Obtaining Comparable Measures of Agency Performance: An Application to U.S. Federal Agencies, 2010-2019[1]

Systematically evaluating the performance of federal agencies in the United States is difficult. The outputs of public sector organizations are difficult to observe, measure, and compare across contexts. Scholars have made important progress measuring comparative agency performance through creative means but critics charge that such measures depend upon questionable self-reports, are limited to specific tasks or contexts that hinder generalizability, or are stymied by disagreements regarding how performance is defined. In this paper, we introduce a new measure of federal agency performance that overcomes many of these difficulties. As a first proof of concept, we generate 460 comparable agency performance estimates for 46 departments and agencies between 2010 and 2019 that vary across agencies and time. We aggregate a vast trove of performance information from dozens of government surveys, employee awards, and other observed measures of performance to generate performance estimates via a multi-rater item response model. We evaluate how well different existing measures of performance contribute to the measuring latent performance, validate the measure with out-of-sample measures of performance from 2020, and explore variation. We conclude with a discussion of how to incorporate new or different performance information and the implications of our findings for the measurement and evaluation of agency performance in the United States and other contexts.

Version 1.0

Comments Welcome

George A. Krause
University of Georgia
gkrause@uga.edu

David E. Lewis
Vanderbilt University
david.lewis@vanderbilt.edu

[1] Paper prepared for presentation at the Public Management Research Conference, Utrecht, Netherlands, June 27-30, 2023. We thank Savannah Farr for excellent research assistance on the GAO-High Risk Program List Data.

Like every American president since Bill Clinton, President Joe Biden has used the authority of the president to direct executive branch agencies to develop clear goals and measure organizational performance.[2] A common feature of modern public sector governance is performance management (Boyne 2010; Kettl 2021; Moynihan 2006, 2008). Among other goals, performance information allows elected officials to hold agencies accountable and allocate attention and resources efficiently (Behn 2002; Yang and Holzer 2006; Poister 2003). Yet, some observers question whether performance measurement can be effective in the public sector (e.g., Heinrich 2002; Radin 2000, 2006; Sanger 2013).

The validity of public sector performance measurement is an important topic since performance is arguably the most important concept in public administration (Andersen, et al. 2016; Rainey 1997). Indeed, we are in what one author calls, "the era of governance by performance management" (Moynihan 2008: 4). Governments across contexts and at all levels have adopted performance measures to inform their budgeting and management processes (e.g., Boyne 2010; Melkers and Willoughby 2005; Moynihan 2006; Poister 2003). 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 comparison across different kinds of programs and agencies (Andrews, et al. 2006; Boyne, et al. 2006). Among the most common complaints about government performance measures is that there is nothing akin to profit or firm value that provides a shorthand measure of comparative performance (Andersen, et al. 2016: 853; Niskanen 1971: 29). In the public sector agencies perform a variety of functions that are hard to observe and hard to connect to changes in outcomes (Wilson

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

1989). While scholars have made important progress measuring comparative agency performance through creative means, existing efforts are often plagued by conceptual and measurement difficulties (Andersen, et al. 2016; Boyne 2010; Boyne et al. 2006). There is a proliferation of measures evaluating different tasks on different dimensions in different parts of agencies, but such measures do not connote overall administrative performance that allow us to compare organizations to one another.

In this paper, we introduce a new measure of U.S. federal agency performance that overcomes many of these difficulties. As a proof of concept, we generate 460 agency performance estimates for 46 departments and agencies between 2010 and 2019 that vary across agencies and time. We describe a way to aggregate a vast trove of subjective and objective performance information at different levels and on different dimensions. We use data from dozens of government surveys, employee awards, and other observed measures of performance to generate performance estimates via a multi-rater item response model.[3] The method provides a means of disentangling how well measures tap into performance and whether it is possible to aggregate different dimensions of performance into one measure (Andrews et al. 2006). We evaluate how well different observed indicators of performance contribute to the measurement of latent performance, validate the measure with out-of-sample measures of performance from 2020, and explore variation. We conclude with a discussion of how to incorporate new or different performance information and the implications of our findings for the measurement and evaluation of agency performance in the United States and other contexts.

CHALLENGES IN COMPARATIVE PERFORMANCE MEASURMENT

Scholars and practitioners have been interested in the systematic measurement of agency performance for some time, with this interest accelerating as part of widespread enthusiasm for the

[3] See Bertelli, et al. (2015) for a similar approach measuring autonomy, satisfaction, and intrinsic motivation in public agencies.

New Public Management (Moynihan 2006; Poister 2003). There is a large literature on why performance management reforms are adopted and whether they contribute to program or organizational improvement (e.g., Kroll and Moynihan 2021; Moynihan 2008; Poister, et al. 2013; Sanger 2013; Wang 2002). Embedded in these evaluations is an important debate about how to meaningfully measure performance in a way that is comparable across contexts. Public organizations can rarely be evaluated with anything like simple private sector metrics such as profit, sales growth, or return on equity (Rainey and Bozeman 2000).[4]

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

[4] 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)

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; see, however, Richardson 2023; Richardson, et al. 2023).

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

While such efforts have helped advance our knowledge and practice of performance measurement, questions remain. Focusing on specific comparable tasks or similar sectors limits making generalizable assessments regarding administrative performance. If we focus on tasks like forecasting or 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 that may or may not be accurate, including self-reports (e.g., Lee and Whitford 2013; Meier, et al. 2015; Richardson, et al. 2023). 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). Government generated agency performance scores such as the U.S. federal government's Program Assessment

Rating Tool (PART) 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. 2023).

What is needed is a measure of organizational performance where the goals are clearly defined and we are clear about the relevant stakeholders (e.g., Republicans and Democrats in government). With such a measure the unit of analysis should be clear (e.g., task, bureau, or agency) and the measure can accommodate and discriminate among various subjective and objective measures (surveys, outputs) on different dimensions of performance (efficacy, satisfaction) in a flexible, reasonable, and transparent way.

DEFINING ELEMENTS OF ADMINISTRATIVE PERFORMANCE

Given the diverse approaches to measuring performance, it is important to be clear conceptually. To begin, we assume that for each agency there is an underlying unobservable latent dimension, agency performance, that is a composite of performance on numerous legally mandated goals or tasks, large and small. To measure it we must rely on various indicators that give us information about the agency on this underlying dimension. The more information we have, the better we can place the agency along this performance dimension.

Of course, not all information is useful or uncontested. Some measures may not reveal much about agreed upon definitions of good performance. Indeed, to measure agency performance, we must first clarify whether measuring performance is even possible given the perspectives of different stakeholders (e.g., Republicans and Democrats). We also need to distinguish contributors to performance from performance itself, disentangle *task* performance from *organizational* performance

at different levels (i.e., performance of a subcomponent versus performance of agency as a whole), account for different dimensions of performance, and clarify the relationship between success and performance.

Different Stakeholder Conceptions of Performance

One difficulty in measuring agency performance is that stakeholders, such as political parties, can disagree about the definition of good performance.[5] This can mean different things. It can mean that parties evaluate performance on different dimensions. For example, one observer may care more about efficacy while another cares more about efficiency (something we discuss further below). More troubling is the possibility that stakeholders accurately observing the same latent performance might classify it differently. For example, a Democrat might suggest that agency actions represent perfect compliance with legal requirements and Republicans would conclude that the same actions do not. We assume here that if stakeholders were able to observe this latent performance dimension perfectly, they would agree on what classifies as good or bad performance. That is, they would agree that an agency is meeting its legal requirements even if they disagree with the agency's legal mandate.

Politicians have policy goals and may prefer that agency officials use their legal authority to pursue some policy goals and not others. This often gets conflated with performance. Agency policy choices influence whether political actors define agency performance as good or bad. When we measure performance, we are not measuring this. Rather, we are interested in evaluating what politicians of different parties or ideological leanings can agree on – the extent to which administrative agencies competently performing their job as prescribed by *legal requirements*. We acknowledge that our approach is limited insofar that there are cases where it can be difficult to distinguish

[5] We focus here are real disagreements in perception rather than efforts to paint performance as good or bad for political or partisan gain.

organizational performance from disagreements over policy goals. We note, however, that legal requirements set a standard of good performance for many government activities.

It is also important to remember that most programs enjoy bipartisan support and many aspects of administrative performance have little to do with policy per se. Indeed, the vast majority of government activities have bipartisan support because they are popular with the public (Bednar and Lewis 2023; Gramlich 2017). This is to be expected since every government activity was supported by majorities in both chambers and the president at the time of enactment. In recent work looking at agency performance ratings by Republicans and Democrats in the United States, there was a strong positive correlation among the ratings (Richardson, et al. 2023). When Democrats thought agencies were performing well, so did Republicans and vice versa. While scholarly attention is naturally drawn to areas of disagreement, there are vast swaths of government activity where there is little or no disagreement, including goals such as effective procurement, safe airports, or an efficient patent system (Richardson 2023).

Measures of Performance vs. Contributors to Performance

Given the difficulty of measuring latent performance, it is common for scholars and practitioners to measure administrative capacity or behaviors that contribute to good performance rather than performance itself (Yang and Holzer 2006: 117). For example, in a social services organization we might measure the number of day care centers funded or employee engagement as measures of performance. In an important sense, neither of these is a measure of performance per se, but we believe that each item measured *contributes* to good performance. Scholars sometimes substitute administrative capacity for performance itself. Higher capacity in the form of more day care centers probably helps the agency achieve its goals. Similarly, an engaged workforce likely increases agency

performance.[6] Neither measure, however, is itself a measure of better health and social welfare in the community. The agency could be performing poorly with a large number of day care centers and high employee engagement.

Being explicit about the relationship between measures of contributors to performance and latent performance can help us properly interpret performance information. First, it helps us prioritize some types of performance related information over others. For example, if we have direct measures of performance ("is your agency performing well?"), these should be prioritized over contributors to good performance (e.g., number of beds funded, employee engagement). Second, it suggests that any one measure of performance is unlikely to be sufficient. Relatedly, administrative capacity is an antecedent for effective administrative performance. Scholars using measures of administrative capacity note that a social services agency that has built capacity in the form of more day care centers or high employee engagement has performed well on an *administrative* task. Information about performance on this task can contribute to our understanding overall performance even though good administrative performance this is not the same as an agency achieving its legally mandated goals of better health and social welfare in the community. When combined with other measures of performance, details about the administrative performance of an agency can be effective at revealing latent overall agency performance.

