

Measuring the Impact of Appointee Vacancies on U.S. Federal Agency Performance¹

In this paper we evaluate the relationship between vacancies in Senate confirmed positions (PAS) and U.S. federal agency performance. We explore this relationship using an innovative new measure of agency performance. This measure is derived from the evaluation of federal executives who work closely with the agencies they assess. The measure is comparable across agencies and avoids many of the limitations of existing measures. We find a robust correlation between vacancies in Senate-confirmed positions and lower evaluations of agency performance, even when accounting for differences in the way Republican and Democratic federal executives perceive performance. We conclude with a discussion of how dysfunction in the U.S. appointment process is influencing federal government performance on key tasks and the implications of our findings for the creation and use of performance measures in a world of partisan differences.

December 22, 2022

Words: 9,837

Mark D. Richardson, Georgetown University
mark.richardson@georgetown.edu

Christopher Piper, Vanderbilt University
christopher.m.piper@vanderbilt.edu

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

¹ We presented an earlier version of this paper at the 2022 Public Management Research Conference, Phoenix, AZ, May 25-28, 2022. We thank the Princeton Survey Research Center for their work fielding this survey. We are grateful to the Partnership for Public Service for their advice and help in designing, supporting, and executing this survey. George Krause, Jennifer Selin, and participants in the PMRC conference provided helpful comments. A seminar audience at the University of Indiana provided many helpful suggestions that helped shape many of the ideas for this paper. The errors that remain are our own.

After sixteen months in office, President Biden had successfully filled 350 out of 799 key executive branch policymaking positions.¹ Prominent media outlets decried the slow pace of presidential nominations and Senate confirmations, with the *New York Times* asking, “Why are so many government positions still vacant?” and the *Washington Post* noting “A Senseless Logjam is Holding Up Nominations” (Bernstein 2022; Risen 2021). Outside observers raised concerns about how vacancies would create leadership gaps, implementation problems and poor agency performance (see, e.g., O’Connell 2021; Stier 2021). Others worried that the slow pace of appointments had adversely affected U.S. policy in areas as diverse as the conflict in Ukraine, the rollout of the infrastructure agenda, and economic policy (see, e.g., Gramer 2022; Hussein et al. 2022).

President Biden’s pace of nominations was by no means an outlier. He was on par with his predecessors (Pager et al. 2021; Stier 2021; Tenpas 2021). During the Trump Administration, for example, critics lambasted the president for the slow pace of executive appointments, but the president was unmoved (Eilperin et al. 2019; Lu 2019). Where Democrats saw vacancy-related performance problems, the Republican president saw a high performing administration even with vacancies. These struggles and concern for their consequences raises the more general question of how vacancies influence federal agency performance. This is an important question since vacancies are a regular feature of modern administration and the stakes for performance can be quite high if they systematically influence performance and policy (Dull and Roberts 2009; O’Connell 2009; Pfiffner 1996; Resh et al. 2021; Rutherford et al. 2018). Further, like the noted differences between Democrats and President Trump, the answer to whether and to what extent vacancies impact agency

¹ Partnership for Public Service, Political Appointee Tracker, May 6, 2022

(<https://ourpublicservice.org/performance-measures/political-appointee-tracker/>).

performance is unresolved in the literature, with some scholars arguing that under certain conditions vacancies may improve performance (e.g., Mendelson 2014) and others arguing vacancies will generally diminish agency performance (e.g., Piper and Lewis 2022).

Evaluating the impact of vacancies on performance has been difficult because appropriate measures of federal agency performance are elusive. Unlike private sector organizations, there is no profit equivalent that provides a shorthand way of comparing performance. This makes assessing the performance of government agencies with and without confirmed appointees challenging. The outputs of federal agencies are difficult to observe and measure across contexts (Wilson 1989). Scholars have made important progress measuring comparative performance through creative means but critics charge that such measures depend upon questionable self-reports, are limited to specific tasks or contexts that limit generalizability, or are hamstrung by partisan disagreements in defining good agency performance (Meier et al. 2015; Nyhan and Marlowe 1995; Boyne and Dahya 2002). It is no surprise that work systematically connecting vacancies to performance is rare (Rutherford et al. 2018; see, however, Piper and Lewis 2022, Wrightman et al. 2022).

In this paper we evaluate the relationship between vacancies in Senate confirmed positions (PAS) and federal agency performance using new measures of performance. These new measures overcome many of the limitations in existing approaches because they are not self-reports and are comparable across agencies, while still measuring performance on core agency missions. We find a robust correlation between vacancies in Senate-confirmed positions and lower evaluations of agency performance by federal executives familiar with agency activities. While there are some agencies where Republicans and Democrats disagree over performance, the effect of vacancies on performance remains substantial after accounting for these differences. We conclude by discussing how appointment process dysfunction is influencing federal government performance and how we

should think about the creation and use of performance information in a world of partisan differences.

The paper makes three key contributions. First, it demonstrates a robust relationship between vacancies in appointed leadership positions and federal agency performance. This provides the strongest evidence to date in the existing scholarly debate on whether vacancies are detrimental to agency performance. Second, it introduces a new way of measuring federal agency performance that can be applied in other contexts. This approach accounts for many of the limitations of existing measures of agency performance and provides a useful method for future scholars. Finally, the paper systematically evaluates the influence of partisanship on perceptions of federal agency performance. We provide some of the first evidence that, despite a few areas of disagreement, Republicans and Democrats generally agree on the performance of federal agencies.

How do Vacancies Influence Performance?

Scholars have long been interested in the relationship between appointed leadership and federal agency performance (see, e.g., MacMahon and Millett 1939, Stanley, Mann, and Doig 1967). While some scholars have focused on the differences in backgrounds between appointed leaders and career professionals (Aberbach and Rockman 2000; Cohen 1998; Krause and O’Connell 2016), others have focused on the systematic effects of filling leadership positions with appointees rather than permanent career professionals (see, e.g., Heclo 1977; Light 1995; Richardson 2019). They have argued that positions filled by appointees experience systematically higher turnover rates and longer vacancies than positions designated to be filled by career professionals (Heclo 1977; Suleiman 2003). These vacancies make it more difficult to engage in long-term planning, generate lower morale, and reduce the incentives of outside stakeholders to invest time and resources in the agency (Piper and Lewis 2022; O’Connell 2009, 2020).

Some scholars, in contrast, have pointed out that vacancies, or at least turnover, can lead to improved agency performance in some contexts (Boyne and Dahya 2002; Boyne et al. 2011). If the career professionals that lead agencies during vacancies are more qualified than the appointees they replace, for example, this could help the agency (Mendelson 2014). Indeed, the career professionals that serve as acting leaders while the president and the Senate negotiate over appointed leaders generally have excellent preparation for their roles. They have long agency experience and significant policy and public management expertise (Lewis 2007). If appointed leaders unnecessarily meddle in agency processes and disrupt efficient routines, long vacant periods of depoliticization may improve the ability of the agency to tackle hard problems.

Efforts to evaluate these claims have been hindered by the difficulty of measuring performance across contexts. The outputs and outcomes of federal agencies are difficult to observe and compare (Wilson 1989). Some scholars have made progress by examining self-reported performance (see, e.g., Moynihan and Pandey 2005; Piper and Lewis 2022), but such efforts raise concerns about the connection between self-reports and actual performance (Meier et al. 2015). The proliferation of government-developed performance measures has provided another means of measuring performance (see, e.g., Boyne et al. 2011; Lewis 2007), but such measures can be subject to manipulation or politicization (see, e.g., Lavertu and Moynihan 2013). Others have focused on proxies for good performance such as concepts measured in surveys of federal employees like the Federal Employee Viewpoint Survey (for a review see Fernandez et al. 2015). These surveys ask about practices we associate with good management (e.g., communication, resources, accountability). The resultant measures have the virtue of being comparable across agencies but it is unclear how well such measures tap underlying concepts and how closely such concepts relate to performance. It is also not clear what organization respondents are evaluating when asked about their agency or organization (Thompson and Siciliano 2021). For example, some agency respondents

may think their organization is the office or subunit in which they work, while others may view their organization as a cabinet department as a whole.

Scholars have employed other creative measures of performance but these are often limited to one type of organization such as city services (Wrightman et al. 2022), law enforcement (Boylan 2004; Hur 2013) or schools (see, e.g., Hill 2005; Meier and Hicklin 2007; Meier and O'Toole 2002; Rutherford 2016) or one type of task that may not be part of an agency's core mission such as budget forecasting (Krause and Douglas 2006; Krause et al. 2006), limiting payment errors (Park n.d.), or responding to FOIA requests (Wood and Lewis 2017). It is not clear whether results in such studies are applicable across agencies to the performance of core agency missions. While the vast majority of government work is ministerial and apolitical and has bipartisan support (e.g., securing loose nuclear weapons, securing ports, rural development), Republicans and Democrats can also disagree about the definition of good performance on some kinds of government activity such as environmental protection or social welfare provision (Boyne and Dahya 2002). This further complicates the development of comparable measures of performance.

Data, Variables, and Methods

To overcome existing measurement problems and effectively evaluate the relationship between vacancies in PAS positions and performance, we introduce a new measure of federal agency performance. We begin by describing how we generated this measure. We regress this new measure of agency performance on data on vacancy lengths. We then evaluate whether this measure is influenced by the partisanship of those evaluating federal agency performance.

Measuring Federal Agency Performance

In 2020, in collaboration with the Partnership for Public Service and academic colleagues, we fielded the *Survey on the Future of Government Service*. This survey targeted all political appointees and senior career managers running agencies, offices, and programs in the executive branch. The

response rate for the survey was 9.1% (1,485 completed surveys out of 16,232) and the participation rate (i.e., the percentage that completed at least part of the survey) was 11.5% (1,861 complete or partial surveys out of 16,232), comparable to most public opinion telephone surveys (AAPOR 2017).

Federal executives working at the highest level of the executive establishment are uniquely situated to observe agency performance, perhaps better than any other population. We leverage this expertise to develop new expert-based measures of agency performance. To narrowly target the expertise of different federal executives, we asked respondents to identify the agencies they worked with the most. Specifically, the survey asked: “Please select the three agencies you have worked with the most in order of how often you work with them.” Each respondent was provided three dropdown menus from which to select agencies, excluding their own.² The menus included more than 200 agencies organized into 1) the Executive Office of the President, 2) the executive departments and their large sub-components, and 3) independent agencies. So, for example, a

² To make sure respondents only evaluate other agencies we need reliable information on their own workplace. We pursued two tacks to identify workplaces. First, we asked respondents to identify their workplace from a dropdown menu that included more than 200 agencies organized by large categories, key sub-components, and catch-all categories. So, for example, a respondent would see in the dropdown menu the Executive Office of the President followed by units like the Council of Economic Advisers, the Office of Management and Budget, but also a catch-all “Other (EOP)”. If the respondent went further, they might see the Department of Agriculture with 16 options, including the Office of the Secretary, key bureaus, and “Other (USDA)” as options. Second, we have information on where people work from the *Federal Yellow Book* which is the source we used to identify and contact our sample.

respondent might scroll past agencies in the Executive Office of the President and see the Department of Agriculture with 13 options, including “Department of Agriculture (All)” and key bureaus inside the USDA such as the Forest Service or Natural Resources Conservation Service.³ They could scroll past other departments and their sub-components and get to a list of independent agencies.

Later in the survey, after asking federal executives about the performance of their own agencies, we then asked them to evaluate the performance of the agencies they had mentioned at the start of the survey, plus two others. So, for example, an executive in the Natural Resources Conservation Service (USDA) might report that they work regularly with the Environmental Protection Agency, the Office of Management and Budget, and the Department of the Interior. This respondent would be asked to evaluate these three agencies and two other agencies they were likely to be familiar with (e.g., other bureaus in the USDA).⁴ Specifically, they were asked, “How would you rate the overall performance of the following agencies in carrying out their missions?” and given options from 1-Not at all effective to 5-Very effective. They were also provided a Don’t know option. Each respondent rated up to 5 agencies.

We use these ratings to generate numerical estimates of agency performance, adjusting for differences in the ability of federal executives to rate performance and differences in the way that federal executives use the 1 to 5 scale. Indeed, some federal executives may not perceive performance clearly but rate agencies anyway. Others may be harder or easier graders (e.g., what

³ This was the pattern for all the departments. Independent agencies and agencies in the Executive Office of the President were listed separately and no sub-components of these other agencies was included.

⁴ For full details of these ratings questions, including screen shots see Appendix A.

level of performance is necessary to be rated a 1, 2, 3, 4 or 5?). We have ratings on 179 agencies from 1,379 raters and 4,555 ratings of federal agency performance overall. We only report data for agencies with at least 5 performance ratings.

