Obtaining Comparable Measures of Organizational Performance: An Application to U.S. Federal Agencies, 2002-2024[1]

Evaluating the comparative performance of United States federal agencies is difficult, particularly since both tasks and missions vary so dramatically. In addition, forces beyond an agency's control (e.g., COVID, an economic downturn, etc.) can determine outcomes even when agencies are performing at a high level. In this paper, we introduce a new approach to measuring organizational performance, something conceptually distinct from, but correlated with, both organizational inputs and outcomes. This measurement approach focuses on how well the internal machinery of agencies is functioning. We analyze a vast trove of subjective and objective performance information and aggregate it using a Bayesian structural equation measurement (BSEM) model. We isolate organizational performance from inputs and outcomes through careful model specification, information from the BSEM models, and model identification through a careful evaluation of different models and diagnostics. Our analysis yields 2,479 organizational performance estimates for 135 U.S. federal departments and agencies spanning 19 years between 2002 and 2024. We explore the validity of these estimates by comparing them with other measures of similar or related concepts. We conclude by discussing the implications of our measurement approach and its usefulness for evaluating organizational performance in diverse and changing contexts.

Keywords: U.S. Federal Agencies, Organizational 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

Data Availability Statement: The data that support this study are openly available in the JPAM Harvard Dataverse at https://doi.org/10.7910/DVN/MDHRH5.

[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, Martin Lodge, Christian Schuster, Manny Teodoro, Weijie Wang, seminar participants at Georgetown University and the London School of Economics and Political Science, and *JPAM* reviewers for extremely helpful comments. We are grateful to Cody Drolc for providing GAO investigations data, William Resh for providing GAO−PAR data, and Mark Richardson for providing 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 Measures of Organizational Performance: An Application to U.S. Federal Agencies, 2002-2024

Evaluating the comparative performance of United States federal agencies is difficult, particularly since both tasks and missions vary so dramatically. In addition, forces beyond an agency's control (e.g., COVID, an economic downturn, etc.) can determine outcomes even when agencies are performing at a high level. In this paper, we introduce a new approach to measuring organizational performance, something conceptually distinct from, but correlated with, both organizational inputs and outcomes. This measurement approach focuses on how well the internal machinery of agencies is functioning. We analyze a vast trove of subjective and objective performance information and aggregate it using a Bayesian structural equation measurement (BSEM) model. We isolate organizational performance from inputs and outcomes through careful model specification, information from the BSEM models, and model identification through a careful evaluation of different models and diagnostics. Our analysis yields 2,479 organizational performance estimates for 135 U.S. federal departments and agencies spanning 19 years between 2002 and 2024. We explore the validity of these estimates by comparing them with other measures of similar or related concepts. We conclude by discussing the implications of our measurement approach and its usefulness for evaluating organizational performance in diverse and changing contexts.

Modern governments are awash in data and activity and yet elected officials rarely have a simple way to compare the performance of one agency to another. Ideally, transition officials would provide new executive and legislative officials with a simple chart or heat map that detailed high and low agency performance. This would allow new leaders to efficiently allocate their management and oversight efforts. Developing an overall picture requires aggregating and filtering a tremendous amount of complex performance information. In the United States federal government, for instance, there are dozens of subjective and objective measures for hundreds of agencies. Public officials need to separate out the helpful from the misleading data (see, e.g., Cheon, Song, McCrea, and Meier 2021; Favero, Walker, and Zhang 2025; Van Ryzin 2006). 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 operating at a high level, but political, economic, or societal events beyond their control can decouple organizational performance from clear changes in outcomes. Without a principled approach to aggregating performance information, public officials fall back on haphazard and informal patterns, increasing the chances they make mistakes.

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; 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).[1] Public sector organizations perform a variety of functions that are hard to observe and hard to connect to changes in outcomes (Wilson 1989). Indeed, such organizations can suffer what looks like poor performance because of events beyond their control. While scholars have made important progress measuring comparative organizational 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 not equate with an aggregate measure of organizational performance.

In this paper, we introduce a new approach to measuring U.S. federal agency performance that overcomes many of these difficulties. Our approach captures *organizational performance* – i.e., how the machinery of agencies is working, something conceptually distinct from, but correlated with, both inputs (e.g., budgets, staffing) and outcomes (i.e., results). This includes the quality of management, execution of core tasks (e.g., human resources, financial management), employee morale, and other correlates of organizational health. 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] We isolate organizational performance – as opposed to inputs or outcomes – through careful choice of

[1] 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). Notably, 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 applied to evaluating public agency characteristics.

which performance measures to include in our models, information from parameter estimates about which indicators load on the appropriate dimension, and model identification through careful evaluation of different models and diagnostics. We generate organizational performance estimates for 135 U.S. federal departments and agencies between 2002 and 2024 that vary across agencies and time. To validate our new measure of organizational performance, we compare our estimates to other measures of similar or related concepts. We conclude by discussing the contribution and limitations our measurement approach and its usefulness for evaluating organizational performance in diverse and changing contexts.

CHALLENGES IN COMPARATIVE PERFORMANCE MEASURMENT

Scholars and practitioners have been interested in the measurement of agency performance for some time, with this interest accelerating as part of widespread enthusiasm for the New Public Management (Moynihan 2008; 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 often perform tasks and expend effort that is hard to observe and this can lead performance measures to be quite removed from what agencies actually do (Nyhan and Marlowe 1995; Smith 2006). Organizational performance can be decoupled from outcomes or results that ensue from administrative activities. For example, it is hard to discern how much credit to give the State Department for success or failure of regional democratization regardless of how well the agency performs. Factors well beyond the control of the agency combine to determine democratization and sometimes their actions bear no fruit immediately. A highly functioning agency may not be rewarded with immediate changes in outcomes. And, by contrast, a poorly functioning agency may be the recipient of fortuitous outcomes. The problem is further complicated by the fact that programs and agencies have different or unclear goals (Chun and Rainey 2005). Ideally, we would be able to observe how agency activities change desired outcomes in a way that allowed direct comparisons. Often, however, this is impossible.

A 'levels of analysis' problem also complicates efforts to measure administrative performance (e.g., Andersen, et al. 2016). Some performance measures are targeted at specific *tasks*. Others are directed at discrete *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. This is particularly the case since scholars and practitioners 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, et al. 2025).

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., Krause and Hong n.d.; Park 2022). Others restrict focus to a single sector such as law enforcement or education (e.g., Boylan 2004; Rutherford 2016). Scholars have also made important advances using subjective assessments in surveys that include comparable questions (Brewer and Selden 2000; Chun and Rainey 2005; Piper and Lewis 2023) and various government generated performance scores (Kroll and Moynihan 2021; Lewis 2007).

Although 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. 2025). 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. 2025).

What is needed is an approach to the measurement of administrative performance that overcomes these challenges. Ideally, the approach would disentangle performance related to administrative operations from factors that might influence performance (e.g., budgets, staffing), as well as those beyond the control of the agencies themselves that may well impact outcomes (e.g., COVID-19, an economic downturn). The unit of analysis and goals should be clear (e.g., task, bureau, or agency) and the measures should accommodate and discriminate among various subjective and objective indicators (e.g., surveys, awards, investigations) on different dimensions of performance (e.g., efficacy, satisfaction) in a flexible, reasonable, and transparent way (see, e.g., Cheon, Song, McCrea, and Meier 2021; Favero, Walker, and Zhang 2025; Van Ryzin 2006). The measure should be broadly acceptable to relevant stakeholders (e.g., Republicans and Democrats in government) and comport with common conceptions of good and bad performance. We turn now to our approach.

DEFINING ELEMENTS OF ORGANIZATIONAL PERFORMANCE

Given the diverse approaches to measuring administrative 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, organizational performance, that is a composite of performance on numerous goals or tasks, large and small. This includes the quality of management, execution of core tasks (e.g., human resources, financial management), employee morale, and other correlates of organizational health. To measure this underlying latent dimension, we must rely on various observable indicators (e.g., average responses to a survey question, agency awards, etc.). Each measure imperfectly reveals information about latent organizational performance. 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 distinctions among the concepts of *inputs*, *performance*, and *outcomes (i.e. results)*. 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 disentangle *task* performance from *aggregate* performance at different levels (i.e., performance of a subcomponent versus performance of the organization as a whole), and account for different dimensions of performance. Hence, our measurement strategy overcomes these limitations by offering a holistic assessment of organizational performance anchored in the effectiveness of administrative operations that is comparable both across agencies and time.

Measuring Performance versus Inputs

Scholars and users of performance measures often conflate organizational performance with either inputs or outcomes even though these concepts are distinct (Yang and Holzer 2006: 117; Rogger and Schuster 2023). Consider the Veterans Health Administration (VHA) as an example. The mission of the VHA is to provide high quality healthcare to veterans. How might we measure the agency's performance? To begin, we might use *inputs* such as count the number of physicians or hospitals funded as measures of performance. In an important sense, however, neither of these is a measure of the health of veterans. We believe that each item measured *contributes* to good performance. The agency could be performing poorly with many physicians and large numbers of healthcare facilities. Higher administrative capacity, in the form of more physicians or facilities funded, is a *precondition* that facilitates the agency in achieving its goals.[3]

Being explicit about the relationship between inputs and 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 others. Second, it helps us understand performance measures in context. Scholars using measures of administrative capacity might argue that VHA officials that have built capacity in the form of more physicians or more facilities have performed well on an *administrative* task. Information about performance on this task can contribute to our understanding of organizational performance even though such performance is not the same as an agency providing excellent healthcare for veterans.

[3] This is not to say that the statutory requirements for the VHA could not include a goal funding more facilities. If the statute specified expanding the VHA network, then the number of facilities, particularly relative to some baseline, could be a measure of performance. The point is that scholars and practitioners can conflate *contributors to* high performance with *actual* high performance. We thank an anonymous reviewer for sharing this insight.

Good Organizational Performance Does Not Always Translate into Success

In the same way that inputs are correlated with, but distinct from, organizational performance, outcomes are also correlated with, but distinct from, organizational performance. 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 area 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. Similarly, while high quality veterans care leads to better outcomes for patients, outcomes are conceptually distinct from administrative performance. Despite receiving quality care, veterans may still die of cancer and other health issues.

The distinction between organizational performance and outcomes is true more generally. Indeed, a nontrivial gap exists between these concepts. This gap can exist because of unforeseen and uncontrollable factors in the environment. It can also emerge 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 simple 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 since many indicators of performance 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 forecasting tasks 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.

Scholars and practitioners have done significant work trying to evaluate performance through the lens of an agency production function, an approach that focuses on the relationship between organizational inputs and outputs (or outcomes). Yet, a production function approach obscures how well internal administrative processes are functioning, treating internal operations as a black box. Nor does a production function approach solve the comparability problem since agency outputs (e.g., air pollution, democracy in the Balkans) are so varied. A focus on inputs and outputs can also shift our attention toward efficiency and cost effectiveness, while ignoring other vital administrative values such as equity, client satisfaction, or output quality (Andersen, et al. 2016; Boyne 2002; Gębczyńska and Brajer-Marczak 2020). In addition, how effective an agency is at fulfilling its mission might be related to factors that cannot be gleaned from the relationship between inputs and outputs (or outcomes).

Although the measurement of public agencies' performance is both contested and complex, we seek to capture its inherent process-based nature by emphasizing a common focus on organizational characteristics relating to the quality of individual processes (e.g., human resources, procurement), and also the collective outputs of those processes (e.g., effective goal setting and accomplishment) that can be compared across agencies and through time.

Different Stakeholder Conceptions of Administrative Performance

In the prior sub-section, we established that our concept of interest is the *organizational performance* of public agencies. Measuring the organizational performance of U.S. federal agencies is complicated by the fact that stakeholders, such as political parties, clientele groups, or citizens, can disagree about the definition of good agency performance. This can mean different things. It can mean that parties evaluate agency 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 agency performance might classify it differently. For example, a Democrat might suggest that an agency is effective and high performing while Republicans would classify the same agency as low performing. We assume here that if stakeholders were able to observe this latent performance dimension perfectly, they would classify it similarly. That is, both parties can look at organizational performance data and determine whether the agency is healthy or sick even when they disagree about what the agency does.

Of course, 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 organizational 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 – *Is the organization healthy and marked by all the relevant characteristics of well-functioning agencies such as high morale, few scandals, low employee turnover, evidence of goal setting and accomplishment, and the like?*

We acknowledge that our approach is limited insofar that cases exist where it can be difficult to distinguish organizational performance from disagreements over policy goals. It is important to remember, however, that most programs enjoy bipartisan support and many aspects of administrative performance have little to do with policy per se (Bednar and Lewis 2024; Gramlich 2017).[4] This is

[4] This is to be expected since most statutory government activity was supported by majorities in both chambers and the president at the time of enactment.

borne out by a recent study showing a strong positive correlation between agency performance ratings by Republicans and Democrats in the United States (Richardson, et al. 2025). When Democrats thought agencies were performing well, so did Republicans and vice versa. Although scholarly attention often focuses on sources of partisan or ideological disagreement, a wide consensus exists for a considerable amount of government activity, particularly activity related to effective operations (Richardson 2024).

Aggregating Performance Information Across Levels

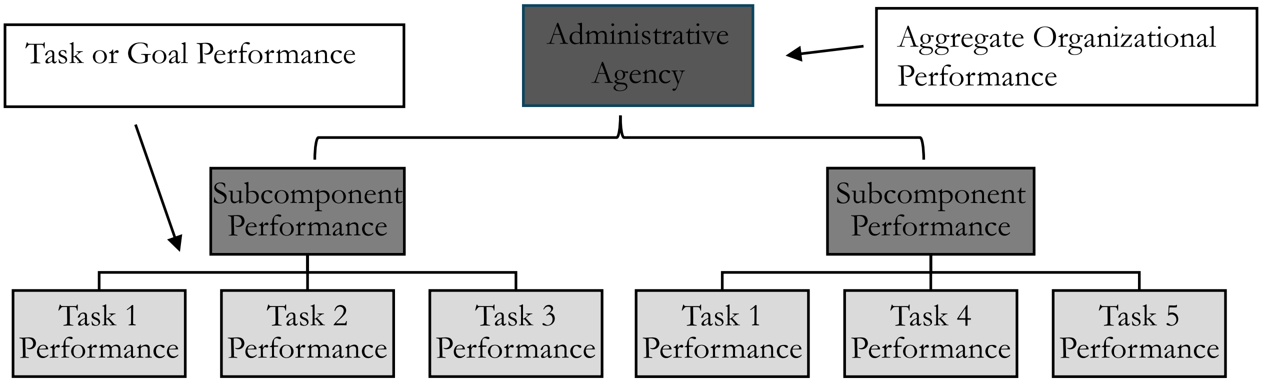
Organizational performance is a composite concept, aggregating performance on numerous *tasks*, large and small. Some of these tasks relate to agency core missions and others to auxiliary statutorily mandated tasks, including internal agency operations and processes like financial management, purchasing, 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 latent organizational performance involves weighted averaging across observable indicators that reflect different tasks or aspects of agency operations (see **Figure 1 below**). Depending upon size, an agency's overall 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 aggregate organizational performance we are implicitly averaging across multiple units (and tasks) within each public agency.

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

[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.

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.

Figure 1. Measuring Department Performance by Aggregating Subcomponent Performance



Similarly, each observable performance indicator provides information about the underlying latent 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 offer a clear signal. We evaluate aggregate agency performance in a manner 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, as well as perform better at mapping an observed measures 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 either a single measure or small set of measures (e.g., employee turnover quality of work unit, caliber of program management). Rather, it uses many different indicators, carefully selected and appropriately weighted to develop organizational performance estimates.

Different Criteria for Evaluating Performance

Evaluations of performance on tasks can encompass 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. For example, a survey of executives might ask, "*How would you rate the overall performance of the VHA in carrying out its mission?*" (i.e., overall performance). By contrast, other measures might tap costs per patient (efficiency), average return visits per patient (effectiveness), or the percentage of veterans satisfied with their treatment (client satisfaction). That is, some measures of performance can measure accomplishment across tasks but are restricted to a single criterion – e.g., evaluating 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 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. Performance measures, 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 organizational performance on different criteria and allow details of the estimation to tell us what measures are best at uncovering this latent construct, and how much they do so.

ORGANIZATIONAL 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 these data are subjective, indicators based upon the perception of persons working in or close to agencies. Other data are objective, presenting counts of good or bad indicators of administrative performance (e.g., presence of award-winning employees, employee turnover). We list data sources in **Table 1**. The sources provide data on 135 agencies for 19 years during the 2002 to 2024 period (**Appendix A** for a full list).