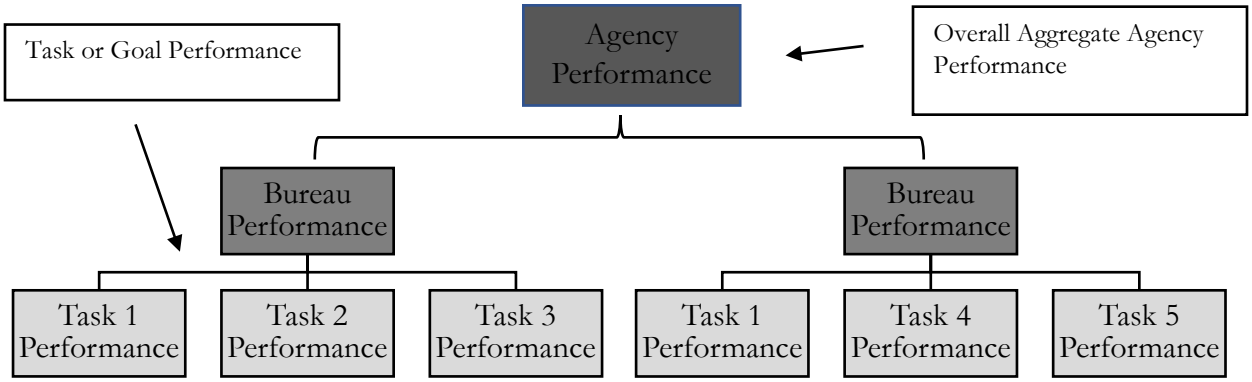
Aggregating Performance Information Across Levels

Agency performance is a composite, aggregating performance on numerous statutorily mandated goals or *tasks*, large and small. Some of these tasks relate to agency core missions and others

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

to auxiliary statutorily mandated tasks, including internal agency management processes like financial management, purchasing, human resources, etc. An agency might very well be performing at a high level on one task (e.g., catching criminals) and poorly on another (e.g., freedom of information requests). Our approach to measuring organizational performance comprises of averaging across performance on these different tasks (Figure 1).

Figure 1. Measuring Department Performance by Aggregating Subcomponent Performance



Depending upon the size of the agency, overall agency performance can also be a composite of the performance of many different agency *subcomponents*. One subcomponent can have high overall performance and another low overall performance. When we measure overall department or agency performance we are implicitly averaging across multiple units (and tasks) within the organization.[7]

Given this complexity, scholars do no observe true performance directly.[8] They observe something analogous to responses to questions on an aptitude test. No one question can reveal true performance but a set of questions properly designed and evaluated can get you closer.[9] In aptitude

[7] Agency performance does not depend upon observability. Agencies can be performing well or poorly on their different tasks whether anyone observes what they do or not.

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

[9] When scholars and practitioners evaluate various performance measures, they see signals of this latent "true" performance. In the same way that individual answers to questions on the three parts of the Grade Record Examinations (GRE) test do not reflect student academic ability directly, performance measures provide a window into something that is hard to observe.

testing, the greater the number of effective questions, the more confident the evaluator. Similarly, each well-defined performance measure provides information about the underlying dimension. Some performance measures help separate *very low* performing agencies from the *low* performing and others *high* performing agencies from *very high* performing. Some measures provide a noisy signal of underlying performance and others a clearer signal. One way to evaluate overall agency performance is to use a method that can incorporate many different measures, accounting for the fact that such measures reflect the complexity of tasks. Some measures will do a better job separating low and high performers. Similarly, some measures will do a better job of mapping an observed output/outcome onto a level of performance. The key is to have a principled way of aggregating this information. Our approach will not infer performance based upon a single measure or small set of individual measures. Rather, it uses a number of different measures, appropriately weighted based upon the informativeness of the measure.

Different Dimensions of Performance

Evaluations of performance on tasks can include performance on different *dimensions* such as efficiency, efficacy, equity, client satisfaction, or other dimensions (Andersen, et al. 2016; Boyne 2002; Gębczyńska and Brajer-Marczak 2020).[10] Some measures tap into performance directly, aggregating across the different dimensions, and others tap into specific dimensions of performance. For example, a survey of executives might ask, "How would you rate the overall performance of the fire department in carrying out its mission?" (i.e., overall performance). By contrast, other measures might tap costs per incident if the task is fire suppression (efficiency), fire deaths per 100,000 population (effectiveness), or percent of fire victims satisfied with fire department response (client satisfaction). Importantly, some measures of organizational performance can measure performance across tasks but

[10] Boyne 2002 identifies 16 different dimensions. For simplicity here we focus on the most common dimensions.

on one dimension. For example, we might evaluate the extent to which an agency is meeting its equity goals across different tasks.

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

Good Performance Does Not Always Mean Success

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

This is true more generally. Quite often, a nontrivial gap exists between agency performance and outcomes. This gap can exist because of unforeseen and uncontrollable factors in the environment. It can also happen because of the complexity of the work. Sometimes the legislature has

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

PERFORMANCE DATA

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

As suggested above, we start with measures that tap agency performance directly. We then supplement this data with other data that measures performance on specific tasks or dimensions. The method we propose here is flexible and can incorporate other forms of performance information.

What we report here is a first attempt at aggregating a large amount of performance information across levels and dimensions. We focus on 46 of the largest and most visible units in the executive branch, starting with a list from Krause and O'Connell (2016) and supplementing with additional agencies that appear regularly in government publications such as the Government Accountability Office's High Risk List.[11] We include a full list of agencies in Appendix A. We imagine future efforts that will add additional agencies, as well as supplement the existing observed indicators, within the same modeling framework.

Objective Data: GAO High Risk List and Employee Awards Data

To begin, we collected data from the GAO's high-risk list.[12] 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.[13] Since its initial publication, GAO has published the list once every Congress (i.e., every two years) and then annually starting in 2010. 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. 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

[11] The list we include in Appendix A includes 61 U.S. federal government agencies. Given gaps in performance data, however, we can reliably estimate performance for 46 U.S. federal agencies.

[12] The GAO is a legislative branch agency in the United States responsible for auditing, evaluating and investigating government agencies. It is non-partisan and insulated from political interference through a 15-year fixed term for its leader, the Comptroller General.

[13] This description comes more or less directly from GAO's own description of the program (https://www.gao.gov/high-risk-list).

their programs off the high-risk list. The list provides a useful cross-agency and temporal source of information about agencies that regularly do well or poorly.

We also make use of data on agencies with employees that are nominated for or winning major awards. Agencies that regularly produce award winning employees are also seeing improvements in programs or efficiency since these criteria determine employee awards. 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.[14] In a given year, an agencies have had up to four employees nominated for performance in different areas and agencies have had up to 3 employees win awards in 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.

Subjective Data: Surveys of Employees and Citizens

Since 2002, the Office of Personnel Management (OPM) has regularly surveyed hundreds of thousands of government employees at different levels about their agencies. OPM has asked federal supervisors and rank-in-file employees about their agencies, including performance overall,

[14] 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).

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.

These surveys have a number of virtues. First, they have a large sample and high response rates.[15] Second, they can be disaggregated to almost all of the agencies on our list.[16] Third, the surveys include a number of performance-related questions asked across time. Finally, the surveys include large enough samples to get agency average responses by different categories of employees—executives/ managers and rank-in-file. In Table 1 we include the questions from the surveys that provide the best performance-related information.[17]

Table 1. Federal Employee Viewpoint Survey Questions

Question 1: How would you rate the overall quality of work done by your work unit? [2002, 2006, 2008, 2010 – 2019]

5 " Very Good"
4 " Good"
3 " Fair"
2 " Poor"
1 " Very Poor"

Question 2: My agency is successful at accomplishing its mission [2010 – 2020]

5 "Strongly Agree"

[15] 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).

[16] Several agencies have opted out of the FEVS and OPM does not report data on some smaller agencies. For example, the following agencies are never included in the FEVS in the 2010-2020 period: Central Intelligence Agency, Federal Deposit Insurance Corporation, Federal Reserve, Office of the Director of National Intelligence, and the U.S. Postal Service. The Department of Veterans Affairs opted out in 2018. We have non-missing values for 460 agencies out of 610 observations between 2010 – 2019. 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.

[17] There are other questions on the FEVS surveys connected in other ways to performance. In Appendix B, for example, we include some questions from 2020 directly related to performance. There are also two other government survey-based sources of performance information. The Merit Systems Protection Board has conducted episodic surveys that include different kinds of performance information since the 1980s. Starting in 2015, the General Services Administration began administering the Customer Satisfaction Survey. This is a survey of tens of thousands of federal users of government human resources, information technology, financial management, and procurement services about the quality of their experience. In effect, agency survey respondents are asked how well their agency does on these key management tasks.

4 "Agree"
3 "Neither Agree nor Disagree"
2 "Disagree"
1 "Strongly Disagree"
X "Do Not Know "

Question 3: Considering everything, how satisfied are you with your organization? [2002, 2004, 2006, 2008, 2010 – 2020]

5 "Very Satisfied"
4 "Satisfied"
3 "Neither Satisfied nor Dissatisfied"
2 "Dissatisfied"
1 "Very Dissatisfied"

Since 2003, the Partnership for Public Service (PPS) has used FEVS data to create a Best Places to Work in Government index. 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."[18] 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. 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.

Starting in 2015, the General Services Administration has surveyed tens of thousands of high-level federal employees (i.e., GS13-15)[19] every year about their experiences with the human resources, financial management, acquisitions, and information technology (IT) functions in their agencies. The

[18] See 2022 Best Places to Work in the Federal Government Rankings (https://bestplacestowork.org/rankings/about, accessed June 19, 2023). Clicking through the links to the rankings themselves provides details on the specific questions used.

[19] 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.

GSA asks these high level employees about the "quality of support and solutions" they receive in these areas. They tap into the internal quality of basic administrative functions within agencies. We include a list of the questions in Table 2.

Table 2. General Services Administration Customer Satisfaction Survey Questions, 2015-2000

Question 1: I am satisfied with the quality of support and solutions I received from the acquisition services function during the last 12 months.

7 "Strongly Agree"
6 "Agree"
5 "Somewhat Agree"
4 "Neither Agree nor Disagree"
3 "Somewhat Disagree"
2 "Disagree"
1 "Strongly Disagree"

Question 2: I am satisfied with the quality of support and solutions I received from the financial management function during the last 12 months.

7 "Strongly Agree"
6 "Agree"
5 "Somewhat Agree"
4 "Neither Agree nor Disagree"
3 "Somewhat Disagree"
2 "Disagree"
1 "Strongly Disagree"

Question 3: I am satisfied with the quality of support and solutions I received from the human resources function during the last 12 months.

7 "Strongly Agree"
6 "Agree"
5 "Somewhat Agree"
4 "Neither Agree nor Disagree"
3 "Somewhat Disagree"
2 "Disagree"
1 "Strongly Disagree"

Question 4: I am satisfied with the quality of support and solutions I received from the IT function during the last 12 months.

