Obtaining Comparable Agency Performance Measures: An Application to U.S. Federal Agencies, 2002-2022[1]

Systematically evaluating the performance of United States federal agencies is difficult. The outputs of public sector organizations are difficult to observe, measure, and compare across contexts. Scholars have made important progress measuring comparative agency performance taking a variety of creative approaches, but critics charge that such measures depend upon questionable self-reports, are limited to specific tasks or contexts that hinder generalizability, or are stymied by disagreements about how performance is defined. In this paper, we introduce a new approach to measuring federal agency performance that overcomes many of these difficulties. We generate comparable agency performance estimates for 139 departments and agencies between 2002 and 2022 that vary across agencies and time. We aggregate a vast trove of government surveys, personnel data, and other performance-related information to generate estimates of latent performance via a Bayesian structural equation measurement (BSEM) model. We evaluate how well different existing measures of performance relate to dimensions of performance, validate our approach with out-of-sample measures of performance, and explore descriptive variation. We conclude with a discussion of how to incorporate new or different performance information and the implications of our findings for the measurement and evaluation of agency performance in the United States and other contexts.

Keywords: Agency, performance, measurement

George A. Krause
University of Georgia
gkrause@uga.edu
ORCID-ID: 0000-0001-6076-2363

David E. Lewis
Vanderbilt University
david.lewis@vanderbilt.edu
ORCID ID: 0000-0002-0803-0074

[1] We presented previous versions of this paper at the Public Management Research Conference, Utrecht, Netherlands, June 27-30, 2023, the annual meeting of the American Political Science Association, Los Angeles, CA, August 31 – September 3, and the annual meeting of the Association for Public Policy Analysis & Management, Atlanta, GA, November 9-11, 2023. We thank Rasmus Broms, Fang-Yi Chiou, Eli Lee, Christian Schuster, Manny Teodoro, and Weijie Wang for helpful comments. We are grateful to Cody Drolc for providing GAO investigations data and Mark Richardson for giving us Office of Personnel Management employment data. Colleagues at the General Services Administration, Office of Management and Budget, and Office of Personnel Management provided critical feedback but are in no way responsible for what we write in this paper. We thank Savannah Farr for excellent research assistance on the GAO High Risk Data. The errors that remain are our own.

Obtaining Comparable Agency Performance Measures:

An Application to U.S. Federal Agencies, 2002-2022

Systematically evaluating the performance of United States federal agencies is difficult. The outputs of public sector organizations are difficult to observe, measure, and compare across contexts. Scholars have made important progress measuring comparative agency performance taking a variety of creative approaches, but critics charge that such measures depend upon questionable self-reports, are limited to specific tasks or contexts that hinder generalizability, or are stymied by disagreements about how performance is defined. In this paper, we introduce a new approach to measuring federal agency performance that overcomes many of these difficulties. We generate comparable agency performance estimates for 139 departments and agencies between 2002 and 2022 that vary across agencies and time. We aggregate a vast trove of government surveys, personnel data, and other performance-related information to generate estimates of latent performance via a Bayesian structural equation measurement (BSEM) model. We evaluate how well different existing measures of performance relate to dimensions of performance, validate our approach with out-of-sample measures of performance, and explore descriptive variation. We conclude with a discussion of how to incorporate new or different performance information and the implications of our findings for the measurement and evaluation of agency performance in the United States and other contexts.

Ideally, new executives and legislators would be provided a simple chart or heat map that detailed high and low agency performance when they transition into office. This would allow them to efficiently allocate their management and oversight efforts. Modern governments are awash in data and activity and yet elected officials rarely have this simple information. Developing an overall picture requires aggregating and filtering a tremendous amount of complex performance information. In the United States federal government there are dozens of subjective and objective measures for hundreds of agencies. Public officials need to separate out the helpful from the misleading data. They also need a principled way to aggregate performance data since diverse measures reveal information about discrete activities and use different criteria (e.g., efficiency, effectiveness, equity, etc.). To complicate matters, agencies can be performing at a high level but political, economic, or societal events beyond their control can decouple high performance from clear changes in outcomes. Without a principled approach to aggregating performance information, officials fall back on haphazard and informal patterns, taxing their already busy schedule and increasing the chances they miss emerging problems.

These challenges are not unique to federal officials in the United States (Rogger and Schuster 2023). Indeed, we are in what one author calls, "the era of governance by performance management" (Moynihan 2008: 4). Governments across contexts and at all levels have adopted performance measures to inform their budgeting and management processes (e.g., Boyne 2010; Melkers and Willoughby 2005; Moynihan 2006; Poister 2003; Rogger and Schuster 2023). Performance measures influence the ways elected officials oversee agencies – from budgets to public hearings – and can drive decision making inside agencies in productive and unproductive ways (Courty and Marschke 2011). While use of performance information has expanded, it has been difficult to find measures that allow for meaningful comparisons *across* different kinds of programs and agencies (Andrews, et al. 2006; Boyne, et al. 2006; Rogger and Schuster 2023). Public organizations can rarely be evaluated with anything like simple private sector metrics such as profit, sales growth, or return on equity that can

facilitate comparative performances assessments (e.g., Andersen, et al. 2016: 853; Niskanen 1971: 29; Rainey and Bozeman 2000).[1] Public sector organizations perform a variety of functions that are hard to observe and hard to connect to changes in outcomes (Wilson 1989). While scholars have made important progress measuring comparative agency performance through creative means, existing efforts are often plagued by conceptual and measurement difficulties (Andersen, et al. 2016; Boyne 2010; Boyne, et al. 2006). There are numerous measures evaluating performance on discrete tasks on different dimensions of performance in distinct parts of agencies but these do equate with an overall measure of agency performance.

In this paper, we introduce a new approach to measuring U.S. federal agency performance that overcomes many of these difficulties. We describe a way to aggregate diverse subjective and objective performance information at different levels. We use data from dozens of different sources, including federal employee surveys, government employment data, and other indicators of performance to generate performance estimates via a Bayesian structural equation measurement (BSEM) model.[2] The method provides a means of distilling voluminous and diverse data and determining which measures are most useful for tapping latent agency performance (Andrews, et al. 2006). The approach also helps us disentangle high organizational performance from observed changes in outcomes or results that are often beyond the control of public agencies. In effect, we create something like an organizational health scan, measuring overall organizational performance without overly relying on measures of success that are beyond an agency's control.[3] Using this approach, we generate agency performance estimates for 139 U.S. departments and agencies between 2002 and 2022 that vary across agencies and time. We evaluate how well different indicators of

[1] Some scholars argue that private sector organizations cannot easily by measured by these metrics either and that the goals of firms are more complicated than such economic performance measures (e.g., Hubbard 2009)
[2] See Bertelli, et al. (2015) for a latent measurement approach measuring autonomy, satisfaction, and intrinsic motivation in public agencies.
[3] We thank Adam Lipton at the Office of Management and Budget for introducing us to this concept and language.

performance contribute to the measuring latent agency performance, assess the face validity of our measures by exploring descriptive variation, and externally validate our measures by comparing them to out-of-sample measures of agency performance. We conclude with the implications of our findings for the measurement and evaluation of agency performance in other types of public sector organizations.

CHALLENGES IN COMPARATIVE PERFORMANCE MEASURMENT

Scholars and practitioners have been interested in the systematic measurement of agency performance for some time, with this interest accelerating as part of widespread enthusiasm for the New Public Management (Moynihan 2006; Poister 2003). There is a large literature on why performance management reforms are adopted and whether they contribute to program or organizational improvement (e.g., Kroll and Moynihan 2021; Moynihan 2008; Poister, et al. 2013; Sanger 2013; Wang 2002). Embedded in these evaluations is an important debate about how to meaningfully measure performance in a way that is comparable across contexts.

Public sector performance is difficult to compare across contexts for many reasons (Nyhan and Marlowe 1995). First, observers note that agencies perform hard to observe tasks and efforts to compare across contexts can lead to measures that are quite distant from what agencies actually do (Nyhan and Marlowe 1995; Smith 2006). This problem is exacerbated by a levels of analysis problem (e.g., Andersen, et al. 2016). Some performance measures are targeted at specific tasks. Others are directed at organizational units such as bureaus that perform many tasks. Still others focus on larger organizations that encompass many smaller units such as an executive agency or department. This makes comparisons across contexts difficult. A third difficulty is that programs and agencies have different or unclear goals (Chun and Rainey 2005). This also makes comparing performance across contexts difficult since there is no natural way of comparing performance in environmental policy to transportation policy or tax policy. Fourth, scholars and practitioners often evaluate performance

using different criteria. Boyne (2002), for example, identifies 16 different performance criteria for evaluation, including equity, efficiency, effectiveness, and satisfaction. It is not clear how to compare a good performance based upon efficiency in one program against good performance on client satisfaction in another program. Finally, stakeholders often disagree on what defines good performance. For example, a Republican and a Democrat looking at the Environmental Protection Agency might define good performance quite differently (e.g., Boyne and Dahya 2002: 181; Nyhan and Marlowe 1995: 335; cf. Richardson 2023; Richardson, et al. 2024).

In response to these concerns, some forms of comparative performance assessment focus on individual task-specific measurable activities like revenue forecasting (e.g., Krause and Douglas 2006) or payment error rates (e.g., Park 2022.). Others restrict focus to a single sector such as law enforcement or education (e.g., Boylan 2004; Meier and O'Toole 2002; Rutherford 2016). For example, a rich literature exists on school performance across contexts. Scholars have also made important advances using subjective assessments in surveys that include comparable questions (e.g., Brewer and Selden 2000; Chun and Rainey 2005; Piper and Lewis 2023) and various government generated performance scores (e.g., Kroll and Moynihan 2021; Lewis 2007; Resh, et al. 2021).

While such efforts have helped advance our knowledge and practice of performance measurement, many questions remain. Focusing on comparable tasks or sectors may limit our ability to generalize to other government activities or components. For example, if we focus on tasks like revenue forecasting or responsiveness to information requests, this means measuring performance on tasks that are not central to most agencies' missions. Similarly, are factors correlated with performance in education or law enforcement generalizable to other public sector contexts like research and development or procurement? When scholars and practitioners use surveys to measure performance across contexts, they rely on subjective evaluations, including self-reports (e.g., Lee and Whitford 2013; Meier, et al. 2015; Richardson, et al. 2024). Moreover, the level of organization evaluated is often

unclear (Thompson and Siciliano 2021), and many survey questions and instruments are designed for purposes other than measuring overall agency performance (Fernandez, et al. 2015; Rogger and Schuster 2023). Government generated agency performance scores can be biased, poorly conceived, and unsuccessfully implemented (e.g., Courty and Marschke 2011; Lavertu and Moynihan 2013; Radin 2000). More generally, what information existing measures convey can vary by stakeholder since different stakeholders may define good performance differently (Andersen, et al. 2016; Boyne and Dahya 2002; cf. Richardson, et al. 2024).

What is needed is an approach to the measurement of organizational performance where the goals are clearly defined and we are clear about the relevant stakeholders (e.g., Republicans and Democrats in government). With such an approach the unit of analysis should be clear (e.g., task, bureau, or agency) and the measures can accommodate and discriminate among various subjective and objective indicators (e.g., surveys, outputs) on different dimensions of performance (e.g., efficacy, satisfaction) in a flexible, reasonable, and transparent way. Ideally, the approach would disentangle fundamental organizational performance from factors beyond the control of the agencies themselves (e.g., COVID-19). Our study seeks to address these challenges by aggregating multiple types and sources of data for a lengthy time period in a way that accounts for differences in the quality of existing data to develop measures of latent agency performance. The estimation method is also flexible enough to allow organizational performance to be disentangled from outcomes.

DEFINING ELEMENTS OF ADMINISTRATIVE PERFORMANCE

Given the diverse approaches to measuring performance, it is important to be clear conceptually. To begin, we start with the simplest assumption – an assumption we relax later – that for each agency there is an underlying unobservable latent dimension, agency performance, that is a composite of performance on numerous legally mandated goals or tasks, large and small. To measure this underlying latent dimension we must rely on various observable indicators (e.g., average responses

to a survey question, agency awards, etc.) that each imperfectly reveal information about the agency's performance on this underlying dimension. The higher the quality of measures we have, the better we can place the agency along this latent performance dimension.

Of course, not all measures are useful or uncontested. Some measures may not reveal much about agreed upon definitions of good performance. We need to start by recognizing the differences between high performance and success. We then must clarify whether measuring performance is even possible given the perspectives of different stakeholders (e.g., Republicans and Democrats). A successful approach must also need to distinguish contributors to performance from performance itself, disentangle *task* performance from *organizational* performance at different levels (i.e., performance of a subcomponent versus performance of agency as a whole), and account for different dimensions of performance. Hence, our measurement strategy aims to overcome these limitations by offering a holistic assessment of organizational performance that is comparable both across agencies and time.

Good Performance Does Not Always Mean Success

Scholars and users of performance measures often conflate good performance with success and poor performance with failure (Boyne 2010: 210-211; Smith 2006: 79-82). For example, economic development in a specific jurisdiction should be correlated with the performance of the economic development bureaucracy in that jurisdiction but not perfectly. As the true performance of the agency improves, so does the expected level of economic development. There are, however, some instances where an agency is performing very well but their level of economic development in that year does not match it. They get lucky or unlucky. For example, it is possible that the regional or world economy experiences a downturn in a particular year.

This is true more generally. Quite often, a nontrivial gap exists between agency performance and outcomes. This gap can exist because of unforeseen and uncontrollable factors in the

environment. It can also happen because of the complexity of the work. Sometimes the legislature has given an agency a very hard task (Netra, et al. 2022). Some agencies have simpler tasks like cutting and mailing checks, others endeavor to solve very hard problems like stopping drug addiction or sending astronauts into space. This distinction between success and performance has an important implication for performance measurement. First, many indicators of performance we employ actually measure either success or results. So, for example, if scholars compare the accuracy of budget forecasts across contexts, a forecast with 0 error is a perfect forecast. Yet, the accuracy of a forecast is somewhat stochastic and high performing budget offices and employees can get it right and wrong. In fact, a lower performing budget office can look better than a higher performing office if they get lucky. Similarly, they may look systematically better if the forecasts are easier in their jurisdiction. As the forecasting example suggests, the larger the number of observations of success and failure, the more confidence we can have in our estimates of latent performance, conditional on some understanding of task complexity.

Different Stakeholder Conceptions of Performance

Measuring agency performance is complicated by the fact that stakeholders, such as political parties, clientele groups, or citizens, can disagree about the definition of good performance. This can mean different things. It can mean that parties evaluate performance on different dimensions. For example, one observer may care more about efficacy while another cares more about efficiency (something we discuss further below). More troubling is the possibility that stakeholders accurately observing the same latent performance might classify it differently. For example, a Democrat might suggest that agency actions represent perfect compliance with legal requirements and Republicans would conclude that the same actions do not. We assume here that if stakeholders were able to observe this latent performance dimension perfectly, they would agree on what classifies as good or bad

performance. That is, these external actors would agree that an agency is meeting its legal requirements even if they disagree with the agency's legal mandate.

Politicians have policy goals and may prefer that agency officials use their legal authority to pursue some policy goals and not others. This often gets conflated with performance. Agency policy choices influence whether political actors define agency performance as good or bad. When we measure performance, we are not measuring agency policy choices that might reflect differences in taste or preference. Rather, we are interested in evaluating what politicians of different parties or ideological leanings can agree on – the extent to which public agencies competently perform their job as prescribed by *legal requirements*. We acknowledge that our approach is limited insofar that there are cases where it can be difficult to distinguish organizational performance from disagreements over policy goals. We note, however, that legal requirements set a standard of good performance for many government activities.

It is also important to remember that most programs enjoy bipartisan support and many aspects of administrative performance have little to do with policy per se. Indeed, the vast majority of government activities have bipartisan support because they are popular with the public (Bednar and Lewis 2024; Gramlich 2017). This is to be expected since every government activity was supported by majorities in both chambers and the president at the time of enactment. This is borne out by a recent study revealing there was a strong positive correspondence involving agency performance ratings for both Republicans and Democrats in the United States (Richardson, et al. 2024). When Democrats thought agencies were performing well, so did Republicans and vice versa. While scholarly attention is naturally drawn to areas of either partisan or ideological disagreement, a considerable amount of government activity reflects consensus regarding effective performance, including goals such as effective procurement, safe airports, or an efficient patent system (Richardson 2024).

Measuring Performance versus Contributors to Performance

Given the difficulty of measuring latent performance, it is common for scholars and practitioners to measure administrative capacity or behaviors that contribute to good performance rather than performance itself (Yang and Holzer 2006: 117; Rogger and Schuster 2023). For example, in a social services organization we might measure the number of day care centers funded or employee engagement as measures of performance. In an important sense, neither of these is a measure of performance per se, but we believe that each item measured *contributes* to good performance. Scholars sometimes substitute administrative capacity for performance itself. Higher capacity, in the form of more day care centers, is a *precondition* that facilitates the agency in achieving its goals. Similarly, an engaged workforce likely increases agency performance.[4] Neither measure, however, is itself a measure of better health and social welfare in the community. The agency could be performing poorly with a large number of day care centers and high employee engagement.