Subjective Data: Surveys of U.S. Federal Employees and Citizens

During 2002-2024, OPM, GSA, and MSPB all surveyed federal employees about factors related to organizational performance. For example, since 2015 the GSA has surveyed tens of thousands of federal employees each year about the quality of services and support that they receive in their agencies in information technology, acquisition, human resources, and financial management. The MSPB and OPM regularly survey employees about the quality of managers in their agencies or the quality of work their organizations deliver. Several outside groups have also conducted federal employee surveys during this period asking performance-related questions.[6] Collectively, there are 37 different surveys of federal employees with 32 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, the number of agencies evaluated, and the number of performance-related questions. We also include the overlapping performance-related questions from

[6] Specifically, we use data from the *Survey on the Future of Government Service* (SFGS), a 2014 and 2020 non-partisan and non-governmental survey of thousands of federal executives (Piper and Lewis 2023; Richardson, et al. 2025). This survey covered hundreds of agencies and included several agency performance measures in those years. They are particularly useful as outside validation of our measures.

the surveys. Federal executives and rank-in-file employees have the most direct information about what is working well or poorly and provide informative measures of organizational performance.

Table 1. U.S. Federal Agency Performance Information, 2002-2024

Source	Title	Years
Subjective		
Office of Personnel Management	FHCS/FEVS	2002-2008 (biannual); 2010-2024 (annual)
Merit Systems Protection Board	Merit Principles Survey	2005, 2007, 2010, 2011, 2016, 2021
Richardson, et al. (2018); Richardson, et al. (2025)	Survey on the Future of Government Service	2014, 2020
General Services Administration	Customer Satisfaction Survey	2015-2024
Partnership or Public Service	Best Places to Work Index	2002-2010 (biannual); 2011-2024 (annual)
National Quality Research Center	American Consumer Satisfaction Index	2011-2024
Objective		
Government Accountability Office	High Risk List	2002-2023 (biannual)
Government Accountability Office	Congressionally Requested Reports (bipartisan)	2002-2023
Office of Personnel Management	Employee Performance Awards	2002-2023
Office of Personnel Management	Employee Turnover Data	2002-2023
Partnership for Public Service	Sammies	2003-2024
Office of Management and Budget	Program Assessment Rating Tool (PART)	2002-2008
Office of Management and Budget	Performance & Accountability Reports (PARS)	2002-2011

Note: Our models only include data from 2002, 2004, 2006, 2008, 2010-2024 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.[7] According to the PPS, "*The index score is calculated using a proprietary weighted formula that looks at responses to three different questions*

[7] 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.

in the federal survey. The more the question predicts intent to remain, the higher the weighting."[8] 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 performance indicator 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 28 different agencies as of 2011.

Objective Data: GAO Analysis, PART Scores, and Employee Award and Turnover Data

The federal government and outside groups have actively collected objective indicators of performance during this period. The GAO, OMB, OPM, 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.[9] We collected counts of programs on the list by agency and year. We also collected data on counts of GAO reports from 2002-2023 resulting from bipartisan requests for GAO investigations.[10] We do so on the assumption that bipartisan requests likely reflect real performance concerns, rather than simple efforts

[8] 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.

[9] This description is based on GAO's own description of the program (https://www.gao.gov/high-risk-list).

[10] We thank Cody Drolc for providing us with this data.

to discredit the presidential administration. 122 of the 135 agencies in our data have been the subject of a GAO investigation, with some exceeding 300 for a given year.

In addition to GAO data, we include two different government performance scores. First, we collected data on all federal programs evaluated during the George W. Bush Administration using the Program Assessment Rating Tool (PART). Between 2002 and 2008, the Bush Administration evaluated the performance of 1,016 programs on four categories of performance (program purpose and design, strategic planning, program management, and program results). We analyze strategic planning and program management scores here since they are closest to the concept of organizational performance. We will later compare them to program results component of the PART scores that reflect administrative outcomes.

Second, we include Performance and Accountability Reports (PAR). The Government Performance and Results Act (GPRA) of 1993 required agencies to set performance goals and document progress toward goals. Between 2002 and 2011, agencies identified more than 20,000 goals and reported progress on these goals (Lee and Whitford 2013; Resh and Cho 2020). We use data provided by Resh and Cho (2020) to generate agency-year averages of goals unmet, met, and exceeded for 27 agencies from 2002 – 2011.[11]

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 (2002 – 2023), high performance award—not rating based (2003 to 2023), individual suggestion/invention award (2002 to 2023), and

[11] We thank William Resh for providing us with this data.

quality step increases (2002 to 2023).[12] 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 each 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 the Air Force and the National Labor Relations Board.

Finally, we collected data from OPM on employee separations, both aggregate agency-year percentages and turnover percentages for subsets of different kinds of employees (e.g., probationary, experienced). We obtained this data from the Office of Personnel Management's Employee Human Resources Integration (EHRI). These data account for the plausible view that high percentages of turnover reflect problems in administrative performance. Because this or any performance measurement approach depends upon the quality and availability of data, we are only able to generate valid performance estimates for 2002, 2004, 2006, 2008, and 2010-2024. Omitted years could not yield valid estimates due to sparseness of data in select years prior to 2010 (2003, 2005, 2007, & 2009).[13]

METHODS

The goal of our measurement strategy is to model the relationship between agencies' latent organizational performance and observed performance indicators of various types. To isolate organizational performance − as opposed to inputs or outcomes − we make careful specification

[12] For descriptions of each type see **Appendix B**.

[13] Initial attempts to generate estimates based on these sparse data years resulted in unusual shifts in theta (θ) estimates, coupled with a sharp rise in the imprecision of the estimates.

choices, glean information from parameter estimates, and conduct a variety of diagnostic tests. These performance indicators consist of process-oriented measures − as opposed to inputs or outcomes. Our model produces a set of numerical estimates that we compare to other measures of the same concept (*convergent validity*) and related concepts (*predictive validity*). These steps together help us isolate operational performance from other conceptions of agency performance.

Model Specification: To begin, we choose subjective and objective measures to include in models that are closest to the concept of organizational performance. We exclude measures of inputs (e.g., budgets, employment) and outcomes (e.g., number of permits, inspections). So, for example, our models include yearly agency average responses to questions like "*My agency is successful at accomplishing its mission*." We also prioritized measures that cover a large number of agencies and years in order to facilitate comparability and yield reliable estimates based on sufficient data. For example, OPM turnover data or data from the FEVS are useful here since these data cover most agencies in all years we examine.

Parameter Estimates: After deciding on an initial specification, we estimated models and use the resulting parameter estimates to determine which measures helped separate low and high performing agencies. One useful feature of the BSEM models is that the parameters provide information about whether the performance measures we include distinguish agencies on the latent dimension. In the same way that a standardized test question asking as "Is blue a good color?" does not help us measure latent academic ability, so some performance measures do not help us measure organizational performance. Some measures offer only modest insight into 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). In such instances, although these measures might offer some limited useful information into actual performance, they will do so in a 'noisy' manner by containing a substantial amount of measurement error and reflected in low standardized factor loadings.[14] We dropped measures that were not helpful predicting latent operational performance.

How Many Dimensions? To isolate organizational performance we need to verify that this concept can be characterized by one dimension. This is something we have explored thoroughly. We used different conceptions of performance and Bayesian Exploratory Factor Analysis (BEFA) as a starting point. We then estimated Bayesian Structural Equation Measurement (BSEM) models accounting for multiple dimensions in different ways and evaluated the comparative fit of different approaches using model diagnostics. The model fit statistics reveal that BSEM models analyzed containing two dimension are better fitting than the reported model based on root mean square error approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI) model fit criteria.[15] We could not, however, reject the one-dimensional model as the best model because of a high correlation between latent constructs in two dimensional models and the similarity of estimates across simpler and more complex models.[16] The posterior medians from the different one and two-dimensional models are correlated at between 0.9859 and 0.9995 and posterior standard deviations

[14] In a latent measurement modeling context, measurement error is formally defined as 1 – (standardized factor loading estimate).^{2} That is, the proportion of variance associated with a given indicator variable that cannot be explained by the latent concept (i.e., organizational performance).

[15] The most parsimonious fitting BSEM models based on these model fit criteria are the one-dimensional model that omits the GSA survey items [*Model 3*], and the two-dimension models predicated on sub-dimensions of the organizational performance indicator variables analyzed in Model 1 [*Models 5 & 6*]. See **Appendix D: Tables D1A** and **D1B**.

[16] One set of two-dimensional BSEM models focused on distinguishing between latent organizational performance versus an outcome-based performance dimensions (*Models 2* & *4*). Another set of analyses focused on modeling two separate 'sub-dimensions' of the *Model 1* specification based on the results from a Bayesian Exploratory Factor Analysis [BEFA] (*Models 5* & *6*). The full information on these respective set of model estimates can be obtained in **Appendix D: Tables D1A** and **D1B**.

are correlated between 0.9828 and 0.9976 (see **Appendix D**: **Tables D2A** & **D2B**). In addition, there was no clear theoretical coherence in the measures that loaded on a second dimension, consistent with the high latent factor correlation among dimensions. Given these considerations, we focus on estimates from a one-dimensional model in the main text and put estimates for two dimensional models in **Appendix D**. We provide more detail on the basic BSEM model below.

Model Identification: While information from multiple measures properly aggregated is better than a single measure, we still need to make sure that the resulting estimates reflect what we think we are measuring (i.e., organizational performance). It is possible, for example, that the inclusion of variables related to employee satisfaction or PART scores may create a measure of some concept other than organizational performance. To address this, we evaluate a variety of different model specifications to determine whether model estimates change appreciably with different specifications.[17] For example, what happens if we exclude certain FEVS questions from the model? Are estimates still similar? We also subject our measurement models to various diagnostic tests to determine 1) whether the indicator variables load on the correct/main dimension (*Average Variance Extracted*, *Construct Reliability*) and 2) are not overly correlated with another latent dimension(s). This allows us to map each indicator variable to only a single latent construct (*Discriminant Validity*), as well as ensure that the latent constructs are empirically distinct from one another (*Nomological Validity*).

Since the organizational performance estimates are very strongly correlated across BSEM models with different specifications, we have confidence that they are isolating organizational performance rather than some other concept. Our estimates are neither sensitive to model specification nor alternative identification choices. In addition, we show below that our estimates

[17] See **Appendix D (**e.g., **Tables D1A**, **D1B**, and **D2)**.

correlate with other measures of organizational performance (e.g., from a different SFGS survey in 2014) or things that should correlate with organizational performance (e.g., COVID−19 response).

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

The Bayesian structural equation 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) one another. The approach also allows post-estimation diagnostics beyond standard model fit statistics. Indeed, the BSEM approach provides information that helps evaluate model identification assumptions by assessing model-based convergent validity, construct reliability, discriminant validity, and nomological validity. A Bayesian estimation approach to structural measurement models is helpful since it allows us to deal with the missing data that naturally arises from using a wide range of data sources.[18] By implementing a BSEM modeling approach, we can cover unique uncertainty estimates for each agency-year observation from the Bayesian posterior distributions.

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

[18] In the reported model (**Model 1**), as well as Models 5, 6, 7, and 8, the number of missing cases on all indicator variables is a total of 26 agency-years contain missing data for the BSEM model (1.049% of full sample of 2,479 agency-year observations), with a low of 7 agency-years – 0.288% of full sample of 2,498 agency-year observations (**Model 2: Appendix D, Table D1A**), and a high of 29 agency-years – 1.171% of full sample of 2,476 agency-year observations (**Model 3: Appendix D, Table D1A**).

$$\boldsymbol{y_i^{*F1} = \upsilon^{F1} + \Lambda_p \eta_{p_i}^{F1} + \varepsilon_i^{F1}} \tag{1}$$

$$\boldsymbol{y_i^{*F2} = \omega^{F2} + \Pi_q^{F2} \theta_{q_i}^{F2} + \zeta_i^{F2}} \tag{2}$$

where υ^{F1}, $\boldsymbol{\omega}^{F2}$ constitute intercept terms for each respective latent trait equation; $\boldsymbol{\eta_p}^{\mathbf{F1}}$, $\boldsymbol{\theta_q}^{\mathbf{F2}}$, represent *p*, *q* -dimensional vectors of observed indicator variables in each measurement equation for each respective latent trait, while $\boldsymbol{\Lambda_p}^{\mathbf{F1}}$, $\boldsymbol{\Pi_q}^{\mathbf{F2}}$ are the corresponding $p \times 1$, $q \times 1$ parameter matrices of factor loadings and $\boldsymbol{\varepsilon}^{\mathbf{F1}}$*,* $\boldsymbol{\zeta}^{\mathbf{F2}}$ constitute the residual vectors for each latent trait equation that are allowed to be correlated. Their corresponding variance-covariance matrix is denoted as $\boldsymbol{\Theta} = \varrho(\varepsilon^{F1}, \zeta^{F2})$. Estimates are generated via the Bayesian posterior density of the parameter distributions for the slope, intercept, and loading parameters ($\boldsymbol{\nu}^{\mathbf{F1}}$, $\boldsymbol{\omega}^{\mathbf{F2}}$; $\boldsymbol{\Lambda_p}^{\mathbf{F1}}$, $\boldsymbol{\Pi_q}^{\mathbf{F2}}$), the variance-covariance parameters ($\boldsymbol{\varepsilon}^{\mathbf{F1}}$, $\boldsymbol{\zeta}^{\mathbf{F2}}$), and the latent variables of interest ($\boldsymbol{\eta_p}^{\mathbf{F1}}$, $\boldsymbol{\theta_q}^{\mathbf{F2}}$). The conjugate non-informative priors for all the free parameters ($\boldsymbol{\nu}^{\mathbf{F1}}$, $\boldsymbol{\omega}^{\mathbf{F2}}$; $\boldsymbol{\Lambda_p}^{\mathbf{F1}}$, $\boldsymbol{\Pi_q}^{\mathbf{F2}}$) are normally distributed with mean zero, and positive infinity variance; the variance-covariance parameters ($\boldsymbol{\varepsilon}^{\mathbf{F1}}$, $\boldsymbol{\zeta}^{\mathbf{F2}}$) 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, ∞).[19] In those instances where only a single latent administrative performance dimension is estimated (such as in the reported **Model 1** in **Table 2**), the BSEM depicted above simplifies to only consisting of equation (1), sans latent factor correlations due to being premised only on a single latent dimension.

We estimated the model 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). Our analysis utilizes multiple imputation to generate plausible values consistent with the observed data through 1,000 draws, which

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

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. The Bayesian posterior median theta (θ) estimates yield point estimates of latent organizational performance, while the Bayesian posterior standard deviation and corresponding 95% credibility intervals provides measures of uncertainty surrounding these point estimates. One of the advantages of the Bayesian approach is that it requires less stringent model identification assumptions compared to a standard frequentist model since the former explicitly accounts for model uncertainty. In our case, each BSEM model estimated in this study relies on an empirical posterior sampling distribution of 1,000 sets of models, as opposed to a single set of estimates for a given model generated from a frequentist modeling approach.

EMPIRICAL RESULTS

Table 2 lists the BSEM model estimates in the form of standardized factor loading coefficients. Each coefficient represents how well the observed indicator correlates with the underlying latent dimension. Each of the 16 coefficient estimates is appropriately signed, substantial, and statistically significant at the $p \leq 0.05$ level. Larger values of the standardized factor coefficients correspond to a greater amount of each indicator's variance being explained by the latent trait. The survey questions related to one's organization as a place to work, agency leadership, and the success of the agency in fulfilling its mission explain the most variance. Those closest to the agency may provide the most revealing information about administrative performance when these measures are properly constructed.

The cases with standardized factor loadings below 0.50 include GSA surveys of users about the quality of Acquisition (0.495) and Informational Technology (0.489) functions in their agencies. The standardized factor loadings for the objective indicators, i.e., agency turnover and PART scores, were also estimated to be less than 0.50. While these latter measures help parse performance, they contribute less than other indicators since they contain nontrivial measurement error with respect to the latent construct of interest – organizational performance. The objective measures may contain higher levels of measurement error than subjective measures because they represent cruder measures of how well the agency is functioning as an organization.[20]

TABLE 2: BSEM Model of Organizational Performance, U.S. Federal Agencies 2002 - 2024

Variables	**Model 1**
Subjective Measures	
FEVS: Fulfilling Agency Mission	0.887***
	(0.008)
FEVS: Quality of Work Unit [2002-2019]	0.801***
	(0.013)
FEVS: Quality of Work Unit [2020-2024]	0.770***
	(0.027)
FHCS: Organization as a Place to Work Compared to Others	0.978***
	(0.019)
MSPB: Satisfaction with Supervisor	0.921***
	(0.016)
MSPB: Satisfaction with Managers Above Supervisor	0.942***
	(0.014)
OPM: Best Places to Work Score [2002-2019]	0.916***
	(0.008)
OPM: Best Places to Work Score [2020-2024]	0.848***
	(0.018)
FHCS: Effective Leadership [2002 & 2004]	0.772***
	(0.047)
GSA Acquisition	0.495***
	(0.038)
GSA Financial Management	0.554***
	(0.034)
GSA Human Capital	0.610***
	(0.031)
GSA Information Technology	0.489***
	(0.036)

[20] This cannot be attributed to data sparseness since the agency turnover measure covers all agencies over the sample period, while some subjective measures (e.g., FHCS, GSA, MSPB) have rather limited temporal coverage.