7 "Strongly Agree"
6 "Agree"
5 "Somewhat Agree"
4 "Neither Agree nor Disagree"
3 "Somewhat Disagree"
2 "Disagree"
1 "Strongly Disagree"

METHODS

Borrowing from a vast literature in social science evaluating the quality of public sector workforces, citizen satisfaction with public services, and other aspects of administrative capacity and performance (e.g., Bertelli, et al. 2015; Dahlstrom and Lapuente 2017; Nistotskaya, et al. 2021; Richardson, et al. 2018; Teorell, et al. 2011), one can imagine a number of experts rating the performance of government agencies (Richardson, et al. 2023).[20] These raters, from their unique vantage point, evaluate a large number of agencies in different years and identify which are doing well or poorly on the dimensions they care about. Given numerous ratings, provided by many different raters, across all the agencies and years, it would be possible to generate estimates of performance, adjusting for the quality of the raters (e.g., Jilke, et al. 2015; Richardson, et al. 2023).

Of course, we rarely have access to human expert raters, at least not consistently across time (see, however, Nistotskaya, et al. 2021; Teorell, et al. 2021). We do, however, have something akin to raters—i.e., regular evaluations of agency performance. These regular evaluations are analogous to a human rater evaluating an agency each year. For example, one rater might be the Partnership for Public Service's Employee Awards. This list implicitly evaluates administrative performance in each agency in a given year through time. Does the agency have an employee (or more than one) that innovated or performed at such a high level that it merits public recognition? The Partnership determines which employees and agencies are on the list and which are not and for what reasons. The awards are a noisy measure—i.e., the rater has some limitations. For example, this only tells us if one or a few programs in an agency are doing well. It does not tell us which programs are doing poorly. The list also does not parse out whether employees and their agencies are on the list because their

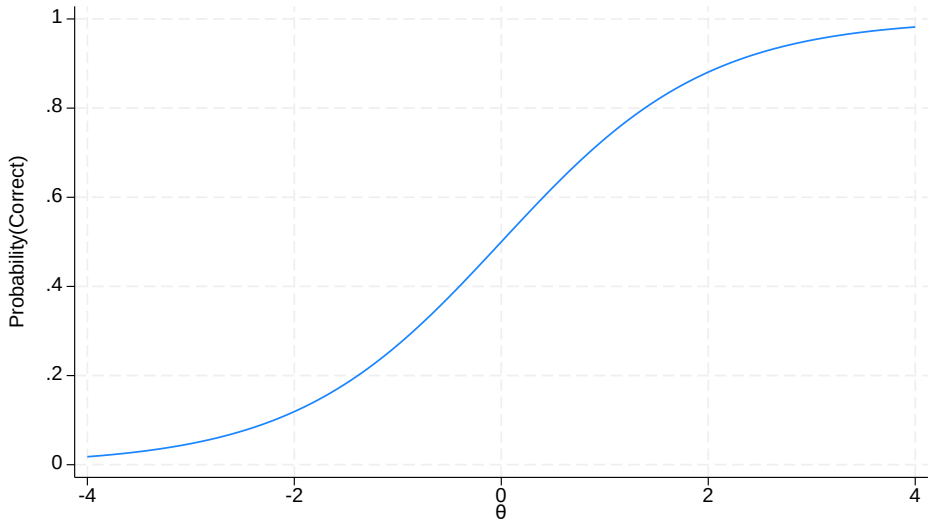
[20] See, particularly, Bertelli, et al. (2015) for a similar approach. In this paper Bertelli and colleagues use aggregated agency responses to federal survey questions to measure agency autonomy, job satisfaction, and intrinsic motivation. The current study forecasts a more granular version of this measurement approach applied to agency performance.

tasks are hard or easy. But, we have other raters, including the group of executives/managers or rank-in-file employees answering questions about their agencies on government surveys. The average response of each group to a key performance question is analogous to a specific rater providing a rating for each agency. Each group-question average effectively "rates" the performance of the agency on a particular task or dimension for a given year. Admittedly, some of the performance information provided by raters is more reliable than others. Fortunately, modern statistical techniques can help us identify empirically which raters do not help parse good from bad performance and how raters map specific measures onto true performance. What is needed are an ample set of measures reflecting performance that can be directly observed to evaluate latent agency performance.

To generate comparable performance measures using these "raters", we employ common models from Item Response Theory (IRT).[21] To illustrate how this works in a simple setting, we assume that there is a latent dimension, θ, that reflects low to high performance. We need a means of recovering where agencies reside on this dimension based upon available performance information. Some agencies, if we could observe performance perfectly, would have a low θ and others a high θ. The goal is to use available data to get accurate placements of agencies on θ.

[21] Much of this discussion comes directly from Raykov and Marcoulides (2018).

Figure 2. Sample Item Characteristic Curve



To do so, we assume that the probability that an agency gets a good, correct, or high score on its performance rating is a function of θ, $P(\theta)$. The function $P(\theta)$ is known as the item characteristic curve (ICC). As θ increases, the probability of a correct answer increases (See Figure 2). Two functions are regularly used to model the item characteristic curve, the normal and the logistic. We use the logistic here and for simplicity describe models based upon the use of binary performance measures (we generalize to ordinal measures below). To begin, the probability of a good performance rating is:

$$P(\theta) = \frac{1}{\{1 + \exp(-x)\}}$$

The relationship between and x and θ is as follows: $x = a(\theta - b)$ so that, after accounting for multiple performance measures, we get:

$$P_j(\theta) = \frac{1}{\{1 + \exp(-a_j\,(\theta - b_j)\}}$$

where $P_j(\theta)$ is the probability of a correct answer on the *j*th performance measure for a given θ. a_j is directly proportional to the steepness of the ICC in its central part, and a_j is referred to as the discrimination parameter. Larger values of a_j mean that a correct response on an observed

performance measure is more responsive to the underlying θ. b_j is a parameter reflecting the fact that some kinds of performance measures pick up variation better at different points on θ. In the same way that some questions on standardized tests distinguish low from moderate performers and other questions distinguish moderate from high performers, some performance measures do the same. b_j reflects this. More formally, b_j is the point on θ where $P_j(\theta) = 0.5$ (i.e., probability of a good rating is 0.5) on the *j*th item. Ideally, the performance measures we use in our IRT model would have high discrimination values and provide information across the appropriate range of θ.

Given polytomous nature of our data with non-identical scales, this simple two-parameter logistic approach can be easily extended by application of the generalized partial credit model (GPCM). The GPCM, originally introduced by Muraki (1992), is a polytomous item response model for discrete ordinal data that permits the mixing of measurement scales within the same IRT modeling framework (e.g., 0-3 scale, 0-1 scale). As a result, this modeling approach allows for performance gradations rather than a simple binary indicator.

The GPCM follows a generalized logistic functional form applied to ordinal response categories. The analytical equation for the GPCM can be stated as follows (e.g., see also de Ayala 2022; Embretson and Reise 2000):

$$P\left(Y_{ij} = k \middle| a_i, b_i, \theta_j\right) = \frac{\exp\left[\sum_{k=0}^{K} a_i\left(\theta_j - b_{ik}\right)\right]}{1 + \sum_{k=0}^{K} k = 0 \ \exp\left[\sum_{k=0}^{K} a_i\left(\theta_j - b_{ik}\right)\right]} \qquad (1)$$

where k is the categorical response observed in a K vector for a given i performance item, a_i denotes the discrimination (slope) parameter of item i, b_{ik} represents the k^{th} step difficulty (location) parameter involving adjacent (or boundary) categories of item i, and θ_j represents the agency j 's latent performance (θ). Given this model depicted in equation (1), our approach is to find a set of parameters that maximizes the likelihood that we obtain the observed performance measures. Across

specifications, we endeavor to find the best fitting model through an analysis of the parameter estimates and model fit statistics. When certain performance measures do not help us place an agency on θ effectively (i.e., a_j is close to 0), this informs model choice. Similarly, we want to include measures that allow us to distinguish performance across a range of values. When comparing different model specifications, overall fit statistics inform choice among models. Specifically, we examine the size and precision of the discrimination and difference parameter estimates for the different performance measures (with standard errors clustered on agency). When comparing among models we use likelihood ratio tests for nested models and use the smaller of the Akaike Information Criterion and Bayesian Information Criterion scores to arbitrate among models.

Generating Administrative Performance Estimates ($\hat{\theta}$) from the GPCM Model

The ultimate goal of this exercise is to generate an estimate of the predicted latent agency performance – i.e, $\hat{\theta}_j$. To do so, we compute the empirical Bayes posterior means and corresponding posterior standard errors (via the posterior standard deviations) to generate values which vary across observation based on available information from item variables. The advantage of this approach is that the $\hat{\theta}_j$ estimates are predicated on the conditional posterior distribution of the latent trait, and hence, allows for variability estimates that are not fixed across the entire sample. The analytical solution for the empirical Bayes mean and corresponding empirical Bayes standard errors generated from equation (1) are given as follows (Stata Corp. 2022, 157-158):

$$E\left(\hat{\theta}_j\right) = \int_{-\infty}^{\infty} \theta_j \omega\left(\theta_j \left| y_j; \hat{a}_i \hat{b}_{ik}\right.\right) d\theta_j, \tag{2}$$

and

$$Var\left(\hat{\theta}_j \left| y_j; \hat{a}_i \hat{b}_{ik}\right.\right) = \int_{-\infty}^{\infty} \left(\theta_j - \hat{\theta}_j\right)^2 \omega\left(\theta_j \left| y_j; \hat{a}_i \hat{b}_{ik}\right.\right) d\theta_j. \tag{3}$$

Based upon the estimates derived from the GPCM (equation 1), we obtain predicted agency performance scores and corresponding variable measures of uncertainty surrounding these estimates generated from equations (2) and (3), respectively. In turn, these estimates will allow one to compare how U.S. federal agency performance systematically varies both across time and space.

EMPIRICAL RESULTS

We begin in Table 3 by presenting the basic results. The three models have slightly different specifications based upon the inclusion or exclusion of the Best Places to Work ratings. We vary the specifications since the Best Places to Work ratings are an index calculated based, in part, on a question from the FEVS that we use in other model specifications. We estimate models with the Best Places to Work ratings but *without* the specific FEVS question (1), models without the Best Places to Work ratings (2), and a model with both (3). The results do not vary too much across specifications.

The table includes the GPCM estimates for both the discrimination (a_j) and difficulty (b_j) parameters for each performance measure. The discrimination parameter (slope) estimates help us see which items do the best differentiating probabilities of a "correct" answer (i.e., a good value on a performance measure). The FEVS performance survey instrument ($2.201 \leq a_i \leq 3.570$), along with OPM's Best Places to Work Index (derived from FEVS survey instruments) ($3.291 \leq a_i \leq 6.391$), are best able to distinguish performance across the sample of observations. Performance assessments of work units by non-supervisory employees are better able to distinguish among agency-year observations than assessments made by executives and supervisory employees (*FEVS Work Unit [Lower Level]*: $1.626 \leq a_i \leq 1.751$; cf. *FEVS Work Unit [Lower Level]*: $0.925 \leq a_i \leq 1.111$).