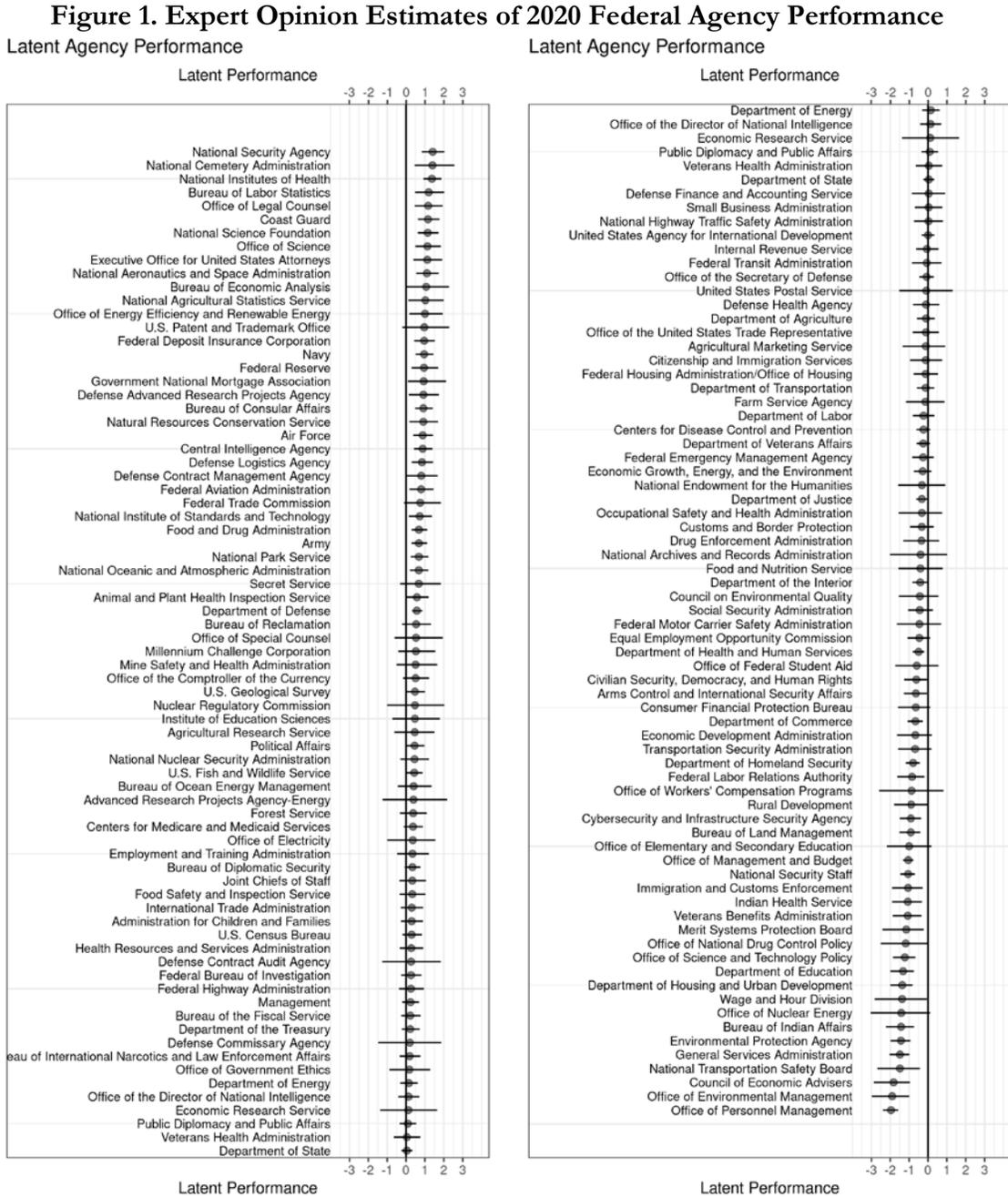
We use these ratings to estimate agency performance using a Bayesian multi-rater item response model. We leave most of the details for Appendix A but note that we estimate a hierarchical model to account for the fact that raters share a workplace and we use informed priors on latent performance to give additional weight to those respondents that reported working closely with the agency being evaluated (as compared to those that evaluated an agency selected at random).⁵

In Figure 1 we include the resulting numerical estimates of federal agency performance (Mean 0.01; SD 0.78; Min -1.96; Max 1.41) and in Appendix B we include the numerical estimates themselves. In the figure the dots represent the estimates and the bars represent the degree of uncertainty about the estimate. The uncertainty can derive from there being few ratings or because there is disagreement among the raters, or both. Indeed, it is possible, for example that Republicans and Democrats rate performance differently. This is something we address in more detail below.

Among the highest performing agencies, according to federal executives, are the National Security Agency, the National Cemetery Administration (VA), and the National Institutes of Health (NIH). The top-10 also includes several technical, science, and military agencies. Among the lowest performing agencies are a part of the Executive Office of the President whose chair was vacant for a significant part of the Trump Presidency (CEA), the agency responsible for environmental cleanup

⁵ In Appendix C we compare hierarchical and non-hierarchical models as well as estimates with informed and naïve priors. The correlations among the estimates from these different models range from 0.97 to 0.99 and so make little substantive difference in the estimates.

in the nation’s nuclear weapons labs whose PAS position was vacant at the time of the survey (OEM), and an agency that has had 9 PAS heads since 2015 (OPM). There is at least *prima facie* evidence that there is a relationship between PAS vacancies and federal agency performance.



Note: Numerical estimates of federal agency performance based upon 4,310 ratings by 1,239 federal executives in response to the question: “How would you rate the overall performance of the following agencies in carrying out their missions?” and given options from 1-Not at all effective to 5-Very effective. Federal executives were also provided a Don’t know option. Ratings were estimated using a Bayesian hierarchical multi-rater item response model with informed priors (based upon those with specific expertise about an agency).

To validate these measures, we compare them to three other measures of agency health and performance (Figure 2). First, we compare the performance ratings to 2020 agency average self-reported performance. The survey asked respondents, “How would you rate the overall performance of [your agency] in carrying out its mission?” The question asks federal executives to evaluate how well their agencies are doing by name (i.e., [your agency] is replaced with the actual name of the respondent’s workplace). Limiting our sample to those agencies with at least 5 respondents leaves us with 91 agencies. Second, we correlate the new measure with COVID performance as evaluated by the Partnership for Public Service. In 2020, the Partnership for Public Service produced scores for agency COVID performance.⁶ The Partnership’s scores are an index created using federal employee responses to the following questions:

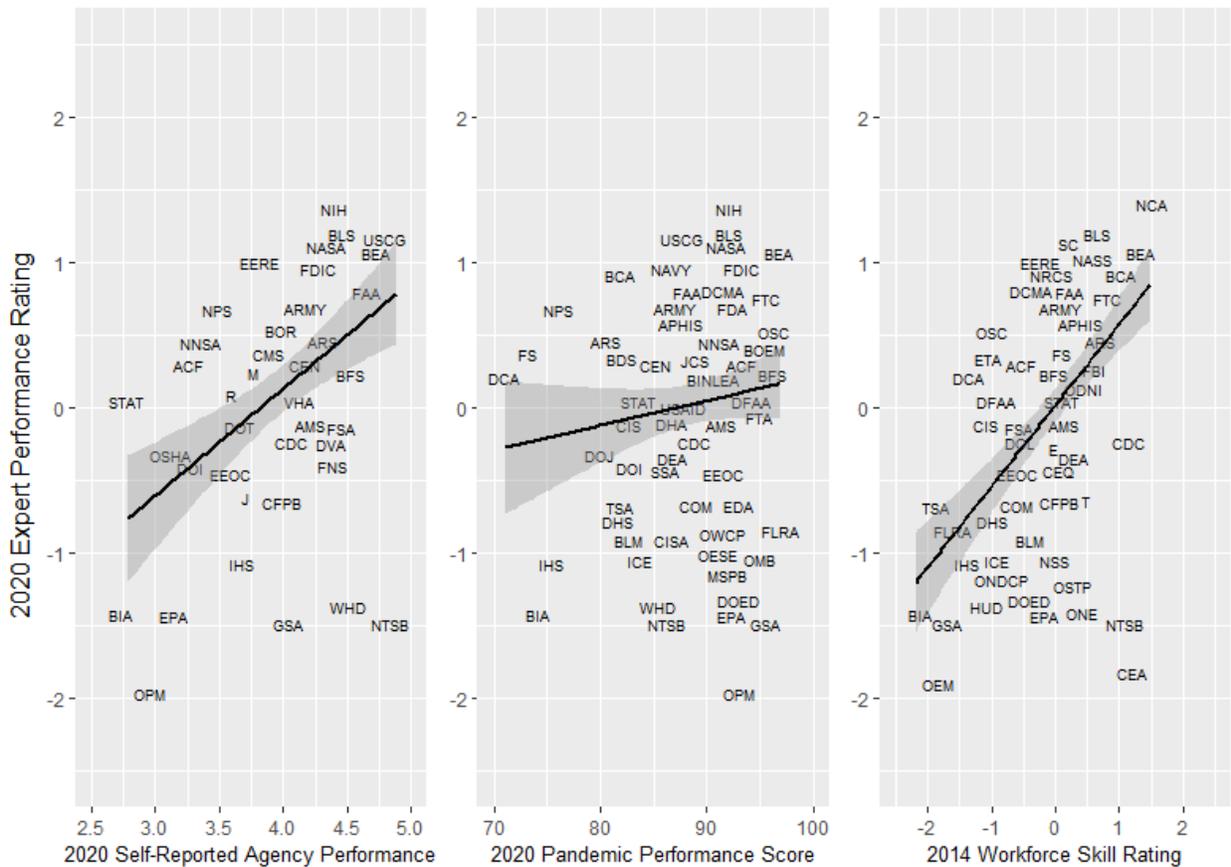
- During the COVID-19 pandemic, my work unit has met the needs of our customers.
- During the COVID-19 pandemic, my work unit has contributed positively to my agency’s performance.
- During the COVID-19 pandemic, my work unit has produced high-quality work.
- During the COVID-19 pandemic, my work unit has adapted to changing priorities.
- During the COVID-19 pandemic, my work unit has successfully collaborated.
- During the COVID-19 pandemic, my work unit has achieved our goals.

The overall score reflects the views of the rank-in-file about how they did during the difficult 2020 pandemic period. Finally, we compare estimates of performance to estimates of workforce skill at the end of the Obama Administration. In 2018, Richardson et al. used a survey of federal executives to ask, “In your view, how skilled are the workforces of the following agencies?” and given options from 1-Not at all skilled to 5-Very skilled. They also provided a Don’t know option. Each

⁶ For details and agency rankings see 2020 Best Places to Work in the Federal Government rankings, COVID: Agency Performance (https://bestplacestowork.org/rankings/?view=category&size=sub&category=covid_sub_agency_perform&).

respondent rated up to 8 agencies, providing thousands of ratings of different agencies. Richardson et al. (2018) used these ratings to develop numerical estimates of workforce skill. We correlate these estimates of workforce skill at the start of the Trump Administration with our new measures of performance.

Figure 2. Comparison of Expert Estimates of Agency Performance vs. 2020 Self-reported Performance, 2020 Best Places to Work Pandemic Performance Scores, and 2014 Workforce Skills Ratings



Note: y-axis is numerical estimate of agency performance based upon executive responses to the question: “How would you rate the overall performance of the following agencies in carrying out their missions?” and given options from 1-Not at all effective to 5-Very effective. Federal executives were also provided a Don’t know option. The x-axis of three graphs are 1) 2020 average responses to question: “How would you rate the overall performance of [your agency] in carrying out its mission?”; 2) a 2020 pandemic performance score from Partnership for Public Service; and 3) a numerical estimate of workforce skill based upon 2014 federal executive responses to question: “In your view, how skilled are the workforces of the following agencies?” (Richardson et al. 2018). Fitted lines for self-reports estimated based upon all agencies with at least 5 respondents in the survey. Some agency acronyms are excluded from the figure to limit overlap among acronyms.

The new expert performance ratings are positively correlated with all three measures, though strongest with federal executive self-reports and previous expert ratings for workforce skill.

Generally, when executives rated their own agencies poorly, outsiders agreed. When federal executives rated their agencies well, so did outsiders. While there are some outliers where agencies rated themselves as performing well but experts did not (e.g., National Transportation Safety Board [NTSB] and Wage and Hour Division [WHD]) and vice versa, the strong correlation suggests that people inside and outside agencies are observing the same performance (corr 0.46). Reassuringly, expert evaluations of the quality of the workforce before the start of the Trump Administration are also correlated positively with 2020 performance (corr 0.67). Notable differences between workforce skills in the Obama Administration and performance in the Trump Administration include agencies in the Executive Office of the President like the Council of Economic Advisers (CEA), National Security Staff (NSS), and Office of Science and Technology Policy (OSTP). Other notable differences include the Centers for Disease Control (CDC) and Prevention and the Environmental Protection Agency (EPA), suggesting these agencies were less successful performing their core missions than their workforce capacity might have predicted at the end of the Obama Administration. The correlation between overall agency evaluations of COVID responses are also positively correlated with overall performance, but only weakly so (corr 0.18). This is not entirely surprising since we survey executives and the Partnership's score is based upon surveys of the entire workforce. We also focus on performance of an agency's core mission and the Partnership is focused on COVID performance. On their face, then, these measures of agency performance appear plausibly related to actual performance.

Appointee Vacancies

To connect survey responses to vacancies, we collected data on vacancies in the Trump Administration. We used the 2016 congressional publication *Policy and Supporting Positions* (i.e., Plum Book) to identify the closest PAS appointee to every agency. We systematically tracked the occupants of each position (including acting officials) from January 20, 2017 until the soft launch of

the survey on June 22, 2020. We counted the number of days the position lacked a Senate-confirmed head. For simplicity, we divided the number of days by 30 to report vacancy length in months. For commissions, we focused on length of time the chair position was vacant.⁷ Out of a maximum of 42 months, the average position was vacant 18 months (SD 13.29), confirming the regularity with which agencies experience vacancies in leadership positions. The raw correlation between vacancies and performance is -0.11, suggesting a modest negative effect on performance. We estimate models that also evaluate the linearity of this relationship since it is possible that vacant months become less or more important as time goes on.

One immediate concern is that poor performance might *cause* vacancies rather than be caused by vacancies. For example, President Trump might have had a more difficult time finding persons to fill leadership positions in agencies that were struggling. To evaluate this possibility, we regress Trump Administration vacancies on Obama Administration workforce skills ratings and other covariates to see whether agencies with lower human capital and, therefore, likely poor performance during the Obama Administration caused Trump Administration vacancies. We cannot reject the null that struggling agencies during the Obama Administration were equally likely to experience vacancies in the Trump Administration.⁸

Controls

To hone in on the effect of vacancies themselves we control for several features of the affected positions. First, we include a control for whether the nearest appointee vacancy is in the

⁷ We have also estimated models omitting the independent commissions and the estimates are substantively similar.

⁸ We note, however, that coefficient is negative, suggesting that higher performing agencies *are* less likely to experience vacancies.

agency itself (0,1; 78%) or a higher level. For example, in our data the Bureau of Ocean Energy Management (BOEM) is headed by a career professional and the closest presidential appointee is the Assistant Secretary for Land and Minerals Management. The Bureau of Land Management (BLM), however, is overseen by the same assistant secretary but the BLM is headed by its own Senate-confirmed appointee. Second, to parse out the unique effect of a vacancy from turnover we control for the turnover in the position, which is the total number of persons that served in that position, acting or confirmed, during the Trump Administration from Inauguration Day until the start of the survey (Mean 1.53; SD 1.1; Min 0, Max 4).⁹

There are a number of agency characteristics that might be correlated with both vacancy length and agency performance. We include controls for a number of pre-treatment covariates, including agency structure, whether the agency implements a policy that President Trump mentioned during the 2016 campaign, agency workforce skills measured during the Obama Administration, employment and agency ideology. We include controls for whether the agency is an agency in the Executive Office of the President (0.05), whether the agency rated is an entire department (e.g., “Department of Agriculture (All)”); 0.11), or in an independent commission (0.08).

⁹ The source of this data is the quarterly edition of the *Federal Yellow Book*. To begin, during the first quarter of 2021, we observe who holds a position. We then track each subsequent quarter to see if there has been a change. If a position is vacant in the first quarter but filled in the second, we do not consider this turnover because it is not a shift from one person to another. It should be noted, however, that because we look at quarters, we may miss some turnover. For example, the Secretary of Commerce position was held by an acting official from January 20, 2017 to January 27, 2017, when Wilbur Ross was confirmed. Because we rely on quarterly data, we do not observe this transition.