Objective Measures	
Agency Turnover (Total Percentage)	−0.085***
	(0.024)
PART Score (Section 2)	0.215**
	(0.100)
PART Score (Section 3)	0.200**
	(0.102)
Model Fit & Diagnostic Statistics	
Comparison Fit Index (CFI)	0.831
	[0.823, 0.840]
Tucker-Lewis Fit Index (TLI)	0.806
	[0.797, 0.816]
Root Mean Square Error of Approximation (RMSEA)	0.052
	[0.050, 0.053]
Average Variance Extracted (Convergent Validity)	0.508
Proportion of Model Variance Explained (Construct Reliability)	0.931

Note: Models estimated with 2,479 observations on 138 agencies in 2002, 2004, 2006, 2008, and 2010–2024. 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. ** $p \leq 0.05$; *** $p \leq 0.01$. The Deviance Information Criterion is 4,219.46. The Bayesian Information Criterion is 4,499.25.

Overall, the model fit statistics and structural measurement model diagnostics reveal that the reported model specification is mixed in terms of model fit. The root mean square approximation (RMSEA) is 0.052 for Model 1. This is close to the threshold of excellent model fit (0.050), while being in the acceptable level (0.05 ≤ 0.10). The comparative fit index (0.831) and Tucker-Lewis fit index (0.806) values fall below the 0.90 threshold for acceptable fit criterion, possibly due to more comprehensive (and less parsimonious) nature of this model specification.[21] Moreover, convergent validity (denoted by average variance extracted) and construct reliability (denoted by proportion of model variance explained) are above acceptable levels of 0.50 (0.508) and 0.80 (0.931), respectively.

[21] These criteria are not immutably fixed since considerations such as our large sample size and model complexity can affect these model fit statistics (e.g., see Shi, Lee, and Maydeu-Olivares 2018). Moreover, analyses of alternative BSEM models reveals that these model fit statistics are much improved and at desirable levels for all but Model 8 (see **Appendix D: Tables D1A & D1B**). Finally, the latent organizational performance estimates generated from Model 1 exhibit exceptionally strong positive correlations with those generated from these alternative models (**Appendix D: Table 2**).

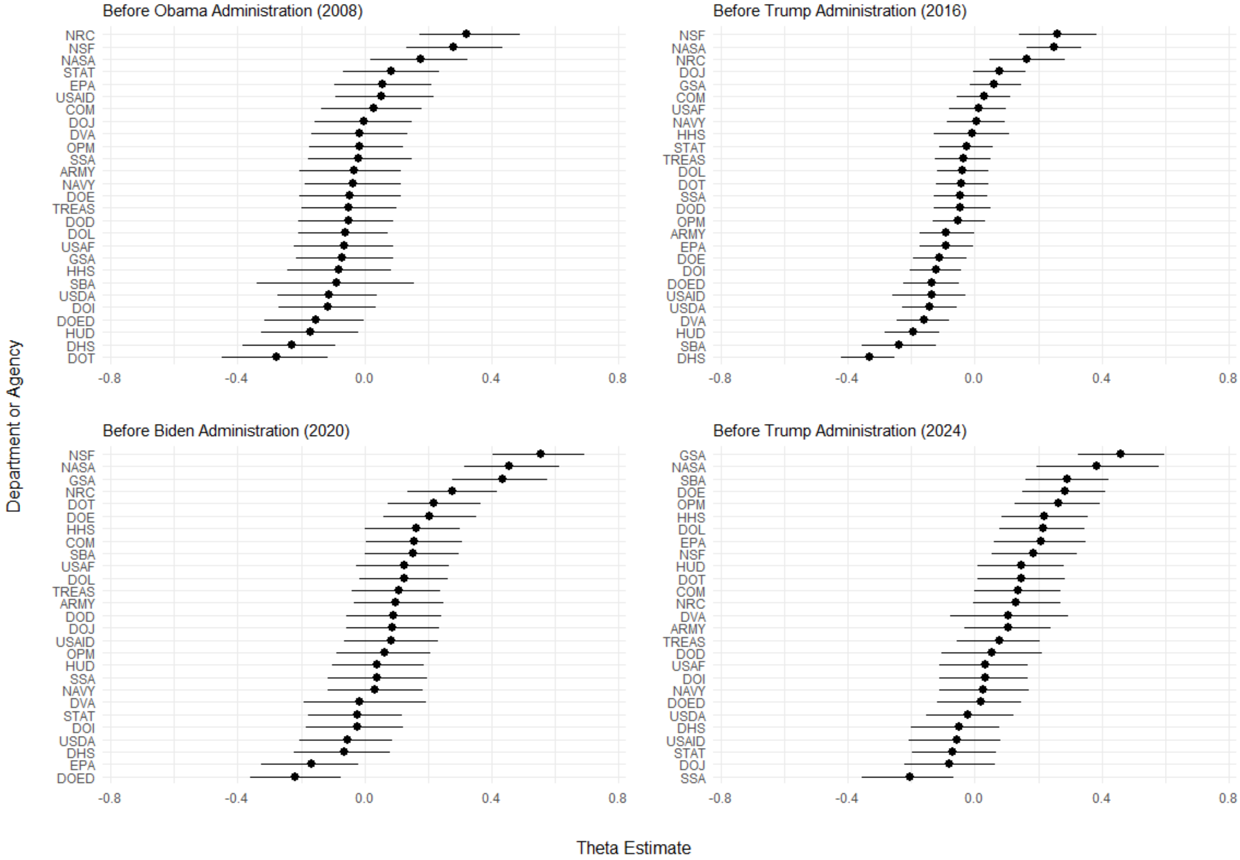
We chose to employ this model for purposes of combining subjective and objective measures in one general model of organizational performance. As noted earlier, the latent organizational performance estimates generated here (based on posterior median) are indistinguishable from those produced by both less and more complex alternative BSEM models. We have been able to aggregate diverse performance information in a way that allows comparison. This approach has potential applicability to other organizations with diverse performance measures and contexts.

We now use the specified model to generate comparable estimates of administrative performance for 135 agencies across 19 years. The estimates provide a type of 'heat map' for decision makers. The organizational performance estimates vary both across agencies and through time. The performance numbers are accompanied by uncertainty estimates since the raw data has errors and our estimates average across tasks, units, and performance criteria. The estimates are not a substitute for a full evaluation of performance. Rather, they offer what can be viewed as a heuristic, something like an 'organizational health scan' that can provide summary aggregate measures of federal agencies' administrative performance for a given year.

Descriptive Patterns of the Organizational Performance Estimates for U.S. Federal Agencies

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 four presidential administrations (i.e., end of 2008, 2016, 2020, and 2024). This information could be helpful in deciding where to allocate time or attention or what kind of person to nominate to lead an agency. During the 2024 transition, for example, President Trump's team might quickly see that some agencies were doing better than others and particular attention might be paid to places like the Social Security Administration and Department of Justice. Both agencies were reporting problems of morale

FIGURE 2: Organizational Performance Estimates, U.S. CFO Act Agencies [Start of Obama, Trump, Biden, Trump Administrations]



Note: The figure includes posterior median estimates and 95% confidence intervals from the end of 2008, 2016, 2020, 2024.

and performance prior to the start of the Trump Administration.[22] By contrast, the Environmental Protection Agency or the Department of Education were struggling prior to the start of the Biden Administration. The low scores for these agencies are hardly surprising given what we know about President Trump's first term efforts to reduce the federal support and reach of both departments. President Trump proposed a 26 percent reduction in EPA funding and an 8 percent cut for Education. These proposals, along with other statements and actions, led to decreases in morale and performance

[22] Among the departments and large agencies, these are the two that were rated the worst places to work in government in 2024. Partnership for Public Service. 2024. *2024 Best Places to Work in the Federal Government*. Partnership for Public Service (https://bestplacestowork.org/rankings/overall/?type=large&subtype=mid&category=overall&). See also, Tom Temin, "Social Security's case backlogs are sliding the wrong way." *Federal News Network*, August 28, 2024.

in both agencies.[23] Both presidents would also see that the National Aeronautics and Space Administration (NASA) and General Services Administration (GSA), two agencies with very different core missions, were doing relatively well.

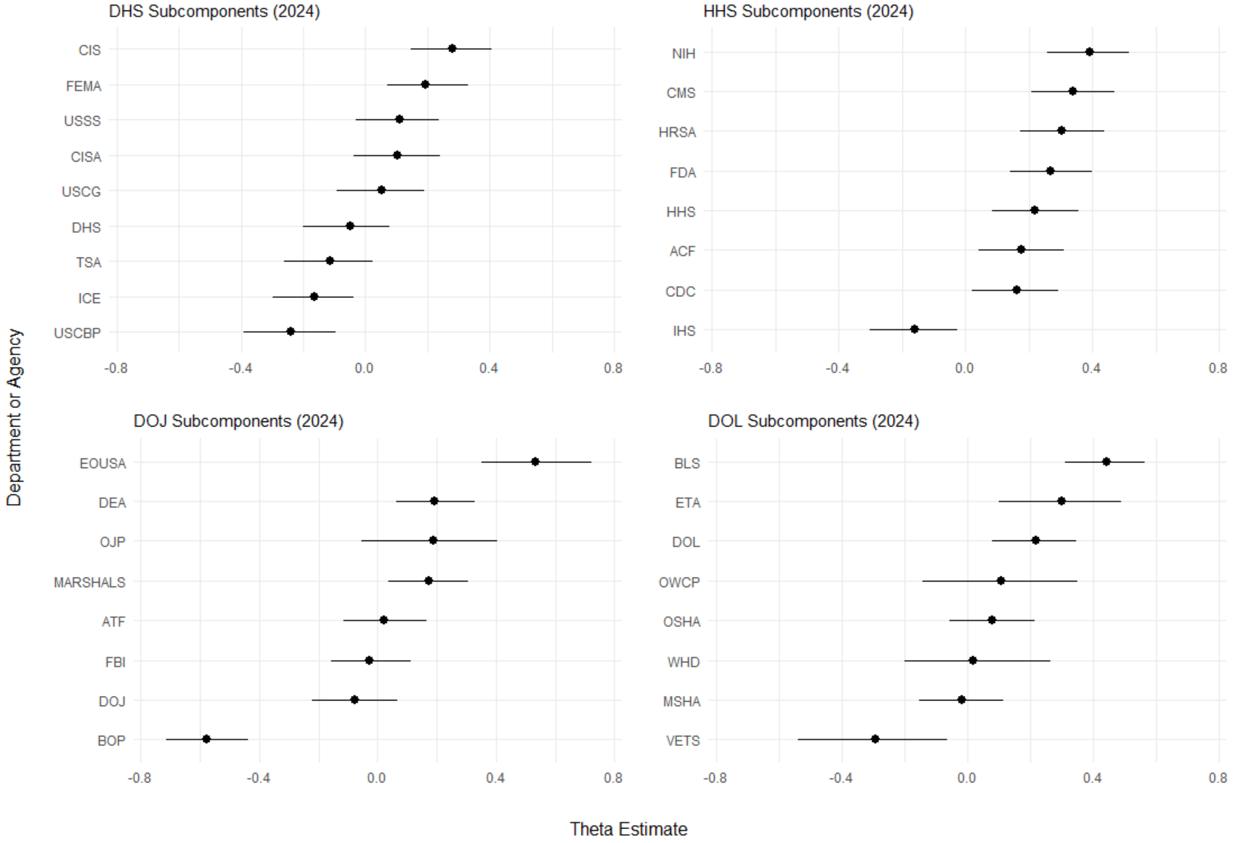
While these estimates appear consistent with common perceptions about the departments, a focus on large entities might obscure the real source of performance problems and successes in the agencies. **Figure 3** includes organizational performance estimates for the major subcomponents of several larger departments. They reveal significant variation. For example, while critics have targeted the Department of Homeland Security across the last three administrations, the data reveal significant variation *within* the department. The Transportation Security Administration, Immigration and Customs Enforcement, and Customs and Border Protection appear to be struggling the most.[24] In the Department of Health and Human Services the National Institutes of Health (NIH) and Centers for Medicare and Medicaid Services (CMS) were estimated to be performing well compared to agencies such as the Indian Health Service (IHS). This is no surprise to those familiar with the department. The CMS and NIH are regularly ranked in the top third of all agencies to work for in government while the IHS has been described as a "never ending crisis."[25] In the Department of Justice, the Bureau of Prisons is estimated to be struggling compared to the rest of the department, something noted by the

[23] 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.

[24] Aaron Blake, "Immigration is now President Obama's worst issue," *Washington Post*, July 31, 2014 (https://www.washingtonpost.com/news/the-fix/wp/2014/07/31/immigration-is-now-president-obamas-worst-issue/); Lesa Jansen and Alan Silverleib, "Obama Unveils Plan to Streamline Government," *CNN*, January 13, 2012 (https://www.cnn.com/2012/01/13/politics/obama-federal-government/index.html).

[25] Andrew Siddons, "The Never-Ending Crisis at the Indian Health Service," *Roll Call*, March 5, 2018 (https://rollcall.com/2018/03/05/the-never-ending-crisis-at-the-indian-health-service/)

FIGURE 3: Organizational Performance Estimates, Subcomponents of Selected Cabinet Departments, 2024



<u>Note:</u> The figure includes posterior median estimates and 95% confidence intervals from the end of 2024 for the Departments of Homeland Security (DHS), Health and Human Services (HHS), Justice (DOJ), and Labor (DOL).

department's inspector general in 2023 and 2024.[26] In the Department of Labor the Bureau of Labor Statistics is estimated to operating well compared to other parts of the department, including the part dealing with veterans' employment. The Veterans Employment and Training Service experienced a dramatic decline in both employee engagement and respect for senior leaders during the Biden

[26] See Office of the Inspector General. Department of Justice. 2023. *Top Management and Performance Challenges Report.* (https://oig.justice.gov/sites/default/files/reports/TMPC-2023.pdf?utm_source=chatgpt.com), Office of the Inspector General. Department of Justice. 2024. *Top Management and Performance Challenges Report* (https://oig.justice.gov/sites/default/files/2024-11/TMPC-2024.pdf?utm_source=chatgpt.com).

Administration and veterans unemployment continues to be a significant problem.[27] In total, the estimates appear to have face validity.

To further explore the estimates, **Table 3** includes a list of the top-10 and bottom-10 agencies across the entire 2002 – 2024 period by average median agency-year performance estimate. Among the high performers are several well-regarded independent agencies as well some science agencies 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 American 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.[28]

Table 3. Average Top and Bottom 10 Performing U.S. Federal Agencies: Average Posterior Median Organizational Performance Estimates, 2002-2024

Department	Agency	Management Performance
Top 10		
Independent	National Aeronautics and Space Administration	0.292
Independent	Federal Trade Commission	0.290
Independent	Federal Energy Regulatory Commission	0.289
Department of the Treasury	Alcohol and Tobacco Tax and Trade Bureau	0.287
Department of Justice	Executive Office of U.S. Attorneys	0.278

[27] Partnership for Public Service. 2024. *2024 Best Places to Work in the Federal Government*. Partnership for Public Service (https://bestplacestowork.org/rankings/overall/?type=large&subtype=mid&category=overall&). Richard Sisk. 2025. "Unemployment for Veterans Spiked More than a Percentage Point to 4.2% in January," *Military News*, February 7, 2025.

[28] 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).