TABLE 3: IRT Model Estimates of U.S. Federal Agency Performance [46 Agencies × 10 Years]

Variable	Model 1	Model 2	Model 3
GSA Acquisition: Discrimination	0.702***	0.679***	0.622***
	(0.247)	(0.220)	(0.216)
Difficulty: 2nd Quartile versus 1st Quartile	−0.494	−0.473	−0.477
	(0.616)	(0.641)	(0.686)
Difficulty: 3rd Quartile versus 2nd Quartile	0.106	0.150	0.104
	(0.595)	(0.609)	(0.661)
Difficulty: 4th Quartile versus 3rd Quartile	0.858	0.890	0.830
	(0.626)	(0.646)	(0.690)
GSA Financial Management: Discrimination	0.456**	0.459**	0.455**
	(0.215)	(0.194)	(0.202)
Difficulty: 2nd Quartile versus 1st Quartile	−0.081	−0.068	−0.085
	(0.697)	(0.698)	(0.689)
Difficulty: 3rd Quartile versus 2nd Quartile	−0.124	−0.092	−0.145
	(0.850)	(0.849)	(0.852)
Difficulty: 4th Quartile versus 3rd Quartile	0.653	0.680	0.613
	(0.920)	(0.913)	(0.915)
GSA Human Capital: Discrimination	0.662**	0.749**	0.660**
	(0.301)	(0.310)	(0.287)
Difficulty: 2nd Quartile versus 1st Quartile	−0.490	−0.491	−0.489
	(0.455)	(0.404)	(0.445)
Difficulty: 3rd Quartile versus 2nd Quartile	0.214	0.224	0.194
	(0.581)	(0.523)	(0.581)
Difficulty: 4th Quartile versus 3rd Quartile	0.611	0.647	0.562
	(0.671)	(0.602)	(0.660)
GSA Information Technology: Discrimination	0.374***	0.378***	0.429***
	(0.107)	(0.094)	(0.113)
Difficulty: 2nd Quartile versus 1st Quartile	−0.063	−0.050	−0.128
	(0.960)	(0.976)	(0.849)
Difficulty: 3rd Quartile versus 2nd Quartile	0.351	0.376	0.282
	(0.825)	(0.813)	(0.722)
Difficulty: 4th Quartile versus 3rd Quartile	0.859	0.882	0.786
	(0.974)	(0.969)	(0.863)
FEVS Work Unit (Upper Level): Discrimination	0.995***	1.111***	0.925***
	(0.284)	(0.303)	(0.256)
Difficulty: 2nd Quartile versus 1st Quartile	−0.689**	−0.669**	−0.681**
	(0.285)	(0.264)	(0.295)
Difficulty: 3rd Quartile versus 2nd Quartile	−0.099	−0.070	−0.102
	(0.266)	(0.246)	(0.271)
Difficulty: 4th Quartile versus 3rd Quartile	0.568*	0.602**	0.544*
	(0.306)	(0.280)	(0.315)
FEVS Work Unit (Lower Level): Discrimination	1.751***	1.626***	1.711***
	(0.517)	(0.426)	(0.435)
Difficulty: 2nd Quartile versus 1st Quartile	−0.711***	−0.680***	−0.707***
	(0.208)	(0.214)	(0.203)
Difficulty: 3rd Quartile versus 2nd Quartile	−0.081	−0.046	−0.085
	(0.188)	(0.191)	(0.180)
Difficulty: 4th Quartile versus 3rd Quartile	0.624***	0.645***	0.603***
	(0.233)	(0.243)	(0.227)
FEVS Performance (Upper Level): Discrimination	2.968***	3.570***	2.346***
	(0.767)	(0.910)	(0.582)
Difficulty: 2nd Quartile versus 1st Quartile	−0.727***	−0.694***	−0.723***
	(0.187)	(0.178)	(0.197)
Difficulty: 3rd Quartile versus 2nd Quartile	−0.098	−0.066	−0.102

	(0.153)	(0.142)	(0.151)
Difficulty: 4th Quartile versus 3rd Quartile	0.685***	0.707***	0.672***
	(0.188)	(0.174)	(0.191)
FEVS Performance (Lower Level): Discrimination	2.689***	2.201***	2.474***
	(0.672)	(0.513)	(0.548)
Difficulty: 2nd Quartile versus 1st Quartile	−0.699***	−0.669***	−0.695***
	(0.175)	(0.186)	(0.172)
Difficulty: 3rd Quartile versus 2nd Quartile	−0.054	−0.016	−0.057
	(0.173)	(0.178)	(0.166)
Difficulty: 4th Quartile versus 3rd Quartile	0.678***	0.706***	0.658***
	(0.208)	(0.219)	(0.202)
OPM Best Places to Work Score: Discrimination	3.291***	_____	6.391***
	(0.606)		(1.935)
Difficulty: 2nd Quartile versus 1st Quartile	−0.588***	_____	−0.574***
	(0.145)		(0.126)
Difficulty: 3rd Quartile versus 2nd Quartile	0.128	_____	0.112
	(0.167)		(0.138)
Difficulty: 4th Quartile versus 3rd Quartile	1.046***	_____	0.968***
	(0.227)		(0.186)
FEVS Organizational Satisfaction (Upper Level): Discrimination	_____	3.795***	2.854***
		(0.759)	(0.548)
Difficulty: 2nd Quartile versus 1st Quartile	_____	−0.660***	−0.687***
		(0.142)	(0.146)
Difficulty: 3rd Quartile versus 2nd Quartile	_____	−0.059	−0.092
		(0.142)	(0.143)
Difficulty: 4th Quartile versus 3rd Quartile	_____	0.650***	0.604***
		(0.176)	(0.179)
FEVS: Organizational Satisfaction (Lower Level): Discrimination	_____	3.096***	6.796***
		(0.633)	(1.616)
Difficulty: 2nd Quartile versus 1st Quartile	_____	−0.641***	−0.661***
		(0.148)	(0.127)
Difficulty: 3rd Quartile versus 2nd Quartile	_____	−0.029	−0.066
		(0.153)	(0.128)
Difficulty: 4th Quartile versus 3rd Quartile	_____	0.693***	0.604***
		(0.186)	(0.154)
SAMMIE Nominations: Discrimination	−0.098	−0.112	−0.092
	(0.112)	(0.112)	(0.110)
Difficulty: 2nd Quartile versus 1st Quartile	−7.639	−6.703	−8.218
	(8.932)	(6.867)	(10.153)
Difficulty: 3rd Quartile versus 2nd Quartile	−8.280	−7.290	−8.891
	(9.156)	(7.061)	(10.312)
Difficulty: 4th Quartile versus 3rd Quartile	−6.849	−6.058	−7.339
	(7.946)	(6.129)	(8.896)
GAO—High Risk Program: Discrimination	−0.136	−0.185	−0.139
	(0.281)	(0.280)	(0.271)
Difficulty: HR Program versus No HR Program	1.420	1.045	1.381
	(3.515)	(2.136)	(3.318)
AIC Statistic	7694.831	8315.606	8976.185
BIC Statistics	7868.343	8501.511	9162.09

Note: Model 1-3 estimates from Generalized Partial Credit Response Models (GPCM). Entries are unstandardized Discrimination and Difficulty parameter estimates. Agency-clustered robust standard errors appear inside parentheses, and probability levels inside brackets. * $p \leq 0.10$; ** $p \leq 0.05$; *** $p \leq 0.01$. N= 460 (46 agencies, 10 years). Log Pseudo-Likelihood is −3805.42, −4112.80, and −4443.09, respectively. LR equivalence tests comparing GPCM v. PCM are 614.49***, 780.57***, and 1031.97***, respectively.

Although the estimated discrimination parameters are smaller for the GSA customer satisfaction data, they are still informative and statistically distinct from the null of being indistinguishable across cases. The GSA customer satisfaction assessments reveal that survey instruments pertaining to Acquisitions and Human Capital (i.e., Personnel) are better able to distinguish performance differences ($0.622 \leq a_i \leq 0.749$) than items concerning Financial Management and Information Technology (IT) ($0.374 \leq a_i \leq 0.459$). The variation among these discrimination parameter estimates corroborates the findings from the likelihood-ratio (LR) tests appearing at the bottom of Table 3. The LR tests reject the null hypothesis that the generalized partial credit model (GPCM) (which contain unique discrimination parameters for each observed indicator) is equivalent to a partial credit rating model (PCM) which contains a single fixed discrimination parameter for all items (i.e., assumes a constant discrimination parameter across measures).

The SAMMIES and GAO-HR list discrimination parameters are negatively signed, but close to zero in both numerical and inferential terms. One possible reason for this lack of discrimination for this pair of items is attributable to differential item functioning (DIF) problems relating to large agencies (i.e., executive departments). Large agencies are more likely to have employees recognized for excellence in U.S. federal public service (SAMMIES) because there are more employees. Similarly, large agencies as more prone to having programs placed on the GAO-high risk list because they have more programs. This problem is further exacerbated by the fact that these larger department agencies are constrained to having at least as many, if not more, occurrences of both counts relative to sub-components that are nested within the larger department (e.g., FEMA can never have more programs on the High Risk list than the Department of Homeland Security).

The difficulty parameters (b_{ik}) represent estimates from these various indicator variables, where an agency has a 50% chance of being in either adjacent category for a given indicator (e.g., 1st Quartile−2nd Quartile). Higher b_{ik} estimates represent higher levels of latent performance (i.e., greater

item response difficulty), while lower values represent lower levels of latent administrative performance (i.e., lower item response difficulty). In all but the case of the GAO-HR list, the difficulty parameters follow a pattern, whereby that increases in the quartiles of indicator variables results in higher *relative* levels of latent administrative performance. The difficulty parameter estimates tend to be much less precise for both SAMMIES nominations and GSA customer satisfaction survey indicator variables. Such imprecision might be attributable to the limited temporal frame for our sample (T = 10 years) or too many or too few categories in the ordinal measures.[22] Nonetheless, these model estimates should be interpreted with caution as a preliminary statistical analyses. Yet, it is worth noting that resulting latent administrative performance measures (empirical Bayes means and corresponding standard errors) generated from **Model 1** are highly correlated (Empirical Bayes Posterior Means: $0.950 \leq \rho \leq 0.991$; Empirical Bayes Posterior Standard Errors: $0.795 \leq \rho \leq 0.982$) with other models estimated both included and excluded from this manuscript.[23] Moreover, the discrimination and difficulty parameter estimates generated from different model specifications (**Models 2** & **3**); omitting both SAMMIES nominations and GAO-High Risk List Programs indicator variables; omitting GSA Customer satisfaction indicator variables, and exclude Defense and military agencies from the sample.

Because the model estimates are quite similar across these three model specifications, from this point onward our attention is focused on interpreting **Model 1** since it is the most parsimonious model specification based on exhibiting the lowest AIC and BIC model fit statistics.

---

[22] This is an issue we plan to further explore in the next iteration of this study.