The inclusion of the structural features means the base category is a sub-component of an executive department or an independent executive agency like the Environmental Protection Agency. Since the president’s decisions about whether to fill vacancies and support the agency more generally could influence performance, we include indicators for whether the department (0,1; 0.65) implemented a policy that was a priority of the president during the 2016 election.¹⁰ We include the 2014 workforce skills rating mentioned above as a measure of the health of the agency at the end of the Obama Administration. Finally, since the persistence of vacancies and presidential attention might be influenced by the ideological contours of what agencies do, we include controls for agency ideology measured prior to the start of the Trump Administration (Richardson et al. 2018). Richardson et al. asked federal executives during the Obama Administration about the ideological leanings of agencies they work with, and whether the “lean liberal, lean conservative, or neither consistently across Democratic and Republican administrations.” They aggregate responses with a method similar to the workforce skill scores described here to generate estimates of agency ideology.¹¹

Methods

To evaluate the relationship between vacancies and performance, we estimate a series of models with OLS. For each hypothesized relationship, we estimate models on the complete set of agencies and the subset of agencies directly led by an appointee. We cluster the standard errors to

¹⁰ We identify policy priorities of the president using the Contract with the American Voter, a campaign document produced by the Trump Campaign in October 2016 (<https://assets.donaldjtrump.com/landings/contract/O-TRU-102316-Contractv02.pdf>).

¹¹ Continuous control variables were scaled so that a one-unit increase represents a one standard deviation increase from the mean value of the variable.

account for the fact that different agencies are not completely independent since many are subcomponents of larger departments. Specifically, we have 17 clusters, one for each executive department, one for the EOP, and one for independent agencies.

Results

We include the initial set of model estimates in Table 1. To begin, there are several interesting results among the controls. First, agencies directly led by appointees are estimated to have lower overall performance, about 0.17 ($\sim 1/5^{\text{th}}$ a standard deviation), although the estimates are imprecise. The estimates suggest that increasing penetration of appointees is correlated with lower agency performance. Second, agencies that had higher leader turnover were estimated to have *higher* performance. Each additional leader during this 42-month period was correlated with a 0.10 higher performance. While scholars generally associate turnover with lower performance, turnover in this case could be a measure of whether agencies have presidential attention at all since such a large number of positions were vacant. The dynamics of turnover and the dynamics of vacancies might be quite different since long vacancies can be correlated with low turnover and vice versa (Rutherford et al. 2018; Piper and Lewis 2022).

Also notable, model estimates reveal that agencies in the EOP, larger departments, and independent commissions had lower performance ratings than other agencies. Agencies in President Trump's EOP were estimated to have three quarters of a point lower performance than bureaus in executive departments or independent administrations (i.e., the base category). This is consistent with early reporting on the performance and staffing of the Trump White House (Lewis and Richardson 2021). Whole departments, when evaluated, were rated between 0.54 to 0.68 lower than other agencies and independent commissions, about 0.18 lower. In addition, agencies that were responsible for implementing policies the president mentioned during the campaign were estimated to have lower performance, about half to a third of a point lower. Not surprisingly, agencies

evaluated to have skilled workforces at the end of the Obama Administration were rated to have higher performance. A standard deviation increase in workforce skill (0.76) is estimated to increase performance by 0.42. Finally, agency ideology is largely uncorrelated with performance.

Table 1: OLS Models of the Effect of Vacancies on Agency Performance

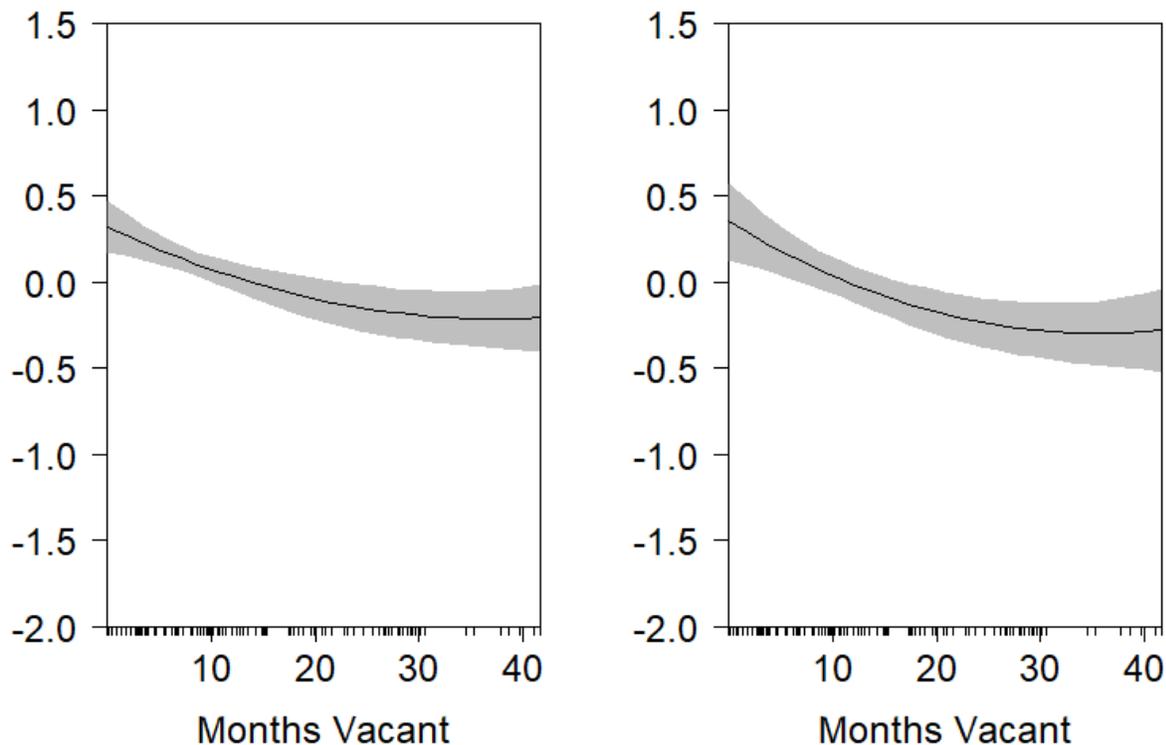
| Agency Performance (-1.96, 1.41) | All Agencies | All Agencies | PAS Head Only | PAS Head Only |
|---|---------------------|---------------------|---------------------|----------------------|
| Months Vacant (0-42) | -0.012 [0.003]** | -0.029 [0.010]** | -0.014 [0.005]** | -0.037 [0.012]** |
| Months Vacant ² | | 0.0004 [0.0002]* | | 0.0005 [0.0002]** |
| <i>Position Controls</i> | | | | |
| Direct PAS Leader (0,1) | -0.177 [0.122] | -0.185 [0.116] | | |
| Leadership Transitions (0-4) | 0.09 [0.034]** | 0.105 [0.035]** | 0.104 [0.041]** | 0.124 [0.042]** |
| <i>Agency Controls</i> | | | | |
| EOP (0,1) | -0.779 [0.088]** | -0.736 [0.082]** | -0.765 [0.094]** | -0.713 [0.082]** |
| Cabinet Department (0,1) | 0.356 [0.067]** | 0.367 [0.058]** | 0.379 [0.075]** | 0.385 [0.064]** |
| Whole Department (0,1) | -0.54 [0.132]** | -0.615 [0.136]** | -0.586 [0.149]** | -0.675 [0.158]** |
| Independent Commission (0,1) | -0.177 [0.074]** | -0.181 [0.070]** | -0.195 [0.078]** | -0.203 [0.073]** |
| Priority Department (0,1) | -0.31 [0.137]** | -0.309 [0.128]** | -0.31 [0.151]* | -0.306 [0.139]** |
| Workforce Skill -- Obama Administration | 0.551 [0.063]** | 0.54 [0.063]** | 0.583 [0.068]** | 0.572 [0.068]** |
| Agency Ideology (L-C) | 0.045 [0.073] | 0.038 [0.074] | 0.008 [0.089] | -0.004 [0.091] |
| Constant | 0.309 [0.152]* | 0.406 [0.131]** | 0.141 [0.064]** | 0.253 [0.052]** |
| R ² | 0.49 | 0.49 | 0.48 | 0.49 |
| N | 125 | 125 | 101 | 101 |

Note: * p<0.1; ** p<0.05. Data: *Survey on the Future of Government Service*, 2020. Dependent variable is agency performance estimate based upon aggregated responses to question: “How would you rate the overall performance of the following agencies in carrying out their missions?” (1-5). Models estimated with Ordinary Least Squares and standard errors are clustered by department. Models 3, 4 estimated using only agencies whose immediate head is a Senate-confirmed political appointee (PAS).

Most importantly for the purposes of the paper, however, the models reveal a robust relationship between vacancy length and perceptions of agency performance. The coefficient estimate for the direct effect of the number of vacant months is negative and statistically significant in all four models. The result is strongest when subsetting to cases where the agency is led directly by a political appointee (Models 3, 4). Collectively, these results indicate that experts rate agency performance more poorly where PAS positions are vacant for long periods. To contextualize the effect, it is useful to evaluate the effect in terms of real vacancy times. For every 12 months a PAS headed position is vacant, the agency's average performance is estimated to decrease by between 0.14 and 0.17 on a scale of -1.96 to 1.41. If the position stays vacant for 24 months or about half the president's term (i.e., about 33% of the data), the estimated effect is between 0.28 and 0.34, or about one third of a standard deviation decrease. Finally, if the position is vacant for the duration of the Trump administration prior to the survey (i.e., about 13% of the data), the estimated performance is about 0.50 lower than other agencies, or about 2/3 of a standard deviation. This is important evidence, using new performance measures, that vacancies are correlated with lower federal agency performance. This confirms some of the theoretical expectations of past work (O'Connell 2009, 2020; Piper and Lewis 2022; Rutherford et al. 2018).

While the average effect of vacancies is notably negative, the effect is non-linear (Models 2, 4; Figure 3). The effect of is largest early in the vacancy and becomes negligible the longer a position is vacant. While we do not observe vacancies that carry over beyond the Trump Administration, there is suggestive evidence that the effect of a vacancy may turn positive if a position is vacant long enough. This may indicate that positions with persistent vacancies perform better if leadership positions were converted to positions filled by permanent career professionals.

Figure 3. Estimated Impact of Vacancies in Senate Confirmed Positions on 2020 Agency Performance Rating



Note: Figure graphs predicted agency performance rating by months most appointed position vacant during the Trump Administration. The first model is based upon estimates from Model 2 in Table 1 and the second model is based upon estimates from Model 4 in Table 1.

In total, the evidence suggests a negative relationship between an agency’s leadership vacancy and performance evaluations by those working most closely with the agency. Obviously, agency performance is multifaceted; the impact of vacancies on performance can vary by context. Still, even in the midst of this complexity, a clear correlation emerges.

How Does Partisanship Shape Government Performance Evaluation?

One concern with public sector performance evaluation is that what a Democrat considers good performance might be considered bad performance by Republicans and vice versa. In other words, Democrats and Republicans can observe the same agency activities and outputs but not agree on whether it is good performance. If this is the case, this has serious ramifications not only for this project but performance management efforts more generally. One reason for confidence in standard performance measurement efforts, however, is that a large percentage of government work is

apolitical, suggesting partisanship should not overwhelm evaluation (Richardson n.d.). In addition, most federal government programs are overwhelmingly popular with the public (Fiorina 1977; Gramlich 2017; Light 2008), suggesting that for most programs partisans agree on good performance when they see it. Still, it is an open question if and how much partisanship influences the evaluation of federal agency performance.

One virtue of our approach to measuring performance is that we know whether federal executives rating agency performance are Republicans, Independents, or Democrats.¹² This allows us to evaluate whether and how much partisanship shapes evaluations of government performance. In Table 2 we show the results of a simple bivariate regression where the raw agency rating (1-Not at all effective to 5-Very effective) is the dependent variable and the partisanship of the rater is the

¹² We have also evaluated whether agencies that are the subject of regular partisan disagreement are rated differently by Democrats and Republicans. In the 2020 survey we asked respondents “How often do Republicans and Democrats in Washington disagree over what [your agency] should do?” [Never, Rarely, Sometimes, Often, Always, Don’t know] We regressed the performance ratings on a measure of latent partisan disagreement derived from responses to these questions about partisan disagreement (see Richardson n.d. for estimation details). Interestingly, the more an agency is subject to partisan disagreement, the lower its rating overall. When Democrats and Republicans disagree, they both seem to think the agency is performing poorly. So, rather than Democrats rating agencies high and Republicans rating agencies low or vice versa, it looks like the raters agree on performance. Importantly, this disagreement also does not appear to decrease the *uncertainty* of the estimates. When we regress the standard deviation of the performance estimates on partisan disagreement, there is no relationship. This is evidence that partisan disagreement does not lead to divergent evaluations. We include this analysis in Appendix D.

independent variable.¹³ The coefficient estimates from the first model show that Democrats rated agency performance about 0.13 lower and Republicans 0.17 higher than independents, on average. The results do not tell us if Democrats and Republicans use the scale similarly, if they rank agencies the same but with an overall shift up or down, or if they see some agencies the same and others differently but they do show that Republican respondents rated Trump-era agencies somewhat higher than Democrats. Importantly, the results in Model 2 show that while Republicans rate agency performance higher than Democrats (and Independents) on average, the effect of partisanship does not differ based upon the ideological leaning of the agency.¹⁴ So, Democrats are no more critical of conservative agencies and Republicans are no more critical of liberal agencies. In effect, what we observe is an intercept shift, where Republicans rate Trump-era agencies somewhat higher and Democrats rate Trump-era agencies somewhat lower.

¹³ In total, there were 2,360 ratings by Democrats, 886 ratings by Republicans and 569 ratings by respondents that were independents or refused to identify with a party in the survey.

¹⁴ We use measures of agency ideology from Richardson et al. (2018). In their 2014 survey, Richardson et al. ask respondents about the ideological leanings of agencies they work with, whether the “lean liberal, lean conservative, or neither consistently across Democratic and Republican administrations.” They aggregate responses with a method similar to the performance scores described here to generate estimates of agency ideology.