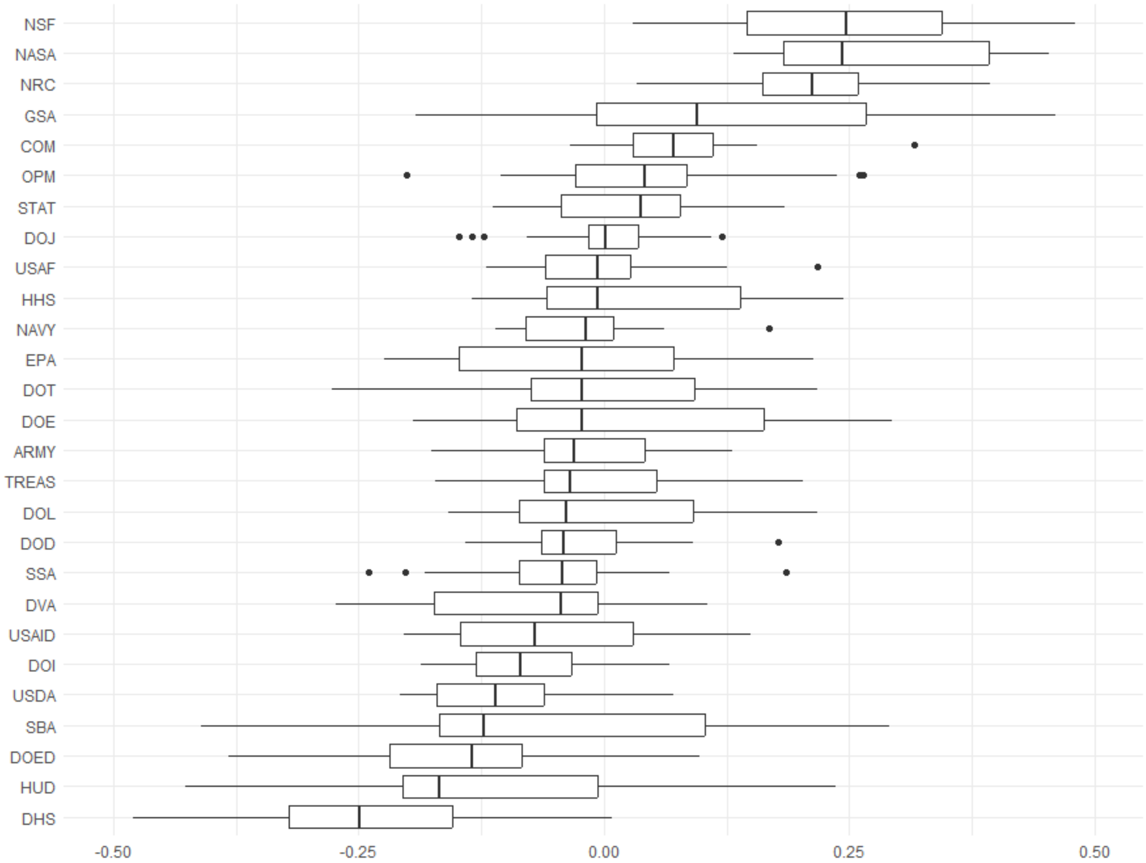
Independent	National Science Foundation	0.272
Department of Transportation	Federal Highway Administration	0.251
Independent	Peace Corps	0.248
Independent	Nuclear Regulatory Commission	0.213
Independent	U.S. International Trade Commission	0.206
Bottom 10		
Department of Health and Human Services	Indian Health Service	-0.219
Department of Justice	Bureau of Prisons	-0.231
Department of Homeland Security		-0.240
Department of Education	Office of Postsecondary Education	-0.257
Department of the Interior	Bureau of Indian Affairs	-0.261
Department of Homeland Security	Customs and Border Protection	-0.306
Independent	U.S. Agency for Global Media	-0.309
Department of Homeland Security	Transportation Security Administration	-0.323
Department of Homeland Security	Immigration and Customs Enforcement	-0.327
Independent	Federal Election Commission	-0.346

Note: BSEM models produce 1,000 posterior distributions for each organizational performance estimate. Table includes annual average of the posterior median estimates for each agency's posterior distributions.

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 4** we graph box plots of the performance estimates for the executive departments and major independent agencies over the 2002-2024 period. A few things stand out. First, some departments and agencies generally performed better across the entire time period. The agencies that stood out in 2024 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.

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 Education seem to regularly be among the low performers. Other agencies such as the Small Business Administration (SBA) and the Department of Transportation fluctuate more. This is reinforced by graphs of agency estimates over time (**Figure 5**). These graphs of estimates show the variation cross-sectionally – e.g., DHS and HUD are on average lower performers—and over time. We note that a dip in the performance of the Department of Veterans Affairs is evident prior to its wait-list scandal in 2014.[29] The efforts President Trump took to redirect the EPA and Department of State are reflected in declines in those agencies during his administration. They notably recover under the Biden Administration.

FIGURE 4: Boxplot of Organizational Performance Estimates of U.S. CFO Act Agencies, 2002-2024



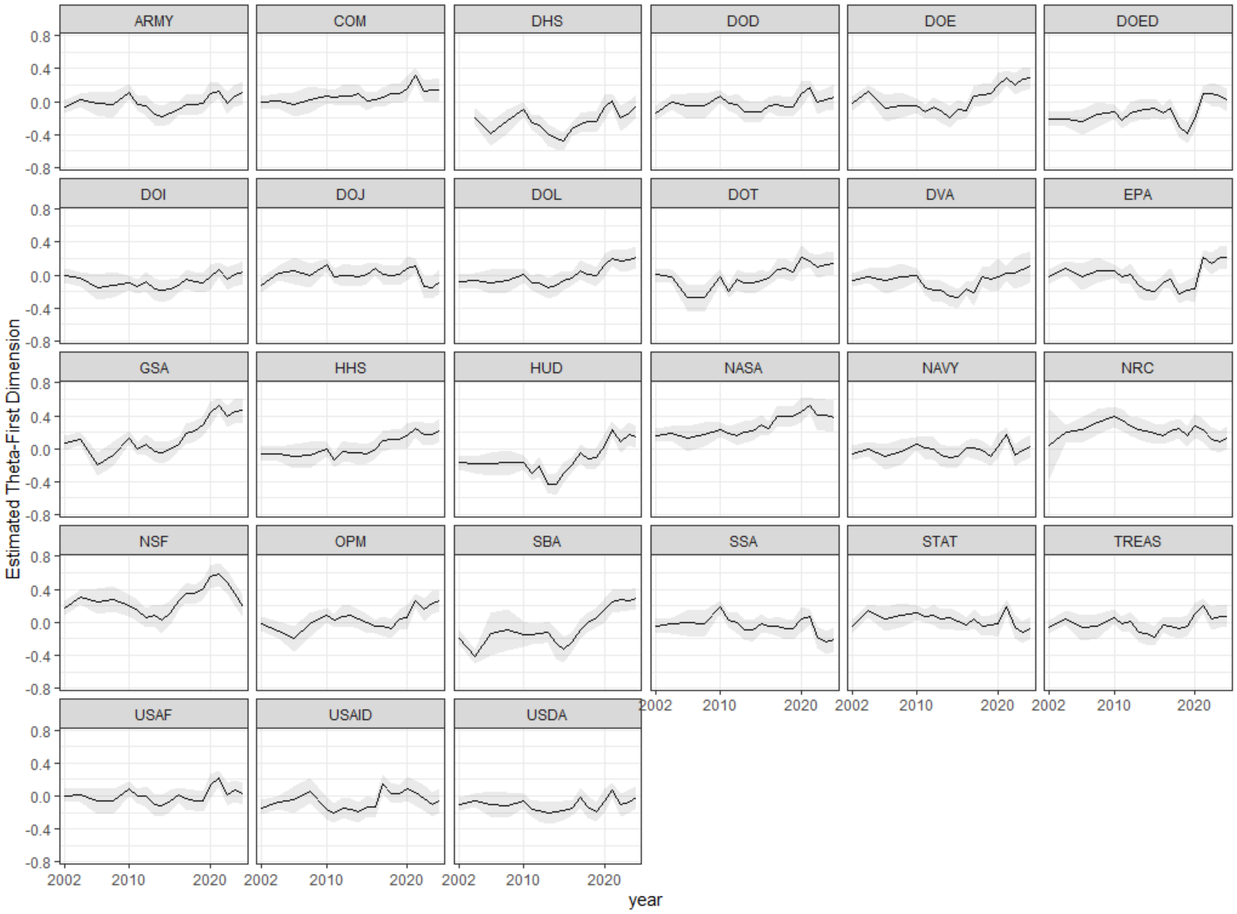
Note: Box plot vertical lines are posterior median estimates. Boxes indicate interquartile range and lines indicate minimum and maximums, excluding clear outliers from distribution (dots).

In total, the descriptive look at the estimates illustrates how the estimates could be useful and demonstrate a significant amount of face validity. Few that follow government closely would be

[29] Scott Bronstein and Drew Griffin. 2014. "A fatal wait: Veterans languish and die on a VA hospital's secret list." *CNN*, April 23, 2014 (https://www.cnn.com/2014/04/23/health/veterans-dying-health-care-delays/).

surprised by the high and low performers and expected patterns of change across time are revealed in the estimates. As we should expect, however, there are also some surprises, cases where we expect change and do not see it and cases where agencies are estimated to be performing worse or better than expected. A useful measure should both strike us as valid and reveal something we did not know. The usefulness of any measure depends upon it telling us something real. The primary way to determine whether a measure tells us something real is whether it seems to correlate with other measures we consider valid.

FIGURE 5: Organizational Performance Estimates of U.S. CFO Act Agencies, 2002-2024



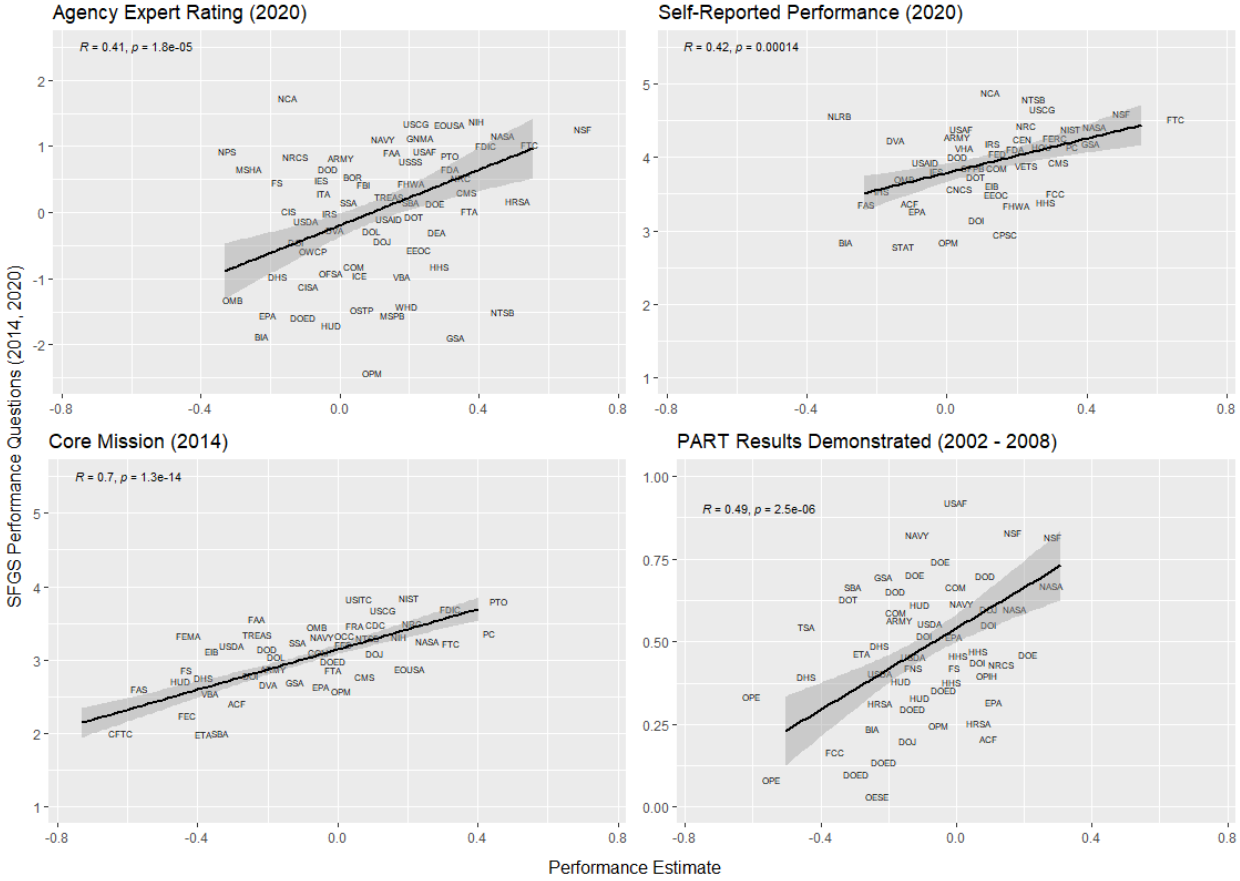
Note: Posterior median estimates and 95% confidence intervals from 2002, 2004, 2006, 2008, 2010-2024.

External Validation with Out-of-Sample Data

We evaluate external validity by comparing our estimates to out-of-sample performance measures excluded from our BSEM model specifications. This includes time-bound measures of the agencies' abilities to accomplish their core missions, useful for *convergent* validity since a FEVS question about core mission is a key element of our latent measure. We also compare our organizational performance estimates to measures of outputs/outcomes more generally. These should also be correlated (i.e., *predictive* validity). Recall, that our estimates are measures of operational performance and are distinct from measures assessing outcome performance. These alternative measures of performance should be correlated with one another, however.

In **Figure 6** we graph the correlations between our estimates and four external performance measures from various years. The top two panels in the figure correlate our estimates 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. 2025). The survey asked a series of questions intended to provide different perspectives on overall agency 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. (2025) generated performance estimates based upon the thousands of ratings federal executives. These scores can be compared to our 2020 estimates.

FIGURE 6: Correlation Between New Organizational Performance Estimates and Other Measures



Note: Panels include correlations between our performance estimates and four outside measures: 1) 2020 elite perceptions of agency performance (Richardson, et al. 2025); 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) program results scores (Gallo and Lewis 2012).

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 and is similar to a FEVS question included in our model estimation. The final panel correlates our performance estimates in 2002, 2004, 2006, and 2008 with average agency PART scores from those same years. dealing with results demonstrated.[30] Specifically, we correlate our estimates with agency average PART scores from the *Results Demonstrated* section of the PART. We include agencies with at least 3 programs evaluated in a year.

The figure reveals a correlation between the 2020 evaluations of federal executives and our 2020 performance estimates, 0.41 ($p = 0.00$) and 0.42 ($p = 0.00$), respectively. As our performance estimates increase, so does the SFGS performance scores for the agency, both the average self-reported performance of agency executives and the agency's reputational score. There are some notable outliers. For example, the Office of Personnel Management (OPM) and the General Services Administration (GSA) do better on our administrative 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 organizational performance estimates. In general, however, higher organizational performance in our estimates is correlated with higher outsider and insider perceptions on agencies' core missions (*convergent validity*). Interestingly, our estimates have a much higher correlation with the other measures. Our measure correlates at 0.70 with agency average responses to questions about the agency's performance on its core mission in 2014. The measures correlate with Bush Administration PART "results demonstrated" scores at 0.49 ($p < 0.01$) (*predictive validity*). Overall, our measure of administrative performance is correlated with other subjective and objective measures of overall performance.

External Validation with Excluded FEVS Data

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

[30] 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.