[23] The somewhat lower empirical Bayes posterior standard error correlations reflect more variable empirical distribution of standard errors generated from models (***Models 2, S1.Model 2, S1.Model 2, and S3.Model 2***) that replace the Best Places To Work Index with the FEVS Organizational Satisfaction instrument since the latter is one of the survey instruments used to construct the former indicator variable.

What is more central to our endeavor is the estimates themselves. In Table 4 we include a list of the highest and lowest values of ($\hat{\theta}$) that resulted from model estimation (for the full list see Appendix C). Among the highest scoring agencies across the 2010 – 2019 period are the National Aeronautics and Space Administration (NASA), the Nuclear Regulatory Commission, and the Federal Trade Commission. All three agencies have regularly performed well on the Partnership for Public Service's Best Places to Work rankings and have received other recognition.[24] For example, the Nuclear Regulatory Commission won several awards for information technology.[25] NASA has long been ranked highest in public opinion surveys and has received recognition from outside groups for its social media activities.[26] The Department of Homeland Security (DHS), the U.S. Agency for Global Media, and the Department of Veterans Affairs are among the lowest scoring agencies. Few scholars looking at this list would be surprised. Congress created the DHS in 2003 and it has been plagued by management problems from the start.[27] The agency includes the Transportation Security

[24] Federal Trade Commission employs have won several high profile awards, including American University's Roger W. Jones Award for Executive Leadership (https://www.washingtonpost.com/brand-studio/wp/2021/11/29/defining-public-service-excellence/) and a Sammie. Jessie Bur, "Awards gala to honor feds' public service," *Federal Times*, October 1, 2018 (https://www.federaltimes.com/management/leadership/2018/10/02/awards-gala-to-honor-feds-public-service-innovation/).

[25] Katie Polit, "Federal CIOs, Tech Teams Shine at FITARA Awards," *MeriTalk: Improving Outcomes of Government IT*, October 4, 2019 (https://www.meritalk.com/articles/federal-cios-tech-teams-shine-at-fitara-awards/); John Curran, "USAID Takes Top Spot at FITARA Awards, GSA's Shive is 'Champion'," *MeriTalk: Improving the Outcomes of Government IT*, March 2, 2023 (https://www.meritalk.com/articles/usaid-takes-top-spot-at-fitara-awards-gsas-shive-is-champion/).

[26] See, for example, J. Baxter Oliphant and Andy Cerda, "Americans feel favorably about many federal agencies, especially the Park Service, Postal Service, and Nasa," Pew Research Center, March 30, 2023 (https://www.pewresearch.org/short-reads/2023/03/30/americans-feel-favorably-about-many-federal-agencies-especially-the-park-service-postal-service-and-nasa/#:~:text=Topping%20the%20list%20are%20the,Services%20(HHS%2C%2055%25); Phil Norris, "20 Government Agencies Doing an Amazing Job on Social Media," Blog: Social Media Strategies Summit, April 27, 2023 (https://blog.socialmediastrategiessummit.com/government-agencies-on-social-media/).

[27] See Chris Strohm, "Report gives DHS mixed grades after one year," Government Executive, March 5, 2004 (https://www.govexec.com/defense/2004/03/report-gives-dhs-mixed-grades-after-one-year/16170/); Katherine Mcintire Peters, "Senators: DHS isn't improving management fast enough," Government Executive, September 30, 2010 (https://www.govexec.com/defense/2010/09/senators-dhs-isnt-improving-management-fast-enough/32461/); Dara Lind, "The Department of Homeland Security is a Total Disaster. It's time to abolish it," Vox, February 17, 2015 (https://www.vox.com/2015/2/17/8047461/dhs-problems); U.S. Congress. House. Committee on Homeland Security.

Administration, which manages airport security[28], and the Federal Emergency Management Agency, an agency with two catastrophic responses to hurricanes in the Gulf Coast and Puerto Rico.[29] It has had widely publicized morale problems for two decades.[30] The U.S. Agency for Global Media has widely been criticized for poor management, including during the Trump Administration where the agency's appointed leader was accused of retaliatory personnel actions and wasting funds.[31] The Department of Veterans Affairs has been plagued by scandal and widely panned for its services to veterans, so much so that Congress has provided veterans the ability to pursue healthcare outside the veterans hospitals.[32]

Subcommittee on Oversight and Management Efficiency. 2015. *Making DHS More Efficient: Industry Recommendations to Improve Homeland Security*. 114th Cong, 1st Sess., September 18, 2015.

[28] For performance problems related to TSA see, Ron Nixon, "Congress's List of Grips with T.S.A. is Long, Like an Airport Security Line," *New York Times*, May 12, 2016 (https://www.nytimes.com/2016/05/13/us/politics/congresss-list-of-gripes-with-tsa-is-long-like-an-airport-security-line.html).

[29] See, for example Jennifer Steinhauer and Eric Lipton, "FEMA, Slow to the Rescue, Now Stumbles in Aid Effort," *New York Times*, September 17, 2005 (https://www.nytimes.com/2005/09/17/us/nationalspecial/fema-slow-to-the-rescue-now-stumbles-in-aid-effort.html); Scott Neuman, "5 years on, failures from Hurricane Maria loom large as Puerto Rico responds to Fiona," *NPR*, September 20, 2022 (https://www.npr.org/2022/09/20/1123846384/puerto-rico-hurricane-fiona-hurricane-maria-anniversary).

[30] U.S. Congress. House of Representatives. Committee on Homeland Security. Subcommittee on Oversight, Management, and Accountability. *Seventeen Years Later: Why is Morale at DHS Still Low?* 116th Cong. 2d Sess. January 14, 2020.

[31] See, for example, Reid Standish, "Waste and Abuse of Power at the Broadcasting Board of Governors," *Foreign Policy*, June 17, 2014 (https://foreignpolicy.com/2014/06/17/waste-and-abuse-of-power-at-the-broadcasting-board-of-governors-according-to-audit/); Elizabeth Williamson, "New Scandals Rock Government's Foreign Broadcasting Service," *New York Times*, July 8, 2019 (https://www.nytimes.com/2019/07/08/us/politics/us-agency-for-global-media-scandals.html); Courtney Buble, "Trump's Global Media CEO Abused Authority and Wasted Funds, Review Finds," *Government Executive*, May 15, 2023 (https://www.govexec.com/oversight/2023/05/trumps-global-media-ceo-abused-authority-and-wasted-funds-review-finds/386361/).

[32] See, for example, Michael D. Shear and Dave Phillipps, "Progress is Slow at V.A. Hospitals in Wake of Crisis," *New York Times*, March 13, 2015 (https://www.nytimes.com/2015/03/14/us/obama-va-hospital-phoenix.html); Maggie Haberman and Nicholas Fandos, "Trump Signs Bill Meant to Restore Trust in V.A.," *New York Times*, June 23, 2017 (https://www.nytimes.com/2017/06/23/us/politics/trump-veterans-accountability-bill.html).

Table 4. Top-5 and Bottom-5 Posterior Mean Estimates of Performance ($\widehat{\theta}$), 2010-2019

Agency	Post. Mean	Post. SE	95% LCL	95% UCL	Class	Low	Low/ Mod	Mod	High/ Mod	High
FTC	1.702	0.548	0.628	2.775	High	0	0	0	0	10
NRC	1.656	0.479	0.718	2.594	High	0	0	0	0	10
NASA	1.602	0.464	0.693	2.511	High	0	0	0	0	10
FERC	1.280	0.458	0.383	2.177	High	0	0	1	1	8
NIST (COM)	1.122	0.382	0.373	1.872	High	0	0	0	1	9
FEMA (DHS)	-1.101	0.388	-1.862	-0.341	Low	9	0	1	0	0
DVA	-1.110	0.409	-1.911	-0.309	Low	8	0	2	0	0
USDA	-1.197	0.397	-1.976	-0.419	Low	9	1	0	0	0
USAGM	-1.403	0.505	-2.392	-0.413	Low	8	2	0	0	0
DHS	-1.707	0.518	-2.722	-0.693	Low	10	0	0	0	0

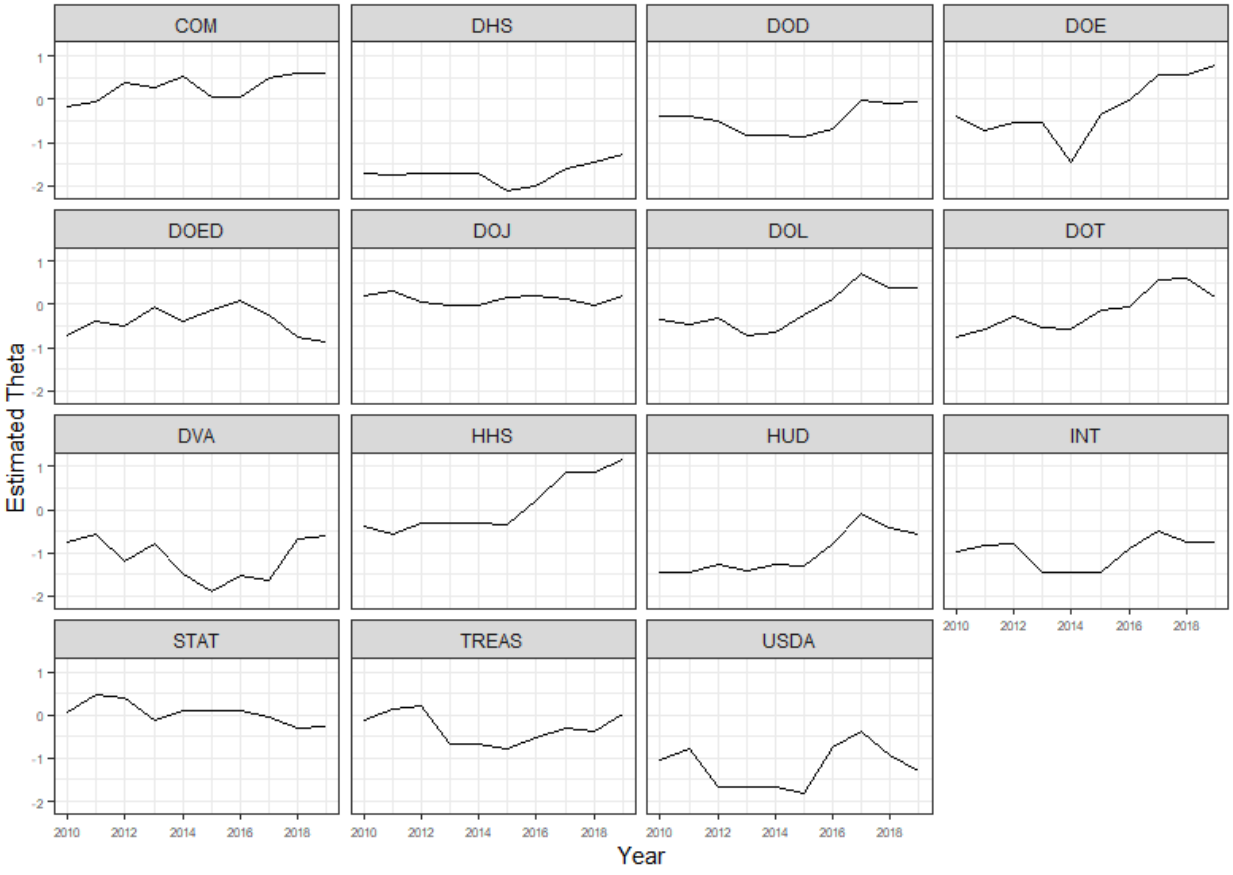
Note: Table lists the top-5 and bottom-5 posterior mean estimates $\widehat{\theta}$, with the posterior estimates of the standard error and 95% confidence bounds. We use these estimates to group agencies into one of 5 categories—low, low/moderate, moderate, high/moderate, and high. The latter columns in the table include counts of the number of years between 2010 - 2019 an agency fell into one of the groupings. A full list of agencies is included in Appendix C.