Table 2. OLS Models of *Raw* Agency Ratings based Upon Self-reported Partisanship of Rater, 2020

| | (1) | | (2) | |
|----------------------------------|-------|---------|-------|---------|
| Democrat rater (0,1) | -0.13 | 0.05** | -0.11 | 0.05* |
| Republican rater (0,1) | 0.17 | 0.05** | 0.16 | 0.05** |
| Agency ideology (L,C) | | | 0.14 | 0.04** |
| Democrat rater*Agency ideology | | | 0.02 | 0.05 |
| Republican rater*Agency ideology | | | 0.05 | 0.06 |
| Constant | 0.03 | 0.04 | 0.01 | 0.04 |
| N | | 3,812 | | 3,662 |
| Adj. R ² | | 0.02 | | 0.04 |
| F-statistic | | 30.30** | | 31.31** |

Note: ** significant at the 0.05 level; * significant at the 0.10 level in two-tailed tests. Dependent variable is agency rating: “How would you rate the overall performance of the following agencies in carrying out their missions?” (1-5). Models estimated via ordinary least squares (OLS). Base category is ratings by respondents that are independent or do not share a party affiliation.

To further unpack the ways that Democrats and Republicans evaluate performance, we generated new agency performance estimates, one set based on the ratings of Democrat federal executives and one set based only on the ratings of Republican executives. To generate the estimates, we replicate the model above, but limit the ratings to Democratic or Republican raters, respectively.¹⁵ We make sure the estimates are on the same scale by fixing the latent performance for

¹⁵ We also generate estimates where the hierarchy structuring the estimates is partisanship rather than workplace. This allows us to evaluate whether Democrats and Republicans are perceiving performance the same way. In the models, there is little difference in the party-level posterior means of the rater slope parameters. The slope for Democrats is 0.06 and for Republicans 0.05 (the scale is standard normal). So, a one unit increase in latent performance affects Republican and Democratic perception of performance almost identically. The party-level posteriors for the rater intercepts are 0.06 for Democrats and 0.33 for Republicans, confirming that Republicans rate Trump-era agencies more highly on average. In principle, it is possible to estimate both department and ideology as hierarchical factors but doing so would require a significantly larger number of ratings across departments and partisans.

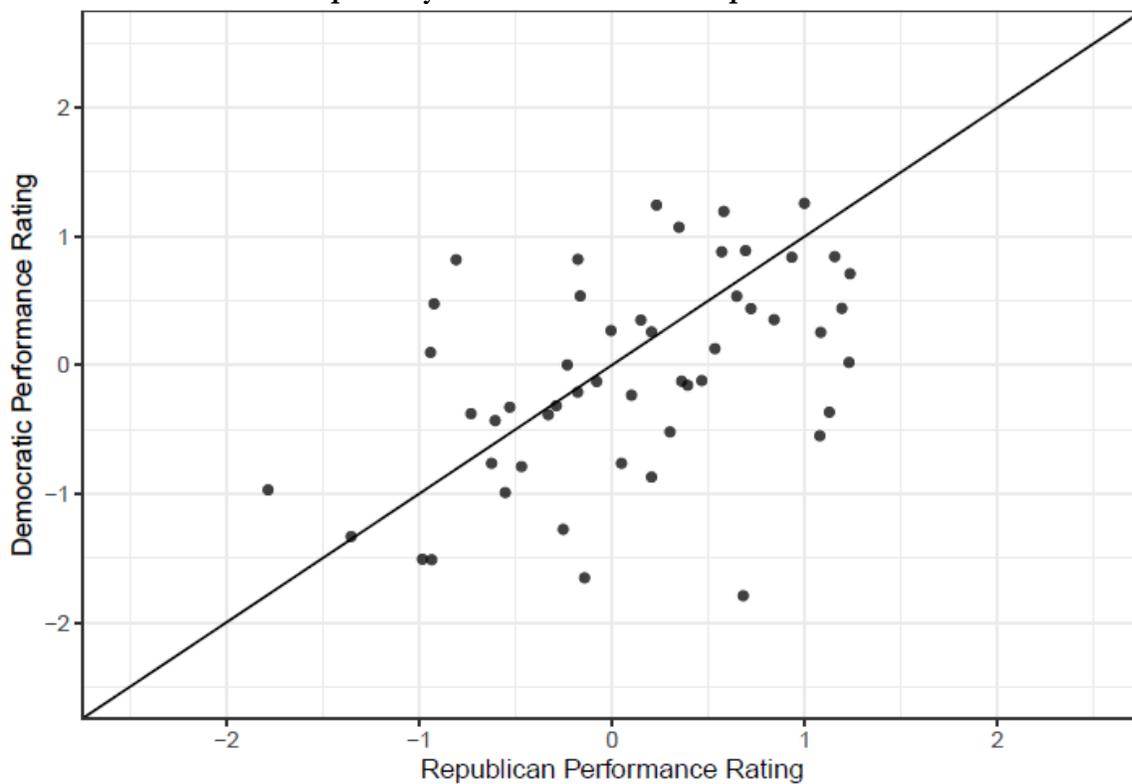
three agencies that span the range of performance.¹⁶ We compare the two sets of ratings to see how much the ratings overlap and identify agencies about which Democrats and Republicans agree and disagree. We generate fewer agency estimates since we have fewer ratings when we separate out Democrats and Republicans but the results are informative.

In Figure 3 we include a scatterplot of these estimates with a 45-degree line for reference. In general, where Republicans rate agencies as high performers, Democrats do the same (corr 0.47).

¹⁶ We used three criteria to identify the agencies for which we fixed their latent performance across parties to recover separate performance ratings by Republicans and Democrats on the same scale for other agencies. First, we wanted agencies whose raw means were very similar across parties suggesting that Republicans and Democrats do indeed perceive performance similarly. Second, we wanted agencies that span the range of performance with one agency at the low end of the scale and one at the high end to locate the ends of the scale. Third, we wanted agencies that many respondents from each party rated to anchor interpersonal comparisons across parties for as many raters as possible. Using these criteria, we selected the Office of Management and Budget, the Office of Personnel Management, and the Coast Guard. Each agency has similar mean performance among Democrats and Republicans. Specifically, the mean rating of OMB on the 1-5 scale was 3.23 among Republicans and 3.12 among Democrats, mean rating of OPM was 2.84 among Republicans and 2.57 among Democrats, and the mean rating of the Coast Guard was 4.35 among Republicans and 4.07 among Democrats. OPM provides an agency on the low end of the scale and the Coast Guard provides an agency on the high end of the scale. Lastly, OMB was rated by 153 Republicans and 206 Democrats, OPM was rated by 125 Republicans and 149 Democrats, and the Coast Guard was rated by 15 Republicans and 12 Democrats. In sum, all three agencies satisfy the first criteria, OMB and OPM satisfy the second, and OPM and Coast Guard satisfy the third.

Where Republicans rate agencies as low performers, Democrats similarly see low performance. This should give us confidence in the regression estimates above showing a relationship between vacancy length and performance. Any differences between how Democrats and Republicans evaluate performance do not lead to materially different their evaluations of underlying performance. This similarity should also provide confidence to those engaged in ongoing efforts to measure federal agency performance.

Figure 3. Scatterplot of Agency Performance Estimates with Estimates Generated Separately for Democrat and Republican Raters

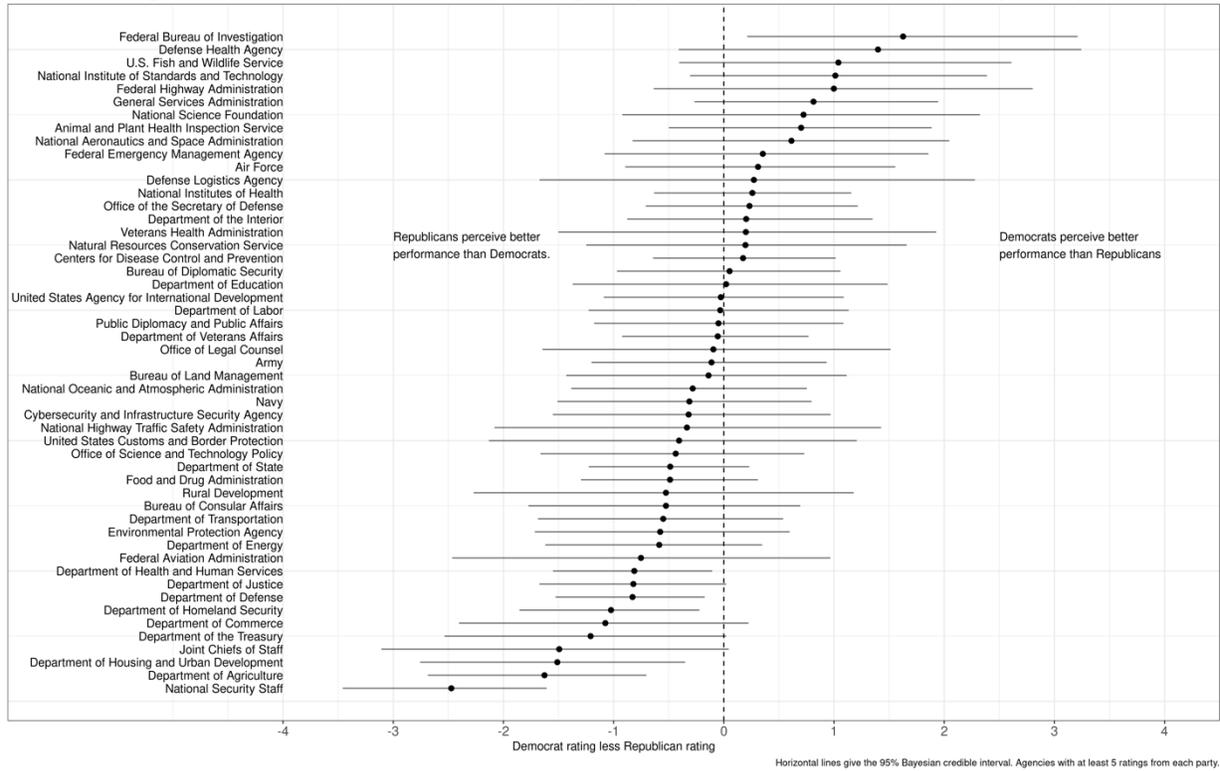


The results thus far suggest that, while Democrats and Republican do sometimes do evaluate performance differently (i.e., some dots are well off the 45-degree line), they observe performance the same way, on average. High performers for Democrats are generally high performers for Republicans and vice versa. While this is generally true, there are cases where raters from the two parties tend to disagree. In Figure 4 we graph the difference in agency performance estimates in

order to identify the agencies where the Democratic and Republican ratings are notably different. In essence, we are just measuring how far off the 45-degree line an agency is from Figure 3 and including the related statistical uncertainty (i.e., do we have high confidence that the ratings really are different?). In this figure, perfect agreement between Democrats and Republicans means points close to zero. Points further to the right are cases where Democrat federal executives rated an agency more highly than Republican federal executives. On the left are cases where the reverse is true, Republicans rated an agency more highly than Democrats.

A few things stand out in the figure. First, more points are to the left of 0, consistent with our earlier finding that Republicans rated agencies slightly more highly during the Trump Administration than Democrats. Second, most of the points are generally centered around 0, suggesting similar ratings by Democrats and Republicans. Finally, there are only a few cases where we are confident that the ratings differ by party (i.e., the 95% interval does not overlap 0). These include the Departments of Agriculture, Defense, Health and Human Services, Housing and Urban Development, and Homeland Security, as well as the Federal Bureau of Investigation (FBI). Given the political debates of the Trump Administration differing opinions about these agencies are understandable. For example, Democrats argued that the FBI was doing its job when investigating the president and Republicans argued that the president was unfairly targeted. Similarly, efforts to move the Economic Research Service out of Washington likely influenced perceptions of the Performance of the Department of Agriculture. Interestingly, many of the cases where partisans disagree are cabinet departments evaluated as a whole, where identifying a core mission is difficult. In cases where there is no one core mission it is easier for factors like partisanship to define perceptions of performance. Large units often also invite evaluations of the performance based upon opinions of visible leaders (e.g., Ben Carson, William Barr) rather than the agency itself.

Figure 4. Difference in Selected Agency Performance Estimates by Party



To return to our original question, we are concerned about whether vacancies in Senate-confirmed leadership positions harm performance. The results with the basic ratings above suggest that vacancies are harmful for performance, but do partisan differences in ratings influence these results? In Table 3 we replicate the models from Table 1, this time excluding the cases where Democrats and Republicans disagree significantly on the ratings (Figure 4).¹⁷ The results are broadly

¹⁷ The creation of separate estimates for Democrats and Republicans makes it possible to estimate models on Democrat ratings and Republican ratings separately and see if the vacancy results are robust. This is difficult for two reasons. First, since there are fewer raters when we subset by party, we get fewer agency performance estimates overall and the cases that remain tend to be larger agencies (e.g., executive departments). Second, given what we know about some of the partisan ratings being determined by partisanship rather than performance, estimating such models would be

consistent with the estimates reported in Table 1. If anything, the relationship between vacancies and federal agency performance emerges more clearly. The coefficient estimates are larger and still estimated precisely, indicating that vacancies in presidentially appointed leadership positions are correlated with lower agency performance, as measured through the opinions of federal executives that work outside these agencies. As before, the coefficients on the interaction effects suggest that the effects of the vacancies taper off the longer a position is vacant. This is important validation that the main finding from the first section of the paper is robust to partisan differences in those providing the performance ratings. The breakdown in the appointments process due to slow nominations and confirmation delays appears to matter for agency performance. Federal executives that evaluate agency performance on core missions rate agencies with PAS vacancies lower than agencies with confirmed appointees in place.

acknowledging that we are not regressing performance on vacancies. Rather, we are regressing partisan perceptions of performance on vacancies. One solution is to estimate separate models on Democrat and Republican ratings but exclude the known cases of partisan disagreement. This leaves us with 37 cases. In Appendix E we include estimates from models of partisan ratings, excluding the cases where the partisan ratings diverge significantly. We include structural controls to account for the fact that larger and more visible units are harder to rate. These models generally confirm what is reported here. We note, however, that in more fully specified models that also include controls for Obama-era performance, these estimates get much smaller and imprecise. With the small N and the limited degrees of freedom, it is difficult to know how much weight to put on these results. That said, we recommend readers apply appropriate caution in interpreting the results and recommend further research of this type.