Being explicit about the relationship between contributors to performance and latent performance can help us properly interpret performance information. First, it helps us prioritize some types of performance related information over others. For example, if we have direct indicators of performance (*"is your agency performing well?"*), these should be prioritized over contributors to good performance (e.g., number of beds funded, employee engagement). Second, it suggests that any one measure of performance is unlikely to be sufficient. Relatedly, administrative capacity is an antecedent for effective administrative performance. Scholars using measures of administrative capacity note that a social services agency that has built capacity in the form of more day care centers or high employee engagement has performed well on an *administrative* task. Information about performance on this task

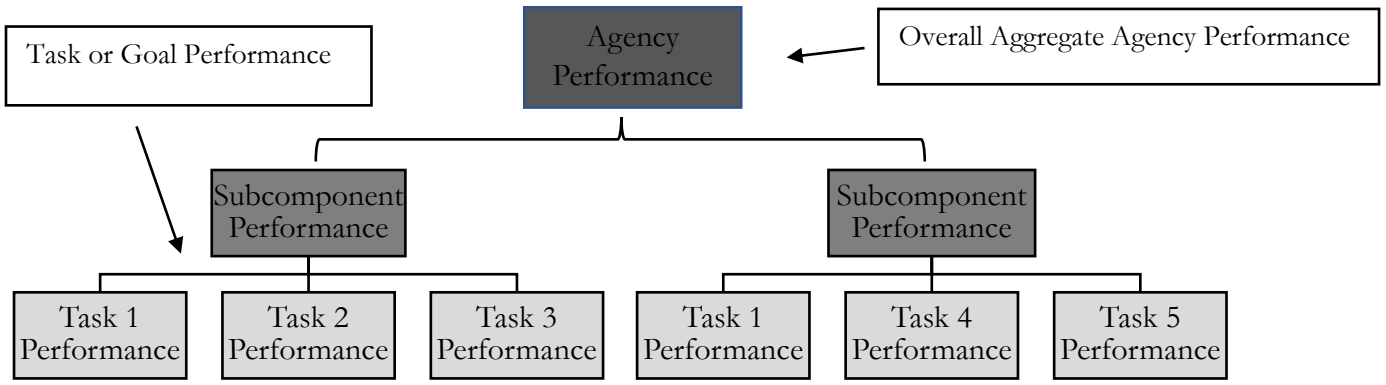
[4] This is not to say that the statutory requirements for a social service agency could not include a goal of building more day care centers. If the statute specified the construction of more day care centers, then the number of day care centers, particularly relative to some baseline, could be a measure of performance. Similarly, a statute could require the agency to improve employee engagement. If so, success in this arena could be a measure of high performance. The point is that scholars and practitioners can conflate *contributors to* high performance with *actual* high performance.

can contribute to our understanding overall performance even though good administrative performance is not the same as an agency achieving its legally mandated goals of better health and social welfare in the community.

Aggregating Performance Information Across Levels

Agency performance is a composite concept, aggregating performance on numerous statutorily mandated goals or *tasks*, large and small. Some of these tasks relate to agency core missions and others to auxiliary statutorily mandated tasks, including internal agency operations and processes like financial management, purchasing, human resources, etc. An agency might be performing at a high level on one task (e.g., catching criminals) and poorly on another (e.g., freedom of information requests). Our approach to measuring organizational performance involves averaging across performance on these different tasks (**Figure 1**).

Figure 1. Measuring Department Performance by Aggregating Subcomponent Performance



Depending upon the size of the agency, overall agency performance can also be a composite of the performance of many different agency *subcomponents*. One subcomponent can have high overall performance and another low overall performance. When we measure overall department or agency performance we are implicitly averaging across multiple units (and tasks) within the organization.

Given this complexity, scholars do not observe true performance directly.[5] They observe something analogous to responses to questions on an aptitude test. No one question can reveal true performance, but a set of questions properly designed and evaluated can get you closer. In aptitude testing, the greater the number of effective questions, the more confident the evaluator. Similarly, each well-defined performance measure provides information about the underlying dimension. Some performance measures help separate *very low* performing agencies from the *low* performing and others *high* performing agencies from *very high* performing. Some measures provide a noisy signal of underlying performance and others a clearer signal. One way to evaluate overall agency performance is to employ a method that can incorporate many different measures, accounting for the fact that such measures reflect the complexity of tasks. Some measures will do a better job separating low and high performers. Similarly, some measures will do a better job of mapping an observed output/outcome onto a level of performance. The key is to have a principled, explicit way of aggregating this information. Our approach will not infer performance based upon a single measure or small set of individual measures. Rather, it uses many different indicators, appropriately weighted based upon the informativeness of each one.

Different Criteria for Evaluating Performance

Evaluations of performance on tasks can include performance on different *criteria* such as efficiency, efficacy, equity, client satisfaction, or other dimensions (Andersen, et al. 2016; Boyne 2002; Gębczyńska and Brajer-Marczak 2020). Some measures tap into performance directly, aggregating across the different criteria, and others tap into specific criteria. For example, a survey of executives might ask, "*How would you rate the overall performance of the fire department in carrying out its mission?*" (i.e., overall performance). By contrast, other measures might tap costs per incident if the task is fire

[5] Agency performance also does not depend upon observability. Agencies can be performing well or poorly on different tasks whether anyone observes them or not.

suppression (efficiency), fire deaths per 100,000 population (effectiveness), or percent of fire victims satisfied with fire department response (client satisfaction). Importantly, some measures of organizational performance can measure performance across tasks but on one criteria. For example, we might evaluate the extent to which an agency is meeting its equity goals across different tasks.

Each performance criterion relates to our overall notions of organizational performance. Agencies that are producing outputs that have the desired effect on outcomes and do so in a way that is cost-effective, generates satisfaction, and treats clients equitably is performing better than one that perhaps accomplished all of these things but wasted funds. Measures of organizational performance, when they are used, are implicitly aggregating evaluations across different performance metrics. When stakeholders report their subjective evaluations of performance, they are themselves usually aggregating across criteria to give an overall rating. Our approach attempts to aggregate evaluations of performance on different criteria and allow details of the estimation to tell us what measures are best at uncovering latent performance and how much they do so.

PERFORMANCE DATA

To develop our measures of performance we collected data from a variety of government and non-profit sources, including the General Services Administration (GSA), the Government Accountability Office (GAO), the Merit Systems Protection Board (MSPB), the Office of Management and Budget (OMB), the Office of Personnel Management (OPM), and the Partnership for Public Service. Some of this data is subjective, indicators based upon the perception of persons working in or close to agencies. Other data is objective, presenting counts of good or bad outputs (e.g., presence of award-winning employees). We list data sources in **Table 1**. The sources in **Table 1** provide data on 139 agencies during the 2002 to 2022 period (**Appendix A** for a full list).

Subjective Data: Surveys of Employees and Citizens

During 2002-2022, the Office of Personnel Management (OPM), the General Services Administration (GSA), and the Merit Systems Protection Board (MSPB) all surveyed federal employees. Several outside groups also conducted federal employee surveys during this period. Collectively, there are 33 different surveys of federal employees with 28 different performance-related questions. Many questions repeat across surveys and years. In **Appendix B** we include a list of surveys of federal employees, the author of the survey (full description in the note), the number of agencies evaluated, and the number of performance-related questions. We also include the overlapping performance-related questions from the surveys.

Table 1. Federal Employee Performance Information, 2002-2022

Source	Title	Years
Objective		
Government Accountability Office	High Risk List	2002-2022 (biannual)
Government Accountability Office	Congressionally Requested Reports (bipartisan)	2002-2020
Office of Personnel Management	Employee Performance Awards	2002-2022
Partnership for Public Service	Sammies	2003-2022
Office of Management and Budget	Program Assessment Rating Tool (PART)	2002-2008
Subjective		
Office of Personnel Management	FHCS/FEVS	2002-2008 (biannual); 2010-2022 (annual)
Merit Systems Protection Board	Merit Principles Survey	2005, 2007, 2010, 2011, 2016, 2021
Richardson, et al. (2018); Richardson, et al. (2024)	Survey on the Future of Government Service	2014, 2020
General Services Administration	Customer Satisfaction Survey	2015-2023
Partnership or Public Service	Best Places to Work Index	2002-2010 (biannual); 2011-2022 (annual)
National Quality Research Center	American Consumer Satisfaction Index	2011-2022

Note: Our models only include data from 2002, 2004, 2006, 2008, 2010-2022 due to available performance data limitations.

Since 2003, the Partnership for Public Service (PPS) has used OPM survey data to create performance indices, including a Best Places to Work in Government index.[6] According to the PPS, "*The index score is calculated using a proprietary weighted formula that looks at responses to three different questions in the federal survey. The more the question predicts intent to remain, the higher the weighting.*"[7] The Partnership also created a 2002 and 2004 Effective Leadership index comprised of answers to 13 different leadership questions on the survey. Component questions for both indices appear in **Appendix B**.

Our final subjective measure of performance is a measure of customer satisfaction. In 1994, the National Quality Research Center at the University of Michigan developed the American customer satisfaction index (ACSI). The ACSI uses customer-survey responses to questions about customer expectations, perceived quality, satisfaction, and complaints, tailored to the public sector context, to create an index of public satisfaction with different agencies. The ACSI provided one aggregate government index rating until 2010, while expanding to as many as 24 different agencies as of 2011.

Objective Data: GAO Reports, PART Scores, and Employee Awards Data

The federal government and outside groups have actively collected objective indicators of performance during this period. The Government Accountability Office, Office of Management and Budget, Office of Personnel Management, and Partnership for Public Service all sought to evaluate or reward agencies for good performance during this period. Starting in 1990, the GAO began publishing a self-initiated report on government activities they considered high risk, called the High-Risk List. The GAO defines high risk as areas of significant weakness in government activities or programs, particularly if the activities involve substantial resources or provide critical services.[8] We collected counts of programs on the list by agency and year during this time period. We also collected data on

[6] The Partnership for Public Service first produced their scores occur in 2003 but these scores were generated using 2002 data. We associate the rankings with the years of the survey.
[7] See 2022 Best Places to Work in the Federal Government Rankings (https://bestplacestowork.org/rankings/about, accessed June 19, 2023). Links to the rankings themselves provides details on the specific questions used.
[8] This description is based on GAO's own description of the program (https://www.gao.gov/high-risk-list).

counts of GAO reports from 2002-2020 resulting from bipartisan requests for GAO investigations.[9] We do so on the assumption that bipartisan requests likely reflect real performance concerns, rather than simple efforts to discredit the presidential administration. Of the 139 agencies in our data, 126 have been the subject of a GAO investigation and some more than 300 for a given year.

We also make use of both government and non-profit data on agencies with employees winning awards. Agencies that regularly produce award winning employees are also seeing improvements in programs or efficiency since these criteria determine employee awards. We obtained government employee performance award data from the Office of Personnel Management (OPM) for four types of awards: high performance award—rating based (2000 – 2022), high performance award—not rating based (2003 to 2022), individual suggestion/invention award (2000 to 2022), and quality step increases (1990 to 2022).[10] Each year since 2001, the Partnership for Public Service has awarded dozens of federal employees Samuel J. Heyman Service to America Medals (also known as "SAMMIES"). In total, more than 700 federal employees working across the executive branch have been awarded this prize. In a given year, agencies have had up to four employees as finalists for performance awards in different areas and agencies have had up to 3 employees win awards for a given year. Among the agencies with the most nominees and winners across this period are the Departments of Commerce, Defense, and Health and Human Services. Some have never had a winner, including agencies like the Department of Education and the National Labor Relations Board.

METHODS

The goal of our measurement strategy is to model the relationship between agencies' latent performance level and observed subjective and objective performance indicators. A natural consequence of this measurement strategy is that some measures will exhibit a stronger connection to

[9] We thank Cody Drolc for providing us with this data.
[10] For descriptions of each type see **Appendix B**.

latent agency performance because the quality of observable indicators varies. Some measures reveal little about actual performance, perhaps because agencies game the measures, the measures are politicized, or the measures are poorly designed (e.g., Andrews et al. 2006; Bertelli and John 2010; Moynihan 2009). Ideally, our measurement strategy would connect latent performance to observed indicators, while accounting for the fact that some indicator measures are more informative than others. It is also possible that there is more than one latent performance dimension, something we explore here. As this suggests, the ultimate success of this approach depends upon the quality and availability of data. Poor data quality or availability limits the ability to produce valid estimates. We are able generate valid estimates for the 2002, 2004, 2006, 2008, and 2010-2022. Valid estimates could not be generated for omitted years due to sparseness of data.[11] As data has become more abundant and of higher quality, our ability to generate valid estimates has improved.

Statistical Methods

We adopt a Bayesian Structural Equation Measurement (BSEM) modeling approach to generate latent agency performance measures. The BSEM modeling approach adopted here begins by employing a Bayesian Exploratory Factor Analysis (BEFA) to empirically evaluate the dimensionality of these observed indicators relating to various aspects of agency performance from multiple data sources. Three criteria were employed in the specification of both the BEFA and BSEM models:

- *Proximity to concept*: We prioritized measures closest to the concept of overall agency performance. So, for example, our models include yearly agency average responses by supervisors (or non-supervisors) to questions like "*My agency is successful at accomplishing its mission.*"

[11] Initial attempts to generate estimates based on these sparse data years resulted in unusual shifts in theta estimates and a sharp rise in the imprecision of the estimates.

- *Coverage*: We also prioritized measures that cover a large number of agencies and/or years. This provides comparability across agencies and years, thus yielding reliable estimates based on sufficient data.
- *Diagnostics*: The development of models was iterative. We used model estimates and fit statistics to compare different specifications.

Next, identification of the BSEM model is predicated on the BEFA analysis to determine the number of dimensions. The latter indicated two latent dimensions, although BSEM model estimates suggest the more robust of the two dimensions is the first dimension. The first dimension consists of indicators measuring performance that reflect the functioning of internal agency operations and processes consistent with agencies fulfilling their core missions. We term this the *management performance dimension*. This is a close-up look at how the agency is doing on both administrative and core tasks. Measures that load on this dimension include performance on core mission, work group work quality, satisfaction with work and organizational environment (*Best Places to Work Index*), effective leadership, satisfaction with supervisors and managerial personnel, and agency performance on functional tasks (acquisitions, human resources, financial management, and IT). These measures come from a variety of different sources—e.g., GSA, MSPB, OPM.

The second dimension, reflecting outcome-related performance that is externally recognized, is comprised of indicators relating to OPM Employee Performance Awards, GAO investigations and high-risk program designations.[12] As we note earlier, management performance can be tightly or loosely related to these outward indicators since some jobs are harder than others and sometimes good performance is not rewarded with good outcomes. Outcome performance, therefore, is likely to yield

[12] The outcome (second) performance dimension observed indicators appearing in **Model 1** (**Table 2**) are adjusted for agency size differences in both OPM and GAO aggregate agency-year counts by dividing through by agency full-time employment equivalents (FTEs) by agency-year observation.

a noisy assessment of agency effectiveness, one that is highly dependent upon external recognition of agency performance, while also likely driven by factors other than management performance.

Generating Latent Administrative Performance Estimates ($\widehat{\theta}$) from the BSEM Model

The Bayesian structural measurement (BSEM) modeling approach is sensible for both practical and statistical purposes. The BSEM model does not restrict estimation to a single dimension of performance. Nor does it assume that multiple latent dimensions are independent of (uncorrelated) with one another. The approach also allows helpful post-estimation diagnostics beyond model fit statistics. Indeed, the BSEM approach provides information that helps evaluate construct reliability, discriminant validity, and nomological validity. A Bayesian approach to SEM estimation is helpful since it allows us to deal with the missing data that naturally arises from using a wide range of data sources.[13] By implementing a BSEM modeling approach, we can cover unique uncertainty estimates for each agency-year observation from the Bayesian posterior distributions.

Our model takes the form of a two-factor confirmatory factor Bayesian structural measurement model with correlated errors. The latent traits for the first and second dimensions of agency performance are defined respectively as y_i^{*F1} and y_i^{*F2}. The Bayesian structural equation measurement (BSEM) model is defined as:

$$y_i^{*F1} = \upsilon^{F1} + \Lambda_p \eta_{p_i}^{F1} + \varepsilon_i^{F1} \quad (1)$$

$$y_i^{*F2} = \omega^{F2} + \Pi_q^{F2} \theta_{q_i}^{F2} + \zeta_i^{F2} \quad (2)$$

where υ^{F1}, ω^{F2} constitute intercept terms for each respective latent trait equation; η_p^{F1}, θ_q^{F2}, represent p, q -dimensional vectors of observed indicator variables in each measurement equation for each respective latent trait, while Λ_p^{F1}, Π_q^{F2} are the corresponding $p \times 1$, $q \times 1$ parameter matrices of factor

[13] In the reported model, a total of 137 agency-years contain missing data for the BSEM model (6.12% of full sample of 2,237 agency-year observations), with a low of 112 agency years – 5.01% of full sample (**Model 2: Appendix D, Table D3**), and a high of 167 agency-years – 7.47% of full sample (**Model 5: Appendix D, Table D3**).

loadings and $\boldsymbol{\varepsilon}^{F1}$, $\boldsymbol{\zeta}^{F2}$ constitute the residual vectors for each latent trait equation that are allowed to be correlated. Their corresponding variance-covariance matrix is denoted as $\boldsymbol{\Theta} = \boldsymbol{\varrho}(\boldsymbol{\varepsilon}^{F1}, \boldsymbol{\zeta}^{F2})$. Estimates are generated via the Bayesian posterior density of the parameter distributions for the slope, intercept, and loading parameters ($\boldsymbol{\nu}^{F1}$, $\boldsymbol{\omega}^{F2}$; $\boldsymbol{\Lambda}_p^{F1}$, $\boldsymbol{\Pi}_q^{F2}$), the variance-covariance parameters ($\boldsymbol{\varepsilon}^{F1}$, $\boldsymbol{\zeta}^{F2}$), and the latent variables of interest ($\boldsymbol{\eta}_p^{F1}$, $\boldsymbol{\theta}_q^{F2}$). The conjugate non-informative priors for all the free parameters ($\boldsymbol{\nu}^{F1}$, $\boldsymbol{\omega}^{F2}$; $\boldsymbol{\Lambda}_p^{F1}$, $\boldsymbol{\Pi}_q^{F2}$) are normally distributed with mean zero, and positive infinity variance; the variance-covariance parameters ($\boldsymbol{\varepsilon}^{F1}$, $\boldsymbol{\zeta}^{F2}$) follow an inverse Wishart distribution containing a mean of 0 (non-binary probit links) or 1 (binary probit links) and a variance of 3; except for the variance parameters that are block diagonal of size 1, and hence follow an inverse gamma distribution with mean set to -1 and variance set equal to zero that is equivalent to a uniform prior on $[0, \infty)$.[14]

This model is estimated with Bayesian Markov Chain Monte Carlo simulation methods, implemented via Gibbs sampling, employing 100,000 iterations, with 2 chains, and 100 intervals employed for thinning using *Mplus* statistical software (Version 8.10). The specific analysis implemented here utilizes multiple imputation to generate plausible values consistent with the observed data through 1,000 draws, which form the basis for the Bayesian posterior distribution for each indicator variable, and more importantly, generate the resulting latent factor estimates based on plausible values for these latent measures by treating the indicator variables as containing missing data on all agency-year observations (Asparouhov and Muthen 2021). Estimation of this model generates 1,000 sets of Bayesian posterior theta/θ (factor score) estimates corresponding to each agency-year observation for both the *management performance* and *outcome performance* latent concepts. The Bayesian posterior median theta/θ estimates yield point estimates of latent agency performance, while the Bayesian posterior standard deviation and corresponding 95% credibility intervals provides measures of uncertainty surrounding these latent agency performance point estimates.