The overall agency averages obscure important changes within agencies over the time period. In Figure 3 we graph the performance estimates for the 15 executive departments. A few things stand out. First, some departments consistently perform well (e.g., Department of Commerce (COM)) and others poorly (Department of Homeland Security (DHS)) across the entire period. Second, estimates increased steadily in some agencies such as the Department of Health and Human Services (HHS). Others, such as the Department of State trended in the opposite direction. Others such as the Department of Energy (DOE) and the Department of Interior (INT) exhibited more variability between years.

An effective measurement approach should be flexible enough to capture *real* variation among agencies and within agencies over time. Figure 4 reveals such variation but additional work is required to determine whether this variation derives from real changes in performance or idiosyncratic factors related to data availability or quality. We are encouraged, however, by variation such as that exhibited by HHS since it had 5 steady years of improvement on FEVS outcomes starting in 2014. In addition,

in 2019, the Department of Agriculture experienced a marked decline in FEVS responses, while COM and Veterans Affairs (DVA) saw employees win major government awards.

Figure 3. Performance Estimates ($\widehat{\theta}$) for Executive Departments by Year, 2010-2019



Convergent Validation

One risk of such an approach is that the dimension θ for which we get numerical estimates is not actually performance at all. This makes validation extremely important. To validate the measure, we use data on performance excluded from our measurement efforts. Specifically, 2020 provided a significant amount of new and unique data on performance that does not exist in other years. We use data from two unique data sources. First, we use data from the *Survey on the Future of Government Service*, a non-partisan and non-governmental survey of thousands of federal executives (Piper and Lewis 2023; Richardson et al. 2023). The survey asked a series of questions intended to provide different

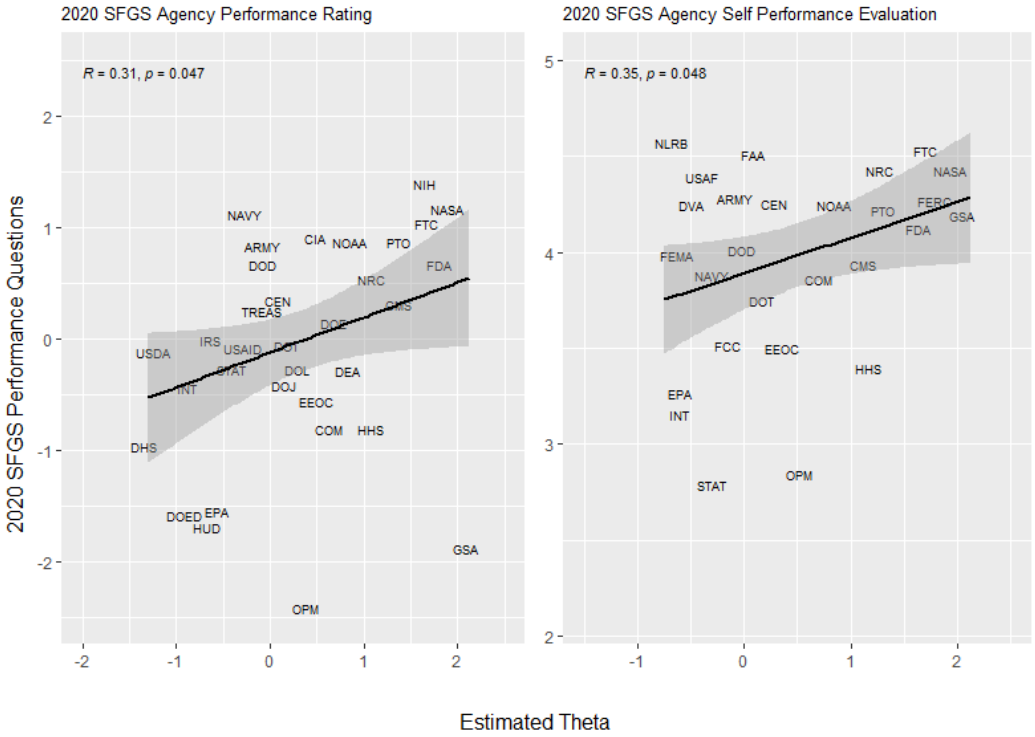
perspectives on performance. Importantly, the survey asked, "How would you rate the overall performance of [your agency] in carrying out its mission?" Respondents were given a sliding scale from 1-Not at all effective to 5-Very effective. They could also indicate a "Don't know" response. Weighted agency average responses to this self-assessment can be compared to our estimates of θ from 2019.

In addition, the survey asked respondents to rate the performance of agencies of other agencies. Specifically, the survey began by asking respondents: "Please select the three agencies you have worked with the most in order of how often you work with them." Each respondent was given a drop-down menu. Later in the survey, respondents were asked "How would you rate the overall performance of the following agencies in carrying out their missions?" and given the list of agencies they provided plus two others. Richardson et al. (2023) generated performance estimates based upon the thousands of ratings federal executives. These scores can be compared to our 2019 estimates.

In Figure 4 we graph the correlations between the 2019 estimated performance, ($\hat{\theta}$), and the responses of federal executives to questions about the performance of their agency and other agencies. The figures reveal a moderate correlation between the evaluations of federal executives about performance and our estimates, 0.31 and 0.35, respectively. As our performance estimates increase, so does the SFGS performance score of the agency, both its reputational score and the average self-reported performance. Some of the gap between the two measures is to be expected since the SFGS targets the subjective assessments of a subpopulation of federal executives and there is a lag between 2019 and 2020. There are some notable outliers. For example, defense and intelligence agencies tend to do better in the SFGS measures than our measure. Interestingly, the Office of Personnel Management (OPM) and the General Services Administration (GSA) do better on our performance

measures than the SFGS measures. This may be due to the emphasis the OPM and the GSA place on surveys they administer (i.e., FEVS, CSS) that play a key role in our estimates.

Figure 4. Correlation Between 2019 Performance Estimates ($\widehat{\theta}$) and 2020 SFGS Performance Questions



Our second unique new source of data comes from a special battery of questions on the 2020 FEVS survey. During the COVID-19 pandemic, the Office of Personnel Management included a series of questions about agency performance that were unique to that year's survey. These questions tap into agency performance directly and are included in Table 4. We can use agency average responses to these questions and compare them to our estimates of θ from 2019.

Table 4. 2020 FEVS Agency Performance Questions

Question 1: Prior to the COVID-19 pandemic, my work unit... [2020 only]
...produced high-quality work.
5 "Always"
4 "Most of the time"
3 "Sometime"
2 "Rarely"
1 "Never"
X "No basis to judge"

Question 2: Prior to the COVID-19 pandemic, my work unit... [2020 only]

...achieved our goals.

5 "Always"
4 "Most of the time"
3 "Sometime"
2 "Rarely"
1 "Never"
X "No basis to judge"

Question 3: During the COVID-19 pandemic, my work unit... [2020 only]

...has produced high quality work.

5 "Always"
4 "Most of the time"
3 "Sometime"
2 "Rarely"
1 "Never"
X "No basis to judge"

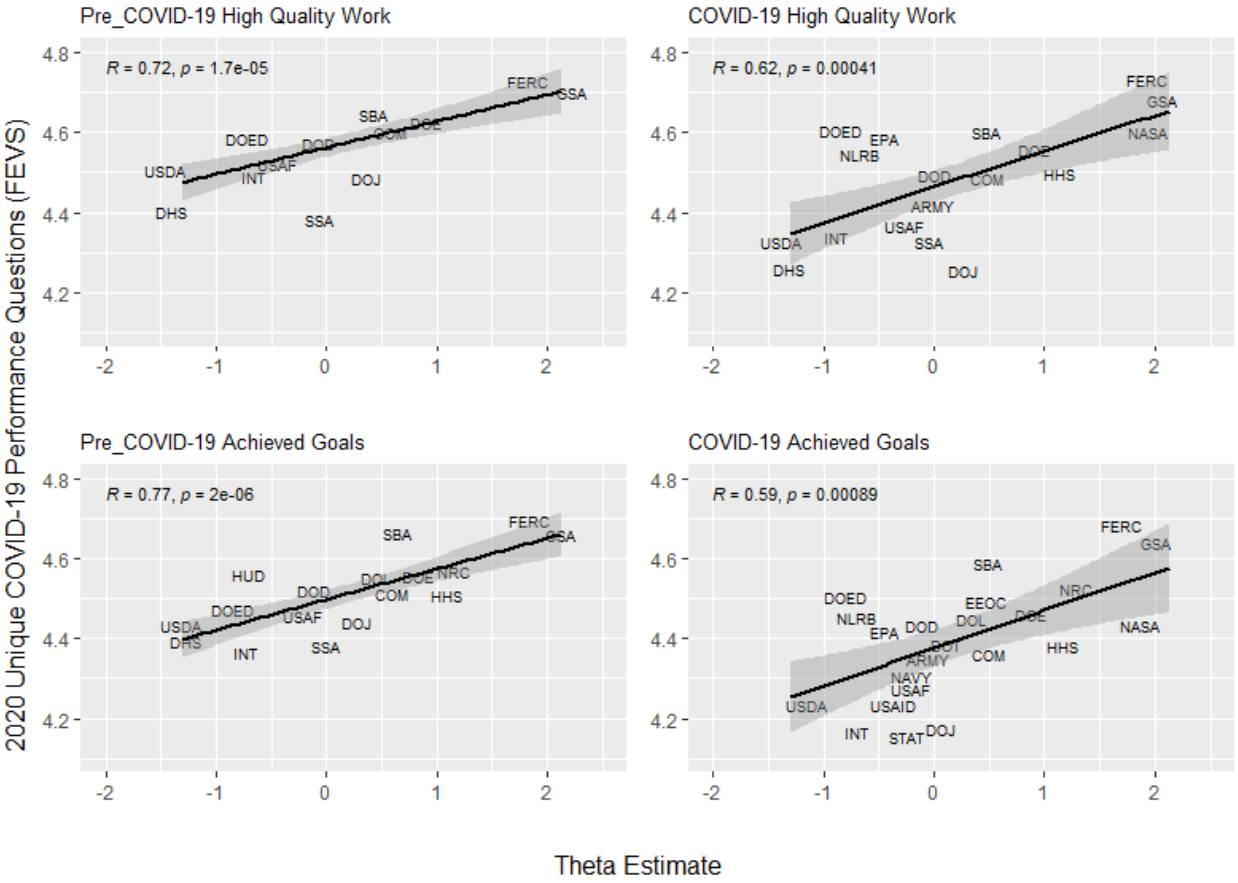
Question 4: During the COVID-19 pandemic, my work unit... [2020 only]

...has achieved our goals.