Table 3. OLS Models of the Effect of Vacancies on Agency Performance, Excluding Cases Where Democrats and Republicans Disagree

| Agency Performance (-2.5, 2.5) | All Agencies | All Agencies | PAS Head Only | PAS Head Only |
|---|---------------------|----------------------|---------------------|----------------------|
| Months Vacant (0-42) | -0.013 [0.004]** | -0.036 [0.010]** | -0.016 [0.005]** | -0.046 [0.013]** |
| <i>Position Controls</i> | | | | |
| Direct PAS Leader (0,1) | -0.216 [0.138] | -0.226 [0.132] | | |
| Leadership Transitions (0-4) | 0.105 [0.033]** | 0.123 [0.033]** | 0.121 [0.041]** | 0.145 [0.039]** |
| <i>Agency Controls</i> | | | | |
| EOP (0,1) | -0.731 [0.099]** | -0.671 [0.088]** | -0.717 [0.109]** | -0.647 [0.093]** |
| Cabinet Department (0,1) | 0.401 [0.075]** | 0.422 [0.064]** | 0.424 [0.088]** | 0.442 [0.076]** |
| Whole Department (0,1) | -0.542 [0.140]** | -0.671 [0.151]** | -0.614 [0.170]** | -0.774 [0.197]** |
| Independent Commission (0,1) | -0.199 [0.078]** | -0.206 [0.070]** | -0.212 [0.085]** | -0.224 [0.076]** |
| Priority Department (0,1) | -0.369 [0.146]** | -0.372 [0.130]** | -0.357 [0.168]* | -0.353 [0.147]** |
| Workforce Skill -- Obama Administration | 0.555 [0.065]** | 0.541 [0.067]** | 0.583 [0.069]** | 0.568 [0.070]** |
| Agency Ideology (L-C) | 0.037 [0.081] | 0.025 [0.083] | -0.019 [0.098] | -0.04 [0.100] |
| Months Vacant ² | | 0.0005 [0.0002]** | | 0.0007 [0.0002]** |
| Constant | 0.368 [0.158]** | 0.500 [0.141]** | 0.151 [0.054]** | 0.304 [0.059]** |
| R ² | 0.49 | 0.50 | 0.48 | 0.50 |
| N | 116 | 116 | 93 | 93 |

Note: * p<0.1; ** p<0.05. Data: *Survey on the Future of Government Service*, 2020. Dependent variable is agency performance estimate based upon aggregated responses to question: “How would you rate the overall performance of the following agencies in carrying out their missions?” (1-5). Models estimated with Ordinary Least Squares and standard errors are clustered by department. Models 3, 4 estimated using only agencies whose immediate head is a Senate-confirmed political appointee (PAS).

In addition, despite concerns that partisanship endangers the performance measurement enterprise more generally, the evidence here suggests that is not the case. While partisanship comes into play, particularly in larger agencies where mission is unclear and leadership is visible and salient,

true latent performance seems to be observed similarly by Democrats and Republicans. While Republicans might have had an easier time seeing the good in Trump Administration agency performance, both Democrats and Republicans appear to see high or low agency performance similarly. In most cases, when Democrats rated agencies more highly, so did Republicans and when Republicans rated an agency more highly, so did Democrats. This does not diminish the importance of the exceptions but suggests that our use of performance measures be tempered by the clarity of what is being measured and its connection to visible political leaders.

Discussion and Conclusion

On May 17, 2022, President Biden renewed his call for Congress to confirm his nominee to lead the Bureau of Alcohol, Tobacco, Firearms, and Explosives (ATF). The agency had been without a permanent director since 2015 (Buble 2022). Biden's claim was that part of the nation's gun problem could be traced to a lack of confirmed leadership in the agency. This paper has sought to determine whether there is a relationship between vacancies and federal agency performance, not just in ATF but government-wide. It tests Biden's claim in a general way with original new data and measures of federal agency performance that are comparable across contexts. It includes a number of key findings. First, consistent with several recent works, the paper shows that vacancies are associated with lower agency performance (O'Connell 2009, 2020; Piper and Lewis 2022; Rutherford et al. 2018). Second, the effect of a vacancy diminishes over time, suggesting that agencies may adjust to operating without confirmed appointed leadership or, at minimum, stem the decline in performance. Finally, the results are robust to partisan differences among raters. Despite differences between Republicans and Democrats, vacancies are correlated with poorer performance even when accounting for these differences.

Several important implications result from these analyses for political appointments and performance measurement and management. First, the results highlight the potential consequences

of the current dysfunction in the appointments process in the United States. As partisan polarization leads to slower nominations and longer and more uncertain confirmations, the average effect will likely be lower agency performance. This is consistent with the arguments by some that speeding up the appointments process or reducing the number of political appointees may have benefits for performance (O’Connell 2021; Stier 2021).

Second, the performance measurement method employed here is generalizable to other contexts, notably cases where persons outside an organization can be surveyed about their impression of the performance of the organization (Richardson et al. 2018). This is particularly useful in contexts where other objective measures of performance (e.g., profit) are difficult to obtain or assess (Wilson 1989). Data on the persons doing the rating can be used to directly evaluate the quality or potential bias among those evaluating performance, reducing concerns about bias in performance measurement (Lavertu and Moynihan 2013).

Finally, these results suggest that practitioners using performance measures can do so reasonably in many contexts. Indeed, since the vast majority of government work is apolitical or popular with the public, Democrats and Republicans observe performance similarly (Fiorina 1977; Gramlich 2017; Light 2008; Richardson n.d.). Nonetheless, the results here suggest that users of performance measures need to be attuned to potential partisan bias in ratings, particularly as agencies get larger and it becomes harder to identify what core task is being evaluated (Lavertu and Moynihan 2013). In our results, raters evaluating whole departments rated them lower than other agencies and they also disagreed the most about these ratings. Together, these findings suggest that performance evaluation is easiest and least subject to bias at lower levels of aggregation and when the performance to be evaluated is clear.

The data underlying this article are available in [Harvard Dataverse Link to come]

References

- Aberbach, Joel D., and Bert A. Rockman. 2000. *In the Web of Politics: Three Decades of the U.S. Federal Executive*. Washington, DC: Brookings.
- Bernstein, Jonathan. 2022. "A Senseless Logjam is Holding Up Nominations." *Washington Post*, February 9, 2022 (https://www.washingtonpost.com/business/a-senseless-logjam-is-holding-up-nominations/2022/02/09/34515dc0-89ad-11ec-838f-0cfd69c9c3c_story.html).
- 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, and Jay Dahya. 2002. "Executive Succession and the Performance of Public Organizations." *Public Administration* 80(1):179-200.
- Boyne, George A., Oliver James, Peter John, Nicolai Petrovsky. 2011. "Top Management Turnover and Organizational Performance: A Test of a Contingency Model." *Public Administration Review* 71(4):572-81.
- Buble, Courtney. 2022. "The White House Repeats Its Call for a Confirmed Guns Agency Director After Buffalo Tragedy." *Government Executive*, May 17, 2022 (<https://www.govexec.com/management/2022/05/white-house-repeats-its-call-confirmed-guns-agency-director-after-buffalo-tragedy/367041/>).
- Cohen, David M. 1998. "Amateur Government." *Journal of Public Administration Research and Theory* 8(4):450-97.
- Dull, Matthew, and Patrick S. Roberts. 2009. "Continuity, Competence, and the Succession of Senate-Confirmed Agency Appointees, 1989-2009." *Presidential Studies Quarterly* 39(3): 432–53.
- Eilperin, Juliet, Josh Dawsey, and Seung Min Kim. 2019. "‘It’s way too many’: As vacancies pile up in Trump administration, senators grow concerned," *Washington Post*, February 4, 2019 (https://www.washingtonpost.com/national/health-science/its-way-too-many-as-vacancies-pile-up-in-trump-administration-senators-grow-concerned/2019/02/03/c570eb94-24b2-11e9-ad53-824486280311_story.html).
- 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.
- Freking, Kevin, "Sluggish pace of confirmations vexes Biden White House," *AP News*, October 24, 2021 (<https://apnews.com/article/joe-biden-donald-trump-barack-obama-ted-cruz-george-w-bush-ec0045ca948d4c2862e9c9a855631b51>)
- Gramer, Robbie. 2022. "3 Years Later, the U.S. Could Finally Send an Ambassador to Ukraine," *Foreign Policy*, April 25, 2022 (<https://foreignpolicy.com/2022/04/25/ukraine-ambassador-biden-bridget-brink-russia/>).

- Hecl, Hugh. 1977. *A Government of Strangers: Executive Politics in Washington*. Brookings Institution.
- Hill, Gregory C. 2005. "The Effects of Managerial Succession on Organizational Performance." *Journal of Public Administration Research & Theory* 15(4):585-97.
- Hur, Yongbeom. 2013. "Turnover, Voluntary Turnover, and Organizational Performance: Evidence from Municipal Police Departments." *Public Administration Quarterly* 37(1):3-35.
- Hussein, Fatima, Ricardo Alonso-Zaldivar, Hope Yen, and Colleen Long. 2022. "Biden agency vacancies to drag on White House priorities," *ABC News*, February 21, 2022 (<https://abcnews.go.com/Business/wireStory/biden-agency-vacancies-drag-white-house-priorities-83036250>).
- 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.
- Lavertu, Stephane, and Donald P. Moynihan. "Agency Political Ideology and Reform Implementation: Performance Management in the Bush Administration." *Journal of Public Administration Research & Theory* 23(3):521-49.
- Lewis, David E. 2007. "Testing Pendleton's Premise: Do Political Appointees Make Worse Bureaucrats?" *Journal of Politics* 69(4):1073-88.
- Lewis, David E., and Mark D. Richardson. 2021. "'The Very Best People: President Trump and the Management of Executive Personnel.'" *Presidential Studies Quarterly* 51(1):51-70.
- Light, Paul. 1995. *Thickening Government: Federal Hierarchy and the Diffusion of Accountability*. Washington, DC: Brookings Institution Press.
- . 2008. *A Government Ill Executed: The Decline of the Federal Service and How to Reverse It*. Cambridge, MA: Harvard University Press.
- Lu, Chris. 2019. "Why Trump's Cabinet vacancies, turnover threaten our government." *USA Today*, April 17, 2019 (<https://www.usatoday.com/story/opinion/2019/04/17/trump-cabinet-vacancies-turnover-threaten-government-column/3465786002/>).

- MacMahon, Arthur W., and John D. Millett. 1939. *Federal Administrators*. New York: Columbia University Press
- Meier, Kenneth J., and Alisa Hicklin. 2007. "Employee Turnover and Organizational Performance: Testing a Hypothesis from Classical Public Administration." *Journal of Public Administration Research & Theory* 18(4):573-90.
- Meier, Kenneth J., and Laurence J. O'Toole Jr. 2002. "Public Management and Organizational Performance: The Effect of Managerial Quality." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 21(4): 629–43.
- 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.
- Mendelson, Nina A. 2014. "The Uncertain Effects of Senate Confirmation Delays in the Agencies." *Duke LJ* 64: 1571.
- Moynihan, Donald P., and Sanjay K. Pandey. 2005. "Testing How Management Matters in an Era of Government by Performance Management." *Journal of Public Administration Research & Theory* 15(3):421-39.
- Nyhan, Ronald C., and Herbert A. Marlowe Jr. 1995. "Performance Measurement in the Public Sector: Challenges and Opportunities." *Public Productivity & Management Review*: 333–48.
- O'Connell, Anne Joseph. 2009. "Vacant Offices: Delays in Staffing Top Agency Positions." *S. Cal. L. Rev.* 82: 913.
- . 2020. "Actings." *Columbia Law Review* 120(3): 613–728.
- . 2021. "The Biden administration still has a lot of vacant positions. It's a growing problem." *Washington Post*, October 28, 2021 (<https://www.washingtonpost.com/politics/2021/10/28/biden-administration-still-has-lot-vacant-positions-its-growing-problem/>).
- Pager, Tyler, Ann E. Marimow, and Laurie McGinley, "Vacancies remain in key Biden administration positions," *Washington Post*, July 10, 2021 (https://www.washingtonpost.com/politics/biden-administration-vacancies/2021/07/10/635b5eba-e0e1-11eb-a501-0e69b5d012e5_story.html).
- 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*.
- Pfiffner, James P. 1996. *The Strategic Presidency: Hitting the Ground Running*. Lawrence: University Press of Kansas.

- Piper, Christopher, and David E. Lewis. 2022. "Do Vacancies Hurt Federal Agency Performance?" *Journal of Public Administration Research & Theory*, forthcoming.
- Resh, William G., Gary E. Hollibaugh, Patrick S. Roberts, and Matthew M. Dull. 2021. "Appointee Vacancies in US Executive Branch Agencies." *Journal of Public Policy* 41(4): 653–76.
- Richardson, Mark D. 2019. "Politicization and Expertise: Exit, Effort, and Investment." *The Journal of Politics* 81(3): 878–91.
- Richardson, Mark D. N.d. "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." *The Journal of Politics* 80(1): 303–8.
- 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.
- Rutherford, Amanda, Jeryl Mumpower, Ricardo A. Bello-Gomez, and Marlisa Griffin. 2018. "Understanding Vacancy Time: A Theoretical Framework Informed by Cross-sector Comparison." *Perspectives on Public Management and Governance* 2(1):3-20.
- Risen, Clay. 2021. "Why Are So Many Government Positions Still Vacant?" *New York Times*, August 31, 2021 (<https://www.nytimes.com/2021/08/31/us/politics/biden-cabinet-appointments.html>).
- Stanley, David T., Dean E. Mann, and Jameson W. Doig. 1967. *Men Who Govern: A Biographical Profile of Federal Political Executives*. Washington, DC: Brookings Institution Press.
- Stier, Max. 2021. "It's time to fix the broken Senate confirmation process," *The Hill*, November 15, 2021 (<https://rollcall.com/2021/11/15/its-time-to-fix-the-broken-senate-confirmation-process/>).
- Suleiman, Ezra N. 2003. *Dismantling Democratic States*. Princeton University Press.
- Tenpas, Kathryn Dunn. 2021. "Biden's confirmations progress at the 300-day mark," *Brookings Institution*, November 24, 2021 (<https://www.brookings.edu/blog/fixgov/2021/11/24/bidens-confirmations-progress-at-the-300-day-mark/>).
- 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.
- Wilson, James Q. 1989. *Bureaucracy*. 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.

Wrightman, G. Breck, Sergio Fernandez, and Amanda Rutherford. 2022. "Job Vacancy and Organizational Performance: Are Senior Managers or Street-level Bureaucrats Missed Most?" *Public Administration Review* 82(4): 660-70.

Appendix A. Details of Estimates of Agency Performance

To avoid the potential bias associated with self-reports, we use federal executive survey respondents as expert informants to tell us about agencies they work with (rather than their own agency). At the beginning, each survey instrument asks respondents identify the agencies they work with the most (other than their own). A screenshot from the survey is included in Figure A1 below. The survey subsequently includes questions about these agencies as a partial way of getting around problems inherent in self-reports. Respondents were provided 5 agencies to rate. The list of agencies was populated in the following steps:

1. Agencies the respondent selected as one of the three agencies they work with most (up to 3)
2. If the respondent worked in an executive department or the Executive Office of the President, two randomly selected bureaus from the same department or EOP that were not selected by Step 1. If there were fewer than two bureaus remaining after eliminating bureaus from Step 1, then all remaining bureaus were selected. (The Federal Energy Regulatory Commission was treated as bureau of the Department of Energy and the U.S. Agency for International Development was treated as a bureau of the Department of State.)
3. Remaining slots were filled by randomly selected executive departments and independent agencies not selected in Step 1.