[14] Additional information and technical details can be obtained from Asparouhov and Muthen (2021).

EMPIRICAL RESULTS

Table 2 lists the BSEM model estimates in the form of standardized factor loading coefficients. They represent how each observed indicator is correlated with the underlying latent management performance and outcome dimensions.[15] Each of the management (first) dimension agency performance indicator estimates are positively signed, substantial, and statistically significant at the $p < 0.01$ level. Larger values of the standardized factor coefficients correspond to a greater amount of each indicator's variance is being explained by a latent trait. Seven of the eight indicator variables are strong predictors of the latent management performance (range between 0.665 [*MSPB: Core Mission (Federal Executives Only*] and 0.963 [*MSPB: Satisfaction with Managers Above Supervisor*]). The only exceptions with standardized factor loadings below 0.50 include the *Best Places to Work Score [2020-2022]* indicator variable (0.480) and the GSA Informational Technology indicator (0.478), where the former is likely the result of limited temporal coverage over the sample period (three years).

TABLE 2: BSEM Model with Correlated Factors — Standardized Factor Loadings of U.S. Federal Agency Performance [2,237 Agency-Year Observations, 2002/2004/2006/2008, 2010-2022]

Variable	1st Dimension	2nd Dimension
FEVS: Fulfilling Agency Mission	0.875*** (0.009)	_____
FEVS: Quality of Work Unit	0.795*** (0.013)	_____
FHCS: Organization as a Place to Work Compared to Others	0.974*** (0.018)	_____
MSPB: Satisfaction with Supervisor	0.936*** (0.011)	_____
MSPB: Satisfaction with Managers Above Supervisor	0.963*** (0.009)	_____
OPM: Best Places to Work Score [2002-2019]	0.908*** (0.008)	_____
OPM: Best Places to Work Score [2020-2022]	0.480*** (0.053)	_____
FHCS: Effective Leadership [2002 & 2004]	0.771*** (0.047)	_____
GSA: Quality of Acquisition Services	0.665*** (0.031)	_____
GSA: Quality of Financial Management Services	0.666***	_____

[15] Additional model sensitivity checks assess different model specifications for these two dimensions and some models with three latent dimensions. They yielded similar Bayesian posterior median theta (θ) estimates. Details of the most credible of these alternative model specifications appear in **Appendix D (Table D3)**.

	(0.031)	
GSA: Quality of Human Capital Services	0.694***	_____
	(0.030)	
GSA: Quality of Information Technology Services	0.478***	_____
	(0.042)	
OPM Innovation Award Annual Frequency (Agency Employment Adjusted)	_____	0.050 (0.057)
OPM Ratings-Based Cash Award Annual Frequency (Agency Employment Adjusted)	_____	0.069 (0.273)
OPM Non-Ratings-Based Cash Award Annual Frequency (Agency Employment Adjusted)	_____	0.060 (0.085)
OPM Quality Step Increase Annual Frequency (Agency Employment Adjusted)	_____	−0.047 (0.085)
GAO High Rish Program Count (Agency Employment Adjusted)	_____	−0.999*** (0.254)
GAO Bipartisan Legislative Investigations (Agency Employment Adjusted)	_____	−0.583 (0.938)
Comparison Fit Index (CFI)	0.920 [0.841, 0.930]	_____
Tucker-Lewis Fit Index (TLI)	1.000 [0.999, 1.000]	_____
Root Mean Square Error of Approximation (RMSEA)	0.003 [0.003, 0.003]	_____
Deviance Information Criterion (DIC) Statistic	52,272.070	_____
Average Variance Extracted	0.471	0.140
Construct Reliability	0.911	0.011
Discriminant Validity	0.471 > 0.011	0.140 > 0.001
Nomological Validity	−0.034 (0.100)	_____

Note: Model estimates generated from 1,000 Bayesian Posterior Empirical Distribution Functions (EDFs) based on 100,000 MCMC iterations with 2 chains using Gibbs Sampling with data missing at random for imputed values. Entries are standardized factor loadings with standard errors inside parentheses, except for Model Fit Statistics content that reports 90% credibility interval values inside brackets. *** $p \leq 0.01$.

For the outcome performance dimension, each of the four *OPM* performance recognition indicators have extremely low standardized factor loadings (ranging between – 0.047 and 0.069), while the *GAO Bipartisan Legislative Investigations* estimate is of a rather sizable magnitude and correct sign (−0.583), but is estimated with considerable imprecision (posterior standard deviation = 0.938). The only informative indicator for the performance dimension is *GAO High Risk Program Count* – which is

both correctly signed and large (−0.999), and also estimated with precision (posterior standard deviation = 0.254). These results corroborate our view noted earlier that outcome-related indicators are prone to yield a noisy assessment of effective agency performance.

The standardized factor loadings for the indicators corresponding to the F2 outcome performance dimension are much weaker, and estimated with less precision, than those of the F1 management performance dimension. All one can conclude from F2 is that these indicators do not comprise a valid latent dimension of performance. This is corroborated by the low Average Variance Extracted and Construct Reliability statistics for the F2 outcome performance dimension denoted at the bottom of **Table 2**, and the meager variability apparent from the Bayesian Posterior estimates. The only empirical leverage offered by the F2 dimension indicators is to differentiate our management-related indicators (F1) from our outcome-related indicators (F2), while revealing that these latent concepts are not measuring the same aspects of performance based on inter-factor correlation (−0.034).

The model fit statistics and structural measurement model diagnostics reveal that the reported model specification yields a superior model fit compared to alternative BSEM models reported in the **Appendix D** (see **Table D3**). The Tucker-Lewis index (TLI) value of 1.00 exceeds 0.95 threshold value, while the root mean square approximation (RMSEA) is 0.003, well below the threshold of excellent model fit (0.050). Although the companion fit index (CFI) has a reasonably high value (0.920), it is estimated with some imprecision based on the 90% confidence interval [0.841, 0.930]. Additional sensitivity checks regarding model specification reported in the **Appendix D** indicate that the Bayesian posterior theta estimates associated with the management (F1) dimension are highly

correlated, thus providing additional credence regarding the durability of these estimates of interest for purposes of evaluating agency performance (see **Tables D1** and **D2**).[16]

Descriptive Patterns of the Agency Performance Estimates

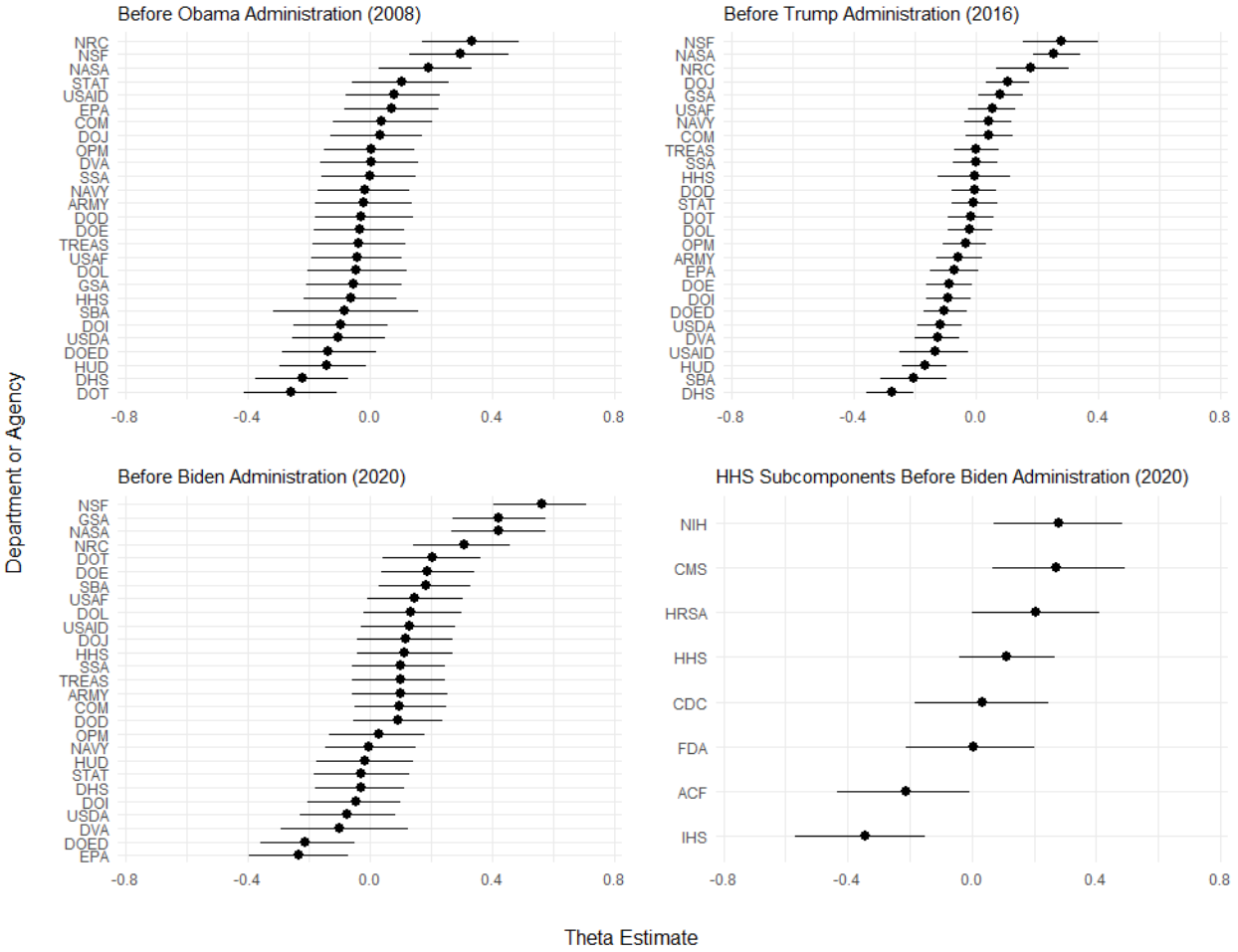
What is more central to our endeavor is the estimates themselves. **Figure 2** displays the Bayesian posterior medians and 95% confidence intervals for the major executive branch departments and agencies (excluding subcomponents) prior to the start of the last three presidential administrations (i.e., end of 2008, 2016, and 2020). It also includes a similar figure for the subcomponents inside the Department of Health and Human Services in 2020 to illustrate variation within larger departments. Such a figure is what we had in mind as something that might be helpful to decision makers. This information could be helpful in deciding where to allocate time or attention or what kind of person to nominate to lead an agency. At minimum, this information would be a signal to dig deeper and investigate the causes of an agency's low rating. During the transition, the president's team could quickly see that some agencies were doing better than others and particular attention might be paid to places like the Environmental Protection Agency or the Department of Education at the end of the Trump Administration. These low agency scores are hardly surprising given what we know about President Trump's efforts to reduce federal support and reach in both departments. The president proposed a 26 percent reduction in EPA funding and an 8 percent cut for education and these agencies saw decreases in morale under the former president.[17] The president's team and newly elected

[16] Bayesian exploratory factor analysis was employed as a diagnostic tool to initially determine the plausible number and type of performance dimensions from these data and various indicators under consideration to evaluate latent agency performance. Preliminary analysis was used as the basis for evaluating several alternative BSEM (confirmatory) model specifications in terms of factor loadings, sufficient indicators per latent dimension, and model fit. This analysis subsequently resulted in five alternative BSEM model specifications, with **Model 1** results being reported in this study and the remaining models appearing in **Table D3**. The Bayesian Posterior estimates from these models are evaluated through correlation analysis in **Tables D1** and **D2**. To summarize, correlations among the Bayesian posterior medians from these alternative models are highly correlated (range between 0.9576 – 0.9968) in the management performance (F1) dimension (**Table D1**). These correlations also remain high for the Bayesian Posterior standard deviation estimates (range between 0.9271 – 0.9972), thus indicating the precision of these BP estimates are similarly high (**Table D2**).

[17] Rebecca Beitsch and Rachel Frazin, "Trump budget slashes EPA funding, environmental programs," *The Hill*, February 10, 2020; Emily Badger, Quoctrung Bui, and Alicia Parlapiano, "The Government Agencies That Became Smaller, and Unhappier Under Trump," *New York Times*, February 1, 2021.

legislators would also see that the National Science Foundation (NSF), National Aeronautics and Space Administration (NASA), and General Services Administration (GSA), three agencies with very different core missions, were doing well.

FIGURE 2: Performance Estimates of CFO Act Agencies, Start of Presidential Administration



Note: The figure includes posterior median estimates and 95% confidence intervals from the end of 2008, 2016, 2020.

Table 3 includes a list of the top-10 and bottom-10 agencies across the entire 2002 – 2022 period by average median agency-year performance estimate. Among the high performers are several science agencies and a few well-regarded independent agencies as well as U.S. Attorneys and the largely evidence-based Federal Highway Administration. Not surprisingly, agencies dealing with immigration and homeland security are among the lowest scoring agencies. In addition, agencies providing services to Native America populations and the U.S. Agency for Global Media are among the low scores. This

is consistent with widespread perceptions and other scholarly research as recent investigations and reports by the Government Accountability Office and Congressional Research Service indicate.[18]

Table 3. Average Top and Bottom 10 Performing Agencies: Average Posterior Median Management Performance Estimates, 2002-2022

Department	Agency	Management Performance
Top 10		
Independent	National Science Foundation	0.293
Independent	National Aeronautics and Space Administration	0.289
Independent	Federal Energy Regulatory Commission	0.277
Independent	Peace Corps	0.273
Department of Justice	U.S. Attorneys	0.272
Independent	Federal Trade Commission	0.267
Department of the Treasury	Alcohol and Tobacco Tax and Trade Bureau	0.253
Department of Transportation	Federal Highway Administration	0.251
Independent	Nuclear Regulatory Commission	0.244
Independent	Federal Deposit Insurance Corporation	0.213
Bottom 10		
Department of Health and Human Services	Indian Health Service	-0.211
Department of Homeland Security		-0.215
Department of the Interior	Bureau of Indian Affairs	-0.242
Department of Homeland Security	Customs and Border Protection	-0.257
Department of Homeland Security	Immigration and Customs Enforcement	-0.281
Independent	Federal Election Commission	-0.298
Department of Homeland Security	Transportation Security Administration	-0.302
Independent	U.S. Agency for Global Media	-0.307
Department of Education	Office of Postsecondary Education	-0.307
Department of Justice	Immigration and Naturalization Service (2002)	-0.429

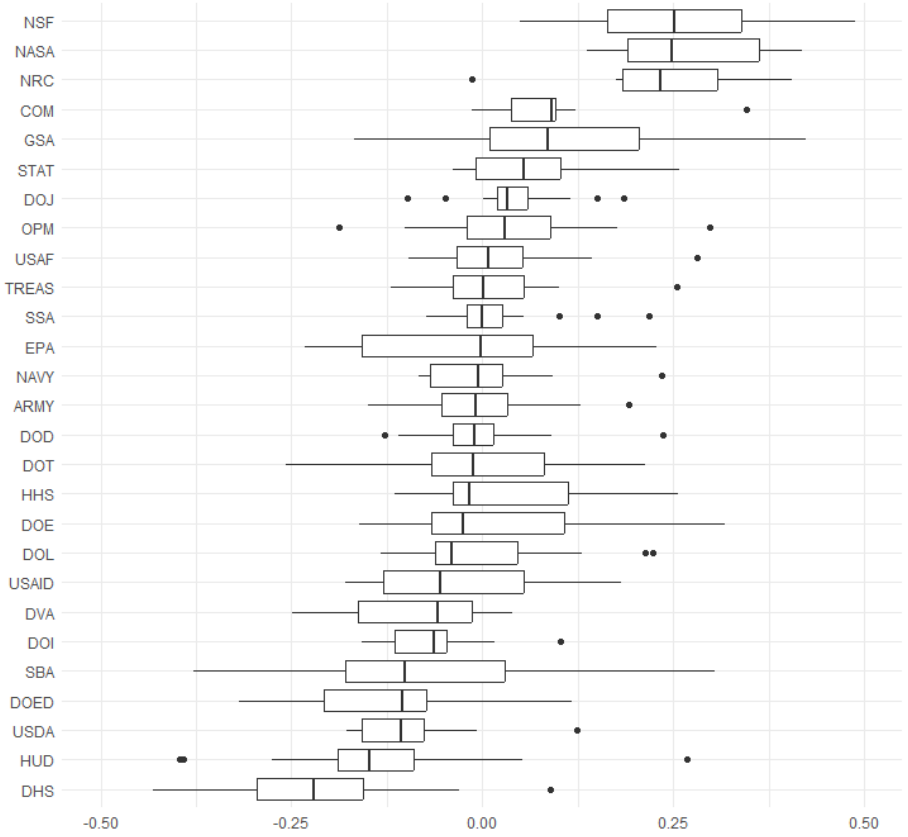
Note: BSEM models produce posterior distributions for each estimated management performance estimate. Table includes average medians of those distributions for each agency.