5 "Always"
4 "Most of the time"
3 "Sometime"
2 "Rarely"
1 "Never"
X "No basis to judge"

When we compare the 2019 performance estimates to the newly added 2020 FEVS questions, the correlations are strong, ranging from 0.59 to 0.77 (Figure 5). The 2019 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 there is, correlates with $\hat{\theta}$. There are fewer consistent outliers and the estimates are tightly organized around a regression line fitted to the data. The Department of Education and the EPA responded more positively to the COVID-19 questions than their performance estimates would indicate while the Department of Justice (DOJ) and the Social Security Administration (SSA) responded more negatively.

Figure 5. Correlation Between 2019 Performance Estimates ($\widehat{\theta}$) and 2020 FEVS COVID-19 Performance Questions



In total, despite the variation, the validation results are encouraging. We would not expect a perfect correlation because of the gap between 2019 and 2020 and because both the SFGS data and FEVS provide one way of revealing performance but not the only one. Indeed, the goal of this paper is to propose a method for aggregating all of the data like the SFGS and FEVS data with other objective and subjective data to produce a better measure. The early internal and external validity of the estimates provides confidence that the approach has promise.

DISCUSSION

President Biden's management agenda, similar to efforts in many countries, places an important emphasis on performance measurement. It encourages agencies to distill key goals from

their missions and measure and report on performance toward those goals. The goals differ by agency and are reported as part of the budget process. While agencies use internal goal setting and performance measurement to compare performance against a historical baseline, agency-specific goals make comparing performance across agencies difficult. Indeed, it is difficult to determine systematically which U.S. federal agencies are performing well and poorly.

As Behn (2003) suggests, decisions about appropriate performance measures should be made with particular purposes in mind—to control, promote, celebrate, etc. The collection of performance information cannot be an end in itself. Rather, it should fulfill the promise of what Moynihan calls "the era of performance management" (Moynihan 2008: 4). Arguably, we need measures that tap the efficacy of specific programs and the meeting of specific agency goals and we need a principled way to tell decision makers where they need to focus their attention across the vast executive establishment in the United States. This paper has attempted to provide a way of aggregating all of this performance information, the specific and the general, the objective and the subjective, to help fulfill the latter goal of providing a roadmap for those managers in the executive and legislative branches seeking to improve performance.

Perhaps the key difficulty with measuring comparative agency performance is the complexity of the enterprise. Scholars have identified dozens of processes, unclear goals, and at least 16 different dimensions of performance. No one measure is likely to satisfy all of the requirements of an effective performance measurement regime. The method and measure we propose and evaluate here, however, is an important step forward in thinking about how to aggregate different performance information without doing too much damage to the task and dimension complexity of such measures. We have assumed throughout that there is such a concept as true latent organizational performance, even while acknowledging that there is high and low performance on different tasks and in different parts of the organization. Agencies can also be good on some dimensions and poor on others. That said, while

noisy, our method and resulting measures hold out hope for a more robust discussion of ways to aggregate different kinds of performance information—both subjective and objective—and let the data help us arbitrate what is useful and what is not.

These latent administrative performance estimates generated from the preliminary analyses conducted in this study are highly promising on three levels. First, these estimates exhibit face validity when comparing these empirical Bayes posterior mean estimate values in relation to the performance item measures used to construct this weighted index model (see Appendix B). Second, these latent administrative performance estimates are robust to alternative model specifications that both omit and include overlapping elements of workplace and organizational satisfaction (see Table 3), poor item predictors (e.g., SAMMIMES and GAO-High Risk List Programs), and Defense and military agencies. Finally, the generated latent administrative performance estimates exhibit convergent validity with multiple out-of-sample survey-based measures (see Figures 4 & 5).

Although these comparable estimates of administrative performance are highly promising, additional work is needed to improve upon the estimation strategy. First, expanding our coverage of agencies and years would improve the power of the statistical estimates, thus reducing the imprecision of the estimates. This is a nontrivial issue in the current design as many of the difficulty parameters are substantively meaningful, albeit some are estimated rather imprecisely since they standard errors are clustered on agency units which only contain ten observations per panel. A related challenge is how to balance the need to limit data sparseness on performance item measures for smaller agencies (i.e., those units housed within executive departments and independent agencies) while expanding the sample of agencies. Although IRT models are ideally suited to handle missingness of data, nonetheless, more data is better. Further, increasing the number of performance item measures increases the chances that the data provide information on more than one dimension. This is a challenge using a

variety of performance data that will increase the prospects for tapping into multiple dimensions of a latent concept (Raykov and Marcoulides 2018).

Finally, differential levels of aggregation among public agencies presents an additional challenge for accurate measurement of administrative performance. This is because a performance estimate for an agency as large as the Department of Defense or the Department of Homeland Security might mean something different than for the Federal Deposit Insurance Corporation with a narrow mission and smaller organizational apparatus. We maintain that meaningful comparisons can be made for two reasons. First, estimates make more sense in the context of comparison. For example, when we compare the estimates of the 15 executive departments, it is informative to see Homeland Security and Agriculture on one end of the scale and Commerce and Justice on the other end. Yet, confidence in such estimates will be further augmented as the bases for comparison expands ─ for example, to include estimates of the U.S. Coast Guard, Transportation Security Administration, U.S. Customs and Border Protection, and the Federal Emergency Management Agency along with our estimate for the Department of Homeland Security. After all, a departmental estimate is derived from an averaging across component performance. The more information we have the better estimates we can produce and the more meaningful become these highly aggregated measures.

These conclusions set an agenda for these efforts moving forward. Specifically, they suggest that future efforts should be directed toward identifying more performance information, ideally in a way that allows us to expand the list of agencies included in model estimation. More and better data would ideally allow us to extend the data forward and backward in time both to improve model estimation and conduct additional out of sample validity tests.

References

Andersen, Lotte Bøgh Andersen, Andreas Boesen, and Lene Holm Pedersen. 2016. "Performance in Public Organizations: Clarifying the Conceptual Space." *Public Administration Review* 76(6):852-62.

Bednar, Nick, and David E. Lewis. 2023. "Presidential Investment in the Administrative State." *American Political Science Review*, forthcoming.

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

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):513-44.

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

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

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

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

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

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

Chun, Young Han, and Hal G. Rainey. 2005. "Goal Ambiguity and Organizational Performance in US Federal Agencies." *Journal of Public Administration Research and Theory* 15(4): 529–57.

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. 203-29.

Dahlström, Carl, and Victor Lapuente. 2017. *Organizing Leviathan: Politicians, Bureaucrats, and the Making of Good Government* (Cambridge: Cambridge University Press).

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

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

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

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

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

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

Hubbard, Graham. 2009. "Measuring Organizational Performance: Beyond the Triple Bottom Line." *Business Strategy and the Environment* 18:177-91.

Jilke, Sebastian, Bart Meuleman, and Steven Van de Walle. 2015. "We Need to Compare, but How? Measurement Equivalence in Public Administration." *Public Administration Review* 75(1):36-48.

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

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

Krause, George A., David E. Lewis, and James W. Douglas. 2006. "Political Appointments, Civil Service Systems, and Bureaucratic Competence: Organizational Balancing and Executive Branch Revenue Forecasts in the American States." *American Journal of Political Science* 50(3):770-87.

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):914-31.

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): 521–49.

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.

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):180–90.

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

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):1004–23.

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

Nistotskaya, Marina, Stefan Dahlberg, Carl Dahlström, Aksel Sundström, Sofia Axelsson, Cem Mert Dalli & Natalia Alvarado Pachon. 2021. The Quality of Government Expert Survey 2020 Dataset: Wave III. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se DOI: 10.18157/qoges2020.

Park, Jungyeon. **n.d**. "How Individual and Organizational Sources of Managerial Capacity Shape Agency Performance: Evidence from the Size of Improper Payment in U.S. Federal Programs." *Working Paper*.

Piper, Christopher, and David E. Lewis. 2023. "Do Vacancies Hurt Federal Agency Performance?" *Journal of Public Administration Research and Theory* 33(2):313-28.

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):625–36.

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

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

Richardson, Mark D. 2023. "The Apolitical Executive Branch." Manuscript, Georgetown University.

Richardson, Mark D., Joshua D. Clinton, and David E. Lewis. 2018. "Elite Perceptions of Agency Ideology and Workforce Skill." *Journal of Politics* 80(1):303-7.

Richardson, Mark D., Christopher Piper, and David E. Lewis 2023. "Measuring the Impact of Appointee Vacancies on U.S. Federal Agency Performance." Paper presented at the 2023 Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 13-16.

Rutherford, Amanda. 2016. "The Effect of Top-Management Team Heterogeneity on Performance in Institutions of Higher Education." *Public Performance & Management Review* 40(1): 119–44.

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

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

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

Teorell, Jan, Carl Dahlström, and Stefan Dahlberg. 2011. *The Quality of Government Expert Survey Dataset*. University of Gothenburg: The Quality of Government Institute (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3569575).

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): 133–56.

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

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

Wood, Abby K., and David E. Lewis. 2017. "Agency Performance Challenges and Agency Politicization." *Journal of Public Administration Research and Theory* 27(4): 581–95.