The Java code used to populate the list of agencies would not run if the respondent was using Internet Explorer. Therefore, the following process was used to populate the list of agencies for these respondents:

1. Agencies the respondent chose as one of the three agencies they worked with most (up to 3)
2. Office of Management and Budget (if not selected in Step 1)
3. Office of Personnel Management (if not selected in Step 1)

These respondents would be offered fewer than 5 agencies to evaluate if they provided fewer than 3 agencies they work with or selected the Office of management and Budget or Office of Personnel Management as an agency they worked with.

Figure A1: Question about Which Agencies Respondents Work with Most

Q4. Please select the three agencies you have worked with the most in order of how often you work with them.

First:
[Drop down menu]

Second:
[Drop down menu]

Third:
[Drop down menu]

Performance: To evaluate whether the independent agency design, in fact, contributes to a more expert workforce both surveys asked respondents the following question (Figure A2):

Figure A2: Screen Shot of (Online) Agency Performance Question

Q38. How would you rate the overall performance of the following agencies in carrying out their missions?

| | Not at all effective | | Very effective | | | Don't know |
|------------|----------------------|---|----------------|---|---|-----------------------|
| | 1 | 2 | 3 | 4 | 5 | |
| [Agency 1] | | | [Slider] | | | <input type="radio"/> |
| [Agency 2] | | | [Slider] | | | <input type="radio"/> |
| [Agency 3] | | | [Slider] | | | <input type="radio"/> |
| [Agency 4] | | | [Slider] | | | <input type="radio"/> |
| [Agency 5] | | | [Slider] | | | <input type="radio"/> |

Estimation of Ratings

We estimate latent performance using a Bayesian multi-rater item response model. Let Y be the I x J ratings matrix with element y_{ij} denoting agency I's evaluation by respondent j. We have ratings for 179 agencies (i.e., I=179), 1,397 raters, and 4,555 ratings. After estimation, we retain only agencies with at least 5 ratings, leaving 142 agencies in the final dataset. We rescaled ratings to be

distributed $N(0,1)$ prior to estimating the model to improve model efficiency (see Stan User’s Guide Version 2.28, Section 24.12).

We assume an evaluation of agency i by respondent j in workplace d is a function of the agency’s latent performance: $y_{ij} \sim N(\alpha_{jd} + \beta_{jd}x_i, \sigma^2)$, where x_i is latent performance. We define workplace as the Executive Office of the President, each executive department, and each independent agency. This response model allows each respondent to have a unique mapping (i.e., location and scale) from their perception of performance to the survey scale. The intercept term, α_{jd} , denotes the performance level that respondent j assigns to an agency with performance level $x_i = 0$ on the latent scale. This parameter allows location shifts in the mapping from x_i to y_{ij} across respondents. The parameter β_{jd} allows the relationship between the scale of x_i and y_{ij} to vary across respondents. The σ^2 parameter assumes a common error in respondents’ perception of latent performance. An assumption of common error is necessary to prevent divergent transitions caused by an agency-specific error approaching zero, which causes model parameters to get “stuck” as the chain explores the posterior distribution.

We assign α_{jd} and β_{jd} a hierarchical prior and assume the parameters for raters who share a workplace are drawn from a common distribution. We have at most 5 ratings per rater and often fewer. The hierarchical prior and the resultant partial pooling across raters improves the precision of the rater parameters and allows us to retain raters who rate only 1 or 2 agencies despite the very limited information we have to estimate their mapping from latent space to the survey scale.

Formally, we assume $\alpha_{jd} \sim N(\mu_{d\alpha}, \sigma_{d\alpha}^2)$ and $\beta_{jd} \sim N(\mu_{d\beta}, \sigma_{d\beta}^2)$ where d indexes workplaces and the α and β subscripts denote the corresponding rater parameter. We also estimate models that do not group raters by workplace to examine the implications of this hierarchical model specification. See the relevant section below.

We used informed priors on latent performance to give additional weight to collective perceptions of respondents who reported that an agency is one of the three agencies that they work with most. Formally, we assume $x_i \sim N(\mu_i, \tau_i^2)$, where μ_i is the mean evaluation of agency i by respondents who report working with that agency and τ_i^2 is the variance of those ratings. We use all respondents who reported working with an agency to construct the informed prior, including raters who do not meet the two-rating threshold. If an agency has no informed prior, we set $\mu_i = 0$ which is the center of the scale. If an agency has fewer than 5 informed ratings, we set $\tau_i^2 = 2.17$, the maximum observed variance of informed ratings for agencies in the model. We also estimate a model without the informed prior, assuming $x_i \sim N(0, 2.25)$ for all agencies. We set the variance to 2.25 so that the prior is near but slightly weaker than the weakest prior used to estimate the informed ratings to facilitate evaluation of the effect of the informed priors on posterior distributions of latent performance. The standard deviations of the posterior distributions for agencies with few ratings are sensitive to the choice of prior. The posterior distribution of latent performance was rescaled after estimation to be distributed $N(0, 1)$ to locally identify the scale. The complete model parameterization with informed priors is:

$$\begin{aligned}
y_{ij} &\sim N(\alpha_{jd} + \beta_{jd}x_i, \sigma^2) \\
x_i &\sim N(\mu_i, \tau_i^2) \\
\alpha_j &\sim N(\mu_{d\alpha}, \sigma_{d\alpha}^2) \\
\beta_j &\sim \text{half-}N(\mu_{d\beta}, \sigma_{d\beta}^2) \\
\mu_{d\alpha} &\sim N(0, 25) \\
\mu_{d\beta} &\sim \text{half-}N(0, 25) \\
\sigma^2 &\sim \text{half-}N(0, 25) \\
\sigma_{d\alpha}^2 &\sim \text{half-}N(0, 25) \\
\sigma_{d\beta}^2 &\sim \text{half-}N(0, 25),
\end{aligned}$$

where μ_i is the mean of y_{ij} calculated using only informed ratings of agency i and τ_i^2 is the variance of y_{ij} calculated using only informed ratings of agency i . The naive model is identical, except for the choice of prior on latent performance: $x_i \sim N(0, 2.25)$.

The β_j , $\mu_{d\beta}$, σ^2 , σ_α^2 , and σ_β^2 parameters have a half-normal prior distribution because we gave them a lower bound of 0. For β_j and $\mu_{d\beta}$, this constraint prohibits larger values of latent performance from mapping to lower ratings on the response scale. This constraint also ensures rotational identification of the latent scale across chains. The σ^2 , σ_α^2 , and σ_β^2 parameters are given a lower bound of 0 because variance parameters are non-negative by definition. The α_{jd} and β_{jd} parameters are given non-centered parameterizations. Non-centered parameterizations tend to be more efficient in hierarchical models. See [Stan User's Guide Version 2.30 Section 25.7](#) for discussion of hierarchical models and non-centered parameterizations.

We used `rstan` version 2.21.3 to estimate the model. We ran 5 chains for 4,000 iterations with the first 1,000 used for warm up. This left 15,000 iterations for inference (3,000 per chain). We used overdispersed initial values for the latent trait parameters - the first chain was initialized at -3, the second at -2, the third at 0, the fourth at 2, and the fifth at 3 - to improve the robustness of inference. Other parameters were initialized at random locations using Stan's default settings. There were no divergent transitions and model diagnostics indicated the chains converged and the parameters have sufficient effective sample sizes to make reliable inferences.

Appendix B. Numerical Estimates of Agency Performance

| Acronym | Agency | Perf. Rating | Standard Deviation | Lower Bound | Upper Bound | # Inf. Ratings | # Ratings |
|---------|---------------------------------------|--------------|--------------------|-------------|-------------|----------------|-----------|
| ACF | Admin. for Children & Families | 0.30 | 0.30 | -0.28 | 0.90 | 18 | 24 |
| ARPA-E | Adv. Res. Projects Agency-Energy | 0.39 | 0.85 | -1.24 | 2.18 | 2 | 5 |
| AMS | Agricultural Marketing Service | -0.11 | 0.56 | -1.34 | 0.91 | 6 | 10 |
| ARS | Agricultural Research Service | 0.46 | 0.54 | -0.62 | 1.51 | 9 | 13 |
| USAF | Air Force | 0.88 | 0.26 | 0.38 | 1.42 | 37 | 40 |
| APHIS | Animal & Plant Health Ins. Service | 0.58 | 0.30 | 0.01 | 1.20 | 28 | 34 |
| T | Arms Control & Int. Security Affairs | -0.64 | 0.31 | -1.28 | -0.05 | 7 | 18 |
| ARMY | Army | 0.68 | 0.21 | 0.29 | 1.11 | 48 | 49 |
| BCA | Bureau of Consular Affairs | 0.91 | 0.24 | 0.48 | 1.42 | 10 | 39 |
| BDS | Bureau of Diplomatic Security | 0.33 | 0.21 | -0.07 | 0.76 | 7 | 42 |
| BEA | Bureau of Economic Analysis | 1.07 | 0.57 | 0.03 | 2.28 | 3 | 9 |
| BIA | Bureau of Indian Affairs | -1.43 | 0.38 | -2.23 | -0.74 | 12 | 17 |
| BINLEA | Bur. of Int. Narc. & Law Enf. Affairs | 0.19 | 0.28 | -0.34 | 0.76 | 3 | 25 |
| BLS | Bureau of Labor Statistics | 1.20 | 0.39 | 0.47 | 2.02 | 13 | 14 |
| BLM | Bureau of Land Management | -0.91 | 0.28 | -1.52 | -0.41 | 16 | 26 |
| BOEM | Bureau of Ocean Energy Management | 0.40 | 0.45 | -0.43 | 1.35 | 4 | 7 |
| BOR | Bureau of Reclamation | 0.53 | 0.39 | -0.22 | 1.32 | 7 | 10 |
| BFS | Bureau of the Fiscal Service | 0.22 | 0.26 | -0.26 | 0.77 | 13 | 13 |
| CDC | Centers for Disease Control & Prev. | -0.24 | 0.19 | -0.61 | 0.13 | 70 | 94 |
| CMS | Cen. for Medicare & Medicaid Serv. | 0.37 | 0.26 | -0.13 | 0.89 | 24 | 34 |
| CIA | Central Intelligence Agency | 0.88 | 0.25 | 0.42 | 1.39 | 22 | 24 |
| CIS | Citizenship and Immigration Services | -0.12 | 0.43 | -0.95 | 0.75 | 5 | 10 |
| J | Civ. Security, Dem., & Human Rights | -0.62 | 0.31 | -1.26 | -0.05 | 5 | 23 |
| USCG | Coast Guard | 1.16 | 0.29 | 0.63 | 1.78 | 25 | 32 |
| CFPB | Consumer Financial Protection Bur. | -0.65 | 0.43 | -1.59 | 0.12 | 8 | 9 |
| CEA | Council of Economic Advisers | -1.82 | 0.48 | -2.87 | -0.96 | 6 | 9 |
| CEQ | Council on Environmental Quality | -0.43 | 0.53 | -1.54 | 0.54 | 11 | 11 |
| CBP | Customs and Border Protection | -0.33 | 0.31 | -0.95 | 0.30 | 16 | 21 |
| CISA | Cyber. & Infrastructure Sec. Agency | -0.91 | 0.29 | -1.50 | -0.37 | 21 | 27 |
| DARPA | Def. Adv. Research Projects Agency | 0.92 | 0.40 | 0.14 | 1.74 | 9 | 14 |
| DCA | Defense Commissary Agency | 0.20 | 0.85 | -1.48 | 1.87 | 1 | 5 |
| DCAA | Defense Contract Audit Agency | 0.26 | 0.78 | -1.26 | 1.84 | 1 | 7 |
| DCMA | Def. Contract Management Agency | 0.80 | 0.42 | -0.02 | 1.67 | 6 | 10 |
| DFAS | Def. Finance & Accounting Service | 0.04 | 0.45 | -0.85 | 0.91 | 8 | 15 |
| DHA | Defense Health Agency | -0.10 | 0.35 | -0.80 | 0.60 | 12 | 15 |
| DLA | Defense Logistics Agency | 0.84 | 0.29 | 0.29 | 1.43 | 10 | 15 |
| USDA | Department of Agriculture | -0.10 | 0.24 | -0.58 | 0.36 | 61 | 61 |
| COM | Department of Commerce | -0.67 | 0.20 | -1.09 | -0.28 | 58 | 62 |
| DOD | Department of Defense | 0.57 | 0.14 | 0.31 | 0.86 | 173 | 175 |