The cross-sectional rankings obscure important changes within agencies over time. Some agencies are doing well, particularly relative to their historical performance and others have a history of excellent or poor performance and one that continues to the present. In **Figure 3** we graph box

[18] See, for example, Government Accountability Office. 2019. "Tribal Programs: Resource Constraints and Management Weaknesses Can Limit Federal Delivery to Tribes." GAO-20-270T, November 19, 2019 (https://www.gao.gov/products/gao-20-270t; Congressional Research Service "U.S. Agency for Global Media: Background, Governance, and Issues for Congress." *CRS Report* R46968, November 17, 2021 (https://sgp.fas.org/crs/row/R46968.pdf).

plots of the performance estimates for the executive departments and major independent agencies over the 2002-2022 period. A few things stand out. First, some departments and agencies generally performed better across the entire time period. The three agencies that stood out in 2020 in **Figure 2** also appear to have performed well during most of this period, though GSA appears to be performing better than normal relative to its historical pattern.

FIGURE 3: Boxplot of BSEM Performance Estimates of CFO Act Agencies, 2002-2022



Note: Box plot vertical lines are posterior median estimates. Boxes indicate interquartile range and lines indicate minimum and maximums, excluding clear outliers from distribution (dots).

Second, some agencies are regularly lower performers than others, while others seem to fluctuate. Notably, the Department of Homeland Security (DHS), the Department of Housing and Urban Development (HUD), and the Department of Agriculture seem to regularly be among the low performers. Other agencies such as the Environmental Protection Agency (EPA) and the Department of Transportation fluctuate more. This is reinforced by graphs of agency estimates over time (**Figure 4**). These graphs of estimates show the variation cross-sectionally – e.g., DHS and HUD are on average

lower performers—and over time. The efforts President Trump took to redirect the EPA and Department of State are reflected in declines in those agencies during his administration.

FIGURE 4: BSEM Performance Estimates of CFO Act Agencies, 2002-2022

Note: Posterior median estimates and 95% confidence intervals from 2002, 2004, 2006, 2008, 2010-2022.

External Validation with Out-of-Sample Data

We evaluate external validity by performing out-of-sample validation tests of these latent performance measures to other performance measures excluded from our BSEM model specifications. To begin, in **Figure 5** we graph the correlations between our performance estimates and four distinct measures of performance from various years. The top two panels in the figure correlate our performance measures with data from the 2020 *Survey on the Future of Government Service* (SFGS), a non-partisan and non-governmental survey of thousands of federal executives (Piper and Lewis 2023; Richardson, et al. 2024). The survey asked a series of questions intended to provide different perspectives on performance. Importantly, the survey asked, "*How would you rate the overall*

performance of [your agency] in carrying out its mission?" Respondents were given a sliding scale from 1-Not at all effective to 5-Very effective. They could also indicate a "*Don't know*" response. Weighted agency average responses to this self-assessment can be compared to our estimates of θ from 2020. In addition, the 2020 survey asked respondents to rate the performance of other agencies. Specifically, the survey began by asking respondents: "*Please select the three agencies you have worked with the most in order of how often you work with them.*" Each respondent was given a drop-down menu. Later in the survey, respondents were asked "*How would you rate the overall performance of the following agencies in carrying out their missions?*" and given the list of agencies they provided plus two others. Richardson, et al. (2023) generated performance estimates based upon the thousands of ratings federal executives. These scores can be compared to our 2020 estimates. The third panel includes a correlation between our 2014 performance estimates and a measure of performance from the 2014 SFGS. In 2014, the SFGS asked respondents whether they agree or disagree with the statement, "*I am confident in the ability of [my agency] to successfully fulfill its core mission.*" (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree, Don't know). This measure nicely fits with our desire to measure performance on key tasks. The final panel correlates our performance estimates in 2002, 2004, 2006, and 2008 with average agency PART scores from those same years. These are numerical federal program performance scores from the George W. Bush Administration for 1,016 programs.[19] Specifically, we correlate our estimates with agency average PART scores for agencies with at least 3 programs evaluated in a year.[20]

The figure reveals a moderate correlation between the 2020 evaluations of federal executives and our 2020 performance estimates, 0.26 ($p = 0.007$) and 0.24 ($p = 0.04$), respectively. As our

[19] Agencies generated these scores via a response to a series of questions about program planning, management, and results. The Office of Management and Budget reviewed each set of scores.

[20] We have also compared our estimates to average agency PART scores using all agencies (even those with only 1 or 2 programs evaluated) and average agency PART scores using only the most reliable PART scores (i.e., scores for agencies whose federal executives in 2007-8 that reported that their agency's scores picked up real differences in program performance; Gallo and Lewis 2012). The correlations are between 0.24 and 0.25.

performance estimates increase, so does the SFGS performance score of the agency, both its reputational score and the average self-reported performance. There are some notable outliers. For example, the Office of Personnel Management (OPM) and the General Services Administration (GSA) do better on our management performance estimates than the SFGS measures. This may be due to the emphasis that both OPM and GSA place on the surveys used in the management performance estimates. Interestingly, our estimates correlate at 0.72 with agency average response to questions about performance on core mission in 2014. The measures correlate with Bush Administration PART scores at 0.37 ($p < 0.01$).

FIGURE 5: Correlation Between BSEM Performance Estimates Outside Measures

Note: Panels include correlations between our performance estimates and four outside measures: 1) 2020 elite perceptions of agency performance (Richardson, et al. 2023); 2) 2020 weighted agency average self-reports to question "I am confident in the ability of [my agency] to successfully fulfill its core mission." (Piper and Lewis 2023); 3) 2014 weighted agency average self-reports to question "I am confident in the ability of [my agency] to successfully fulfill its core mission" (Richardson 2019); 4) 2002 – 2008 Program Assessment Rating Tool (PART) scores (Gallo and Lewis 2012).

Another unique new source of data comes from a special battery of questions on the 2020 Federal Employee Viewpoint (FEVS) survey. During the COVID-19 pandemic, the Office of Personnel Management included a series of questions about agency performance that were unique to that year's survey. These questions tap into agency performance before the pandemic and during the pandemic and are as follows:

- *Question 1: Prior to the COVID-19 pandemic, my work unit...produced high-quality work.*
- *Question 2: Prior to the COVID-19 pandemic, my work unit...achieved our goals.*
- *Question 3: During the COVID-19 pandemic, my work unit...has produced high quality work.*
- *Question 4: During the COVID-19 pandemic, my work unit...has achieved our goals.*

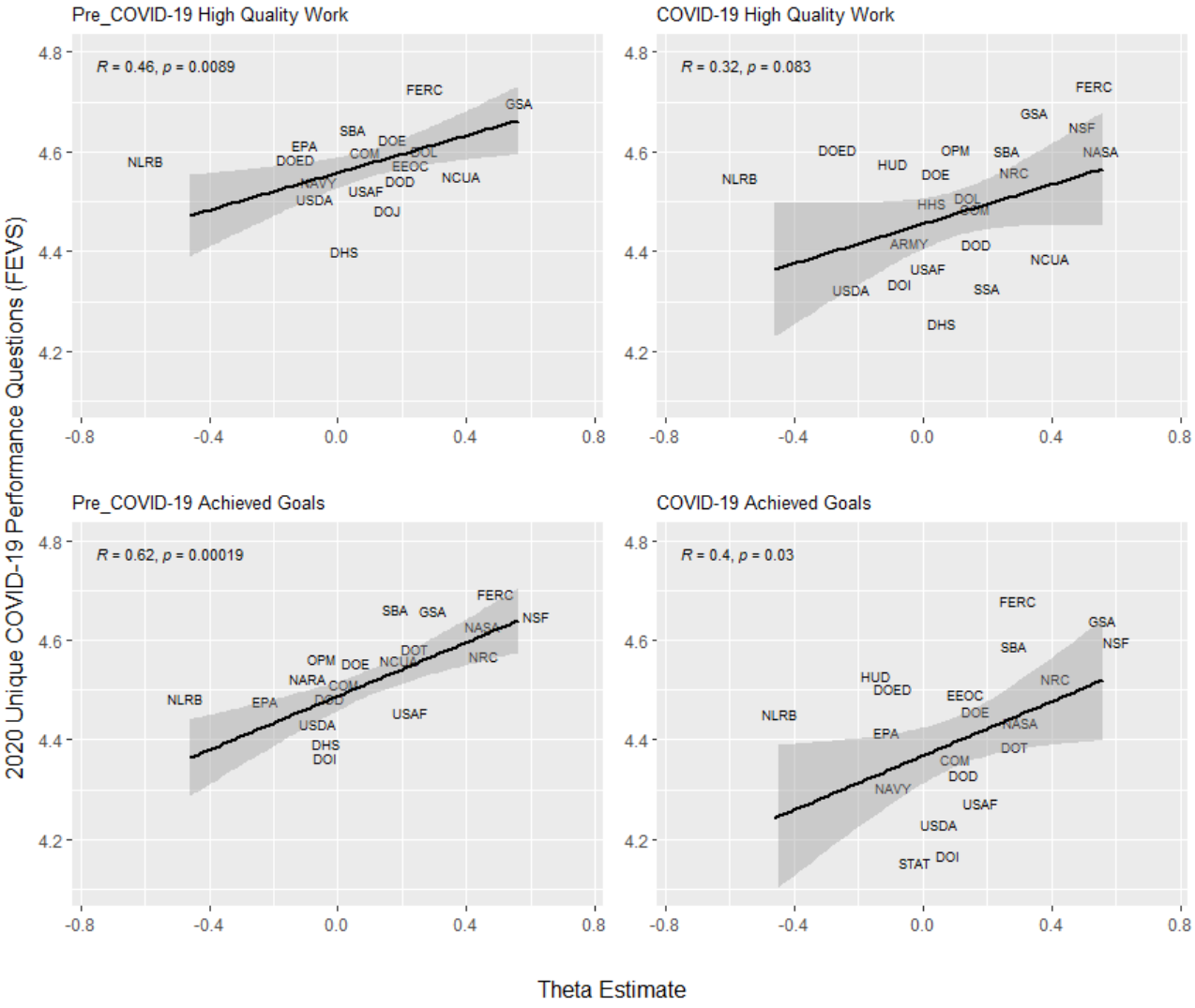
The response categories are *5 "Always"; 4 "Most of the time"; 3 "Sometime"; 2 "Rarely"; 1 "Never"; X "No basis to judge"*. We compare agency average responses to these questions to our estimates from 2020.

When we compare the 2020 performance estimates to the newly added 2020 FEVS questions, the correlations appearing in **Figure 6** are strong, ranging from 0.32 ($p = 0.083$) to 0.62 ($p < 0.001$). The 2020 management performance estimates are a reasonably good predictor of how agencies respond to questions about their performance before and during the COVID-19 pandemic. It is important to note that the agency average responses to the FEVS questions do not vary much, primarily between 4 and 5 on a 5-point scale. Still, what variation that exists, correlates with our estimates. There are fewer consistent outliers and the estimates are tightly organized around a regression line fitted to the data. Notably, the correlations are higher between our estimates and agency assessments of their performance *before* COVID.

In total, despite the variation, the validation results are encouraging for the performance estimates. We would not expect a perfect correlation because both the SFGS data and FEVS provide one way of revealing performance but not the only one. Indeed, the goal of this essay is to propose a method for aggregating data like the SFGS and FEVS data with other objective and subjective data to

produce better insights regarding agency performance measurement. The early internal and external validity of the estimates provides confidence that the approach has promise.

FIGURE 6: Correlation Between 2020 Performance Estimates and 2020 FEVS COVID-19 Questions



Note: Panels include correlations between our performance estimates and four outside measures: 1) Prior to the COVID-19 pandemic, my work unit...produced high-quality work; 2) Prior to the COVID-19 pandemic, my work unit…achieved our goals; 3) During the COVID-19 pandemic, my work unit...has produced high quality work; 4) During the COVID-19 pandemic, my work unit...has achieved our goals.

DISCUSSION

President Biden's management agenda, similar to efforts in many countries, places an important emphasis on performance measurement.[21] It encourages agencies to distill key goals from their missions and measure and report on performance toward those goals. The goals differ by agency and are reported as part of the budget process. While agencies use internal goal setting and performance measurement to compare performance against a historical baseline, agency-specific goals make comparing performance across agencies difficult. Indeed, it is difficult to determine systematically which U.S. federal agencies are performing well and poorly.

As Robert Behn (2003) suggests, decisions about appropriate performance measures should be made with particular purposes in mind—to control, promote, celebrate, etc. The collection of performance information cannot be an end in itself. Rather, it should fulfill the promise of what Moynihan calls "the era of performance management" (Moynihan 2008: 4). Arguably, students of public administration need measures that tap the efficacy of specific programs and the meeting of specific agency goals, and also need a principled way to tell decision makers where they need to focus their attention across the vast executive establishment. Without a principled approach to aggregate performance information, performance data risk being analyzed in a haphazard or selective manner, giving a biased portrait of agency performance.

This paper has attempted to provide a way of aggregating performance information to provide a roadmap for those managers in the executive and legislative branches seeking to improve performance. Perhaps the key difficulty with measuring comparative agency performance is the complexity of the enterprise. Scholars have identified dozens of processes, unclear goals, and different criteria for evaluating performance. No one measure is likely to satisfy all the requirements of an

[21] Donald Kettl, "Why Biden's Management Agenda is a Big Deal," *Government Executive Magazine*, November 19, 2021 (https://www.govexec.com/management/2021/11/why-bidens-presidential-management-agenda-big-deal/186989/).

effective performance measurement regime. The method and measures we propose and evaluate here, however, constitute an important step forward in thinking about how to aggregate different performance information. We have assumed throughout that there is true latent organizational performance, even while acknowledging that there is high and low performance on different tasks and in different parts of the organization. Agencies can also be good on some dimensions and poor on others. That said, while noisy, our method and resulting measures hold out hope for a more robust discussion of ways to aggregate different kinds of performance information—both subjective and objective—and let the data help us arbitrate what is useful and what is not.

The performance estimates we have generated are promising on two levels. First, they exhibit face validity when comparing these estimates to agency reputations. Second, the estimates are robust to alternative model specifications, poor item predictors (e.g., SAMMIES and GAO-High Risk List Programs), the exclusion of small agencies or Defense and military agencies. Finally, the performance estimates exhibit convergent validity with multiple out-of-sample measures, showing reasonable correlation with other one-off measures of organizational performance.

While these estimates are promising, what is perhaps more exciting is how they can be expanded as new and better data emerges and as scholars adopt a similar approach in different contexts. There should be widespread interest, including from the president, but also from governors, legislators, and the public in comparative agency performance. Government agencies implement programs that voters themselves support and have been enacted with the approval of legislative majorities. They provide essential services in including income security, health care, and public safety. At a fundamental level, the efficacy of these services is what governance and elections are about. Better tools can help managers from the president down to advance the efficacy of government and improve accountability.

References

Andersen, Lotte Bøgh Andersen, Andreas Boesen, and Lene Holm Pedersen. 2016. "Performance in Public Organizations: Clarifying the Conceptual Space." *Public Administration Review* 76(6): 852-862.

Andrews, Rhys, George A. Boyne, and Richard M. Walker. 2006. "Subective and Objective Measures of Organizational Performance: An Empirical Exploration." In George A. Boyne, Kenneth J. Meier, Laurence J. O'Toole, Jr., and Richard M. Walker, eds. *Public Service Performance: Perspectives on Measurement and Management* (Cambridge: Cambridge University Press), pp. 14-34.

Asparouhov, Timar, and Bength Muthen. 2021. "Bayesian Analysis of Latent Variable Models Using Mplus." Version 5. September 18, 2021. *Retrieved: October 25, 2023.* https://www.statmodel.com/download/BayesAdvantages18.pdf.

Bednar, Nick, and David E. Lewis. 2024. "Presidential Investment in the Administrative State." *American Political Science Review* 118(1): 442-457.

Behn, Robert D. 2003. "Why Measure Performance? Different Purposes Require Different Measures." *Public Administration Review* 63(5): 586–606.

Bertelli, Anthony M., and Peter John. 2010. "Government Checking Government: How Performance Measures Expand Distributive Politics." *Journal of Politics* 72(2): 545-558.

Bertelli, Anthony M., Dyana P. Mason, Jennifer M. Connolly, and David A. Gastwirth. 2015. "Measuring Agency Attributes with Attitudes Across Time: A Method and Examples Using Large-Scale Federal Surveys." *Journal of Public Administration Research and Theory* 25(2): 513-544.

Boylan, Richard T. 2004. "Salaries, Turnover, and Performance in the Federal Criminal Justice System." *The Journal of Law and Economics* 47(1): 75–92.

Boyne, George A. 2002. "Theme: Local Government: Concepts and Indicators of Local Authority Performance: An Evaluation of the Statutory Frameworks in England and Wales." *Public Money & Management* 22(2): 17-24.

Boyne, George A. 2010. "Performance management: does it work?" in R. Walker and George A. Boyne, eds. *Public Management and Performance: Research Directions* (Cambridge: Cambridge University Press), 207-26.

Boyne, George, and Jay Dahya. 2002. "Executive Succession and the Performance of Public Organizations." *Public Administration* 80(1): 179-200.

Boyne, George A., Kenneth J. Meier, Laurence J. O'Toole Jr., and Richard M. Walker, eds. 2006. *Public Service Performance: Perspectives on Measurement and Management*. Cambridge: Cambridge University Press.

Brewer, Gene A., and Sally Coleman Selden. 2000. "Why Elephants Gallop: Assessing and Predicting Organizational Performance in Federal Agencies." *Journal of Public Administration Research and Theory* 10(4):685-711.

Chun, Young Han, and Hal G. Rainey. 2005. "Goal Ambiguity and Organizational Performance in US Federal Agencies." *Journal of Public Administration Research and Theory* 15(4): 529–557.

Courty, Pascal, and Gerald Marschke. 2011. "Measuring Government Performance: An Overview of Dysfunctional Responses," in James J. Heckman, Carolyn J. Heinrich, Pascal Courty, Gerald

Marschke, and Jeffrey Smith, eds., *The Performance of Performance Standards* (Kalamazoo, MI: W.E. Upjohn Institute for Employment Research), pp. 203-229.

De Ayala, R.J. 2022. *The Theory and Practice of Item Response Theory*. Second Edition. New York: Guilford Press.

Embretson, Susan E., and Steven P. Reise. 2000. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum and Associates.