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

Appendix A. List of Agencies

OKCODE	Acronym	Name
1	USDA	Department of Agriculture
2	COM	Department of Commerce
3	DOD	Department of Defense
4	ARMY	Department of the Army
5	USAF	Department of the Air Force
6	NAVY	Department of the Navy
7	DOED	Department of Education
8	DOE	Department of Energy
9	HHS	Department of Health and Human Services
11	DHS	Department of Homeland Security
12	HUD	Department of Housing and Urban Development
13	INT	Department of the Interior
14	DOJ	Department of Justice
15	DOL	Department of Labor
16	STAT	Department of State
17	DOT	Department of Transportation
18	TREAS	Department of Treasury
19	DVA	Department of Veterans Affairs
20	CIA	Central Intelligence Agency
21	EPA	Environmental Protection Agency
22	FEMA	Federal Emergency Management Agency (Pre-2003)
23	GSA	General Services Administration
24	NASA	National Aeronautics and Space Administration
25	SBA	Small Business Administration
26	SSA	Social Security Administration
27	USAID	U.S. Agency for International Development
28	USIA/BBG/USAGM	U.S. Agency for Global Media
29	OMB	Office of Management and Budget
30	USTR	Office of the U.S. Trade Representative
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
51	CMS	Centers for Medicare and Medicaid Services
52	DEA	Drug Enforcement Administration
53	FAA	Federal Aviation Administration
54	FDA	Food and Drug Administration
55	FEMA	Federal Emergency Management Agency (in DHS)
56	IRS	Internal Revenue Service
57	NHTSA	National Highway Traffic Safety Administration
58	NIH	National Institutes of Health
59	NIST	National Institute of Standards and Technology
60	NOAA	National Oceanic and Atmospheric Administration
61	PTO	Patent and Trademark Office
70	PBGC	Pension Benefit Guarantee Corporation
71	USPS	U.S. Postal Service
72	OPM	Office of Personnel Management
73	OSTP	Office of Science and Technology Policy
74	ODNI	Office of the Director of National Intelligence
75	NSC	National Security Council
76	NEC	National Economic Council
77	HSC	Homeland Security Council
78	FDIC	Federal Deposit Insurance Corporation
79	CBP	Customs and Border Protection (in DHS)
80	USCS	Customs Service (Pre-2003)

Appendix B: Raw Data and Estimates, with Missing Data (2019)

		Empirical Bayes		GSA				FEVS Work Unit		FEVS Performance		FEVS	SAMMIES	GAO
Name	Year	Posterior Mean	Posterior SE	Proc.	Fin. Mgt	HR	IT	Upper-Level	Lower-Level	Upper-Level	Lower-Level	BPTW	Noms.	High Risk Prog
USDA	2019	-1.302	0.387	1	0	0	2	0	1	0	0	0	1	0
COM	2019	0.605	0.268	1	0	1	3	2	3	2	3	2	2	1
DOD	2019	-0.059	0.241	1	2	1	0	1	1	2	2	1	1	1
ARMY	2019	0.001	0.252	.	.	.	.	1	1	2	2	1	0	0
USAF	2019	-0.319	0.255	.	.	.	.	1	0	2	2	0	0	0
NAVY	2019	-0.256	0.253	.	.	.	.	0	1	1	2	1	1	0
DOED	2019	-0.861	0.294	2	2	0	0	3	1	0	0	0	0	0
DOE	2019	0.804	0.285	2	2	1	2	3	3	3	2	2	1	1
HHS	2019	1.155	0.335	1	1	1	2	2	3	3	3	3	2	1
DHS	2019	-1.283	0.383	2	1	2	1	0	0	0	0	0	2	1
HUD	2019	-0.573	0.259	0	1	1	3	2	2	0	0	1	0	1
INT	2019	-0.753	0.278	1	1	1	2	1	1	0	0	1	1	1
DOJ	2019	0.196	0.246	3	3	3	3	1	1	2	2	1	1	1
DOL	2019	0.378	0.254	2	1	2	1	2	2	3	2	1	0	0
STAT	2019	-0.273	0.243	2	3	2	1	2	2	0	1	1	1	1
DOT	2019	0.164	0.245	2	1	2	3	1	2	2	2	1	1	1
TREAS	2019	0.032	0.242	2	2	3	1	1	2	2	1	1	0	1
DVA	2019	-0.596	0.437	0	1	0	1	.	.	.	.	1	3	1
CIA	2019	0.487	0.468	.	.	.	.	.	.	.	.	2	0	1
EPA	2019	-0.455	0.250	0	0	1	2	3	3	0	0	1	1	1
GSA	2019	2.125	0.571	3	3	3	3	3	3	3	3	3	0	1
NASA	2019	1.953	0.528	3	3	3	2	3	3	3	3	3	2	1
SBA	2019	0.533	0.263	2	3	3	3	2	2	3	2	1	0	0
SSA	2019	-0.140	0.241	3	3	3	2	0	0	2	1	1	0	1

USAID	2019	-0.300	0.244	3	3	1	2	1	2	0	1	1	1	0
USAGM	2019	-0.875	0.313	.	.	.	.	2	0	1	0	0	0	0
EEOC	2019	0.408	0.268	.	.	.	.	3	3	2	1	2	0	0
FCC	2019	-0.197	0.252	.	.	.	.	3	3	0	1	1	0	0
FERC	2019	1.785	0.538	.	.	.	.	3	3	3	3	3	0	0
FTC	2019	1.785	0.538	.	.	.	.	3	3	3	3	3	0	0
NLRB	2019	-0.598	0.273	.	.	.	.	3	3	0	0	0	0	0
NRC	2019	1.200	0.343	3	1	3	1	3	3	3	3	2	0	0
SEC	2019	1.747	0.526	.	.	.	.	3	3	3	3	3	0	1
CEN	2019	0.191	0.257	.	.	.	.	2	2	1	2	2	0	1
CMS	2019	1.228	0.376	.	.	.	.	3	3	3	2	3	0	1
DEA	2019	0.710	0.291	.	.	.	.	3	2	3	2	2	0	0
FAA	2019	0.107	0.254	.	.	.	.	1	2	2	2	1	1	0
FDA	2019	1.747	0.526	.	.	.	.	3	3	3	3	3	0	1
FEMA	2019	-0.596	0.273	.	.	.	.	1	1	1	1	0	0	1
IRS	2019	-0.552	0.269	.	.	.	.	1	1	1	0	1	0	1
NIH	2019	1.757	0.529	.	.	.	.	3	3	3	3	3	1	0
NIST	2019	1.757	0.529	.	.	.	.	3	3	3	3	3	1	0
NOAA	2019	0.941	0.320	.	.	.	.	3	2	3	3	2	2	0
PTO	2019	1.356	0.408	.	.	.	.	1	3	3	3	3	0	0
OPM	2019	0.510	0.261	1	1	2	0	3	3	2	2	2	0	0
ODNI	2019	0.517	0.470	.	.	.	.	.	.	.	.	2	0	0

Appendix C. Summary Performance by Agency, 2010-2019

	Empirical Bayes				Performance Class	Performance Years				
Agency	Posterior Mean	Posterior SE	95% LCL	95% UCL		Low Count	Low−Moderate Count	Moderate Count	High−Moderate Count	High Count
Agriculture (USDA)	-1.197	0.397	-1.976	-0.419	Low	9	1	0	0	0
Commerce (COM)	0.277	0.259	-0.230	0.784	Moderate	0	0	6	2	2
Defense (DOD)	-0.467	0.268	-0.992	0.059	Low−Moderate	4	3	3	0	0
Army (ARMY)	-0.310	0.271	-0.841	0.220	Moderate	4	0	6	0	0
Air Force (USAF)	-0.116	0.255	-0.616	0.384	Moderate	0	0	10	0	0
Navy (NAVY)	-0.182	0.256	-0.684	0.321	Moderate	1	2	7	0	0
Education (DOED)	-0.394	0.262	-0.906	0.119	Low−Moderate	1	1	8	0	0
Energy (DOE	-0.208	0.285	-0.767	0.351	Moderate	4	0	3	0	3
HHS (HHS)	0.083	0.270	-0.447	0.612	Moderate	1	0	6	0	3
Home. Sec. (DHS)	-1.707	0.518	-2.722	-0.693	Low	10	0	0	0	0
Hous. & Urban (HUD)	-0.998	0.360	-1.704	-0.292	Low	8	1	1	0	0
Interior (INT)	-0.983	0.339	-1.648	-0.318	Low	10	0	0	0	0
Justice (DOJ)	0.123	0.250	-0.367	0.613	Moderate	0	0	10	0	0
Labor (DOL)	-0.114	0.261	-0.626	0.398	Moderate	2	1	5	1	1
State (STAT)	0.051	0.251	-0.441	0.543	Moderate	0	0	9	1	0
Transportation (DOT)	-0.158	0.262	-0.671	0.354	Moderate	4	0	4	0	2

Treasury (TREAS)	-0.311	0.260	-0.820	0.198	Moderate	4	0	6	0	0
Veterans Affairs (DVA)	-1.110	0.409	-1.911	-0.309	Low	8	0	2	0	0
CIA	0.485	0.468	-0.433	1.403	Moderate	0	0	10	0	0
EPA	-0.439	0.266	-0.959	0.081	Low−Moderate	4	2	4	0	0
GSA	0.616	0.331	-0.031	1.264	High−Moderate	0	0	6	0	4
NASA	1.602	0.464	0.693	2.511	High	0	0	0	0	10
SBA	-0.537	0.341	-1.206	0.133	Low Moderate	4	3	2	0	1
SSA	-0.091	0.252	-0.585	0.403	Moderate	0	0	9	1	0
USAID	-0.870	0.370	-1.596	-0.144	Low	7	0	2	0	1
USAGM	-1.403	0.505	-2.392	-0.413	Low	8	2	0	0	0
EEOC	-0.271	0.320	-0.898	0.356	Moderate	2	1	6	1	0
FCC	-0.224	0.257	-0.727	0.279	Moderate	1	1	8	0	0
FERC	1.280	0.458	0.383	2.177	High	0	0	1	1	8
FTC	1.702	0.548	0.628	2.775	High	0	0	0	0	10
NLRB	-0.269	0.321	-0.898	0.361	Moderate	2	2	5	0	1
NRC	1.656	0.479	0.718	2.594	High	0	0	0	0	10
SEC	0.264	0.395	-0.509	1.038	Moderate	4	0	2	0	4
CEN (TREAS)	-0.256	0.275	-0.795	0.282	Moderate	3	1	6	0	0
CMS (HHS)	0.586	0.322	-0.045	1.216	High−Moderate	1	1	1	2	5
DEA (DOJ)	0.999	0.377	0.259	1.738	High	0	0	0	0	10

FAA (DOT)	-0.060	0.265	-0.579	0.459	Moderate	2	1	6	0	1
FDA (HHS)	0.582	0.324	-0.053	1.216	High−Moderate	0	0	7	0	3
FEMA (DHS)	-1.101	0.388	-1.862	-0.341	Low	9	0	1	0	0
IRS (TREAS)	-0.612	0.297	-1.194	-0.030	Low	7	0	3	0	0
NIH (HHS)	1.061	0.370	0.336	1.786	High	0	0	0	1	9
NIST (COM)	1.122	0.382	0.373	1.872	High	0	0	0	1	9
NOAA (COM)	0.114	0.273	-0.420	0.649	Moderate	1	0	6	0	3
PTO (COM)	1.085	0.387	0.326	1.843	High	0	1	1	0	8
OPM	0.177	0.257	-0.328	0.681	Moderate	0	0	7	1	2
ODNI	0.515	0.470	-0.406	1.437	Moderate	0	0	10	0	0
Total Average	**-0.000204**	**0.339**	**-0.664**	**0.664**						