Personnel Management included a series of questions about agency performance that were unique to that year's survey. While the data come from the 2020 FEVS, a survey we use in our estimation, we did not include responses to these survey questions in our models. 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.*

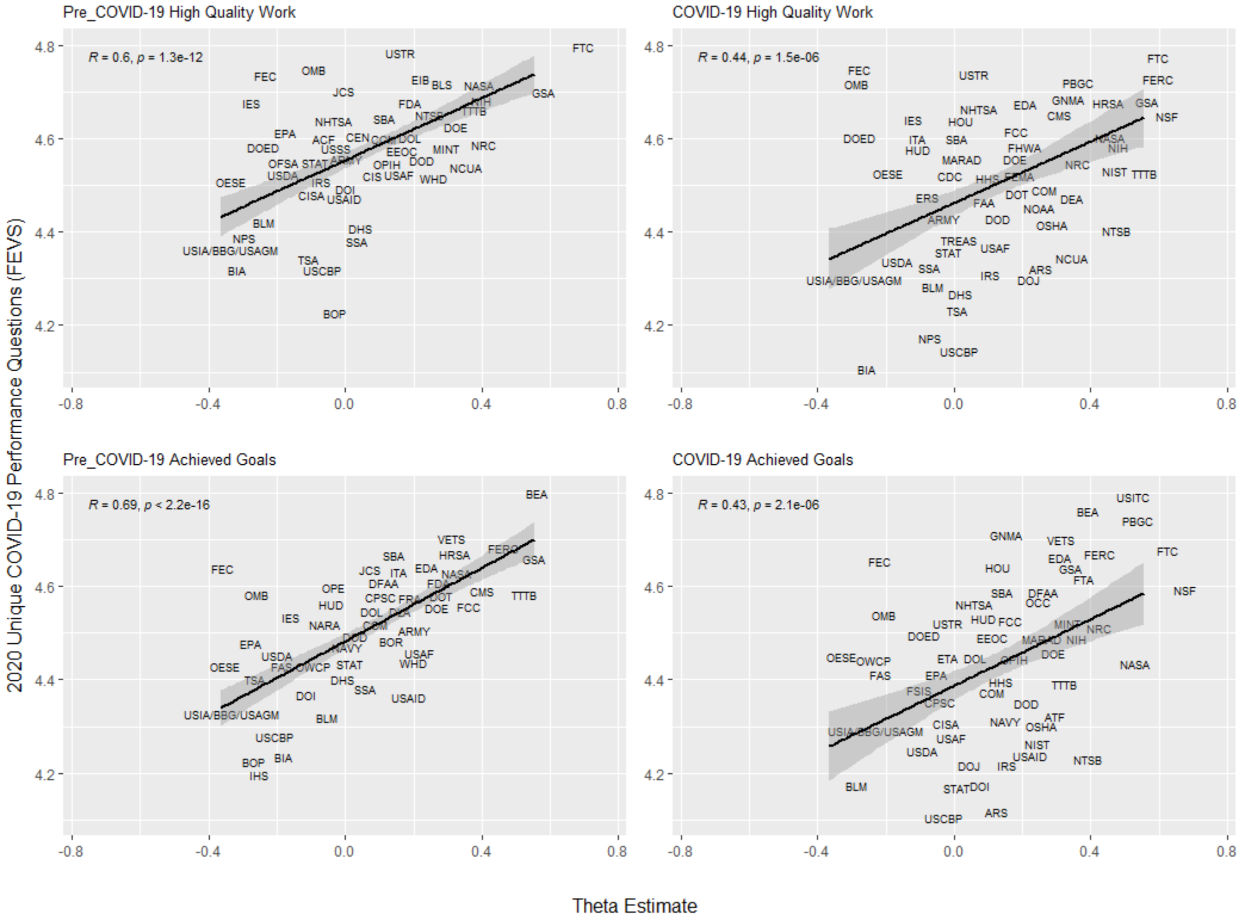
The response categories are *5 "Always"; 4 "Most of the time"; 3 "Sometime"; 2 "Rarely"; 1 "Never"; X "No basis to judge"*. We compare agency average responses to these questions to our estimates from 2020 to see whether our estimates correlate with goal achievement and the quality of agency work. These are useful tests of *convergent* validity since goal achievement and work quality are components of our measure of organizational performance.

When we compare the 2020 performance estimates to the newly added 2020 FEVS questions, the correlations appearing in **Figure 7** are strong, ranging from 0.43 ($p = 0.00$) to 0.69 ($p < 0.00$). The 2020 administrative 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 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 these set of agency 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 our project is to aggregate data like the SFGS and FEVS data with other objective and subjective data to produce better organizational performance measures. The convergent and predictive validity of the estimates provides confidence that the approach has promise.

FIGURE 7: Correlation Between 2020 Organizational Performance Estimates and 2020 FEVS COVID-19 Questions



Note: Panels include correlations between our performance estimates and four FEVS survey measures unique to the 2020 survey: "Prior to the COVID-19 pandemic, my work unit...produced high-quality work"; "Prior to the COVID-19 pandemic, my work unit...achieved our goals"; "During the COVID-19 pandemic, my work unit...has produced high quality work"; "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".

DISCUSSION

With the advent of each new presidential administration, a fresh team must determine which agencies are performing at a high level and which are struggling. This team and their counterparts in

Congress also need the ability to determine whether changes the administration has made to government, including dramatic changes to agency structure and personnel, have led to higher performance. The administration cannot control many of the external factors that influence outcomes but they can shape *organizational performance*, including the quality of management, execution of core tasks (e.g., human resources, financial management), employee morale, and other correlates of organizational health and well-being. It is difficult to measure administrative performance across agencies and time without a principled way of aggregating voluminous amounts of data. This problem is one that affects performance measurement for both the public and private sectors. When it comes to measuring organizational performance, Richard, et al. (2009: 737-738) note:

> *"Performance measurement is further complicated by the availability of the data needed to construct the measures and the need to carefully specify how the data and measures relate to other constructs in a model and to one another… There is little agreement between researchers on either an accepted definition of performance or the appropriate structural form of the relationships between measures."*

This paper has attempted to provide a systematic way of aggregating organizational performance information to provide a roadmap for those managers in the executive and legislative branches seeking to measure and improve agency 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 effective performance measurement regime. Both the measures and statistical method we propose and evaluate here, however, constitute an important step forward in thinking about how to aggregate different performance information. We assume that there is true latent administrative 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.

Several notes of caution are in order. First, the reliability of our estimates depends upon the quality of available data. In the context of the U.S. federal government there is more data for some agencies than others and in some years than others. For example, we are limited in our ability to generate estimates prior to 2010. Similarly, should a new administration reduce the quality or amount of performance data, or introduce bias into the data, this will make the efforts like this one more difficult.

In addition, most of the available performance information is survey data. Respondents to government surveys may be better or worse equipped to observe performance and their own perceptions may be shaped by implicit agency benchmarks or political views (e.g., Meier and O'Toole 2012; Meier, et al. 2015).[31] While some scholars rightly question the use of subjective measures, subjective measures load highly in our models. We note, however, that our measures come from different sources, subjects, and instruments and are less subject to common source bias. The helpfulness of the subjective measures in our models is likely since they are aligned with the key concept.[32] This is particularly the case since we adjust for measurement error in our BSEM models,

[31] Scholars using our estimates that are concerned about implicit benchmarks may consider estimating models with agency fixed effects to isolate within-agency changes in performance.

[32] Private sector management research on organizational performance measures underscores the limits of objective measures, as well as the validity of subjective measures (e.g., Singh, Darwish, and Potocnik 2016). In the realm of public management research, Schacter (2010: 562) concludes that the distinction between objective versus subjective measures should be guided by the same measurement principle: "*How well does the indicator help the agency move toward attaining the underlying conceptual goal?*" – in our case, measuring organizational performance.

partly by including objective measures alongside subjective measures. The BSEMs also account for measurement error by separating the variance common to all the indicators of a particular construct from the variance unique to a particular measure (Lee 2007).

Second, to generate estimates for government agencies we must average across tasks, units, and criteria. Yet, agencies have multiple, competing, and often unclear goals (Chun and Rainey 2005). Indeed, better performance on one goal may lead to poorer performance on another. In addition, *effective* performance may not equate with *efficient* performance or *customer satisfaction*. At a fundamental level, the process of aggregating might produce biased or inaccurate assessments of how well the entire organization is functioning from a holistic perspective. This is a consideration to keep in mind. There are inherent challenges in aggregating across so many different kinds of measures. That said, when public officials make evaluations informally of how well an organization is performing, they are implicitly aggregating a lot of different performance information, including stories they read, reports consumed, personal experiences, etc. If aggregation is unavoidable in assessing organizational performance, it should be done in a principled and transparent manner. The alternative is to focus on single measures of organizational performance. This is unlikely to be adequate for accurately characterizing *overall* organizational performance. This is particularly the case since many available measures may not be correlated with the concept of interest. Our approach provides a transparent way of aggregating available information and assessing which measures are useful (and which are not) for measuring the concept of interest (i.e., organizational performance).

Third, the usefulness of estimates like ours depends upon their credibility with government leaders and other stakeholders. We have assumed that inputs are generally unbiased and that Republicans and Democrats would agree on good administrative performance if they saw it, even if they disagreed on agency mission. However, increasing political polarization may bring even measures such as employee satisfaction or good procurement outcomes into political contestation. If this happens, the measures would lose credibility and become less useful. Such an eventuality, however, reinforces the importance of relying on multiple different measures collated in a statistically sound and transparent manner. Indeed, one benefit of this approach is that we can account for some amount of error in model estimation and evaluate its effects in validation.

With these caveats, the agency performance estimates we have generated are promising on two levels. First, these estimates contain face validity compared to the perceptions of agency performance of informed observers. Second, the estimates are robust to alternative model specifications, poor item predictors, and alternative model identification choices. Finally, these performance estimates exhibit convergent and predictive validity with subjective and objective out-of-sample measures, showing reasonable correlation with other 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 a presidential administration, 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, 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 bureaucratic accountability.

References

Andersen, Lotte Bøgh, 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. "Subjective 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.

Cheon, Ohbet, Miyeon Song, Austin M. McCrea, and Kenneth J. Meier. 2021. "Health Care in America: The Relationship Between Subjective and Objective Assessments of Hospitals." *International Public Management Journal* 24(5):596-622.

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.

Favero, Nathann Richard M. Walker, and Jiasheng Zhang. 2025. "A Dynamic Study of Citizen Satisfaction: Replicating and Extending Van Ryzin's "Testing the Expectancy Disconfirmation Model of Citizen Satisfaction with Local Government" *Public Management Review* 27(6): 1588-1606.

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.

Gallo, Nick, and David E. Lewis. 2012. "The Consequences of Presidential Patronage for Federal Agency Performance." *Journal of Public Administration Research and Theory* 22(2): 219-243.

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/).

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

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* 25(1): 53–74.

Krause, George A., and Ji Hyeun Hong. n.d. "Organizational Adaptation, Task Complexity, and Effective Administration of Unemployment Programs in the American States." *Journal of Policy Analysis and Management* (*Early View*: https://doi.org/10.1002/pam.70024).

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, Sik-Yum. 2007. *Structural Equation Modeling: A Bayesian Approach*. New York: John Wiley & Sons.

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., and Laurence J. O'Toole, Jr. 2012. "Subjective Organizational Performance and Measurement Error: Common Source Bias and Spurious Relationships." *Journal of Public Administration Research and Theory* 23(2):429-56.

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.

Resh, William G., and Heejin Cho. 2020. "Revisiting James Q. Wilson's *Bureaucracy*: Appointee Politics and Outcome Observability." Manuscript, Georgia State University. Available at SSRN: https://ssrn.com/abstract=3444698 or http://dx.doi.org/10.2139/ssrn.3444698.

Richard, Pierre J., Timothy M. Devinney, George S. Yip, and Gerry Johnson. 2009. "Measuring Organizational Performance: Towards Methodological Best Practice." *Journal of Management* 35(3): 718-804

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*, 86(3): 1110-4.

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 2025. "Measuring the Impact of Appointee Vacancies on U.S. Federal Agency Performance." *Journal of Politics* 87(2):680-95.

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.

Schachter, Hindy Lauer. 2010. "Objective and Subjective Performance Measures: A Note on Terminology." *Administration & Society* 42(5): 550-567.

Shi, Dexin, Taehun Lee, and Alberto Maydeu-Olivares. 2018. "Understanding the Model Size Effect on SEM Fit Indices." *Educational and Psychological Measurement* 79(2): 310-334.

Singh, Satwinder, Tamer K. Darwish, and Kristina Potocnik. 2016. "Measuring Organizational Performance: A Case for Subjective Measures." *British Journal of Management* 27(1): 214-224.

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.

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.

Van Ryzin, Gregg G. 2006. "Testing the Expectancy Disconfirmation Model of Citizen Satisfaction with Local Government.*" Journal of Public Administration Research and Theory* 16(4):599-611.

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: What Government Agencies Do and Why They Do It.* New York: Basic Books.

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, 2002−2024”

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

118	ETA	Employment and Training Administration (DOL)
119	MSHA	Mine Safety and Health Administration (DOL)
120	OSHA	Occupational Safety and Health Administration (DOL)
121	OWCP	Office of Workers Compensation Programs (DOL)
122	VETS	Veterans Employment and Training Service (DOL)
123	WHD	Wage and Hour Division (DOL)
124	FHWA	Federal Highway Administration (DOT)
125	FMCSA	Federal Motor Carrier Safety Administration (DOT)
126	FRA	Federal Railroad Administration (DOT)
127	FTA	Federal Transit Administration (DOT)
128	MARAD	Maritime Administration (DOT)
129	NCA	National Cemetery Administration (DVA)
130	VBA	Veterans Benefits Administration (DVA)
131	VHA	Veterans Health Administration (DVA)
134	ONDCP	Office of National Drug Control Policy (EOP)
135	ACF	Administration for Children and Families (HHS)
136	CDC	Centers for Disease Control and Prevention (HHS)
137	HRSA	Health Resources and Services Administration (HHS)
138	IHS	Indian Health Service (HHS)
139	GNMA	Government National Mortgage Association (HUD)
140	HOU	Office of Housing/Federal Housing Administration (HUD)
141	OPIH	Office of Public and Indian Housing (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 (DOI)
161	BLM	Bureau of Land Management (DOI)
162	BOEM/MMS	Bureau Ocean Energy Management/Minerals Management (DOI)
163	BOR	Bureau of Reclamation (DOI)
164	FWS	Fish and Wildlife Service (DOI)
165	NPS	National Park Service (DOI)
166	USGS	U.S. Geological Survey (DOI)
177	OCC	Office of the Comptroller of the Currency (TREAS)
178	AMS	Agricultural Marketing Service (USDA)

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

Note: There are 137 agencies in the dataset. This assumes that FEMA and the Coast Guard are the same agency before and after their incorporation into the Department of Homeland Security. Otherwise, the number is 139. The number of years per agency varies from 1 to 19 and we average 18.10 years per agency. The years include 2002, 2004, 2006, 2008, 2010-2024. We are able to generate organizational performance estimates for 135 out of 137 agencies. We do not generate organizational performance estimates for the Army Corps of Engineers or the U.S. Postal Service due a lack of data.[1]

[1] To construct our list of agencies, we started with the agencies in Krause and O'Connell (2016). To this list we added 9 agencies that had listings on the Government Accountability Office (GAO) high-risk list but were not in the Krause and O'Connell dataset. We subsequently expanded the list to include the major subcomponents of every executive department. We also added major units of the Executive Office of the President and some of the smaller independent agencies excluded from the first list. We dropped a few agencies for which we could not get performance information, including the intelligence agencies and some of the smaller units in the Executive Office of the President, namely the National Security Council, National Economic Council, and Homeland Security Council. For a full discussion see *Codebook for Krause-Lewis Performance Measurement Dataset*.

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 indicators (e.g., presence of award-winning employees, employee turnover).

Subjective Data: Surveys of Employees and Citizens (2002-2024)

During the 2002 – 2024 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 37 different surveys of federal employees with 32 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 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, 2002-2024

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	117	8
2020	SFGS	125	4
2020	GSA	79	4
2021	MSPB	53	4
2021	FEVS	120	6
2021	GSA	81	4
2022	FEVS	119	5
2022	GSA	87	4
2023	FEVS	30	5
2023	GSA	88	4
2024	FEVS	84	5
2024	GSA	84	8

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.[2] 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."[3] 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.[4] The Partnership generously provided this data for 2020 – 2024. 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*

[2] 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.

[3] 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.

[4] 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.

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

During the 2002 to 2024 period, the 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)[5] 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.[6] 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 – 2024.

[5] 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.

[6] 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.

Government surveys of federal employees have a number of virtues. First, they have large samples and high response rates.[7] Second, they can be disaggregated to almost all of the agencies on our list.[8] Third, the surveys include a number of performance-related questions asked across time. In **Table B3** we include all a table that lists all the performance related questions by survey and year in order to illustrate the overlap.[9] Fourth, government employees are often closest to agency actions and have the most information about agency operations. 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. 2025). For the latter type of questions the authors asked respondents to identify the agencies that they worked

[7] 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).

[8] 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 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.

[9] We include a record of surveys and question wording back to 1996 in Table B3 but use only surveys and questions from 2002 – forward in our analysis.

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. 2025).

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 28 different agencies.

Table B3. Performance Related Survey Questions for Federal Employees, 1996-2024

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, 1996-2024 [continued]

Question	*2014*	*2015*	*2015*	*2016*	*2016*	*2016*	*2017*	*2017*	*2018*	*2018*	*2019*	*2019*	*2020*	*2020*	*2020*	*2021*	*2021*	*2021*	*2022*	*2022*
#	*SFGS*	*FEVS*	*GSA*	*MSPB*	*FEVS*	*GSA*	*FEVS*	*GSA*	*FEVS*	*GSA*	*FEVS*	*GSA*	*FEVS*	*GSA*	*SFGS*	*MSPB*	*FEVS*	*GSA*	*FEVS*	*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, 1996-2024 [continued]

Question #	*2023 FEVS*	*2023 GSA*	*2024 FEVS*	*2024 GSA*
1				
2				
3				
4	x		x	
5				
6				
7				
8				
9				
10				
11				
12				
13				
14	x		x	
15				
16	x		x	
17				
18				
19	x		x	
20				
21		x		x
22		x		x
23		x		x
24		x		x
25				
26				
27				
28				
29				
30				
31	x		x	
32				

Table B3. Performance Related Survey Questions for Federal Employees, 1996-2024 [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"? This is also: Considering everything, how satisfied are you with your organization?
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 satisfied with the quality of support and solutions I received from the acquisition services function during the last 12 months
22	I am satisfied with the quality of support and solutions I received from the financial management function during the last 12 months
23	I am satisfied with the quality of support and solutions I received from the human resources function during the last 12 months
24	I am satisfied 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.[10] 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.[11] 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.[12] 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.[13]

To supplement this data, we collected data on counts of GAO reports from 2002-2023 that resulted from bipartisan requests for GAO investigations.[14] Each Congress, members request hundreds of GAO investigations of federal activities. These requests come from individual members

[10] The GAO is a non-partisan legislative branch agency in the United States responsible for auditing, evaluating and investigating government agencies.