| | | | | | | | |
|-------|-------------------------------------|-------|------|-------|-------|-----|-----|
| DOED | Department of Education | -1.32 | 0.31 | -1.99 | -0.74 | 25 | 26 |
| DOE | Department of Energy | 0.15 | 0.23 | -0.30 | 0.62 | 49 | 52 |
| HHS | Dep. of Health & Human Services | -0.50 | 0.16 | -0.82 | -0.20 | 135 | 137 |
| DHS | Department of Homeland Security | -0.79 | 0.19 | -1.18 | -0.43 | 83 | 87 |
| HUD | Dep. of Housing & Urban Dev. | -1.36 | 0.30 | -1.99 | -0.82 | 36 | 36 |
| DOJ | Department of Justice | -0.32 | 0.16 | -0.63 | -0.01 | 102 | 109 |
| DOL | Department of Labor | -0.24 | 0.29 | -0.81 | 0.34 | 28 | 30 |
| STAT | Department of State | 0.04 | 0.15 | -0.25 | 0.33 | 131 | 136 |
| INT | Department of the Interior | -0.41 | 0.21 | -0.84 | -0.02 | 45 | 47 |
| TREAS | Department of the Treasury | 0.22 | 0.24 | -0.23 | 0.72 | 36 | 37 |
| DOT | Department of Transportation | -0.12 | 0.23 | -0.59 | 0.33 | 46 | 50 |
| DVA | Department of Veterans Affairs | -0.25 | 0.18 | -0.60 | 0.11 | 58 | 60 |
| DEA | Drug Enforcement Administration | -0.34 | 0.49 | -1.31 | 0.59 | 7 | 8 |
| EDA | Economic Development Admin. | -0.67 | 0.47 | -1.65 | 0.20 | 4 | 8 |
| E | Econ. Growth, Energy, & the Env. | -0.28 | 0.23 | -0.75 | 0.18 | 23 | 41 |
| ERS | Economic Research Service | 0.12 | 0.76 | -1.38 | 1.65 | 5 | 10 |
| ETA | Employment & Training Admin. | 0.34 | 0.43 | -0.47 | 1.21 | 11 | 12 |
| EPA | Environmental Protection Agency | -1.43 | 0.27 | -1.98 | -0.94 | 51 | 51 |
| EEOC | Equal Emp. Opportunity Commission | -0.46 | 0.31 | -1.09 | 0.11 | 12 | 14 |
| EOUSA | Executive Office for U.S. Attorneys | 1.13 | 0.39 | 0.40 | 1.92 | 10 | 14 |
| FSA | Farm Service Agency | -0.14 | 0.51 | -1.16 | 0.87 | 16 | 23 |
| FAA | Federal Aviation Administration | 0.80 | 0.33 | 0.18 | 1.46 | 16 | 22 |
| FBI | Federal Bureau of Investigation | 0.26 | 0.27 | -0.27 | 0.82 | 25 | 28 |
| FDIC | Federal Deposit Insurance Corp. | 0.96 | 0.28 | 0.43 | 1.52 | 7 | 10 |
| FEMA | Federal Emergency Mgt Agency | -0.26 | 0.29 | -0.84 | 0.30 | 29 | 34 |
| FHWA | Federal Highway Administration | 0.26 | 0.33 | -0.38 | 0.95 | 9 | 15 |
| HOU | Fed. Housing Admin./Ofc of Housing | -0.12 | 0.33 | -0.76 | 0.53 | 9 | 19 |
| FLRA | Federal Labor Relations Authority | -0.84 | 0.37 | -1.63 | -0.18 | 3 | 6 |
| FMCSA | Fed Motor Carrier Safety Admin. | -0.45 | 0.59 | -1.66 | 0.70 | 3 | 6 |
| FED | Federal Reserve | 0.95 | 0.36 | 0.28 | 1.70 | 11 | 12 |
| FTC | Federal Trade Commission | 0.75 | 0.50 | -0.11 | 1.84 | 8 | 8 |
| FTA | Federal Transit Administration | -0.07 | 0.40 | -0.85 | 0.71 | 8 | 13 |
| FDA | Food and Drug Administration | 0.68 | 0.21 | 0.29 | 1.12 | 50 | 65 |
| FNS | Food and Nutrition Service | -0.40 | 0.60 | -1.56 | 0.78 | 6 | 7 |
| FSIS | Food Safety and Inspection Service | 0.33 | 0.35 | -0.34 | 1.03 | 18 | 22 |
| FS | Forest Service | 0.37 | 0.36 | -0.32 | 1.09 | 16 | 21 |
| GSA | General Services Administration | -1.48 | 0.27 | -2.04 | -0.98 | 56 | 61 |
| GNMA | Govt National Mortgage Association | 0.93 | 0.51 | 0.10 | 2.12 | 2 | 9 |
| HRSA | Health Resources & Services Admin. | 0.27 | 0.32 | -0.33 | 0.91 | 15 | 20 |
| ICE | Immigration & Customs Enforcement | -1.06 | 0.42 | -1.91 | -0.28 | 13 | 15 |
| HIS | Indian Health Service | -1.07 | 0.40 | -1.91 | -0.32 | 7 | 13 |

| | | | | | | | |
|-------|--|-------|------|-------|-------|-----|-----|
| IES | Institute of Education Sciences | 0.47 | 0.64 | -0.72 | 1.79 | 5 | 6 |
| IRS | Internal Revenue Service | -0.06 | 0.30 | -0.63 | 0.55 | 19 | 21 |
| ITA | International Trade Administration | 0.30 | 0.30 | -0.29 | 0.90 | 18 | 24 |
| JCS | Joint Chiefs of Staff | 0.33 | 0.36 | -0.35 | 1.06 | 14 | 17 |
| M | Management | 0.24 | 0.23 | -0.20 | 0.69 | 10 | 43 |
| MSPB | Merit Systems Protection Board | -1.15 | 0.56 | -2.41 | -0.23 | 4 | 6 |
| MCC | Millennium Challenge Corporation | 0.52 | 0.50 | -0.41 | 1.54 | 6 | 6 |
| MSHA | Mine Safety and Health Administration | 0.50 | 0.54 | -0.51 | 1.65 | 4 | 6 |
| NASA | National Aeronautics & Space Admin. | 1.10 | 0.31 | 0.54 | 1.73 | 22 | 28 |
| NASS | National Agricultural Statistics Service | 1.02 | 0.47 | 0.12 | 1.99 | 10 | 17 |
| NARA | National Archives & Records Admin. | -0.39 | 0.75 | -2.01 | 1.01 | 4 | 5 |
| NCA | National Cemetery Administration | 1.40 | 0.54 | 0.45 | 2.56 | 0 | 8 |
| NEH | Nat. Endowment for the Humanities | -0.32 | 0.63 | -1.58 | 0.92 | 5 | 5 |
| NHTSA | Nat. Highway Traffic Safety Admin. | 0.03 | 0.39 | -0.74 | 0.80 | 9 | 13 |
| NIST | National Institute of Standards & Tech | 0.72 | 0.31 | 0.14 | 1.37 | 15 | 27 |
| NIH | National Institutes of Health | 1.37 | 0.24 | 0.92 | 1.87 | 59 | 74 |
| NNSA | National Nuclear Security Admi. | 0.45 | 0.38 | -0.30 | 1.22 | 15 | 17 |
| NOAA | Nat. Oceanic & Atmospheric Admin. | 0.67 | 0.25 | 0.20 | 1.18 | 37 | 48 |
| NPS | National Park Service | 0.68 | 0.24 | 0.25 | 1.19 | 17 | 28 |
| NSF | National Science Foundation | 1.15 | 0.28 | 0.62 | 1.72 | 37 | 43 |
| NSA | National Security Agency | 1.41 | 0.30 | 0.83 | 2.03 | 7 | 7 |
| NSS | National Security Staff | -1.05 | 0.20 | -1.47 | -0.69 | 96 | 100 |
| NTSB | National Transportation Safety Board | -1.49 | 0.57 | -2.68 | -0.42 | 7 | 10 |
| NRCS | Natural Resources Conservation Ser. | 0.91 | 0.39 | 0.18 | 1.69 | 18 | 25 |
| NAVY | Navy | 0.96 | 0.24 | 0.51 | 1.45 | 42 | 45 |
| NRC | Nuclear Regulatory Commission | 0.47 | 0.77 | -1.00 | 2.03 | 5 | 5 |
| OSHA | Occupational Safety & Health Admin. | -0.32 | 0.58 | -1.55 | 0.75 | 8 | 9 |
| OE | Office of Electricity | 0.35 | 0.63 | -0.99 | 1.55 | 4 | 6 |
| OESE | Office of Elementary & Secondary Ed. | -1.00 | 0.59 | -2.19 | 0.18 | 3 | 7 |
| EERE | Ofc of Energy Eff. & Renewable Ener. | 1.00 | 0.45 | 0.17 | 1.93 | 12 | 14 |
| OEM | Office of Environmental Management | -1.90 | 0.51 | -2.98 | -0.99 | 8 | 8 |
| OFSA | Office of Federal Student Aid | -0.59 | 0.58 | -1.73 | 0.55 | 4 | 8 |
| OGE | Office of Government Ethics | 0.18 | 0.55 | -0.88 | 1.27 | 6 | 7 |
| OLC | Office of Legal Counsel | 1.17 | 0.38 | 0.45 | 1.93 | 12 | 15 |
| OMB | Office of Management and Budget | -1.04 | 0.14 | -1.33 | -0.78 | 180 | 415 |
| ONDCP | Office of National Drug Control Policy | -1.18 | 0.64 | -2.50 | 0.03 | 5 | 6 |
| ONE | Office of Nuclear Energy | -1.41 | 0.79 | -3.05 | 0.12 | 5 | 6 |
| OPM | Office of Personnel Management | -1.96 | 0.21 | -2.39 | -1.57 | 57 | 318 |
| SC | Office of Science | 1.13 | 0.34 | 0.49 | 1.84 | 19 | 20 |
| OSTP | Office of Science & Technology Policy | -1.23 | 0.30 | -1.85 | -0.67 | 36 | 40 |
| OSC | Office of Special Counsel | 0.53 | 0.64 | -0.61 | 1.94 | 4 | 7 |

| | | | | | | | |
|-------|---|-------|------|-------|-------|----|----|
| OCC | Off. of the Comptroller of Currency | 0.50 | 0.35 | -0.15 | 1.23 | 8 | 8 |
| ODNI | Office of the Dir. of Nat. Intelligence | 0.13 | 0.28 | -0.41 | 0.70 | 11 | 13 |
| OSD | Office of the Secretary of Defense | -0.07 | 0.19 | -0.45 | 0.30 | 60 | 62 |
| USTR | Office of the U.S. Trade Representative | -0.11 | 0.35 | -0.82 | 0.57 | 26 | 26 |
| OWCP | Ofc of Workers' Compensation Prog. | -0.87 | 0.85 | -2.60 | 0.82 | 4 | 6 |
| P | Political Affairs | 0.45 | 0.26 | -0.04 | 0.97 | 13 | 26 |
| R | Public Diplomacy and Public Affairs | 0.09 | 0.22 | -0.34 | 0.54 | 12 | 39 |
| RD | Rural Development | -0.89 | 0.45 | -1.79 | -0.02 | 20 | 21 |
| USSS | Secret Service | 0.67 | 0.54 | -0.31 | 1.84 | 4 | 8 |
| SBA | Small Business Administration | 0.04 | 0.37 | -0.69 | 0.76 | 18 | 18 |
| SSA | Social Security Administration | -0.44 | 0.34 | -1.06 | 0.26 | 14 | 18 |
| TSA | Transportation Security Administration | -0.68 | 0.45 | -1.58 | 0.18 | 8 | 11 |
| CEN | U.S. Census Bureau | 0.29 | 0.26 | -0.19 | 0.84 | 15 | 26 |
| FWS | U.S. Fish and Wildlife Service | 0.43 | 0.22 | 0.02 | 0.87 | 31 | 39 |
| USGS | U.S. Geological Survey | 0.47 | 0.26 | 0.00 | 1.00 | 27 | 29 |
| PTO | U.S. Patent and Trademark Office | 0.96 | 0.63 | -0.19 | 2.29 | 3 | 9 |
| USAID | U.S. Agency for Int. Development | 0.00 | 0.17 | -0.34 | 0.34 | 56 | 74 |
| USPS | United States Postal Service | -0.09 | 0.71 | -1.52 | 1.30 | 4 | 5 |
| VBA | Veterans' Benefits Administration | -1.07 | 0.39 | -1.87 | -0.34 | 13 | 18 |
| VHA | Veterans' Health Administration | 0.04 | 0.36 | -0.65 | 0.75 | 13 | 17 |
| WHD | Wage and Hour Division | -1.37 | 0.70 | -2.82 | -0.06 | 5 | 6 |

Appendix C. Comparison of Hierarchical and Non-hierarchical Models of Agency Performance, 2020

We have also estimated models assuming no hierarchical structure and compared these estimates to the hierarchical estimates. The estimates are quite similar.

Non-hierarchical model: Estimation details

We assume an evaluation of agency i by respondent j is a function of the agency's latent performance: $y_{ij} \sim N(\alpha_j^* + \beta_j x_i, \sigma^2)$, where x_i is latent performance. This response model allows each respondent to have a unique mapping (i.e., location and scale) from their perception of performance to the survey scale. The intercept term, α_j^* , denotes the performance level that respondent j assigns to an agency with performance level $x_i = 0$ on the latent scale. This parameter allows location shifts in the mapping from x_i to y_{ij} across respondents. The parameter β_j allows the relationship between the scale of x_i and y_{ij} to vary across respondents. The σ^2 parameter assumes a common error in respondents perception of latent performance. An assumption of common error is necessary to prevent divergent transitions caused by an agency-specific error approaching zero, which causes model parameters to get “stuck” as the chain explores the posterior distribution.

We used informed priors on latent performance to give additional weight to collective perceptions of respondents who reported that an agency is one of the three agencies that they work with most. Formally, we assume $x_i \sim N(\mu_i, \tau_i^2)$, where μ_i is the mean evaluation of agency i by respondents who report working with that agency and τ_i^2 is the variance of those ratings. We use all respondents who reported working with an agency to construct the informed prior, including raters who do not meet the two-rating threshold. If an agency has no informed prior, we set $\mu_i = 0$ which is the center of the scale. If an agency has fewer than 5 informed ratings, we set $\tau_i^2 = 2.17$, the maximum observed variance of informed ratings for agencies in the model. We also estimate a model without the informed prior, assuming $x_i \sim N(0, 2.25)$ for all agencies. We set the variance to 2.25

so that the prior is near but slightly weaker than the weakest prior used to estimate the informed ratings to facilitate evaluation of the effect of the informed priors on posterior distributions of latent performance. The standard deviations of the posterior distributions for agencies with few ratings are sensitive to the choice of prior. The posterior distribution of latent performance was rescaled after estimation to be distributed $N(0, 1)$ to locally identify the scale. The complete model parameterization with informed priors is:

$$\begin{aligned}
y_{ij} &\sim N((\alpha_j + \theta_\alpha) + \beta_j x_i, \sigma^2) \\
x_i &\sim N(\mu_i, \tau_i^2) \\
\alpha_j &\sim N(0, \sigma_\alpha^2) \\
\beta_j &\sim \text{half-}N(0, \sigma_\beta^2) \\
\theta_\alpha &\sim N(0, 25) \\
\sigma^2 &\sim \text{half-}N(0, 25) \\
\sigma_\alpha^2 &\sim \text{half-}N(0, 25) \\
\sigma_\beta^2 &\sim \text{half-}N(0, 25),
\end{aligned}$$

where μ_i is the mean of y_{ij} calculated using only informed ratings of agency i and τ_i^2 is the variance of y_{ij} calculated using only informed ratings of agency i . The naive model is identical, except for the choice of prior on latent performance: $x_i \sim N(0, 2.25)$.