Fernandez, Sergio, William G. Resh, Tima Moldogaziev, and Zachary W. Oberfield. 2015. "Assessing the Past and Promise of the Federal Employee Viewpoint Survey for Public Management Research: A Research Synthesis." *Public Administration Review* 75(3): 382–94.

Gębczyńska, Alicja, and Renata Brajer-Marczak. 2020. "Review of Selected Performance Measurement Models Used in Public Administration." *Administrative Sciences* 10(4): 99-119.

Gramlich, John. 2017. "Few Americans support cuts to most government programs, including Medicaid," Pew Research, May 26, 2017 (https://www.pewresearch.org/fact-tank/2017/05/26/few-americans-support-cuts-to-most-government-programs-including-medicaid/).

Heinrich, Carolyn J. 2002. "Outcomes–Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness." *Public Administration Review* 62(6): 712-725.

Hubbard, Graham. 2009. "Measuring Organizational Performance: Beyond the Triple Bottom Line." *Business Strategy and the Environment* 18: 177-191.

Kettl, Donald F. 2021. *Politics of the Administrative Process*, 8th ed. Washington, DC: CQ Press.

Krause, George A., and James W. Douglas. 2006. "Does Agency Competition Improve the Quality of Policy Analysis? Evidence from OMB and CBO Fiscal Projections." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 25(1): 53–74.

Krause, George A., David E. Lewis, and James W. Douglas. 2006. "Political Appointments, Civil Service Systems, and Bureaucratic Competence: Organizational Balancing and Executive Branch Revenue Forecasts in the American States." *American Journal of Political Science* 50(3): 770-787.

Krause, George A., and Anne Joseph O'Connell. 2016. "Experiential Learning and Presidential Management of the U.S. Federal Bureaucracy: Logic and Evidence from Agency Leadership Appointments." *American Journal of Political Science* 60(4): 914-931.

Kroll, Alexander, and Donald P. Moynihan. 2021. "Tools of Control? Comparing Congressional and Presidential Performance Management Reforms." *Public Administration Review* 81(4): 599–609.

Lavertu, Stéphane, and Donald P. Moynihan. 2013. "Agency Political Ideology and Reform Implementation: Performance Management in the Bush Administration." *Journal of Public Administration Research and Theory* 23(3): 521–549.

Lee, Soo-Young, and Andrew B. Whitford. 2013. "Assessing the Effects of Organizational Resources on Public Agency Performance: Evidence from the U.S. Federal Government." *Journal of Public Administration Research and Theory* 23(July): 687-712.

Lewis, David E. 2007. "Testing Pendleton's Premise: Do Political Appointees Make Worse Bureaucrats?" *Journal of Politics* 69(4): 1073-1088.

Meier, Kenneth J., Søren C. Winter, Laurence J. O'Toole, Jr., Nathan Favero, Simon Calmar Andersen. 2015. "The Validity of Subjective Performance Measures: School Principals in Texas and Denmark." *Public Administration* 93(4): 1084–1101.

Melkers, Julia, and Katherine Willoughby. 2005. "Models of Performance-Measurement Use in Local Governments: Understanding Budgeting, Communication, and Lasting Effects." *Public Administration Review* 65 (2): 180–190.

Moynihan, Donald P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Washington, DC: Georgetown University Press.

Moynihan, Donald P. 2009. "Through a Glass, Darkly: Understanding the Effects of Performance Regimes." *Public Performance and Management Review* 32(4): 592-603.

Netra, Søren, Sørensen, Peter, and Nejstgaard, Camilla Hansen. 2022. "Does Public Managers' Type of Education Affect Performance in Public Organizations? a Systematic Review." *Public Administration Review* 82(6): 1004–1023.

Niskanen, William A. 1971 [2007]. Bureaucracy & Representative Government. New Brunswick, NJ: Aldine Transaction.

Park, Jungyeon. 2022. "How Individual and Organizational Sources of Managerial Capacity Shape Agency Performance: Evidence from the Size of Improper Payment in U.S. Federal Programs." Essay in the Ph.D. Dissertation *Understanding Negative Performance Management in U.S. Federal Agencies*. University of Georgia. https://esploro.libs.uga.edu/esploro/outputs/9949467728802959.

Piper, Christopher, and David E. Lewis. 2023. "Do Vacancies Hurt Federal Agency Performance?" *Journal of Public Administration Research and Theory* 33(2): 313-328.

Poister, Theodore H. 2003. *Measuring Performance in Public and Nonprofit Organizations*. San Francisco, CA: Jossey-Bass.

Poister, Theodore H., Obed Q. Pasha, and Lauren Hamilton Edwards. 2013 "Does Performance Management Lead to Better Outcomes? Evidence from the U.S. Public Transit Industry." *Public Administration Review* 73(4): 625–636.

Radin, Beryl A. 2000. "The Government Performance and Results Act and the Tradition of Federal Management Reform: Square Pegs in Round Holes?" *Journal of Public Administration Research and Theory* 10(1): 111–135.

Rainey, Hal G., and Barry Bozeman. 2000. "Comparing Public and Private Organizations: Empirical Research and the Power of the A Priori." *Journal of Public Administration Research and Theory* 10(April): 447-469.

Richardson, Mark D. 2019. "Politicization and Expertise: Exit, Effort, and Investment." *Journal of Politics* 81(3): 878-891.

Richardson, Mark D. 2024. "Characterizing Agencies' Political Environments: Partisan Agreement and Disagreement in the U.S. Executive Branch." *Journal of Politics*, forthcoming.

Richardson, Mark D., Joshua D. Clinton, and David E. Lewis. 2018. "Elite Perceptions of Agency Ideology and Workforce Skill." *Journal of Politics* 80(1): 303-307.

Richardson, Mark D., Christopher Piper, and David E. Lewis 2024. "Measuring the Impact of Appointee Vacancies on U.S. Federal Agency Performance." *Journal of Politics*, forthcoming.

Rogger, Daniel, and Christian Schuster, eds. 2023. *The Government Analytics Handbook*. Washington, DC: World Bank.

Rutherford, Amanda. 2016. "The Effect of Top-Management Team Heterogeneity on Performance in Institutions of Higher Education." *Public Performance & Management Review* 40(1): 119–144.

Sanger, Mary Byrna. (2013). "Does Measuring Performance Lead to Better Performance?" *Journal of Policy Analysis and Management*, *32*(1): 185–203.

Smith, Peter C. 2006. "Quantitative Approaches Towards Assessing Organizational Performance," in Boyne et al., eds. *Public Service Performance: Perspectives on Measurement and Management* (Cambridge: Cambridge University Press), pp. 75-91.

Stata Corporation. 2022. *Stata Item Response Theory Reference Manual: Release 18*. College Station, TX: Stata Press.

Thompson, James R., and Michael D. Siciliano. 2021. "The 'Levels' Problem in Assessing Organizational Climate: Evidence From the Federal Employee Viewpoint Survey." *Public Personnel Management* 50(1): 133–156.

Wang, XiaoHu. 2002. "Assessing Performance Measurement Impact: A Study of U.S. Local Governments." *Public Performance & Management Review* 26(1): 26–43.

Wilson, James Q. 1989. *Bureaucracy*. New York: Basic Books.

Wood, Abby K., and David E. Lewis. 2017. "Agency Performance Challenges and Agency Politicization." *Journal of Public Administration Research and Theory* 27(4): 581–595.

Yang, Kaifeng, and Marc Holzer. 2006. "The Performance–Trust Link: Implications for Performance Measurement." *Public Administration Review* 66(January-February): 114-126.

Supplementary Appendix for

"Obtaining Comparable Measures of Agency Performance:

An Application to U.S. Federal Agencies, 2002−2022"

Contents

Appendix A. List of Agencies

OKCODE	Acronym	Name
1	USDA	Department of Agriculture
2	COM	Department of Commerce
3	DOD	Department of Defense
4	ARMY	Department of the Army
5	USAF	Department of the Air Force
6	NAVY	Department of the Navy
7	DOED	Department of Education
8	DOE	Department of Energy
9	HHS	Department of Health and Human Services
11	DHS	Department of Homeland Security
12	HUD	Department of Housing and Urban Development
13	INT	Department of the Interior
14	DOJ	Department of Justice
15	DOL	Department of Labor
16	STAT	Department of State
17	DOT	Department of Transportation
18	TREAS	Department of Treasury
19	DVA	Department of Veterans Affairs
20	CIA	Central Intelligence Agency
21	EPA	Environmental Protection Agency
22	FEMA	Federal Emergency Management Agency (Pre-2003)
23	GSA	General Services Administration
24	NASA	National Aeronautics and Space Administration
25	SBA	Small Business Administration
26	SSA	Social Security Administration
27	USAID	U.S. Agency for International Development
28	USIA/BBG/USAGM	U.S. Agency for Global Media
29	OMB	Office of Management and Budget (in EOP)
30	USTR	Office of the U.S. Trade Representative (in EOP)
33	CSPC	Consumer Product Safety Commission
34	EEOC	Equal Employment Opportunity Commission
35	FCC	Federal Communications Commission
37	FEC	Federal Election Commission
38	FERC	Federal Energy Regulatory Commission
40	FED	Federal Reserve
41	FTC	Federal Trade Commission
43	NLRB	National Labor Relations Board
44	NTSB	National Transportation Safety Board
45	NRC	Nuclear Regulatory Commission

49	SEC	Securities and Exchange Commission
50	CEN	Bureau of the Census (in COMM)
51	CMS	Centers for Medicare and Medicaid Services (in HHS)
52	DEA	Drug Enforcement Administration (in DOJ)
53	FAA	Federal Aviation Administration (in DOT)
54	FDA	Food and Drug Administration (in HHS)
55	FEMA	Federal Emergency Management Agency (in DHS since 2003)
56	IRS	Internal Revenue Service (in TREAS)
57	NHTSA	National Highway Traffic Safety Administration (in DOT)
58	NIH	National Institutes of Health (in HHS)
59	NIST	National Institute of Standards and Technology (in COMM)
60	NOAA	National Oceanic and Atmospheric Administration (in COMM)
61	PTO	Patent and Trademark Office (in COMM)
70	PBGC	Pension Benefit Guarantee Corporation
71	USPS	U.S. Postal Service
72	OPM	Office of Personnel Management
73	OSTP	Office of Science and Technology Policy (in EOP)
78	FDIC	Federal Deposit Insurance Corporation
79	CBP	Customs and Border Protection (in DHS since 2003)
82	BEA	Bureau of Economic Analysis (in COMM)
83	EDA	Economic Development Administration (in COMM)
84	ITA	International Trade Administration (in COMM)
85	CIS	Citizenship and Immigration Services (in DHS since 2003)
86	CISA	Cybersecurity and Infrastructure Agency (in DHS since 2003)
87	ICE	Immigration and Customs Enforcement (in DHS since 2003)
88	TSA	Transportation Security Administration (in DHS since 2003)
89	USCG	U.S. Coast Guard (in DHS since 2003)
90	USSS	U.S. Secret Service (in DHS since 2003)
91	DARPA	Defense Advanced Research Projects Agency (in DOD)
94	DCMA	Defense Contract Management Agency (in DOD)
95	DFAA	Defense Finance and Accounting Service (in DOD)
97	DLA	Defense Logistics Agency (in DOD)
98	JCS	Joint Chief of Staffs (in DOD)
108	IES	Institute of Education Sciences (in DOED)
109	OESE	Office of Elementary and Secondary Education (in DOED)
110	OFSA	Office of Federal Student Aid (in DOED)
111	BOP	Bureau of Prisons (in DOJ)
112	EOUSA	Executive Office of U.S. Attorneys (In DOJ)
113	FBI	Federal Bureau of Investigation (in DOJ)
114	MARSHALS	U.S. Marshals Service (in DOJ)

115	OJP	Office of Justice Programs (in DOJ)
117	BLS	Bureau of Labor Statistics (in DOL)
118	ETA	Employment and Training Administration (in DOL)
119	MSHA	Mine Safety and Health Administration (in DOL)
120	OSHA	Occupational Safety and Health Administration (in DOL)
121	OWCP	Office of Workers Compensation Programs (in DOL)
122	VETS	Veterans Employment and Training Service (in DOL)
123	WHD	Wage and Hour Division (in DOL)
124	FHWA	Federal Highway Administration (in DOT)
125	FMCSA	Federal Motor Carrier Safety Administration (in DOT)
126	FRA	Federal Railroad Administration (in DOT)
127	FTA	Federal Transit Administration (in DOT)
128	MARAD	Maritime Administration (in DOT)
129	NCA	National Cemetery Administration (in DVA)
130	VBA	Veterans Benefits Administration (in DVA)
131	VHA	Veterans Health Administration (in DVA)
134	ONDCP	Office of National Drug Policy (in EOP)
135	ACF	Administration for Children and Families (in HHS)
136	CDC	Centers for Disease Control and Prevention (in HHS)
137	HRSA	Health Resources and Services Administration (in HHS)
138	IHS	Indian Health Service (in HHS)
139	GNMA	Government National Mortgage Association (in HUD)
140	HOU	Office of Housing/Federal Housing Administration (in HUD)
141	OPIH	Office of Public and Indian Housing (in HUD)
143	CFPB	Bureau of Cons Fin Prot/Consumer Financial Protection Bureau
144	CFTC	Commodity Futures Trading Commission
145	CNCS	Corporation for National and Community Service
146	DFC/OPIC	Development Finance Corp/Overseas Private Investment Corp
147	EIB	Export-Import Bank
150	MCC	Millenium Challenge Corporation
151	MSPB	Merit Systems Protection Board
152	NARA	National Archives and Records Administration
154	NSF	National Science Foundation
159	PC	Peace Corps
160	BIA	Bureau of Indian Affairs (in DOI)
161	BLM	Bureau of Land Management (in DOI)
162	BOEM/MMS	Bureau Ocean Energy Management/Minerals Management (in DOI)
163	BOR	Bureau of Reclamation (in DOI)
164	FWS	Fish and Wildlife Service (in DOI)
165	NPS	National Park Service (in DOI)
166	USGS	U.S. Geological Survey (in DOI)

177	OCC	Office of the Comptroller of the Currency (in TREAS)
178	AMS	Agricultural Marketing Service (in USDA)
179	APHIS	Animal and Plant Health Inspection Service (in USDA)
180	ARS	Agricultural Research Service (USDA)
181	ERS	Economic Research Service (in USDA)
182	FAS	Foreign Agricultural Service i(in USDA)
183	FNS	Food and Nutrition Service (In USDA)
184	FS	Forest Service (in USDA)
186	FSIS	Food and Safety Inspection Service (in USDA)
188	NRCS	Natural Resources Conservation Service (in USDA)
193	USCG	U.S. Coast Guard (in DOT pre-2003)
194	INS	Immigration and Naturalization Service (in DOJ)
196	OPE	Office of Postsecondary Education (in DOED)
197	ATF	Bureau of Alcohol, Tobacco, and Firearms (in DOJ)
200	ESA	Employment and Standards Administration (in DOL)
201	ACE	Army Corps of Engineers (in DOD)
202	NCUA	National Credit Union Administration
203	USITC	U.S. International Trade Commission

Appendix B. Raw Subjective and Objective Data Used in BSEM Models

To develop our measures of performance we collected data from a variety of government and non-profit sources, including the General Services Administration (GSA), the Government Accountability Office (GAO), the Merit Systems Protection Board (MSPB), the Office of Management and Budget (OMB), the Office of Personnel Management (OPM), and the Partnership for Public Service. Some of this data is subjective, indicators based upon the perception of persons working in or close to agencies. Other data is objective, presenting counts of good or bad outputs (e.g., presence of award-winning employees).

Subjective Data: Surveys of Employees and Citizens, 2002 - 2022

During the 2002 – 2022 period, the Office of Personnel Management (OPM), Merit Systems Protection Board (MSPB), and General Services Administration (GSA) surveyed federal employees regularly. Several outside groups also conducted federal employee surveys during this period. In total, there are 33 different surveys of federal employees with 28 different performance-related questions. Many questions repeat across surveys and years. **Table B1** lists the surveys, the author of the survey (full description in the note), the number of agencies evaluated, and the number of performance-related questions.

Most prominently, the Office of Personnel Management conducted surveys episodically after its creation in 1978, including a series of surveys as part of the National Performance Review in 1998-2000. Starting in 2002, however, the agency has regularly surveyed hundreds of thousands of government employees at different levels about their agencies. OPM has asked federal supervisors and rank-in-file employees about their agencies, including performance overall, performance on specific tasks, and other features of agency work. The OPM conducted these surveys, originally titled the Federal Human Capital Survey (FHCS) and later Federal Employee Viewpoint Survey (FEVS), every two years until 2010 when they began conducting them annually.

Table B1. Surveys of Federal Employees with Performance Information, 2002-2022

Survey	Source	# Agencies	# Questions
2002	FHCS	49	5
2004	FHCS	59	4
2005	MSPB	57	5
2006	FHCS	109	3
2007	MSPB	61	2
2008	FHCS	106	3
2010	MSPB	59	4
2010	FEVS	107	5
2011	MSPB	60	4
2011	FEVS	109	5
2012	FEVS	95	5
2013	FEVS	96	5
2014	FEVS	77	5
2014	SFGS	114	1
2015	FEVS	75	5
2015	GSA	23	4
2016	MSPB	24	4
2016	FEVS	95	5
2016	GSA	24	4
2017	FEVS	92	5
2017	GSA	24	4
2018	FEVS	94	5
2018	GSA	24	4
2019	FEVS	92	5
2019	GSA	84	4
2020	FEVS	31	8
2020	SFGS	125	4
2020	GSA	79	4
2021	MSPB	53	4
2021	FEVS	30	6
2021	GSA	81	4
2022	FEVS	30	5
2022	GSA	87	4

Note: Survey sources are Office of Personnel Management (OPM): Federal Human Capital Survey (FHCS), Federal Employee Viewpoint Survey (FEVS); Merit Systems Protection Board Survey (MSPB); General Services Administration (GSA) Customer Satisfaction Survey (CSS); Non-profit and Academic Partners: Survey on the Future of Government Service (SFGS).