[11] This description is based on GAO's own description of the program (https://www.gao.gov/high-risk-list).

[12] Among the 135 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.

[13] 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.

[14] We thank Cody Drolc for providing us with this data.

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 the 135 agencies in our data, 122 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 evaluated the performance of 1,016 programs on four categories of performance (program purpose and design, strategic planning, program management, and program results). We analyze strategic planning and program management scores here since they are closest to the concept of operational performance. This provides data on 120 agencies.

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 (between 15 and 46 agencies per year, depending upon the number of programs evaluated).

We also include data from Performance and Accountability Reports (PAR) between 2002 and 2011. The Government Performance and Results Act (GPRA) of 1993 required agencies to set performance goals and document progress toward goals. Between 2002 and 2011, agencies identified more than 20,000 goals and reported progress on these goals (Lee and Whitford 2013; Resh and Cho 2020). We use data provided by Resh and Cho (2020) to generate agency-year averages of goals unmet, met, and exceeded for 27 agencies from 2002 – 2011.

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 employee performance award data from the Office of Personnel Management (OPM) for four types of awards: high performance award—rating based (2002 – 2023)[15], high performance award—not rating based (2003 to 2023), individual suggestion/invention award (2002 to 2023)[16], and quality step increases (2002 to 2023).[17]

[15] 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/).

[16] 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).

[17] 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

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.[18] 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.

Finally, we collected data from OPM on employee separations between 2002 - 2023, both aggregate agency-year percentages and turnover percentages for subsets of different kinds of employees (e.g., probationary, experienced). We obtained this data from the Office of Personnel Management's Employee Human Resources Integration (EHRI).

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/).

[18] 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.

Table C1. Comprehensive Listing of Agency Management Performance Estimates from BSEM Model 1

Table C1: Raw Data and Estimates, with Missing Data (2024)

Name	BP	BP	Ag.	Qual.	Qual	Org	Satis	Satis	BPTW	BPTW	Eff.	GSA	GSA	GSA	GSA	Turn	PART	PART
	Median	SD	Miss	Work	Co	Comp	Sup	Sup		Post-	Lead	Acq.	FM	HC	IT	Pct	Sec 2	Sec 3
	θ			Unit	Work	Others		Above		2019								
DOD	0.054	0.081	4.067		4.235							4.90	5.21	4.63	4.84			
ARMY	0.106	0.069	4.138		4.267					70.30		4.71	5.26	4.64	4.83			
USAF	0.033	0.068	4.096		4.195					67.00		4.65	5.31	4.32	4.86			
NAVY	0.026	0.071	4.046		4.244					68.10		4.47	5.14	4.49	4.70			
DOED	0.018	0.068	3.892		4.445					65.90		4.56	4.89	4.28	5.66			
DOE	0.284	0.067	4.237		4.459					77.60		5.26	5.39	4.77	5.57			
HHS	0.220	0.070	4.159		4.417					76.30		4.99	5.21	4.65	5.65			
DHS	-0.049	0.069	3.889		4.086					65.10		4.97	5.13	4.48	5.49			
HUD	0.147	0.069	4.053		4.366					70.50		4.22	5.51	5.31	5.72			
DOI	0.033	0.072	3.918		4.267					70.00		4.82	5.06	4.45	5.42			
DOJ	-0.079	0.071	3.813		4.100					61.30		5.18	5.38	4.84	5.32			
DOL	0.217	0.067	4.151		4.387					71.60		5.35	5.22	5.36	5.81			
STAT	-0.069	0.069	3.903		4.193					62.80		4.66	5.04	4.46	5.06			
DOT	0.146	0.070	4.084		4.343					70.40		5.13	5.24	4.94	5.58			
TREAS	0.077	0.070	3.977		4.295					69.10		4.91	5.45	4.73	5.24			
DVA	0.106	0.097								72.10		4.62	5.21	4.18	5.65			
EPA	0.210	0.073	4.195		4.444					79.90		4.40	5.00	3.82	5.71			
GSA	0.460	0.069	4.393		4.536					85.00		5.37	5.75	5.30	5.94			
NASA	0.383	0.099								81.60		5.21	5.55	5.32	5.58			
SBA	0.291	0.068	4.205		4.442					78.40		5.35	5.25	5.13	5.87			
SSA	-0.203	0.071	3.652		3.996					54.20		5.05	5.51	4.94	5.50			
USAID	-0.057	0.073	3.872		4.207					63.00		5.40			5.10			
USAGM	-0.103	0.075	3.792		4.194					65.20								
OMB	0.290	0.075	4.189		4.481					81.00								
USTR	-0.229	0.121								55.60								
CPSC	0.269	0.122								79.00								
EEOC	0.187	0.072	4.122		4.450					73.70								
FCC	0.225	0.124								76.80								
FEC	-0.209	0.121								56.70								
FERC	0.441	0.078	4.434		4.543					85.00								

FTC	0.332	0.074	4.233		4.669					77.80								
NLRB	-0.128	0.077	3.719		4.394					58.80								
NTSB	0.135	0.122								72.60								
NRC	0.129	0.070	4.161		4.346					68.90		5.53	4.82	4.11	5.86			
SEC	0.380	0.118								84.20								
CEN	0.127	0.067	4.044		4.374					70.90		4.71	5.27	4.66	5.72			
CMS	0.339	0.068	4.280		4.495					79.60		5.63	5.21	5.02	5.83			
DEA	0.193	0.071	4.090		4.347					77.60		5.10	5.41	4.76	5.33			
FAA	0.120	0.069	4.071		4.327					69.10		5.06	5.18	4.80	5.59			
FDA	0.267	0.068	4.251		4.475					79.10		4.80	4.78	4.70	5.77			
FEMA	0.194	0.068	4.130		4.309					74.70		5.10	5.35	4.71	6.05			
IRS	0.033	0.069	3.940		4.275					68.10		4.68	5.35	4.64	5.22			
NHTSA	0.214	0.107								76.00		4.00	5.21	5.67	5.34			
NIH	0.392	0.068	4.317		4.490					81.40		5.27	5.71	5.36	5.80			
NIST	0.288	0.068	4.207		4.429					80.40		4.87	5.59	4.78	5.79			
NOAA	0.060	0.071	4.109		4.322					73.40		4.70	3.86	4.06	5.39			
PTO	0.174	0.070	4.091		4.391					72.80		5.36	5.40	5.06	4.81			
PBGC	0.616	0.075	4.632		4.676					90.10								
OPM	0.265	0.070	4.180		4.441					77.10		5.20	5.20	5.41	5.51			
FDIC	-0.031	0.120								65.20								
USCBP	-0.239	0.072	3.568		3.961					60.30		4.88	5.06	4.64	5.29			
BEA	0.471	0.118								88.40								
EDA	-0.542	0.118								40.60								
ITA	-0.018	0.070	4.086		4.383					66.50		4.33	4.13	3.60	4.74			
CIS	0.280	0.068	4.150		4.366					77.80		5.23	5.70	5.43	5.87			
CISA	0.102	0.070	3.999		4.342					71.70		5.23	5.10	4.65	5.32			
ICE	-0.165	0.068	3.679		4.162					62.70		4.81	5.04	3.53	5.73			
TSA	-0.112	0.070	4.006		3.928					60.70		4.76	4.94	4.51	4.98			
USCG	0.054	0.072	4.136		4.303					75.90		4.24	4.63	3.80	4.13			
USSS	0.109	0.069	4.258		4.279					66.20		4.65	4.91	4.31	5.41			
DARPA	0.287	0.117								79.90								
DCMA	0.144	0.074	4.149		4.334					71.50								
DFAA	0.194	0.074	4.207		4.389					72.20								
DLA	-0.011	0.074	4.029		4.276					61.00								
JCS	0.070	0.121								69.30								
IES	0.167	0.118								73.70								
OESE	0.079	0.120								69.90								

OFSA	-0.156	0.101								60.10		4.40	4.66	3.83	5.72			
BOP	-0.579	0.070	3.227		3.629					41.00		5.05	4.95	4.71	5.10			
EOUSA	0.532	0.094								82.20		5.72	6.26	5.78	5.81			
FBI	-0.028	0.068	3.931		4.216					63.80		4.72	5.30	4.45	5.15			
USM	0.172	0.067	4.173		4.302					71.20		5.02	5.41	4.92	5.36			
OJP	0.187	0.116								75.00								
BLS	0.441	0.067	4.423		4.545					83.10		5.27	5.16	5.63	5.77			
ETA	0.301	0.101								75.10		5.36	5.30	5.23	6.05			
MSHA	-0.020	0.070	3.905		4.243					66.90		4.69	4.38	5.14	5.23			
OSHA	0.077	0.067	4.086		4.268					70.60		3.74	4.71	4.94	5.85			
OWCP	0.106	0.124								71.70								
VETS	-0.294	0.117								52.50								
WHD	0.019	0.118								67.10								
FHWA	0.345	0.067	4.245		4.450					81.20		5.36	5.36	5.45	5.75			
FMCSA	0.312	0.101								76.30		5.58	5.77	4.95	5.37			
FRA	0.200	0.072	3.996		4.448					78.50		5.24	5.67	4.32	5.64			
FTA	0.377	0.069	4.335		4.482					80.60		5.43	5.40	5.31	5.70			
MARAD	0.001	0.098								60.80		5.12	5.17	4.94	5.27			
NCA	0.096	0.098								75.80		3.16	4.90	4.98	5.40			
VBA	0.237	0.102								71.70		5.03	5.66	5.02	5.74			
VHA	0.074	0.095								71.90		4.51	5.17	4.04	5.64			
ACF	0.176	0.065	4.123		4.398					73.60		4.65	5.46	4.38	5.79			
CDC	0.164	0.069	4.069		4.472					74.00		4.79	5.19	4.35	5.65			
HRSA	0.304	0.068	4.291		4.516					76.20		4.63	5.49	4.96	5.70			
IHS	-0.158	0.071	3.763		4.095					66.30		3.65	5.05	3.86	5.26			
GNMA	-0.041	0.123								64.10								
HOU	0.187	0.070	4.216		4.419					76.60		3.77	4.44	4.92	5.72			
OPIH	0.045	0.073	3.978		4.288					69.60								
CFPB	0.101	0.121								71.00								
CFTC	0.176	0.076	4.107		4.434					73.50								
CNCS	0.246	0.116								77.90								
DFC	0.047	0.115								68.40								
EIB	-0.572	0.121								39.40								
MCC	0.142	0.121								72.50								
MSPB	-0.150	0.119								59.00								
NARA	0.071	0.076	4.031		4.395					67.00								
NSF	0.186	0.070	4.136		4.445					68.80		5.56	5.63	4.14	5.97			

PC	0.133	0.122								72.60								
BIA	-0.128	0.070	3.775		4.087					63.50		4.73	5.28	4.03	5.37			
BLM	-0.079	0.068	3.815		4.213					66.80		4.63	4.76	4.09	5.37			
BOEM	0.117	0.113								71.80								
BOR	0.180	0.067	4.161		4.353					76.10		4.37	5.06	4.97	5.40			
FWS	0.048	0.070	3.914		4.342					75.60		4.23	4.60	4.48	5.59			
NPS	-0.110	0.071	3.778		4.189					63.80		4.59	5.15	4.13	5.15			
USGS	0.150	0.074	4.168		4.440					76.70		4.17	4.96	3.37	5.65			
OCC	0.185	0.072	4.188		4.348					78.40		3.50	5.26	4.77	5.42			
AMS	-0.082	0.132										4.54	4.92	4.44	4.75			
ARS	-0.199	0.135										4.06	4.78	3.98	5.04			
FAS	0.020	0.073	4.155		4.400					67.60		4.09	4.24	3.67	4.80			
FNS	0.236	0.071	4.307		4.423					77.30		4.82	5.25	4.27	4.87			
FS	-0.166	0.077	3.725		4.163					62.20								
FSIS	0.111	0.069	4.070		4.235					71.70		5.05	5.17	5.18	5.14			
NRCS	-0.288	0.140										4.02	4.39	3.81	5.01			
OPE	0.104	0.115								71.00								
ATF	0.023	0.072	3.956		4.269					67.10		4.81	5.59	4.28	4.97			
MINT	0.317	0.100								75.50		5.43	5.91	5.11	5.41			
TTTB	0.372	0.119								83.70								
NCUA	0.279	0.074	4.255		4.372					80.00								
USITC	0.249	0.119								78.30								

Note: Empty cells represent missing data in 2024. BP: Bayesian Posterior Estimates: (Median, Standard Deviation).

Table C2: Summary Performance by Agency (2002, 2004, 2006, 2008, 2010-2024)

Agency	Dept	ID	Avg. BP Med.	Avg. BP SD	Avg. BLCI 2.5%	Avg. BUCI 97.5%	Performance Class	Low Count	Low Mod Count	Mod Count	Mod High Count	High Count
									Quintile			
							Average Rank	1st	2nd	3rd	4th	5th
Department of Agriculture	USDA	1	-0.107	0.060	-0.226	0.013	*Low-Moderate*	6	10	2	1	0
Department of Commerce	COM	2	0.079	0.060	-0.041	0.195	*Moderate-High*	0	1	5	11	2
Department of Defense	DOD	3	-0.026	0.061	-0.146	0.093	*Moderate*	0	11	4	3	1
Department of the Army	DOD	4	-0.018	0.061	-0.137	0.102	*Moderate*	2	7	5	5	0
U.S. Air Force	DOD	5	0.000	0.061	-0.123	0.120	*Moderate*	0	8	7	3	1
Department of the Navy	DOD	6	-0.020	0.062	-0.142	0.102	*Moderate*	0	8	9	1	1
Department of Education	DOED	7	-0.129	0.060	-0.248	-0.012	*Low*	8	7	1	3	0
Department of Energy	DOE	8	0.037	0.061	-0.082	0.156	*Moderate-High*	1	8	1	4	5
Department of Health and Human Services	HHS	9	0.035	0.061	-0.087	0.154	*Moderate-High*	0	9	2	3	5
Department of Homeland Security	DHS	11	-0.240	0.062	-0.363	-0.119	*Low*	13	4	1	0	0
Department of Housing & Urban Development	HUD	12	-0.123	0.061	-0.244	-0.004	*Low*	11	3	0	3	2
Department of the Interior	INT	13	-0.078	0.060	-0.198	0.039	*Low-Moderate*	4	10	4	1	0
Department of Justice	DOJ	14	-0.002	0.060	-0.120	0.116	*Moderate*	0	4	10	5	0
Department of Labor	DOL	15	0.005	0.060	-0.111	0.120	*Moderate*	1	9	3	2	4
Department of State	STAT	16	0.023	0.060	-0.093	0.139	*Moderate-High*	0	6	4	8	1
Department of Transportation	DOT	17	-0.007	0.060	-0.127	0.111	*Moderate*	3	5	3	6	2
Department of the Treasury	TREAS	18	-0.010	0.060	-0.128	0.110	*Moderate*	1	9	2	6	1
Department of Veterans Affairs	DVA	19	-0.076	0.068	-0.209	0.057	*Low-Moderate*	7	3	7	2	0
Environmental Protection Agency	IND	21	-0.015	0.060	-0.136	0.102	*Moderate*	5	3	4	4	3
General Services Administration	IND	23	0.160	0.060	0.041	0.276	*High*	1	2	3	5	8
National Aeronautics and Space Administration	IND	24	0.292	0.065	0.166	0.421	*High*	0	0	0	1	18
Small Business Administration	IND	25	-0.053	0.073	-0.196	0.090	*Low-Moderate*	5	7	1	2	4
Social Security Administration	IND	26	-0.045	0.061	-0.168	0.072	*Low-Moderate*	3	7	7	1	1
U.S. Agency for International Development	IND	27	-0.063	0.066	-0.194	0.067	*Low-Moderate*	4	8	3	4	0
U.S. Agency for Global Media	IND	28	-0.309	0.082	-0.472	-0.151	*Low*	14	3	2	0	0
Office of Management and Budget	EOP	29	0.153	0.075	0.006	0.301	*High*	2	1	0	5	11
Office of the U.S. Trade Representative	EOP	30	-0.115	0.101	-0.312	0.082	*Low*	7	6	3	0	3
Consumer Product Safety Commission	IND	33	0.015	0.091	-0.165	0.191	*Moderate*	0	6	5	7	1
Equal Employment Opportunity Commission	IND	34	-0.015	0.079	-0.172	0.137	*Moderate*	4	6	4	2	3
Federal Communications Commission	IND	35	0.036	0.092	-0.144	0.216	*Moderate-High*	2	3	7	2	5

