategic annual review between the agency and OMB. Again, OMB guidance calls on agencies to provide evaluation evidence alongside performance information as part of this process (OMB 2015).

Overall, there is therefore clear evidence of a sustained effort, driven largely by the OMB, to integrate program evaluation and performance management. This effort has drawn on a variety of tactics, including rhetorical encouragement on the merits of such an integration; education 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 have made it impossible for agencies to shrug off the other efforts as window-dressing.

Conclusion

In examining the relationship between program evaluation and performance management, this paper has offered empirical evidence for a series of previously advanced claims (Blalock 1999; Blalock and Barnow 2001; Heinrich 2007; Moynihan 2013; Newcomer and Brass 2015; Nielsen and Eljer 2008). First, we show that managers exposed to program evaluations are more likely to use performance data. The practical implication here is that program evaluations can indeed draw attention to performance generally, facilitating a behavioral goal of the performance management movement. Second, we offer evidence of an interactive effect between the two approaches. Here, the practical implication is that there are benefits in trying to connect the two approaches. Finally, we 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.

By comparing the relationships program evaluation and performance management at two different points in time the case analysis also offers some insights on the design of knowledge production in government. The Bush and Obama administration invested unprecedented effort to
connect the two. 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 2015). At the same time, real progress has been made. The most obvious implication therefore is that 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 organically. It first required the development of a consensus that designing such connections was possible and desirable, as well as a period of sustained effort. A second implication is that it took a powerful actor to drive such connections on a governmentwide 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 without the sustained effort of this actor.

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 US over the course of decades. But if such a separation exists, the case offers some insights into how to bridge it.
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>
<th>(9)</th>
<th>(10)</th>
<th>(11)</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>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. GPRA</td>
<td>2000</td>
<td>1.47 (1.63)</td>
<td>0-4</td>
<td>0.17</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. GPRAMA</td>
<td>2013</td>
<td>2.41 (1.01)</td>
<td>0-5</td>
<td>0.36</td>
<td>NA</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Program Evaluation</td>
<td>Pooled</td>
<td>0.52 (0.50)</td>
<td>0-1</td>
<td>0.11</td>
<td>0.27</td>
<td>0.27</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. PE Implementation Problems</td>
<td>2013</td>
<td>0 (1) -1.61 -2.82</td>
<td>-0.11</td>
<td>NA</td>
<td>0.08</td>
<td>NA</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. PE Support Problems</td>
<td>2013</td>
<td>0 (1) -1.65 -2.36</td>
<td>-0.20</td>
<td>NA</td>
<td>-0.09</td>
<td>NA</td>
<td>0.52</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Missing Link to Action</td>
<td>Pooled</td>
<td>2.48 (1.12)</td>
<td>1-5</td>
<td>-0.32</td>
<td>-0.13</td>
<td>-0.07</td>
<td>-0.07</td>
<td>0.43</td>
<td>0.47</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Leadership Commitment</td>
<td>Pooled</td>
<td>3.64 (1.10)</td>
<td>1-5</td>
<td>0.38</td>
<td>0.20</td>
<td>0.28</td>
<td>0.08</td>
<td>-0.17</td>
<td>-0.33</td>
<td>-0.26</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Accountability to Results</td>
<td>Pooled</td>
<td>3.78 (0.98)</td>
<td>1-5</td>
<td>0.37</td>
<td>0.08</td>
<td>0.24</td>
<td>0.06</td>
<td>-0.09</td>
<td>-0.13</td>
<td>-0.18</td>
<td>0.42</td>
<td>0.42</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>10. Learning Routine</td>
<td>Pooled</td>
<td>3.39 (1.20)</td>
<td>1-5</td>
<td>0.37</td>
<td>0.18</td>
<td>0.25</td>
<td>0.08</td>
<td>-0.08</td>
<td>-0.20</td>
<td>-0.20</td>
<td>0.42</td>
<td>0.42</td>
<td>0.36</td>
<td>1.00</td>
</tr>
<tr>
<td>11. Discretion</td>
<td>Pooled</td>
<td>3.12 (1.09)</td>
<td>1-5</td>
<td>0.34</td>
<td>0.20</td>
<td>0.27</td>
<td>0.07</td>
<td>-0.19</td>
<td>-0.30</td>
<td>-0.21</td>
<td>0.48</td>
<td>0.29</td>
<td>0.38</td>
<td>1.00</td>
</tr>
<tr>
<td>12. SES</td>
<td>Pooled</td>
<td>0.19 (0.39)</td>
<td>0-1</td>
<td>0.09</td>
<td>0.27</td>
<td>0.23</td>
<td>0.12</td>
<td>-0.01</td>
<td>-0.08</td>
<td>-0.05</td>
<td>0.17</td>
<td>0.09</td>
<td>0.09</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Note: “NA” indicates impossible correlations, e.g., when one variable is only available for 2000 and the other only for 2013, or when one variable represents a subset of another.
Table 2: The Direct and Interactive Effects of Program Evaluation Utilization on Performance Information Use