Since 2003, the Partnership for Public Service (PPS) has used OPM survey data to create a Best Places to Work in Government index.[1] The specific questions they use are the following:

Q43: I recommend my organization as a good place to work. (Q. 43)
Q68: Considering everything, how satisfied are you with your job? (Q. 68)
Q70: Considering everything, how satisfied are you with your organization? (Q. 70)

According to the PPS, "The index score is calculated using a proprietary weighted formula that looks at responses to three different questions in the federal survey. The more the question predicts intent to remain, the higher the weighting."[2] We collected data on all the rankings for agencies in our dataset using data publicly available on the web, including pages captured through the *Wayback Machine* (archive.org), a digital archive of the web.[3] The Partnership also created a 2002 and 2004 Effective Leadership index comprised of answers to 13 different leadership questions on the survey. We also include this measure and include a list of the component questions in **Table B2**.

Table B2. List of Questions Included in Partnership for Public Service Effective Leadership Index, 2002 and 2004

1. Overall, how good a job do you feel is being done by your immediate supervisor/team leader?
2. Supervisors/team leaders in my work unit provide employees with the opportunity to demonstrate their leadership skills
3. Employees have a feeling of personal empowerment and ownership of work processes
4. Discussions with my supervisor/team-leader about my performance are worthwhile
5. I have a high level of respect for my organization's senior leaders
6. In my organization, leaders generate high levels of motivation and commitment in the workforce
7. My organization's leaders maintain high standards of honesty and integrity
8. Complaints, disputes or grievances are resolved fairly in my work unit
9. Arbitrary action, personal favoritism and coercion for partisan political purposes are not tolerated
10. I can disclose a suspected violation of law, rule or regulation without fear of reprisal
11. Supervisors/team leaders in my work unit support employee development
12. Satisfaction with involvement in decisions that affect work

[1] The Partnership first produced the scores in 2003 but used 2002 data to do so. We associate the rankings with the years of the survey.
[2] See 2022 Best Places to Work in the Federal Government Rankings (https://bestplacestowork.org/rankings/about, accessed June 19, 2023). Links to the rankings themselves provides details on the specific questions used.
[3] Given the overlap between Q70 in the index and the individual FEVS question, we do not include Q70 in models including the Best Places to Work scores. Best Places to Work data up to 2019 and after 2020 are not comparable because the way the PPS aggregated positive responses to survey questions changed.

13. Satisfaction with the information received from management on what's going on in the organization

During the 2002 to 2022 period, Merit Systems Protection Board also conducted 6 federal employee surveys: 2005, 2007, 2010, 2011, 2016, and 2021. The samples for these surveys tend to be smaller than OPM surveys but still in the tens of thousands of employees. MSPB's questions focus more on prohibited personnel practices, but the surveys also regularly include performance-related questions. They provide an important source of subjective performance information.

Starting in 2015, the General Services Administration began surveying tens of thousands of high-level federal employees (i.e., GS13-15)[4] about their experiences with the human resources, financial management, acquisitions, and information technology (IT) functions in their agencies. The GSA asks high-level employees about the "quality of support and solutions" they receive in these areas.[5] The questions tap into the internal quality of basic administrative functions within agencies. GSA provides summaries of agency average responses to questions as part of the budget process. We obtained from GSA the average responses (but not the data itself) for 23 agencies for the 2015-2018 period and 79 or more agencies from 2019 – 2022.

Government surveys of federal employees have a number of virtues. First, they have large samples and high response rates.[6] Second, they can be disaggregated to almost all of the agencies on our list.[7] Third, the surveys include a number of performance-related questions asked across time. In

[4] On the standard federal pay scale, the general schedule (GS), grades range from 1 to 15. Only employees working in jobs that could be generally filled by appointees or in specific occupations (adjudication, physicians, etc.) can generally earn more. So, employees in GS13-15 are very senior. The GSA reports this data for 23 executive agencies, including all of the executive departments and the largest independent agencies.

[5] Specifically, GSA asks respondents whether they agree or disagree with the following statement, "I am satisfied with the quality of support and solutions I received from the [*acquisition services, financial management, human resources, IT*] function during the last 12 months." 1-Strongly disagree to 7-Strongly agree.

[6] For example, in 2021, 292,520 federal employees completed the FEVS survey out of 938,638 for a response rate of 33.8 percent. See U.S. Office of Personnel Management. 2021. *Federal Employee Viewpoint Survey Results: Technical Report* (https://www.opm.gov/fevs/reports/technical-reports/technical-report/technical-report/2021/2021-technical-report.pdf, p. 14).

[7] Several agencies have opted out of the FEVS and OPM does not report data on some smaller agencies. For example, the intelligence agencies have never participated. The Department of Veterans Affairs opted out in 2018. Starting in 2020, the OPM significantly reduced the available agency information in the FEVS so that data was no longer available for many

Table B3 we include all a table that lists all the performance related questions by survey and year in order to illustrate the overlap. Finally, the surveys include large enough samples to get reliable agency average responses, including by different categories of employees—executives/ managers and rank-in-file.

In 2014 and 2020 a group of academics, along with non-profit partners, conducted surveys of federal *executives*, generating performance information for 110 - 125 agencies. The surveys include self-reported performance information and information derived from questions asking federal executives to evaluate *other* agencies (Richardson, et al. 2018; Richardson 2019, Richardson, et al. 2023). For the latter type of questions the authors asked respondents to identify the agencies that they worked with most frequently (other than their own). They then asked respondents to evaluate the performance of these agencies on core missions (Richardson, et al. 2018; Richardson, et al. 2023).

Our final subjective measure of performance is a measure of customer satisfaction. In 1994, the National Quality Research Center at the University of Michigan developed the American customer satisfaction index (ACSI). The ACSI uses customer-survey responses to questions about customer expectations, perceived quality, satisfaction, and complaints, tailored to the public sector context, to create an index of public satisfaction with different agencies. Prior to 2011, the ACSI provided one aggregate government index rating. Starting in 2011, however, the ACSI rated as many as 24 different agencies.

smaller agencies and subcomponents. In addition, after 2020, the index is not comparable to earlier indices since the way the PPS aggregated positive responses to survey questions changed.

Table B3. Performance Related Survey Questions for Federal Employees, 1996-2022

Question #	*1996 MSPB*	*1998 NPR*	*1999 NPR*	*2000 NPR*	*2000 MSPB*	*2002 FHCS*	*2004 FHCS*	*2005 MSPB*	*2006 FHCS*	*2007 MSPB*	*2008 FHCS*	*2010 MSPB*	*2010 FEVS*	*2011 MSPB*	*2011 FEVS*	*2012 FEVS*	*2013 FEVS*	*2014 FEVS*
1	x																	
2	x																	
3	x																	
4		x	x	x		x	x		x		x		x		x	x	x	x
5		x	x	x	x	x	x		x		x		x		x	x	x	x
6					x													
7					x			x				x		x				
8					x			x				x		x				
9					x													
10					x													
11					x													
12						x												
13						x	x											
14						x	x		x		x		x		x	x	x	x
15								x										
16								x		x		x	x	x	x	x	x	x
17								x				x		x				
18										x								
19													x		x	x	x	x
20																		
21																		
22																		
23																		
24																		
25																		
26																		
27																		
28																		
29																		
30																		
31																		
32																		

Table B3. Performance Related Survey Questions for Federal Employees, 1996-2022 [continued]

Question #	*2014 SFGS*	*2015 FEVS*	*2015 GSA*	*2016 MSPB*	*2016 FEVS*	*2016 GSA*	*2017 FEVS*	*2017 GSA*	*2018 FEVS*	*2018 GSA*	*2019 FEVS*	*2019 GSA*	*2020 FEVS*	*2020 GSA*	*2020 SFGS*	*2021 MSPB*	*2021 FEVS*	*2021 GSA*	*2022 FEVS*	*2022 GSA*
1																				
2																				
3																				
4		x			x		x		x		x		x				x		x	
5		x			x		x		x		x									
6																				
7				x												x				
8				x												x				
9																				
10																				
11																				
12																				
13																				
14		x			x		x		x		x		x		x		x		x	
15																				
16		x		x	x		x		x		x		x			x	x		x	
17				x												x				
18																				
19		x			x		x		x		x		x				x		x	
20	x																			
21			x			x		x		x		x		x				x		x
22			x			x		x		x		x		x				x		x
23			x			x		x		x		x		x				x		x
24			x			x		x		x		x		x				x		x
25													x							
26													x							
27													x							
28													x							
29															x					
30															x					
31																	x		x	
32																	x			

Table B3. Performance Related Survey Questions for Federal Employees, 1996-2022 [continued]

Question #	Question Wording
1	A private sector company could perform the work of my organization just as effectively as government does.
2	The work performed by my work unit provides the public a worthwhile return on their tax dollars
3	Overall, how would you rate the quality of the work performed by: Your current coworkers in your immediate work group
4	Overall, how good a job do you feel is being done by your immediate supervisor
5	How would you rate the overall quality of work being done in your work group/by your work unit?
6	Overall, how would you rate the quality of work performed by: the larger organization that includes your work unit?
7	Overall, I am satisfied with my supervisor
8	Overall, I am satisfied with managers above my immediate supervisor
9	A private sector company could perform just as effectively as my work
10	Overall productivity of: Your work unit
11	Overall productivity of: Your organization
12	I believe my organization can perform its function as effectively as any private sector provider.
13	How would you rate your organization as an organization to work for compared to other organizations?
14	Considering everything, how would you rate your overall satisfaction in your organization? In 2002 includes "at the present time"?
15	My agency produces high quality products and services
16	My agency/organization is successful in accomplishing its mission
17	My work unit produces high quality products and services
18	Overall, how would you rate your immediate supervisor's performance as a supervisor?
19	Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor/team leader?
20	I am confident in the ability of [my agency] to successfully fulfill its core mission
21	I am satisifed with the quality of support and solutions I received from the acquisition services function during the last 12 months
22	I am satisifed with the quality of support and solutions I received from the financial management function during the last 12 months
23	I am satisifed with the quality of support and solutions I received from the human resources function during the last 12 months
24	I am satisifed with the quality of support and solutions I received from the IT function during the last 12 months
25	Prior to the COVID-19 pandemic, my work unit... Produced high quality work[2020 only]
26	Prior to the COVID-19 pandemic, my work unit…achieved our goals [2020 only]
27	During the COVID-19 pandemic, my work unit… has produced high quality work [2020 only]
28	During the COVID-19 pandemic, my work unit… has achieved our goals [2020 only]
29	How would you rate the overall performance of [your agency] in carrying out its mission?"
30	[My agency] is an effectively managed, well-run organization.
31	Employees in my work unit produce high-quality work
32	Employees in my work unit achieve our goals

Objective Data: GAO Reports, PART Scores, and Employee Awards Data

To add objective data, we collected data from the GAO's high-risk list.[8] Starting in 1990, the GAO began publishing a self-initiated report on government activities they considered high risk. The GAO defines high risk as areas of significant weakness in government activities or programs, particularly if the activities involve substantial resources or provide critical services.[9] Since its initial publication, GAO published a report in 1992 and then has published the list once every Congress (i.e., every two years) starting in 1995. The list includes programs specific to individual agencies (e.g., the prison system, flood insurance) or activities that span many agencies (e.g., human capital management). Some agencies have several programs on the list and some have none.[10] Some agencies, often with the help of Congress or the administration, have been successful responding to the GAO's concerns and have succeeded in getting their programs off the high-risk list. The list provides a cross-agency and temporal source of information about agencies that regularly do well or poorly.[11]

To supplement this data, we collected data on counts of GAO reports from 1990-2020 that resulted from bipartisan requests for GAO investigations.[12] Each Congress, members request hundreds of GAO investigations of federal activities. These requests come from individual members or groups of members, on and off the committees with jurisdiction. We organize counts of the number of reports by agency year, limiting the relevant data to investigations requested by members from both parties as a measure of performance. We do so on the assumption that bipartisan requests likely reflect real performance concerns, rather than simple efforts to discredit the presidential administration. Of

[8] The GAO is a non-partisan legislative branch agency in the United States responsible for auditing, evaluating and investigating government agencies.
[9] This description is based on GAO's own description of the program (https://www.gao.gov/high-risk-list).
[10] Among the 139 agencies in our dataset, excluding government-wide programs, 63 agencies had programs on the high-risk list. It is difficult to determine whether agencies never on the list are omitted because they were performing well or because GAO never considered them worthy of evaluation. Thus, agencies never on the list are treated as missing data.
[11] We assume that programs on the list in consecutive two-year periods were on the list in the year between publication of the list. If a program dropped off the list between publication of the lists, we assume the program was on the list until the publication of the new list where it was absent.
[12] We thank Cody Drolc for providing us with this data.

the 139 agencies in our data, 126 have been the subject of a GAO investigation and some more than 300 for a given year.

During the George W. Bush Administration, the Office of Management and Budget (OMB) collected systematic performance information on federal programs. The OMB used the Program Assessment Rating Tool (PART) to evaluate program performance. Between 2002 and 2008, the Bush Administration graded 1,016 federal programs on a scale from 0 to 100. We calculate agency year average PART scores as a measure of performance. This provides data on 120 agencies, with agency average ratings varying between 34 and 93.[13]

We also calculate agency year averages using only scores for agencies where federal executives reported that the scores were somewhat effective at disentangling performance. Specifically, we use data from a 2007-8 survey of federal executives. The survey asked federal executives "*To what extent did the PART pick up real differences in program performance among programs in your agency?*" [Almost always reflected real differences (2.62%), generally reflected real differences (14.94%), sometimes reflected real differences (26.58%), rarely reflected real differences (22.70%), PART scores have no connection to real performance (14.18%), don't know (18.99%)]. We calculate agency year averages for agencies where more than half reported that PART scores almost always, generally, or sometimes reflect real differences among programs in their agencies. This provides data on 611 programs and 70 agencies overall. This works out to data on 15 and 46 agencies per year, depending upon the number of programs evaluated.

We also make use of government and non-profit data on agencies with employees winning awards. Agencies that regularly produce award winning employees are also seeing improvements in programs or efficiency since these criteria determine employee awards. We obtained government

[13] We also calculate agency year averages using only scores for agencies where federal executives reported that the scores were somewhat effective at disentangling performance (Gallo and Lewis 2012). We include full details in **Appendix B**.

employee performance award data from the Office of Personnel Management (OPM) for four types of awards: high performance award—rating based (2000 – 2022)[14], high performance award—not rating based (2003 to 2022), individual suggestion/invention award (2000 to 2022)[15], and quality step increases (1990 to 2022).[16]

Each year since 2001, the Partnership for Public Service has awarded dozens of federal employees Samuel J. Heyman Service to America Medals (also known as "SAMMIES"). In total, more than 700 federal employees working across the executive branch have been awarded this prize. These awards recognize extraordinary agency leadership that resulted in high agency performance—effective program implementation, unusual innovation, and effective responses to complex problems. Nominees are evaluated based upon the significance and impact of the candidate, how well they foster innovation, their demonstrated leadership, and the extent to which they embody excellence in public service.[17] In a given year, agencies have had up to four employees as finalists for performance awards in different areas and agencies have had up to 3 employees win awards for a given year. Among the agencies with the most nominees and winners across this period are the Departments of Commerce,

[14] These agency awards are based upon high performance ratings that effectively distinguish performance among employees. Agencies can also give cash awards unconnected to ratings for special actions or service to employees that "contribute to the efficiency, economy, or other improvement of government operations." (https://www.opm.gov/combined-federal-campaign/running-a-local-campaign/running-a-local-campaign/awards-and-recognition/).

[15] As described by on agency, these awards are "lump-sum cash payments (minus applicable taxes) that recognize individuals or groups who adopt and implement written suggestions or develop inventions that significantly improve the efficiency or effectiveness of Government operations, and that support or enhance accomplishment of strategic plan or mission goals and objectives of the agency, Department, or Federal Government." (https://directives.sc.egov.usda.gov/RollupViewer.aspx?hid=17055).

[16] According to OPM, a quality step increase is "an additional within-grade increase (WGI) used to recognize and reward General Schedule (GS) employees at any grade level who display outstanding performance. A QSI has the effect of moving an employee through the GS pay range faster than by periodic step increases alone." (https://www.opm.gov/policy-data-oversight/pay-leave/pay-administration/fact-sheets/quality-step-increase/).

[17] This is drawn more or less directly from the Partnership for Public Service website about the awards (https://servicetoamericamedals.org/about/selection-process-and-committee). There is also a category for lifetime achievement. We exclude lifetime achievement award winners since their award is not for performance in a specific year, or even necessarily a specific agency.

Defense, and Health and Human Services. Some have never had a winner, including agencies like the Department of Education and the National Labor Relations Board.