Federal Election Commission	IND	37	-0.346	0.097	-0.537	-0.155	***Low***	17	1	1	0	0
Federal Energy Regulatory Commission	IND	38	0.289	0.073	0.143	0.429	***High***	0	0	0	4	11
Federal Reserve Board	IND	40	-0.001	0.222	-0.440	0.429	***Moderate***	0	0	12	0	0
Federal Trade Commission	IND	41	0.290	0.079	0.135	0.444	***High***	0	0	1	2	16
National Labor Relations Board	IND	43	-0.079	0.084	-0.243	0.083	***Low-Moderate***	4	7	7	1	0
National Transportation Safety Board	IND	44	0.150	0.094	-0.036	0.333	***High***	0	1	4	5	9
Nuclear Regulatory Commission	IND	45	0.213	0.075	0.067	0.361	***High***	0	0	1	3	15
Securities and Exchange Commission	IND	49	0.135	0.084	-0.033	0.298	***High***	3	2	2	4	8
Bureau of the Census	COM	50	0.035	0.064	-0.088	0.161	***Moderate-High***	0	4	8	6	1
Centers for Medicare and Medicaid Services	HHS	51	0.070	0.072	-0.070	0.211	***Moderate-High***	5	1	1	5	7
Drug Enforcement Administration	DOJ	52	0.114	0.070	-0.022	0.250	***High***	0	1	1	13	4
Federal Aviation Administration	DOT	53	-0.016	0.066	-0.145	0.112	***Moderate***	3	7	1	6	2
Food and Drug Administration	HHS	54	0.103	0.065	-0.027	0.234	***High***	0	2	5	4	8
Federal Emergency Management Agency	DHS	55	-0.112	0.075	-0.257	0.037	***Low***	8	4	1	2	3
Internal Revenue Service	TREAS	56	-0.053	0.063	-0.178	0.071	***Low-Moderate***	5	6	4	3	1
National Highway Traffic Safety Administration	DOT	57	-0.055	0.128	-0.306	0.192	***Low-Moderate***	5	2	7	4	1
National Institutes of Health	HHS	58	0.189	0.066	0.058	0.318	***High***	0	0	2	8	9
National Institutes of Standards & Technology	COM	59	0.180	0.071	0.040	0.317	***High***	0	0	1	7	11
National Oceanic & Atmospheric Administration	COM	60	0.050	0.064	-0.077	0.176	***Moderate-High***	0	4	4	10	1
Patent and Trademark Office	COM	61	0.170	0.064	0.043	0.293	***High***	1	2	1	1	14
Pension Benefit Guarantee Corporation	IND	70	0.177	0.092	-0.004	0.353	***High***	0	4	5	2	8
Office of Personnel Management	IND	72	0.043	0.060	-0.076	0.160	***Moderate-High***	1	4	4	7	3
Office of Science and Technology Policy	EOP	73	0.002	0.222	-0.429	0.442	***Moderate***	0	4	8	6	0
Federal Deposit Insurance Corporation	IND	78	0.197	0.098	0.001	0.388	***High***	0	4	1	2	12
U.S. Customs and Border Protection	DHS	79	-0.306	0.066	-0.437	-0.176	***Low***	14	3	1	0	0
Bureau of Economic Analysis	COM	82	0.117	0.180	-0.232	0.469	***High***	0	0	13	2	4
Economic Development Administration	COM	83	-0.094	0.132	-0.355	0.162	***Low-Moderate***	5	2	5	4	3
International Trade Administration	COM	84	-0.083	0.073	-0.226	0.061	***Low-Moderate***	6	7	4	1	1
Citizenship and Immigration Services	DHS	85	0.042	0.070	-0.097	0.178	***Moderate-High***	0	5	4	7	2
Cybersecurity and Infrastructure Agency	DHS	86	-0.034	0.078	-0.187	0.116	***Low-Moderate***	1	1	1	3	0
Immigration and Customs Enforcement	DHS	87	-0.327	0.076	-0.476	-0.178	***Low***	15	3	0	0	0
Transportation Security Administration	DHS	88	-0.323	0.075	-0.473	-0.174	***Low***	14	3	1	0	0
U.S. Coast Guard	DHS	89	0.099	0.067	-0.034	0.229	***Moderate-High***	1	0	2	10	5
U.S. Secret Service	DHS	90	-0.124	0.075	-0.272	0.023	***Low***	6	4	3	5	0
Defense Advanced Research Projects Agency	DOD	91	0.004	0.214	-0.415	0.424	***Moderate***	0	4	13	1	1
Defense Contract Management Agency	DOD	94	-0.082	0.077	-0.231	0.071	***Low-Moderate***	6	6	1	3	3

Defense Finance and Accounting Service	DOD	95	0.020	0.078	-0.134	0.172	*Moderate-High*	3	6	3	2	5
Defense Logistics Agency	DOD	97	0.047	0.074	-0.099	0.190	*Moderate-High*	0	3	7	7	2
Joint Chiefs of Staff	DOD	98	-0.005	0.158	-0.320	0.300	*Moderate*	1	2	12	4	0
Institute of Education Sciences	DOED	108	-0.028	0.137	-0.294	0.243	*Low-Moderate*	4	2	9	1	3
Office of Elementary and Secondary Education	DOED	109	-0.202	0.099	-0.397	-0.007	*Low*	11	3	2	2	1
Office of Federal Student Aid	DOED	110	-0.132	0.079	-0.287	0.022	*Low*	10	3	5	1	0
Bureau of Prisons	DOJ	111	-0.231	0.067	-0.363	-0.099	*Low*	14	4	1	0	0
Executive Office of the U.S. Attorneys	DOJ	112	0.278	0.070	0.138	0.415	*High*	0	1	0	2	16
Federal Bureau of Investigation	DOJ	113	0.057	0.083	-0.109	0.217	*Moderate-High*	2	0	2	13	2
U.S. Marshals Service	DOJ	114	0.073	0.068	-0.063	0.205	*Moderate-High*	1	1	3	11	3
Office of Justice Programs	DOJ	115	-0.009	0.096	-0.199	0.179	*Moderate*	6	5	1	2	5
Bureau of Labor Statistics	DOL	117	0.187	0.068	0.053	0.318	*High*	0	1	3	5	10
Employment and Training Administration	DOL	118	-0.093	0.080	-0.251	0.064	*Low-Moderate*	11	2	1	1	4
Mine Safety and Health Administration	DOL	119	-0.027	0.073	-0.170	0.115	*Low-Moderate*	1	7	7	4	0
Occupational Safety and Health Administration	DOL	120	0.006	0.070	-0.131	0.142	*Moderate*	2	8	1	5	3
Office of Workers Compensation Programs	DOL	121	-0.123	0.080	-0.282	0.031	*Low*	9	1	0	4	0
Veterans Employment and Training Service	DOL	122	0.003	0.132	-0.261	0.257	*Moderate*	2	3	8	3	3
Wage and Hour Division	DOL	123	0.004	0.078	-0.151	0.157	*Moderate*	1	2	8	2	1
Federal Highway Administration	DOT	124	0.251	0.068	0.116	0.383	*High*	0	0	0	2	17
Federal Motor Carrier Safety Administration	DOT	125	0.048	0.105	-0.157	0.254	*Moderate-High*	1	1	5	10	2
Federal Railroad Administration	DOT	126	0.126	0.102	-0.075	0.326	*High*	0	0	5	6	8
Federal Transit Administration	DOT	127	0.072	0.123	-0.174	0.309	*Moderate-High*	1	3	7	3	5
Maritime Administration	DOT	128	0.018	0.127	-0.226	0.266	*Moderate-High*	0	4	9	6	0
National Cemetery Administration	DVA	129	0.100	0.097	-0.091	0.288	*Moderate-High*	0	0	6	8	5
Veterans Benefits Administration	DVA	130	-0.090	0.071	-0.232	0.049	*Low-Moderate*	9	2	3	3	2
Veterans Health Administration	DVA	131	-0.079	0.071	-0.218	0.060	*Low-Moderate*	5	6	6	2	0
Office of National Drug Control Policy	EOP	134	-0.010	0.223	-0.449	0.428	*Moderate*	0	4	13	1	0
Administration for Children and Families	HHS	135	-0.022	0.087	-0.192	0.149	*Moderate*	2	7	6	2	2
Centers for Disease Control and Prevention	HHS	136	0.094	0.065	-0.033	0.220	*Moderate-High*	1	2	2	8	6
Health Resources and Services Administration	HHS	137	0.124	0.094	-0.060	0.309	*High*	1	4	4	2	8
Indian Health Service	HHS	138	-0.219	0.069	-0.355	-0.084	*Low*	15	4	0	0	0
Government National Mortgage Association	HUD	139	-0.064	0.154	-0.364	0.239	*Low-Moderate*	4	4	7	3	1
Federal Housing Administration	HUD	140	-0.070	0.095	-0.258	0.117	*Low-Moderate*	6	6	2	0	5
Office of Public and Indian Housing	HUD	141	-0.082	0.113	-0.306	0.135	*Low-Moderate*	8	2	4	3	2
Bureau of Consumer Financial Protection	IND	143	0.037	0.120	-0.199	0.272	*Moderate-High*	1	2	4	2	4
Commodity Futures Trading Commission	IND	144	-0.030	0.084	-0.200	0.133	*Low-Moderate*	4	4	4	3	4

Corporation for National & Community Service	IND	145	0.004	0.091	-0.174	0.182	*Moderate*	1	5	5	6	2
Development Finance Corporation	IND	146	0.178	0.111	-0.040	0.394	*High*	0	2	3	3	11
Export-Import Bank	IND	147	-0.165	0.109	-0.380	0.050	*Low*	9	4	2	3	1
Millennium Challenge Corporation	IND	150	-0.042	0.106	-0.251	0.163	*Low-Moderate*	5	3	1	7	1
Merit Systems Protection Board	IND	151	0.120	0.087	-0.051	0.288	*High*	2	0	4	5	8
National Archives and Records Administration	IND	152	-0.148	0.080	-0.305	0.006	*Low*	11	2	2	4	0
National Science Foundation	IND	154	0.272	0.067	0.139	0.403	*High*	0	0	1	4	14
Peace Corps	IND	159	0.248	0.118	0.013	0.477	*High*	0	1	1	3	14
Bureau of Indian Affairs	INT	160	-0.261	0.076	-0.410	-0.111	*Low*	16	3	0	0	0
Bureau of Land Management	INT	161	-0.163	0.065	-0.292	-0.035	*Low*	11	6	2	0	0
Bureau of Ocean Energy Management	INT	162	0.089	0.080	-0.066	0.246	*Moderate-High*	1	4	4	4	6
Bureau of Reclamation	INT	163	0.014	0.071	-0.127	0.152	*Moderate*	2	6	3	4	4
Fish and Wildlife Service	INT	164	-0.007	0.067	-0.139	0.123	*Moderate*	0	8	6	5	0
National Park Service	INT	165	-0.191	0.064	-0.316	-0.065	*Low*	13	5	0	1	0
U.S. Geological Survey	INT	166	0.068	0.067	-0.066	0.200	*Moderate-High*	0	2	3	13	1
Office of the Comptroller of the Currency	TREAS	177	0.162	0.069	0.026	0.296	*High*	0	0	1	7	11
Agricultural Marketing Service	USDA	178	-0.008	0.121	-0.249	0.226	*Moderate*	1	7	7	3	1
Animal and Plant Health Inspection Service	USDA	179	-0.033	0.120	-0.271	0.200	*Low-Moderate*	3	5	6	3	1
Agricultural Research Service	USDA	180	-0.047	0.084	-0.210	0.118	*Low-Moderate*	4	7	4	3	1
Economic Research Service (USDA)	USDA	181	0.047	0.146	-0.244	0.331	*Moderate-High*	1	0	10	2	5
Foreign Agricultural Service	USDA	182	-0.190	0.084	-0.355	-0.028	*Low*	10	3	6	0	0
Food and Nutrition Service	USDA	183	0.002	0.121	-0.236	0.237	*Moderate*	4	3	5	5	2
Forest Service	USDA	184	-0.210	0.073	-0.355	-0.069	*Low*	13	5	1	0	0
Food Safety and Inspection Service	USDA	186	-0.031	0.071	-0.172	0.107	*Low-Moderate*	4	7	1	6	1
Natural Resources Conservation Service	USDA	188	-0.064	0.074	-0.212	0.079	*Low-Moderate*	4	6	6	3	0
Immigration and Naturalization Service	DOJ	194	-0.456	0.048	-0.546	-0.356	*Low*	1	0	0	0	0
Office of Postsecondary Education	DOED	196	-0.257	0.096	-0.443	-0.067	*Low*	12	2	0	2	3
Bur. of Alcohol, Tobacco, Firearms, & Exp.	DOJ	197	0.040	0.068	-0.093	0.173	*Moderate-High*	0	7	2	6	3
U.S. Mint	TREAS	198	-0.014	0.080	-0.173	0.143	*Moderate*	4	4	5	2	4
Alcohol and Tobacco Tax and Trade Bureau	TREAS	199	0.287	0.090	0.108	0.462	*High*	0	0	1	2	16
Employment Standards Administration	DOL	200	-0.135	0.087	-0.311	0.034	*Low*	2	2	0	0	0
National Credit Union Administration	IND	202	0.094	0.082	-0.068	0.254	*Moderate-High*	0	1	4	11	3
International Trade Commission	IND	203	0.206	0.092	0.024	0.387	*High*	0	2	3	4	10
Total Average			-0.005	0.086	-0.174	0.162						

Note: BP: Bayesian Posterior Estimates: (Median, Standard Deviation, and 95% Bayesian Credibility Intervals: 2.5% Lower CI & 97.5% Upper CI).

Appendix D

Appendix D compares different model specifications to assess how various performance measures tap into latent operational performance. It also evaluates alternative model identification restrictions. We vary the number of dimensions (one versus two). We include different sets of variables capturing outcome-based performance for the 2nd latent performance dimension (**Models 2** & **4**)[19], as well as disaggregate the 1st dimension by creating sub-dimensions of organizational performance informed by the results of Bayesian exploratory factor analysis (BEFA) (**Models 5** & **6**). In addition, we consider variations of the single dimension BSEM **Model 1**, that omits indicator variables (**Models 3 & 7**), and also add an additional indicator variable (**Model 8**) We conclude by discussing the various model diagnostic tests briefly covered in the manuscript.[20]

Tables D1A and **D1B** compare the factor loading estimates in the manuscript (**Model 1**) to other single dimension BSEM models with different specifications (**Models 3**, **7**, & **8**). The estimates are substantively identical for common covariates that appear across these model specifications. They are also nearly identical for those appearing in two dimensional BSEM models (**Models 2**, **4**, **5**, & **6**).

It is also worth pointing out that BSEM models with a separate dimension (reflecting results or outcomes) in **Models 2** & **4** do not suggest a coherent or unique latent 2nd dimension. We infer this based on both the low convergent validity (Average Variance Extracted [AVE] statistic well below the 0.50 threshold desired value) and low construct reliability (Construct Reliability [CR] statistic falls

[19] **Model 2** contains the same 1st dimension model specification as **Model 1**; whereas **Model 4** has the same 1st dimension model specification as **Model 3**. Unlike **Model 1**, **Model 3** omits all GSA core function survey indicator variables from the latent operational performance (1st) dimension.

[20] More information on these diagnostics (e.g., see Fornell and Larcker 1981).

far below the 0.80 threshold desired value).[21] As a matter of fact, the 2nd dimension capturing outcome-based performance in **Models 2** and **4** are dominated by measurement error variance – as evinced by the exceedingly subpar AVE and CR statistics that range from a low of 0.162 (**Model 2**: AVE) to a high of 0.239 (**Model 2**: CR).[22] Thus, we rule out this pair of two-dimensional BSEM models as being valid when it comes to the latent construct being accounted for by this alternative model identification.