<table>
<thead>
<tr>
<th></th>
<th>Models</th>
<th>2000</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Program Evaluation</td>
<td>0.705*</td>
<td>0.628*</td>
<td>0.694*</td>
</tr>
<tr>
<td></td>
<td>(3.49)</td>
<td>(3.17)</td>
<td>(2.79)</td>
</tr>
<tr>
<td>GPRA</td>
<td>0.097</td>
<td>0.132</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.67)</td>
<td>(1.47)</td>
<td></td>
</tr>
<tr>
<td>GPRAMA</td>
<td></td>
<td></td>
<td>0.610*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(5.95)</td>
</tr>
<tr>
<td>GPRA x Program Evaluation</td>
<td>-0.048</td>
<td>(-0.41)</td>
<td></td>
</tr>
<tr>
<td>GPRAMA x Program Evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Link to Action</td>
<td>-0.675*</td>
<td>-0.670*</td>
<td>-0.671*</td>
</tr>
<tr>
<td></td>
<td>(-4.78)</td>
<td>(-4.71)</td>
<td>(-4.77)</td>
</tr>
<tr>
<td>Leadership Commitment</td>
<td>0.365*</td>
<td>0.355*</td>
<td>0.353*</td>
</tr>
<tr>
<td></td>
<td>(2.73)</td>
<td>(2.61)</td>
<td>(2.60)</td>
</tr>
<tr>
<td>Accountability to Results</td>
<td>0.576*</td>
<td>0.584*</td>
<td>0.585*</td>
</tr>
<tr>
<td></td>
<td>(4.34)</td>
<td>(4.42)</td>
<td>(4.44)</td>
</tr>
<tr>
<td>Learning Routine</td>
<td>0.448*</td>
<td>0.442*</td>
<td>0.443*</td>
</tr>
<tr>
<td></td>
<td>(3.83)</td>
<td>(3.79)</td>
<td>(3.77)</td>
</tr>
<tr>
<td>Discretion</td>
<td>0.585*</td>
<td>0.570*</td>
<td>0.571*</td>
</tr>
<tr>
<td></td>
<td>(4.89)</td>
<td>(4.74)</td>
<td>(4.73)</td>
</tr>
<tr>
<td>SES</td>
<td>-0.218</td>
<td>-0.304</td>
<td>-0.299</td>
</tr>
<tr>
<td></td>
<td>(-0.88)</td>
<td>(-1.16)</td>
<td>(-1.12)</td>
</tr>
<tr>
<td>n</td>
<td>1,321</td>
<td>1,321</td>
<td>1,321</td>
</tr>
<tr>
<td>R²</td>
<td>0.242</td>
<td>0.243</td>
<td>0.243</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.238</td>
<td>0.239</td>
<td>0.238</td>
</tr>
</tbody>
</table>

Notes: coefficients are unstandardized; t statistics in parentheses; standard errors adjusted for 23 agency clusters 2013 and for 17 agency clusters in 2000; ^ p<0.1, * p<0.05
Figure 1: How Program Evaluation Can Foster Performance Information Use

Model fit for hypothesized measurement model: \( \chi^2(11)=36.97, CFI=0.99, TLI=0.98, RMSEA=0.05, pclose=0.58, SRMR=0.02 \)
Model fit for hypothesized structural model: \( \chi^2(13)=40.03, CFI=0.99, TLI=0.98, RMSEA=0.04, pclose=0.71, SRMR=0.02 \)
Note: Hypothesized model shown in bold; all shown coefficients are significant at \( p<0.05 \); error variances only shown for main variables; standard errors adjusted for 23 agency clustered; \( n=1,020 \)
Figure 2: How Program Evaluation-related (PE) Problems Constrain Performance Information Use

Model fit for measurement as well as structural model: $\chi^2(60)= 195.12$, CFI=0.96, TLI=0.95, RMSEA=0.06, pclose=0.01, SRMR=0.04
Note: All shown coefficients are significant at p<0.05; standard errors adjusted for 23 agency clustered; n=547
## Appendix

<table>
<thead>
<tr>
<th>Variable</th>
<th>Operationalization</th>
</tr>
</thead>
</table>
| Performance Information Use,  | 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?  
  $\alpha=0.87$                                                                 | 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), $\alpha=0.92$   | 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), $\alpha=0.60$ | Additive index of involvement in cross-agency priority goals, agency priority goals, and data-driven reviews, rescaled to 5 point scale  
  $\alpha=0.77$                                                                 |                                                                                                                                                                                                                                                                                                           | Involvement in cross-agency priority goals, $\alpha=0.77$  
  (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.  
  $\alpha=0.73$                                                                 | Involvement in agency priority goals, $\alpha=0.73$  
  (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.  
  $\alpha=0.73$                                                                 | Involvement in data-driven reviews  
  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.                                                                                                                                                                                                 |

35
<table>
<thead>
<tr>
<th>Program Evaluation</th>
<th>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)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program Evaluation Implementation (IP) and Support Problems (SP)</td>
<td>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)?</td>
</tr>
<tr>
<td><strong>IP</strong></td>
<td><strong>SP</strong></td>
</tr>
<tr>
<td>IP 1: Concerns about the credibility (validity or reliability) of study results</td>
<td>0.89</td>
</tr>
<tr>
<td>IP 2: Difficulty obtaining study results in time to be useful</td>
<td>0.87</td>
</tr>
<tr>
<td>IP 3: Concerns that the evaluation did not address issues that are important to decision-makers</td>
<td>0.82</td>
</tr>
<tr>
<td>IP 4: Difficulty distinguishing between the results produced by the program and the results caused by other factors</td>
<td>0.77</td>
</tr>
<tr>
<td>IP 5: Difficulty generalizing the results to other persons or locations</td>
<td>0.70</td>
</tr>
<tr>
<td>SP 1: Lack of ongoing top executive commitment or support for using program evaluation to make program or funding decisions</td>
<td>0.08</td>
</tr>
<tr>
<td>SP 2: Lack of ongoing Congressional commitment or support for using program evaluation to make program or funding decisions</td>
<td>-0.06</td>
</tr>
<tr>
<td>SP 3: Lack of resources to implement the evaluation findings</td>
<td>0.04</td>
</tr>
</tbody>
</table>

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