Appendix C. Comprehensive Listing of Agency Management Performance Estimates from BSEM Model 1

Table C1: Raw Data and Estimates, with Missing Data (2022)

	Bayesian Estimates		Government Surveys OPM, MSPB					PPS		GSA				OPM Personnel Data				GAO	
Name	Post Median	Post. SD	Ag. Mission	Qual. Work Unit	Org Comp Others	Satis Sup	Satis Sup Abov	BPTW Post-2019	Eff. Lead	GSA Acq.	GSA FM	GSA HC	GSA IT	OPM Innov.	OPM Cash Rating	OPM Cash No Rating	OPM Qual. Step	GAO High Risk	Bipart Leg Req. GAO
USDA	-0.12	0.08	3.85					62.00		4.28	4.74	4.33	4.82	24	1853	101936	780	1	
COM	0.12	0.08	4.10					70.60		4.87	5.04	4.54	5.53	0	25341	43015	392	1	
DOD	0.02	0.08	4.02					64.43		4.74	5.05	4.66	4.27	1204	450732	287823	23547	8	
ARMY	0.03	0.08	4.02					63.80		4.70	5.11	4.71	4.28	370	125683	73449	11651		
USAF	0.09	0.08	4.08					65.20		4.97	5.23	4.66	4.43	1	108086	27429	5177		
NAVY	-0.07	0.08	3.96					61.80		4.47	4.84	4.54	3.91	772	147120	131068	3602		
DOED	0.08	0.08	4.00					68.30		4.68	5.36	4.42	5.51	0	2565	1271	194	1	
DOE	0.21	0.08	4.19					73.70		5.13	5.32	4.47	5.49	0	9207	8551	805	4	
HHS	0.14	0.08	4.12					74.30		4.65	5.11	4.55	5.58	0	56394	23671	6602	11	
DHS	-0.11	0.08	3.72					54.90		4.78	4.84	4.50	5.39	22	104504	218055	1947	4	
HUD	0.05	0.08	4.02					69.50		3.91	5.10	4.98	5.64	2	5187	6678	352	1	
DOI	-0.05	0.08	3.83					65.20		4.70	5.07	4.27	5.27	2	39827	25376	2457	6	
DOJ	-0.05	0.08	3.74					55.30		4.91	5.14	4.76	5.39	34	34731	52987	8634	3	
DOL	0.21	0.08	4.09					68.50		5.12	5.30	5.20	5.74	254	10596	5036	739	1	
STAT	-0.04	0.08	3.88					61.80		4.69	5.06	4.45	4.86	0	5348	0	382	2	
DOT	0.14	0.08	4.04					68.30		5.06	5.26	4.92	5.37	0	7965	19844	329	2	
TREAS	0.07	0.08	3.91					67.20		5.00	5.31	4.91	5.07	112	52958	63792	3231	6	
DVA	-0.01	0.11						68.40		4.58	5.19	3.93	5.55	0	236053	65401	1201	5	
EPA	0.09	0.08	4.04					75.60		4.46	5.12	4.43	5.75	0	284	16621	375	1	
GSA	0.42	0.08	4.29					81.00		5.62	5.80	5.48	5.81	0	8611	2867	122	1	
NASA	0.35	0.11						84.30		5.26	5.53	5.24	5.53	1	14405	5265	674	2	
SBA	0.29	0.08	4.18					76.60		5.34	5.19	5.35	5.92	1	4308	836	155	1	
SSA	-0.01	0.08	3.68					53.90		5.11	5.40	5.08	5.47	0	27720	39619	1387	1	

USAID	-0.03	0.10	3.89					66.10						0	2813	2532	179		
USAGM	-0.06	0.18						62.80						0	0	0			
OMB	0.03	0.18						69.70						0	0	0		2	
USTR	-0.11	0.19						57.90						0	0	0		1	
CPSC	-0.06	0.19						63.20						0	0	0			
EEOC	0.08	0.10	4.03					71.70						0	0	0			
FCC	0.07	0.19						73.20						0	0	0			
FEC	-0.07	0.19						61.60						0	0	0			
FERC	0.31	0.10	4.30					80.30						0	29	1522	75		
FED	0.02	0.21												0	0	0		2	
FTC	0.02	0.10	3.96					67.30						0	0	0			
NLRB	-0.11	0.18						59.70						0	0	0			
NTSB	0.02	0.20						70.90						0	0	0			
NRC	0.18	0.08	4.14					66.50		5.53	5.33	4.02	5.86	0	2097	776	55		
SEC	0.17	0.19						82.20						0	0	0		1	
CEN	0.04	0.10						69.70		4.79	4.99	4.32	5.62	0	10674	4001	3	0	
CMS	0.41	0.11						79.20		5.52	5.66	5.32	5.82	0	4551	3447	247	3	
DEA	0.11	0.11						68.00		4.98	5.19	4.56	5.42	0	5452	3017	229	1	
FAA	0.15	0.11						67.60		5.08	5.19	4.88	5.35	0	4	15990	82	0	
FDA	0.04	0.11						77.70		4.61	4.75	4.54	5.62	0	14668	5710	790	4	
IRS	0.13	0.10						66.30		4.97	5.31	4.88	5.00	17	46498	50903	2958	3	
NHTSA	-0.04	0.11						69.10		4.09	4.73	5.00	5.14	0	480	242	29		
NIH	0.28	0.11						80.00		4.82	5.49	5.13	5.84	0	14751	3688	1743	2	
NIST	0.13	0.10						75.70		4.72	5.11	4.70	5.87	0	0	0			
NOAA	0.02	0.10						68.80		4.79	4.87	4.43	5.42	0	5885	8431	173	0	
PTO	0.31	0.11						72.00		5.16	5.53	5.30	5.68	0	3214	28708	130		
PBGC	0.23	0.18						87.60						0	0	0		0	
USPS	0.01	0.21												0	0	0		2	
OPM	0.18	0.08	4.03					71.20		5.31	5.11	5.25	5.51	0	1841	1527	84	1	
OSTP	0.11	0.22												0	0	0		1	

FDIC	0.01	0.18						68.50						0	0	0		2	
USCBP	-0.01	0.10						51.20		4.82	4.90	4.81	5.20	0	42219	14954	273		
BEA	-0.01	0.21												0	0	0			
EDA	0.02	0.18						69.90						0	219	75	8		
ITA	-0.08	0.10						69.90		4.68	4.82	4.11	4.77	0	1029	647	58		
CIS	0.17	0.11						69.20		4.94	5.13	4.97	5.84	0	14082	10189	258		
CISA	-0.17	0.11						65.20		4.62	4.58	3.71	5.33	0	1940	1824	51		
ICE	-0.08	0.10						52.00		4.89	4.83	4.01	5.67	0	18450	820	131		
TSA	-0.28	0.10						45.20		4.49	4.41	4.08	4.74	0	0	0			
USCG	-0.35	0.11						71.80		4.17	3.69	4.27	4.34	1	5589	2987	255		
USSS	-0.04	0.11						58.40		4.66	4.77	4.58	5.31	0	4703	1718	320		
DARPA	0.04	0.21												0	133	9	13		
DCMA	0.04	0.19						72.10						0	6685	3947	80		
DFAA	0.11	0.18						77.10						16	6776	12578	42	1	
DLA	0.05	0.19						71.10						45	20505	11200	909		
JCS	-0.23	0.11						57.30		4.53	4.60	4.22	4.08	0	695	46	39		
IES	0.04	0.19						72.10						0	97	15	12		
OESE	0.05	0.18						70.30						0	150	20	37		
OFSA	0.07	0.10						63.70		4.64	5.73	3.89	5.57	0	979	449	41	0	
BOP	-0.10	0.11						35.50		4.84	4.77	4.68	5.17	0	10524	11503	6677	1	
EOUSA	0.45	0.11						74.00		5.45	5.94	5.39	5.87	0	4308	2939	655		
FBI	-0.23	0.11						57.20		4.11	4.47	4.09	5.12	0	0	0			
MARSHALS	0.10	0.11						66.50		4.79	5.14	4.69	5.42	0	3431	1911	319	0	
OJP	0.41	0.11						73.00		5.25	6.00	5.64	5.07	0	392	218	45		
BLS	0.36	0.10						80.80		5.25	5.48	5.35	5.63	38	1609	750	56		
ETA	0.38	0.11						70.00		5.68	5.59	5.14	6.00	13	815	214	34	1	
MSHA	0.01	0.11						53.20		4.87	4.88	4.81	5.42	0	1408	79	26		
OSHA	0.32	0.11						71.20		5.26	5.49	5.35	5.72	0	1454	1073	81		
OWCP	0.19	0.11						62.10		4.75	5.31	5.23	5.88	0	902	455	114		
VETS	0.01	0.19						70.10						11	200	140	7		

WHD	-0.02	0.11						61.90		4.33	4.76	4.75	5.72	126	1084	412	168		
FHWA	0.40	0.11						79.40		5.33	5.64	5.56	5.55	0	2424	1723	46		
FMCSA	0.16	0.10						68.40		5.29	5.09	4.90	5.15	0	988	239	19		
FRA	0.27	0.11						69.50		5.50	5.45	4.78	5.66	0	798	1140	26		
FTA	0.22	0.10						74.20		4.96	5.41	4.88	5.66	0	495	199	12	0	
MARAD	-0.06	0.10						62.20		4.50	5.17	4.13	4.95	0	651	0	17		
NCA	0.19	0.11						73.70		4.57	5.02	5.64	5.48	0	593	2369	5		
VBA	0.19	0.11						67.30		5.03	5.48	4.73	5.54	0	9009	6690	336		
VHA	-0.07	0.10						68.20		4.45	5.15	3.73	5.54	0	217999	52699	657	1	
ONDCP	0.01	0.21												0	0	0		1	
ACF	-0.22	0.10						68.60		3.87	4.34	4.13	5.58	0	1240	754	30		
CDC	0.06	0.10						72.70		4.81	5.13	4.28	5.55	0	8683	3820	1830	1	
HRSA	0.25	0.11						80.80		4.63	5.36	5.25	5.76	0	1811	321	217		
IHS	-0.20	0.11						62.80		3.98	4.87	3.93	5.06	0	5217	2608	1281	1	
GNMA	-0.06	0.19						63.60						0	118	154	8		
HOU	0.09	0.19						75.50						1	725	990	55	0	
OPIH	-0.01	0.19						67.00						1	458	253	38		
CFPB	-0.02	0.18						66.00						0	0	0		1	
CFTC	-0.05	0.18						64.80						0	0	0		1	
CNCS	-0.01	0.18						66.10						0	0	0			
DFC	0.07	0.18						74.10						0	0	0			
EIB	-0.17	0.19						55.00						0	0	0			
MCC	0.03	0.19						70.10						0	0	0			
MSPB	0.02	0.19						70.00						0	0	0			
NARA	0.00	0.10	3.94					66.20						0	0	0			
NSF	0.49	0.08	4.46					82.80		5.66	5.43	5.57	6.09	0	1077	1631	108		
PC	0.06	0.18						72.20						0	0	0			
BIA	-0.07	0.11						57.90		4.68	5.22	3.89	5.17	0	2272	650	67	1	
BLM	-0.20	0.11						61.40		4.05	4.96	3.65	5.17	0	6370	7047	201	2	
BOEM	0.39	0.11						78.30		4.92	6.10	5.41	5.34	0	424	95	85	2	

BOR	0.17	0.11						72.50		4.85	5.35	4.99	5.21	0	4023	1828	379		
FWS	-0.11	0.10						70.60		4.24	4.60	4.23	5.49	0	6374	3541	442		
NPS	-0.03	0.11						59.00		4.88	5.15	4.10	4.98	0	10258	8128	246		
USGS	-0.10	0.11						70.80		4.76	4.98	3.35	5.47	2	5749	1428	606		
OCC	0.02	0.18						69.60						0	0	0		2	
AMS	-0.12	0.11								4.42	4.75	4.52	4.43	0	0	1718	12		
APHIS	0.00	0.22												0	4	4152	33		
ARS	-0.27	0.11								3.91	4.37	4.32	4.68	0	43	5176	108		
ERS	-0.01	0.21												0	2	243	5		
FAS	-0.64	0.11								3.46	4.35	2.10	4.61	0	79	876	6		
FNS	-0.05	0.11								4.64	5.49	4.12	3.99	0	507	1163	13		
FS	-0.26	0.10						54.10		4.06	4.59	4.05	4.93	0	18	43247	238	0	
FSIS	0.01	0.11						65.10		4.58	5.29	4.61	4.59	0	3	21930	128	1	
NRCS	-0.01	0.11								4.83	5.07	4.35	4.78	4	5	8919	20		
OPE	0.12	0.18						77.50						0	134	52	13		
ATF	-0.04	0.10						68.50		4.40	5.02	4.26	5.31	34	2752	1562	131		
MINT	0.33	0.11						62.50		5.80	5.67	5.00	5.54	85	0	1941	25		
TTTB	0.19	0.19						84.10						1	414	112	11		
NCUA	0.14	0.10	4.10											0	0	0		1	
USITC	0.10	0.19						76.60						0	0	0			

Note: Empty cells represent missing data. Different raw data is available in different years (e.g., agency average survey results). We omit one column of data for the Best Places to Work Ranking before 2020 for simplicity but the rankings before and after that point are not comparable because a change in methodology.

Table C2: Summary Performance by Agency, 2002-2022

			Bayesian Estimates				Performance Class	Performance Years				
Agency	Dept	ID	Post. Mean	Post. SE	LCL 95	UCL 95	Average Rank	Low Count [Bottom Quintile]	Low Mod. Count [2nd Quintile]	Mod. Count [3rd Quintile]	Mod. High Count [4th Quintile]	High Count [Top Quintile]
Department of Agriculture	USDA	1	-0.095	0.057	-0.209	0.020	Low-Moderate	6	9	1	1	0
Department of Commerce	COM	2	0.081	0.058	-0.030	0.196	Moderate-High	0	0	3	13	1
Department of Defense	DOD	3	-0.007	0.057	-0.119	0.106	Moderate	0	8	6	2	1
Department of the Army	DOD	4	-0.004	0.059	-0.122	0.111	Moderate	1	6	5	4	1
U.S. Air Force	DOD	5	0.024	0.059	-0.091	0.141	Moderate-High	0	7	3	5	2
Department of the Navy	DOD	6	-0.003	0.060	-0.118	0.118	Moderate	0	7	5	4	1
Department of Education	DOED	7	-0.122	0.058	-0.231	-0.006	Low	8	7	0	2	0
Department of Energy	DOE	8	0.027	0.058	-0.088	0.139	Moderate-High	1	8	1	4	3
Dept of Health & Human Services	HHS	9	0.023	0.059	-0.093	0.138	Moderate-High	0	8	3	4	2
Department of Homeland Security	DHS	11	-0.215	0.059	-0.333	-0.102	Low	12	3	0	1	0
Dept of Housing & Urban Develop.	HUD	12	-0.139	0.058	-0.252	-0.027	Low	10	4	1	1	1
Department of the Interior	INT	13	-0.067	0.057	-0.182	0.044	Low-Moderate	3	10	3	1	0
Department of Justice	DOJ	14	0.044	0.057	-0.071	0.153	Moderate-High	0	2	3	10	2
Department of Labor	DOL	15	0.005	0.058	-0.109	0.119	Moderate	1	8	2	4	2
Department of State	STAT	16	0.060	0.058	-0.053	0.173	Moderate-High	0	3	3	8	3
Department of Transportation	DOT	17	-0.003	0.057	-0.117	0.110	Moderate	3	4	3	4	3
Department of the Treasury	TREAS	18	0.009	0.057	-0.105	0.120	Moderate	0	7	4	5	1
Department of Veterans Affairs	DVA	19	-0.085	0.062	-0.207	0.039	Low-Moderate	5	6	4	2	0
Environmental Protection Agency	IND	21	-0.030	0.058	-0.145	0.082	Low-Moderate	5	3	3	5	1
Fed Emergency Management Agency	IND	22	-0.192	0.039	-0.282	-0.120	Low	1	0	0	0	0
General Services Administration	IND	23	0.144	0.058	0.030	0.258	High	1	1	4	4	7
National Aeronautics & Space Admin.	IND	24	0.289	0.059	0.173	0.405	High	0	0	0	0	17
Small Business Administration	IND	25	-0.063	0.073	-0.207	0.078	Low-Moderate	6	6	0	2	3

Social Security Administration	IND	26	0.020	0.058	-0.099	0.130	Moderate	0	5	7	3	2
U.S. Agency for Intl Development	IND	27	-0.040	0.067	-0.170	0.092	Low-Moderate	5	6	0	5	1
U.S. Agency for Global Media	IND	28	-0.307	0.094	-0.495	-0.127	Low	14	2	1	0	0
Office of Management & Budget	EOP	29	0.150	0.092	-0.034	0.330	High	2	1	0	4	10
Office of the United States Trade Rep.	EOP	30	-0.104	0.115	-0.331	0.126	Low-Moderate	5	6	4	0	2
Consumer Product Safety Commission	IND	33	0.003	0.105	-0.202	0.212	Moderate	0	7	4	5	1
Equal Employment Opportunity Com.	IND	34	-0.029	0.083	-0.193	0.134	Low-Moderate	4	5	3	4	1
Federal Communications Commission	IND	35	0.025	0.107	-0.183	0.237	Moderate-High	1	3	6	6	1
Federal Election Commission	IND	37	-0.298	0.113	-0.525	-0.078	Low	12	3	2	0	0
Federal Energy Regulatory Commission	IND	38	0.277	0.077	0.126	0.428	High	0	0	0	3	10
Federal Reserve Board	IND	40	-0.004	0.213	-0.419	0.413	Moderate	0	2	10	0	0
Federal Trade Commission	IND	41	0.267	0.093	0.082	0.448	High	0	1	2	0	14
National Labor Relations Board	IND	43	-0.085	0.093	-0.266	0.096	Low-Moderate	4	6	6	1	0
National Transportation Safety Board	IND	44	0.140	0.110	-0.079	0.350	High	0	0	4	5	8
Nuclear Regulatory Commission	IND	45	0.244	0.074	0.097	0.390	High	0	0	1	0	16
Securities & Exchange Commission	IND	49	0.098	0.086	-0.070	0.263	High	3	1	3	3	7
Bureau of the Census	COM	50	0.038	0.064	-0.087	0.165	Moderate-High	0	5	4	5	3
Ctrs. for Medicare & Medicaid Services	HHS	51	0.063	0.077	-0.087	0.213	Moderate-High	5	1	1	3	7
Drug Enforcement Administration	DOJ	52	0.121	0.074	-0.021	0.268	High	0	1	1	7	8
Federal Aviation Administration	DOT	53	-0.008	0.067	-0.140	0.121	Moderate	3	6	1	3	4
Food & Drug Administration	HHS	54	0.079	0.066	-0.054	0.207	Moderate-High	0	2	3	8	4
Internal Revenue Service	TREAS	56	-0.031	0.063	-0.157	0.090	Low-Moderate	4	7	2	3	1
Nat. Highway Traffic Safety Admin.	DOT	57	-0.098	0.135	-0.364	0.165	Low-Moderate	6	4	7	0	0
National Institutes of Health	HHS	58	0.171	0.066	0.039	0.300	High	0	0	1	7	9
Nat. Institutes of Standards & Technology	COM	59	0.160	0.075	0.014	0.310	High	0	0	0	6	11
Nat. Oceanic & Atmospheric Admin.	COM	60	0.041	0.063	-0.084	0.166	Moderate-High	1	3	4	7	2
Patent & Trademark Office	COM	61	0.186	0.064	0.058	0.309	High	1	2	0	3	11
Pension Benefit Guarantee Corporation	IND	70	0.100	0.107	-0.113	0.306	High	0	3	5	2	7
U.S. Postal Service	IND	71	-0.002	0.210	-0.415	0.406	Moderate	0	2	15	0	0