In **Models 5** and **6** we disaggregate the measures used to estimate **Model 1** as a single dimension BSEM model. This involves altering both the model specification and identification restrictions by creating a pair of separate sub-dimensions of **Model 1**. The estimates corresponding to common indicators appearing on the 1st dimension across **Models 1**, **5**, & **6** are substantively identical. The 2nd dimension estimates fare poorly in **Model 5**, as evinced by exhibiting diagnostic statistics well below desired threshold values stated above (AVE = 0.336, CR = 0.600).[23] The latent constructs contain substantial overlap (i.e., are not sufficiently distinct from one another), and hence raise concerns regarding nomological validity due to high inter-factor correlations between latent constructs (**Model 5**: 0.501, **Model 6**: 0.509). Yet, discriminant validity is met based on the AVE statistics exceeding the square inter-factor correlation in all four two dimension models.[24] Because we

[21] $AVE = \dfrac{\sum_{i=1}^{k} SFL_k^2}{\sum_{i=1}^{k} SFL_k^2 + \sum_{i=1}^{k} RV}$; $CR = \dfrac{\left[\sum_{i=1}^{k} SFL_k\right]^2}{\left[\sum_{i=1}^{k} SFL_k\right]^2 + \sum_{i=1}^{k} RV}$.

[22] In addition, the AVE and construct reliability statistics for the latent operational performance (1st) dimension are somewhat lower in these model specifications compared to **Model 1**.

[23] The AVE statistic for the 2nd (sub-) dimension in *Model 6* is strong (0.917), but the CR statistic falls short of the 0.80 threshold (0.758).

[24] $DV : AVE > \rho^2_{F_1,F_2}$, where $\rho^2_{F_1,F_2}$ is the squared latent factor correlation between latent constructs/dimensions.

wish to balance model parsimony against model complexity, we report the **Model 1** estimates as the basis of our analysis in the manuscript and elsewhere in this **Appendix** document.

This decision to focus on **Model 1** estimates is further buttressed by the exceptionally strong positive bivariate correlations for the posterior median and standard deviation estimates involving the common latent operational performance (1st) dimension among **Models 1** through **8**. Since the posterior median estimates constitute the point estimate measures of agency performance proposed in this study, we are encouraged by the high correlations among all the performance estimates. The correlations range between 0.9890 (**Model 4**) and 0.9995 (**Models 2 & 8**). Similarly, the posterior standard deviation estimates are also highly correlated with the reported **Model 1** estimates (low correlation: **Model 3** = 0.9887, high correlation: **Model 2** = 0.9973). This 'bounds' interpretation of the posterior median and standard deviation factor score estimates offer ancillary evidence that our latent measures of organizational performance are insensitive to various model specification and identification choices.

TABLE D1A: Alternative BSEM Models and Model Fit and Diagnostics: MODELS 1−4

Standardized Factor Loadings of U.S. Federal Agency Operational Performance [2,476 – 2,498 Agency-Year Observations, 2002/2004/2006/2008, 2010-2024]

	MODEL 1		MODEL 2		MODEL 3		MODEL 4	
Variable	**1st Dimension**	**2nd Dimension**	**1st Dimension**	**2nd Dimension**	**1st Dimension**	**2nd Dimension**	**1st Dimension**	**2nd Dimension**
FEVS: Fulfilling Agency Mission	0.887*** (0.008)	———	0.888*** (0.008)	———	0.895*** (0.008)	———	0.895*** (0.008)	———
FEVS: Quality of Work Unit [2002-2019]	0.801*** (0.013)	———	0.801*** (0.013)	———	0.803*** (0.013)	———	0.803*** (0.013)	———
FEVS: Quality of Work Unit [2020-2024]	0.770*** (0.027)	———	0.768*** (0.028)		0.801*** (0.024)		0.802*** (0.024)	
FHCS: Organization as a Place to Work Compared to Others	0.978*** (0.019)	———	0.975*** (0.017)	———	0.975*** (0.018)	———	0.974*** (0.018)	———
MSPB: Satisfaction with Supervisor	0.921*** (0.016)	———	0.921*** (0.016)	———	0.901*** (0.021)	———	0.898*** (0.022)	———
MSPB: Satisfaction with Managers Above Supervisor	0.942*** (0.014)	———	0.942*** (0.014)	———	0.919*** (0.019)	———	0.917*** (0.019)	———
OPM: Best Places to Work Score [2002-2019]	0.916*** (0.008)	———	0.917*** (0.008)	———	0.919*** (0.007)	———	0.919*** (0.007)	———
OPM: Best Places to Work Score [2020-2024]	0.848*** (0.018)	———	0.846*** (0.019)	———	0.878*** (0.017)	———	0.879*** (0.059)	———
FHCS: Effective Leadership [2002 & 2004]	0.772*** (0.047)	———	0.775*** (0.046)	———	0.776*** (0.046)	———	0.779*** (0.045)	———
GSA Acquisition	0.495*** (0.038)	———	0.496*** (0.037)	———	———	———	———	———
GSA Financial Management	0.554*** (0.034)	———	0.555*** (0.034)	———	———	———	———	———
GSA Human Capital	0.610*** (0.031)	———	0.611*** (0.031)	———	———	———	———	———
GSA Information Technology	0.489*** (0.036)	———	0.489*** (0.036)	———	———	———	———	———
Agency Turnover (Total Percentage)	−0.085*** (0.024)	———	−0.087*** (0.025)	———	−0.087*** (0.025)	———	−0.087*** (0.025)	———
PART Score (Section 2)	0.215** (0.100)	———	0.219** (0.099)	———	0.214** (0.098)	———	0.204** (0.104)	———
PART Score (Section 3)	0.200**	———	0.195**	———	0.197**	———	0.189*	———

	(0.102)		(0.104)		(0.103)		(0.110)	
OPM Innovation Award Annual Count (AE Adjusted)	———	———	———	0.112*** (0.023)	———	———	———	0.102*** (0.011)
OPM Ratings-Based Cash Award Annual Count (AE Adjusted)	———	———	———	−0.012 (0.021)	———	———	———	−0.010 (0.021)
OPM Ratings-Based Non-Cash Award Annual Count (AE Adjusted)	———	———	———	0.999*** (0.001)	———	———	———	0.999*** (0.001)
OPM Quality Step Increase Annual Count (AE Adjusted)	———	———	———	0.999*** (0.001)	———	———	———	0.999*** (0.000)
GAO High Risk Program Count (AE Adjusted)	———	———	———	−0.532*** (0.176)	———	———	———	−0.246 (0.347)
GAO Bipartisan Legislative Investigations (AE Adjusted)	———	———	———	0.175*** (0.020)	———	———	———	0.176*** (0.020)
PART Score (Section 4)	———	———	———	0.222 (0.512)	———	———	———	−0.567** (0.231)
Model Fit & Diagnostic Statistics								
Comparison Fit Index (CFI)	0.831 [0.823, 0.830]	———	0.904 [0.862, 0.917]	———	0.921 [0.911, 0.932]	———	0.931 [0.880, 0.947]	———
Tucker-Lewis Fit Index (TLI)	0.806 [0.797, 0.816]	———	0.903 [0.860, 0.916]	———	0.905 [0.893, 0.918]	———	0.923 [0.867, 0.941]	———
Root Mean Square Error of Approximation (RMSEA)	0.052 [0.050, 0.053]	———	0.046 [0.042, 0.056]	———	0.043 [0.040, 0.046]	———	0.048 [0.042, 0.065]	———
Deviance Information Criterion	4,219.46	———	75,684.36	———	1,652.21	———	73,129.43	———
Bayesian Information Criterion	4,499.25	———	76,114.90	———	1,862.48	———	73,472.88	———
Average Variance Extracted	0.508	———	0.444	0.162	0.584	———	0.423	0.202
Construct Reliability	0.931	———	0.896	0.239	0.931	———	0.875	0.181
Discriminant Validity	———	———	0.444 > 0.00078	0.162 > 0.00078	———	———	0.423 > 0.00044	0.202 > 0.00044
Nomological Validity	———	———	0.028 (0.025)	———	———	———	0.021 (0.024)	———
Sample of Observations	2,479		2,498		2,476		2,495	

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 ≤ 0.10 ** p ≤ 0.05 *** p ≤ 0.01.

TABLE D1B: Alternative BSEM Models and Model Fit and Diagnostics: MODELS 1, 5−8

Standardized Factor Loadings of U.S. Federal Agency Operational Performance

[2,479 Agency-Year Observations, 2002/2004/2006/2008, 2010-2024]

	Model 1		Model 5		Model 6		Model 7		Model 8	
Variable	**1st Dimension**	**2nd Dimension**	**1st Dimension (1a)**	**2nd Dimension (1b)**	**1st Dimension (1a)**	**2nd Dimension (1b)**	**1st Dimension**	**2nd Dimension**	**1st Dimension**	**2nd Dimension**
FEVS: Fulfilling Agency Mission	0.887*** (0.008)	———	0.893*** (0.008)	———	0.896*** (0.007)	———	0.887*** (0.008)	———	0.887*** (0.008)	———
FEVS: Quality of Work Unit [2002-2019]	0.801*** (0.013)	———	0.803*** (0.013)	———	0.803*** (0.012)	———	0.800*** (0.013)	———	0.801*** (0.013)	———
FEVS: Quality of Work Unit [2020-2024]	0.770*** (0.027)	———	0.808*** (0.024)	———	0.798*** (0.025)	———	0.769*** (0.028)	———	0.768*** (0.028)	———
FHCS: Organization as a Place to Work Compared to Others	0.978*** (0.019)	———	0.973*** (0.017)	———	0.974*** (0.017)	———	0.973*** (0.019)	———	0.973*** (0.016)	———
MSPB: Satisfaction with Supervisor	0.921*** (0.016)	———	0.897*** (0.021)	———	0.900*** (0.020)	———	0.921*** (0.016)	———	0.922*** (0.016)	———
MSPB: Satisfaction with Managers Above Supervisor	0.942*** (0.014)	———	0.916*** (0.019)	———	0.919*** (0.018)	———	0.943*** (0.014)	———	0.944*** (0.014)	———
OPM: Best Places to Work Score [2002-2019]	0.916*** (0.008)	———	0.921*** (0.008)	———	0.919*** (0.008)	———	0.917*** (0.008)	———	0.916*** (0.008)	———
OPM: Best Places to Work Score [2020-2024]	0.848*** (0.018)	———	0.891*** (0.015)	———	0.879*** (0.016)	———	0.847*** (0.019)	———	0.846*** (0.019)	———
FHCS: Effective Leadership [2002 & 2004]	0.772*** (0.047)	———	0.777*** (0.046)	———	0.775*** (0.047)	———	0.778*** (0.047)	———	0.774*** (0.046)	———
GSA Acquisition	0.495*** (0.038)	———	———	0.759*** (0.024)	———	0.764*** (0.023)	0.495*** (0.037)	———	0.496*** (0.037)	———
GSA Financial Management	0.554*** (0.034)	———	———	0.846*** (0.022)	———	0.827*** (0.022)	0.553*** (0.034)	———	0.554*** (0.034)	———
GSA Human Capital	0.610*** (0.031)	———	———	0.701*** (0.027)	———	0.706*** (0.027)	0.610*** (0.031)	———	0.611*** (0.031)	———
GSA Information Technology	0.489*** (0.036)	———	0.464*** (0.037)	———	———	0.465*** (0.037)	0.489*** (0.036)	———	0.489*** (0.036)	———
Agency Turnover (Total Percentage)	−0.085*** (0.024)	———	−0.088*** (0.025)	———	−0.086*** (0.025)	———	−0.089*** (0.024)	———	−0.085*** (0.025)	———
PART Score (Section 2)	0.215** (0.100)	———	———	0.787*** (0.084)	———	0.786*** (0.083)	———	———	0.218*** (0.100)	———
PART Score (Section 3)	0.200** (0.102)	———	———	0.753*** (0.082)	———	0.754*** (0.081)	———		0.202** (0.102)	———

GAO−PARs	———	———	———	———	———	———	———	———	−0.252** (0.124)	———
OPM Innovation Award Annual Count (AE Adjusted)	———	———	———	———	———	———	———	———	———	———
OPM Ratings-Based Cash Award Annual Count (AE Adjusted)	———	———	———	———	———	———	———	———	———	———
OPM Ratings-Based Non-Cash Award Annual Count (AE Adjusted)	———	———	———	———	———	———	———	———	———	———
OPM Quality Step Increase Annual Count (AE Adjusted)	———	———	———	———	———	———	———	———	———	———
GAO High Rish Program Count (AE Adjusted)	———	———	———	———	———	———	———	———	———	———
GAO Bipartisan Legislative Investigations (AE Adjusted)	———	———	———	———	———	———	———	———	———	———
PART Score (Section 4)	———	———	———	———	———	———	———	———	———	———
Model Fit & Diagnostic Statistics										
Comparison Fit Index (CFI)	0.831 [0.823, 0.830]	———	0.936 [0.927, 0.945]	———	0.937 [0.928, 0.947]	———	0.838 [0.830, 0.846]	———	0.835 [0.826, 0.840]	———
Tucker-Lewis Fit Index (TLI)	0.806 [0.797, 0.816]	———	0.925 [0.915, 0.936]	———	0.927 [0.917, 0.938]	———	0.810 [0.801, 0.820]	———	0.813 [0.803, 0.823]	———
Root Mean Square Error of Approximation (RMSEA)	0.052 [0.050, 0.053]	———	0.032 [0.030, 0.034]	———	0.032 [0.029, 0.034]	———	0.058 [0.057, 0.060]	———	0.048 [0.046, 0.049]	———
Deviance Information Criterion	4,219.46	———	3,791.93	———	3,782.02	———	4,568.32	———	4,027.03	———
Bayesian information Criterion	4,499.25	———	4,076.74	———	4,066.75	———	4,813.33	———	4,324.28	———
Average Variance Extracted	0.508	———	0.549	0.336	0.539	0.917	0.574	———	0.482	———
Construct Reliability	0.931	———	0.921	0.600	0.911	0.758	0.943	———	0.920	———
Discriminant Validity	———	———	0.549 > 0.251001	0.336 > 0.251001	0.539 > 0.259081	0.917 > 0.259081	———	———	———	———
Nomological Validity	———	———	0.501*** (0.038)	———	0.509*** (0.038)	———	———	———	———	———
Sample of Observations	2,479		2,479		2,479		2,479		2,479	

<u>Note:</u> 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. ** $p \leq 0.05$ *** $p \leq 0.01$.

Table D2. Alternative BSEM Model Specification Estimates and Correspondence with Model 1 [Reported] Bayesian Posterior Estimates

Table D2A. Correlation of Bayesian Posterior Median Estimates (Models 1−8)

	Model 1 (Reported)	Model 2	Model 3	Model 4	Model 5 (F1a)	Model 6(F1a)	Model 7	Model 8
Model 1	1	________	________	________	________	________	________	________
Model 2	**0.9995**	1	________	________	________	________	________	________
Model 3	**0.9891**	0.9892	1	________	________	________	________	________
Model 4	**0.9890**	0.9891	0.9995	1	________	________	________	________
Model 5 (F1a)	**0.9951**	0.9952	0.9969	0.9968	1	________	________	________
Model 6 (F1a)	**0.9943**	0.9943	0.9982	0.9981	0.9994	1	________	________
Model 7	**0.9963**	0.9965	0.9859	0.9861	0.9911	0.9903	1	________
Model 8	**0.9995**	0.9995	0.9891	0.9889	0.9950	0.9942	0.9963	1

Table D2B. Correlation of Bayesian Posterior Standard Deviation Estimates (Models 1−8)

	Model 1 (Reported)	Model 2	Model 3	Model 4	Model 5 (F1a)	Model 6(F1a)	Model 7	Model 8
Model 1	1	________	________	________	________	________	________	________
Model 2	**0.9973**	1	________	________	________	________	________	________
Model 3	**0.9887**	0.9889	1	________	________	________	________	________
Model 4	**0.9888**	0.9891	0.9974	1	________	________	________	________
Model 5 (F1a)	**0.9924**	0.9926	0.9945	0.9943	1	________	________	________
Model 6 (F1a)	**0.9916**	0.9918	0.9962	0.9960	0.9991	1	________	________
Model 7	**0.9916**	0.9920	0.9828	0.9828	0.9856	0.9850	1	________
Model 8	**0.9972**	0.9976	0.9887	0.9887	0.9925	0.9918	0.9917	1

Appendix References

Fornell, Claes, and David F. Larcker. 1981. "Evaluating structural equation models with unobservable variables and measurement error." *Journal of Marketing Research* 18(1):39–50.

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.

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.

Resh, William G., and Heejin Cho. 2020. "Revisiting James Q. Wilson's *Bureaucracy*: Appointee Politics and Outcome Observability." Manuscript, Georgia State University. Available at SSRN: https://ssrn.com/abstract=3444698 or http://dx.doi.org/10.2139/ssrn.3444698.

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

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 2025. "Measuring the Impact of Appointee Vacancies on U.S. Federal Agency Performance." *Journal of Politics* 87(2):680-95.