Office of Personnel Management	IND	72	0.035	0.058	-0.077	0.151	Moderate-High	1	4	3	7	2
Office of Science & Technology Policy	EOP	73	0.075	0.212	-0.337	0.488	Moderate-High	0	0	2	15	0
Federal Deposit Insurance Corporation	IND	78	0.213	0.102	0.012	0.414	High	0	2	3	0	12
U.S. Customs & Border Protection	DHS	79	-0.257	0.066	-0.388	-0.128	Low	11	1	1	3	0
Bureau of Economic Analysis	COM	82	0.012	0.198	-0.379	0.392	Moderate	0	0	19	1	1
Economic Development Admin.	COM	83	-0.104	0.150	-0.392	0.193	Low-Moderate	4	1	7	4	1
International Trade Administration	COM	84	-0.111	0.078	-0.263	0.039	Low-Moderate	7	8	1	1	0
Citizenship & Immigration Services	DHS	85	0.050	0.074	-0.095	0.193	Moderate-High	0	5	3	3	5
Cybersecurity & Infrastructure Agency	DHS	86	-0.125	0.135	-0.390	0.140	Low	2	1	1	0	0
Immigration & Customs Enforcement	DHS	87	-0.281	0.076	-0.431	-0.133	Low	12	1	2	1	0
Transportation Security Administration	DHS	88	-0.302	0.077	-0.455	-0.151	Low	13	1	1	1	0
U.S. Coast Guard	DHS	89	0.073	0.069	-0.060	0.207	Moderate-High	2	2	0	4	8
U.S. Secret Service	DHS	90	-0.143	0.080	-0.301	0.012	Low	5	7	2	2	0
Def. Adv. Research Projects Agency	DOD	91	0.005	0.211	-0.413	0.417	Moderate	0	0	16	1	0
Defense Contract Management Agency	DOD	94	-0.110	0.092	-0.288	0.071	Low-Moderate	6	6	1	4	0
Defense Finance & Accounting Service	DOD	95	-0.016	0.094	-0.199	0.167	Moderate	3	5	4	5	0
Defense Logistics Agency	DOD	97	0.044	0.090	-0.137	0.220	Moderate-High	0	3	3	11	0
Joint Chiefs of Staff	DOD	98	-0.003	0.163	-0.328	0.314	Moderate	2	2	4	9	0
Institute of Education Sciences	DOED	108	-0.053	0.153	-0.350	0.247	Low-Moderate	5	0	10	1	1
Office of Elementary & Secondary Ed.	DOED	109	-0.203	0.118	-0.435	0.025	Low	10	3	3	1	0
Office of Federal Student Aid	DOED	110	-0.116	0.081	-0.276	0.042	Low	8	2	5	2	0
Bureau of Prisons	DOJ	111	-0.114	0.068	-0.249	0.021	Low	7	7	2	0	1
Executive Office of the U.S. Attorneys	DOJ	112	0.272	0.070	0.136	0.408	High	0	1	0	0	16
Federal Bureau of Investigation	DOJ	113	0.080	0.085	-0.089	0.244	Moderate-High	1	0	3	7	6
U.S. Marshals Service	DOJ	114	0.082	0.072	-0.063	0.219	Moderate-High	1	2	0	9	5
Office of Justice Programs	DOJ	115	0.013	0.104	-0.193	0.214	Moderate	4	4	3	3	3
Bureau of Labor Statistics	DOL	117	0.157	0.073	0.013	0.300	High	0	1	2	4	10
Employment & Training Admin.	DOL	118	-0.123	0.091	-0.305	0.056	Low	10	3	2	1	1
Mine Safety & Health Administration	DOL	119	0.009	0.079	-0.146	0.164	Moderate	1	4	5	6	1
Occupational Safety & Health Admin.	DOL	120	0.017	0.077	-0.133	0.166	Moderate	2	6	3	4	2

Ofc. of Workers Compensation Prog.	DOL	121	-0.145	0.097	-0.336	0.045	Low	9	2	0	0	1
Vets Employment & Training Service	DOL	122	0.003	0.149	-0.293	0.290	Moderate	1	2	9	4	1
Wage & Hour Division	DOL	123	-0.005	0.093	-0.189	0.175	Moderate	1	2	4	5	0
Federal Highway Administration	DOT	124	0.251	0.070	0.115	0.390	High	0	0	0	1	16
Federal Motor Carrier Safety Admin.	DOT	125	0.035	0.112	-0.188	0.252	Moderate-High	1	2	5	6	3
Federal Railroad Administration	DOT	126	0.143	0.109	-0.073	0.357	High	0	0	5	4	8
Federal Transit Administration	DOT	127	0.031	0.132	-0.231	0.286	Moderate-High	1	3	8	2	3
Maritime Administration	DOT	128	0.024	0.136	-0.246	0.289	Moderate-High	0	3	8	4	2
National Cemetery Administration	DVA	129	0.093	0.098	-0.101	0.284	High	1	0	3	7	6
Veterans Benefits Administration	DVA	130	-0.094	0.067	-0.227	0.037	Low-Moderate	9	2	2	3	1
Veterans' Health Administration	DVA	131	-0.084	0.067	-0.215	0.047	Low-Moderate	6	6	4	1	0
Office of National Drug Control Policy	EOP	134	0.001	0.211	-0.412	0.419	Moderate	0	2	14	1	0
Administration for Children & Families	HHS	135	-0.050	0.093	-0.234	0.132	Low-Moderate	3	8	2	4	0
Ctrs. for Disease Control & Prevention	HHS	136	0.092	0.067	-0.043	0.221	Moderate-High	1	2	1	6	7
Health Resources & Services Admin.	HHS	137	0.098	0.100	-0.101	0.296	High	1	1	5	3	7
Indian Health Service	HHS	138	-0.211	0.069	-0.348	-0.078	Low	14	2	1	0	0
Government National Mortgage Assoc.	HUD	139	-0.087	0.171	-0.419	0.246	Low-Moderate	5	4	7	1	0
Ofc of Housing/Fed. Housing Admin.	HUD	140	-0.110	0.117	-0.340	0.120	Low-Moderate	5	6	3	3	0
Office of Public & Indian Housing	HUD	141	-0.105	0.134	-0.371	0.153	Low-Moderate	6	3	6	2	0
Consumer Financial Protection Bureau	IND	143	0.029	0.137	-0.244	0.292	Moderate-High	1	2	3	2	3
Commodity Futures Trading Com.	IND	144	-0.043	0.101	-0.240	0.153	Low-Moderate	4	6	3	1	3
Corp. for Nat. & Community Service	IND	145	-0.003	0.107	-0.213	0.206	Moderate	2	3	5	4	3
Development Finance Corp (OPIC)	IND	146	0.182	0.127	-0.068	0.426	High	0	2	3	2	10
Export-Import Bank	IND	147	-0.112	0.125	-0.355	0.137	Low	7	3	4	1	2
Millennium Challenge Corporation	IND	150	-0.025	0.123	-0.269	0.214	Low-Moderate	3	3	4	5	1
Merit Systems Protection Board	IND	151	0.163	0.097	-0.026	0.357	High	0	0	4	4	9
National Archives & Records Admin.	IND	152	-0.153	0.084	-0.319	0.012	Low	11	2	3	1	0
National Science Foundation	IND	154	0.293	0.067	0.160	0.424	High	0	0	0	3	14
Peace Corps	IND	159	0.273	0.132	0.015	0.531	High	0	1	1	3	12
Bureau of Indian Affairs	INT	160	-0.242	0.080	-0.399	-0.087	Low	13	4	0	0	0

Bureau of Land Management	INT	161	-0.151	0.066	-0.283	-0.023	Low	11	4	1	1	0
Bureau of Ocean Energy Mgt (MMS)	INT	162	0.083	0.082	-0.080	0.242	Moderate-High	0	5	1	7	4
Bureau of Reclamation	INT	163	0.007	0.075	-0.141	0.155	Moderate	0	8	3	4	2
Fish & Wildlife Service	INT	164	-0.021	0.070	-0.156	0.118	Low-Moderate	1	7	5	4	0
National Park Service	INT	165	-0.161	0.065	-0.289	-0.035	Low	11	4	1	1	0
U.S. Geological Survey	INT	166	0.051	0.071	-0.090	0.190	Moderate-High	1	3	2	8	3
Ofc of the Comptroller of the Currency	TREAS	177	0.163	0.078	0.007	0.315	High	0	0	3	2	12
Agricultural Marketing Service	USDA	178	-0.009	0.123	-0.252	0.232	Moderate	1	5	7	3	1
Animal & Plant Health Inspect Service	USDA	179	-0.051	0.132	-0.310	0.208	Low-Moderate	2	7	6	2	0
Agricultural Research Service	USDA	180	-0.033	0.080	-0.188	0.124	Low-Moderate	3	5	4	5	0
Economic Research Service (USDA)	USDA	181	0.026	0.153	-0.274	0.328	Moderate-High	1	2	7	1	6
Foreign Agricultural Service	USDA	182	-0.201	0.092	-0.381	-0.021	Low	10	3	3	1	0
Food & Nutrition Service	USDA	183	-0.017	0.121	-0.255	0.217	Moderate	4	4	4	3	2
Forest Service	USDA	184	-0.180	0.070	-0.318	-0.041	Low	9	6	2	0	0
Food Safety & Inspection Service	USDA	186	-0.015	0.072	-0.159	0.123	Moderate	4	6	2	3	2
Natural Res Conservation Service	USDA	188	-0.025	0.069	-0.163	0.110	Low-Moderate	3	4	4	5	1
Immigration & Naturalization Service	DOJ	194	-0.429	0.045	-0.508	-0.335	Low	1	0	0	0	0
Office of Postsecondary Education	DOED	196	-0.307	0.113	-0.534	-0.088	Low	12	1	2	2	0
Bur. of Alc, Tobacco, Firearms, & Expl	DOJ	197	0.054	0.071	-0.087	0.195	Moderate-High	0	7	0	4	5
U.S. Mint	TREAS	198	-0.014	0.088	-0.186	0.159	Moderate	4	2	6	3	2
Alcohol & Tobacco Tax & Trade Bur	TREAS	199	0.253	0.109	0.037	0.463	High	0	0	1	1	15
Employment Standards Administration	DOL	200	-0.111	0.083	-0.280	0.051	Low	2	2	0	0	0
National Credit Union Administration	IND	202	0.094	0.084	-0.070	0.258	High	0	2	0	11	4
International Trade Commission	IND	203	0.163	0.108	-0.053	0.375	High	0	2	3	4	8
Total Average			0.001	0.093	-0.182	0.182						

Appendix D. Alternative BSEM Model Specification Estimates and Correspondence with Model 1 [Reported] Bayesian Posterior Estimates

Table D1. Correlation of Posterior Median Estimates, Models 1 -5

	Model 1 (Reported)	Model 2	Model 3	Model 4	Model 5
Model 1	1				
Model 2	0.9579	1			
Model 3	0.9593	0.9966	1		
Model 4	0.9946	0.9593	0.9542	1	
Model 5	0.9965	0.9576	0.9562	0.9968	1

Table D2. Correlation of Posterior Standard Deviations, Models 1 -5

	Model 1 (Reported)	Model 2	Model 3	Model 4	Model 5
Model 1	1				
Model 2	0.9271	1			
Model 3	0.9279	0.9978	1		
Model 4	0.9972	0.9276	0.9271	1	
Model 5	0.9970	0.9272	0.9266	0.9972	1

TABLE D3: Alternative BSEM Models and Model Fit and Diagnostics:

Standardized Factor Loadings of U.S. Federal Agency Performance
[2,237 Agency-Year Observations, 2002/2004/2006/2008, 2010-2022]

	MODEL 1		MODEL 2		MODEL 3		MODEL 4		MODEL 5	
Variable	**1st Dimension**	**2nd Dimension**	**1st Dimension**	**2nd Dimension**	**1st Dimension**	**2nd Dimension**	**1st Dimension**	**2nd Dimension**	**1st Dimension**	**2nd Dimension**
FEVS: Fulfilling Agency Mission	0.875*** (0.009)	————	0.874*** (0.009)	————	0.874*** (0.009)	————	0.876*** (0.010)		0.875*** (0.009)	
FEVS: Quality of Work Unit	0.795*** (0.013)	————	0.795*** (0.013)	————	0.795*** (0.013)	————	0.797*** (0.015)		0.795*** (0.013)	
FHCS: Organization as a Place to Work Compared to Others	0.974*** (0.018)	————	0.971*** (0.019)	————	0.971*** (0.019)	————	0.978*** (0.018)		0.977*** (0.018)	
MSPB: Satisfaction with Supervisor	0.936*** (0.011)	————	0.937*** (0.011)	————	0.937*** (0.011)	————	0.937*** (0.012)		0.936*** (0.011)	
MSPB: Satisfaction with Managers Above Supervisor	0.963*** (0.009)	————	0.958*** (0.009)	————	0.959*** (0.009)	————	0.964*** (0.009)		0.963*** (0.008)	
OPM: Best Places to Work Score [2002-2019]	0.908*** (0.008)	————	0.912*** (0.008)	————	0.913*** (0.008)	————	0.908*** (0.008)		0.908*** (0.008)	
OPM: Best Places to Work Score [2020-2022]	0.480*** (0.053)	————	0.467*** (0.072)	————	0.467*** (0.076)	————	0.478*** (0.053)		0.475*** (0.055)	
FHCS: Effective Leadership [2002 & 2004]	0.771*** (0.047)	————	0.775*** (0.047)	————	0.775*** (0.046)	————	0.769*** (0.048)		0.769*** (0.046)	
GSA Acquisition	0.665*** (0.031)	————	————	————	————	0.633*** (0.063)	0.668*** (0.033)		0.663*** (0.031)	
GSA Financial Management	0.666*** (0.031)	————	————	————	————	0.740*** (0.046)	0.669*** (0.033)		0.666*** (0.031)	
GSA Human Capital	0.694*** (0.030)	————	————	————	————	0.665*** (0.059)	0.696*** (0.032)		0.692*** (0.030)	
GSA Information Technology	0.478*** (0.042)	————	————	————	————	0.568*** (0.078)	0.480*** (0.043)		0.477*** (0.042)	

PART Score (Reliable Component)	_____	_____	_____	_____	_____	_____	0.336*** (0.106)	_____	_____	_____
PART Score (Total)	_____	_____	_____	_____	_____	_____	_____	_____	0.388*** (0.082)	_____
OPM Innovation Award Annual Count (AE Adjusted)	_____	0.050 (0.057)	_____	0.002 (0.018)	_____	0.140*** (0.016)	_____	0.000 (0.002)	_____	0.009*** (0.016)
OPM Ratings-Based Cash Award Annual Count (AE Adjusted)	_____	0.069 (0.273)	_____	0.934*** (0.054)	_____	0.100*** (0.023)	_____	0.000 (0.417)	_____	0.935*** (0.059)
OPM Ratings-Based Non-Cash Award Annual Count (AE Adjusted)	_____	0.060 (0.085)	_____	0.282*** (0.029)	_____	0.070*** (0.022)	_____	0.060 (0.085)	_____	0.282*** (0.030)
OPM Quality Step Increase Annual Count (AE Adjusted)	_____	−0.047 (0.085)	_____	0.226*** (0.032)	_____	0.063 (0.250)	_____	0.492*** (0.195)	_____	0.226*** (0.032)
GAO High Risk Program Count (AE Adjusted)	_____	−0.999*** (0.254)	_____	−0.174*** (0.035)	_____	−0.999*** (0.000)	_____	−0.102 (0.768)	_____	−0.175*** (0.035)
GAO Bipartisan Legislative Investigations (AE Adjusted)	_____	−0.583 (0.938)	_____	−0.070*** (0.026)	_____	−0.990*** (0.000)	_____	0.999 (0.521)	_____	−0.070*** (0.026)
Comparison Fit Index (CFI)	0.920 [0.841, 0.930]	_____	0.879 [0.707, 0.953]	_____	0.876 [0.871, 0.882]	_____	0.916 [0.756, 0.931]	_____	0.863 [0.724, 0.915]	_____
Tucker-Lewis Fit Index (TLI)	1.000 [0.999, 1.000]	_____	0.999 [0.999, 1.000]	_____	0.862 [0.856, 0.869]	_____	0.999 [0.998, 0.999]	_____	0.991 [0.982, 0.994]	_____
Root Mean Square Error of Approximation (RMSEA)	0.003 [0.003, 0.003]	_____	0.003 [0.002, 0.005]	_____	0.053 [0.051, 0.054]	_____	0.004 [0.003, 0.005]	_____	0.009 [0.007, 0.014]	_____
Deviance Information Criterion (DIC) Statistic	52,272.070	_____	66,059.742	_____	84,755.536	_____	69,713.465	_____	87,122.477	_____

Average Variance Extracted	0.471	0.140	0.446	0.126	0.429	0.329	0.475	0.131	0.476	0.111
Construct Reliability	0.911	0.011	0.862	0.167	0.853	0.117	0.917	0.188	0.917	0.149
Discriminant Validity	0.471 > 0.011	0.140 > 0.001	0.471 > 0.000036	0.126 > 0.000036	0.471 > 0.00846	0.140 > 0.00846	0.475 > 0.0071	0.131 > 0.0071	0.476 > 0.000049	0.111 > 0.000049
Nomological Validity	−0.034 (0.100)	_____	−0.006 (0.029)	_____	−0.092*** (0.028)	_____	0.084 (0.198)	_____	0.007 (0.028)	_____

Note: Model estimates generated from 1,000 Bayesian Posterior Empirical Distribution Functions (EDFs) based on 100,000 MCMC iterations with 2 chains using Gibbs Sampling with data missing at random for imputed values. Entries are standardized factor loadings with standard errors inside parentheses, except for Model Fit Statistics content that reports 90% credibility interval values inside brackets. *** $p \leq 0.01$.