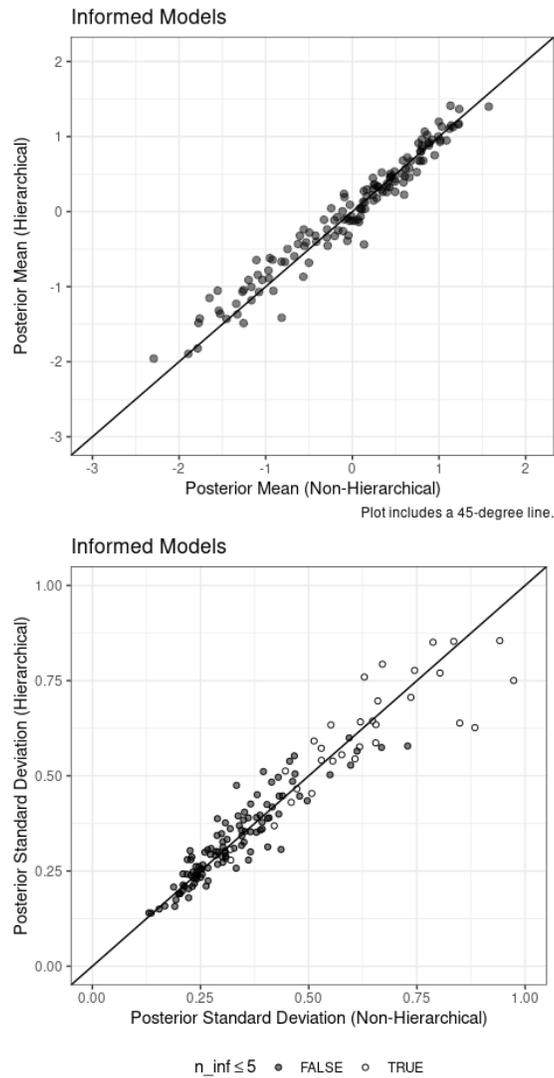
The β_j , σ^2 , σ_α^2 , and σ_β^2 parameters have a half-normal prior distribution because we gave them a lower bound of 0. For β_j , this constraint prohibits larger values of latent performance from mapping to lower ratings on the response scale. This constraint also ensures rotational identification of the latent scale across chains. The σ^2 , σ_α^2 , and σ_β^2 parameters are given a lower bound of 0 because variance parameters are non-negative by definition. The α_j parameter is given a non-centered parameterization, with grand mean θ_α . Non-centered parameterizations tend to be more efficient in hierarchical models. In the initial exposition of the model above, we used $\alpha_j^* = \alpha_j + \theta_\alpha$ to simplify presentation. See [Stan User's Guide Version 2.30 Section 1.11](#) for discussion of both the constraint on β_j and non-centered parameterization of α_j for item response models.

We used `rstan` version 2.21.3 to estimate the model. We ran 4 chains for 4,000 iterations with the first 1,000 used for warm up. This left 12,000 iterations for inference (3,000 per chain). We used overdispersed initial values for the latent trait parameters - the first chain was initialized at -3, the second at -2, the third at 2, and the fourth at 3 - to improve the robustness of inference. Other parameters were initialized at random locations using Stan's default settings. There were no divergent transitions and model diagnostics indicated the chains converged and the parameters have sufficient effective samples sizes to make reliable inferences.

Comparing Hierarchical and Non-hierarchical Models

In Figure C1 we compare the posterior means and standard deviations for the hierarchical and non-hierarchical models. Overall, there are not large differences in posterior means. The posterior means are highly correlated at 0.98. The performance ratings on the lower end of the scale, namely below 0, tend to be closer to zero in the hierarchical model than the non-hierarchical model as evidenced by most of the posterior means above the 45-degree line in that portion of the scale. In other words, the lower performing agencies get higher scores in the hierarchical models. Similarly, the standard deviations are similar across models with correlations of 0.94. The hierarchical model yields smaller standard deviations for 51% of agencies in the informed models.

Figure C1. Comparison of Posterior Means and Standard Deviations for Hierarchical and Non-hierarchical Models



Comparing Rater Parameters

Figure C2 compares the posterior means and standard deviations for the rater parameters between hierarchical and non-hierarchical models. The top row shows that the slope parameters (i.e., β 's) tend to be larger in the hierarchical models. The posterior means of the slope parameters are larger in the hierarchical model for 97% of raters and the median absolute difference is 0.23. Therefore, there is a stronger relationship between the latent performance and the survey scale in the hierarchical model. The standard deviations of the posterior distributions of slope parameters

also tend to be larger in hierarchical model. The standard deviations are larger for 80% of raters and the median absolute difference is 0.058.

The top row of Figure C3 shows that the extreme intercept parameters (i.e., α 's) tend to be closer to zero in the hierarchical models as evidenced by ratings below zero tending to be below the 45-degree line and ratings above zero tending to be above the line. Recall the raw ratings are rescaled to be distributed $N(0,1)$ before the Bayesian models are estimated, meaning 0 is the mean of the observed ratings. Therefore, raters with extreme intercepts would assign an agency with latent performance of zero a value closer to zero in the hierarchical model than the non-hierarchical model. In other words, these extreme raters now evaluate performance more similarly to other raters, as we would expect if the hierarchical model makes raters' parameters more similar within groups. Similar to the slope parameters, the standard deviations are larger for 51% of the raters in the hierarchical models than the nonhierarchical informed models.

Overall, the hierarchical model has two effects on rater parameters. First, there is a stronger relationship between latent performance and assigned performance in the hierarchical models (due to the larger slope parameters). Second, raters with extreme thresholds use the scale more similarly to other raters (due to the more similar intercept parameters).

Figure C2. Comparison of Posterior Means and Standard Deviations for Rater Slope Parameters for Hierarchical and Non-hierarchical Models

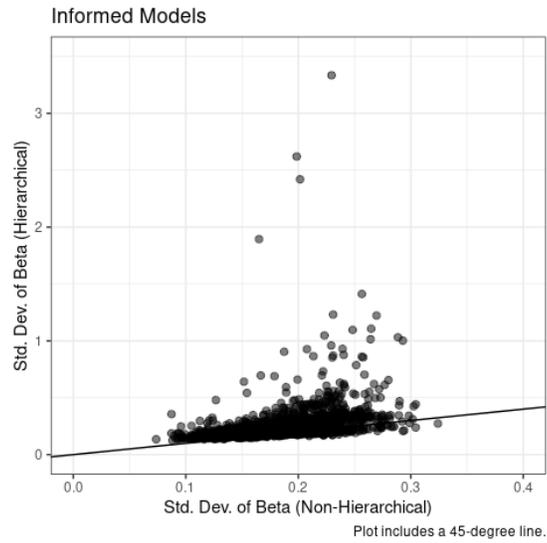
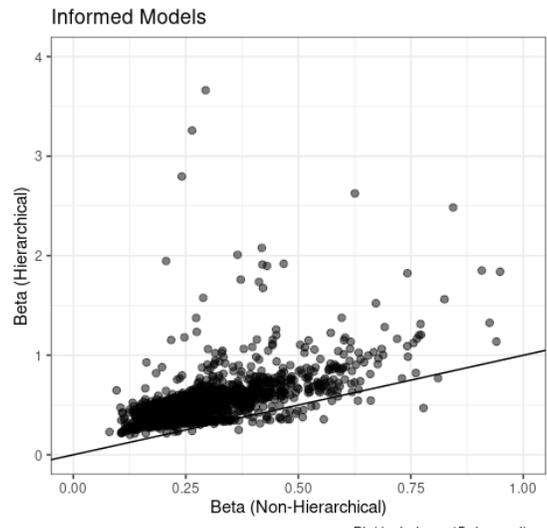
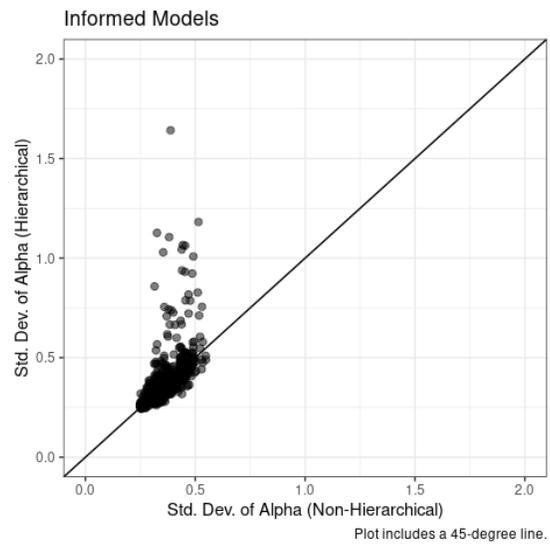
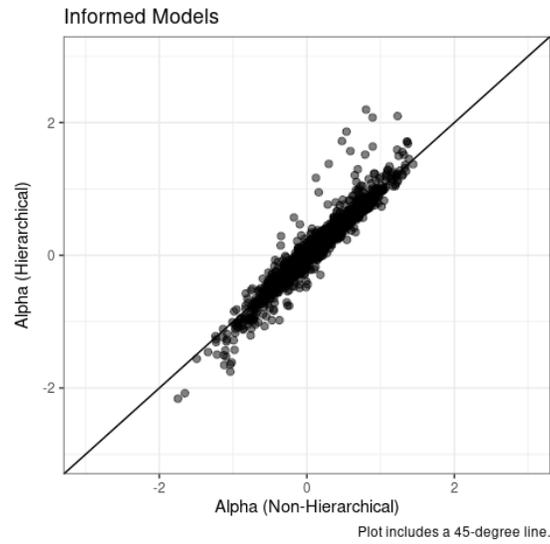


Figure C3. Comparison of Posterior Means and Standard Deviations for Rater Intercept Parameters for Hierarchical and Non-hierarchical Models



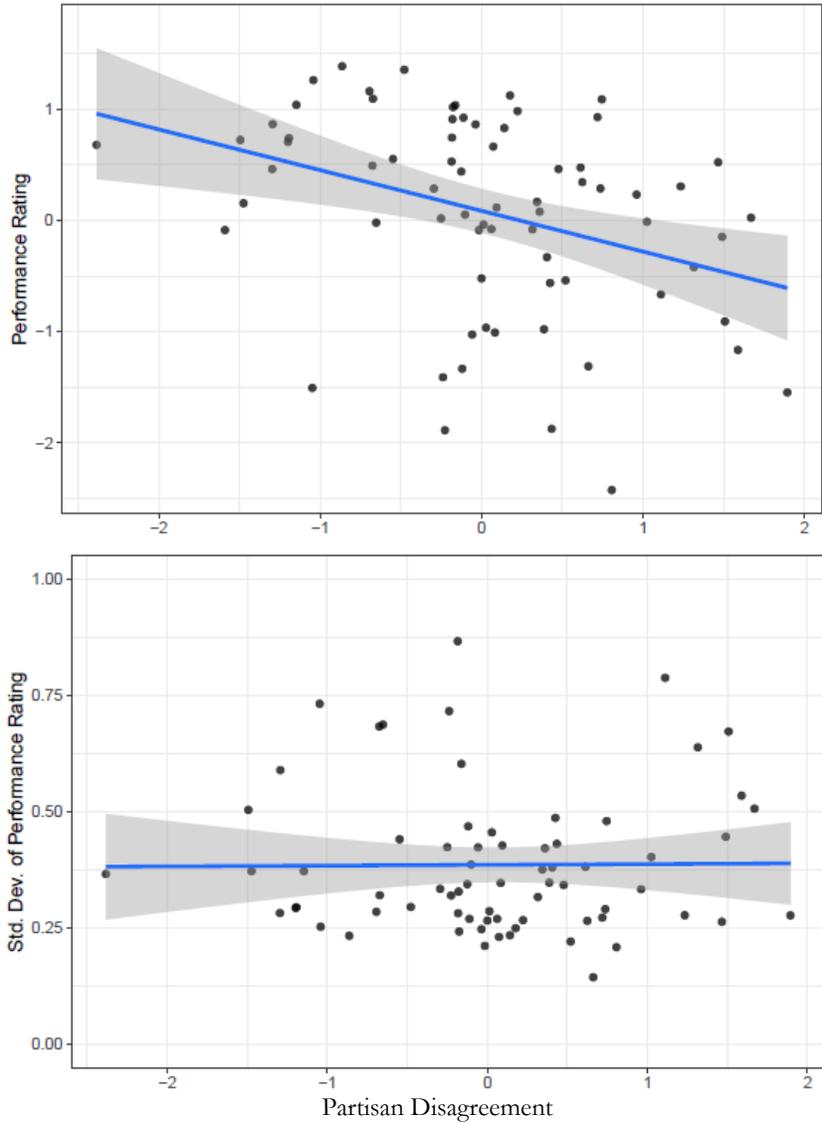
Appendix D. Partisan Disagreement and Agency Performance Evaluations

As a way of evaluating the impact of party on performance evaluations we compare our agency performance ratings to a measure of partisan disagreement to see whether agencies that are subject to regular partisan disagreement are rated differently by Democrats and Republicans. In the 2020 survey we asked respondents “How often do Republicans and Democrats in Washington disagree over what [your agency] should do?” [Never, Rarely, Sometimes, Often, Always, Don’t know] We regress the performance ratings on the normalized agency average responses to this question about partisan disagreement and graph these estimates in Figure 2.¹⁸ Interestingly, the more an agency is subject to partisan disagreement, the lower its rating overall. When Democrats and Republicans disagree, they both seem to think the agency is performing poorly. So, rather than Democrats rating agencies high and Republicans rating agencies low or vice versa, it looks like the raters agree on performance.

Importantly, this disagreement also does not appear to decrease the *uncertainty* of the estimates. If Democrats and Republicans saw performance differently when they disagreed, this should increase the uncertainty of the estimates. For example, if Republicans and Democrats always gave 1s and 5s, respectively, when they disagreed and always gave similar scores when they agreed, we should see less precision in the estimates when they disagree. The bottom panel in Figure D1 regresses the standard deviation of the performance estimates on partisan disagreement and shows no relationship. This is evidence that partisan disagreement does not lead to divergent evaluations, at least for most agencies.

¹⁸ We rescaled the agency average responses to the partisan disagreement question to be distributed $N(0,1)$.

Figure D1. Estimates of Agency Performance Ratings based Upon Partisan Disagreement About Agency Work, 2020



Note: The top panel includes a scatter plot and fitted line of agency performance ratings (y-axis) and agency average responses to the following question about partisan disagreement: “How often do Republicans and Democrats in Washington disagree over what [your agency] should do?” [Never, Rarely, Sometimes, Often, Always, Don’t know] (x-axis). The bottom panel includes a scatter plot and fitted line of the standard deviation of agency performance ratings (y-axis) and agency average responses to the partisan disagreement question (x-axis).

Appendix E. Models of Performance Estimated with Democrat Performance Ratings and Republican Performance Ratings Separately

| Agency Performance (-2.5, 2.5) | Democrats | Republicans | Democrats | Republicans |
|---|---------------------|---------------------|---------------------|---------------------|
| Months Vacant (0-42) | -0.027 [0.007]** | -0.026 [0.009]** | -0.007 [0.008] | -0.006 [0.011] |
| <i>Agency Controls</i> | | | | |
| EOP (0,1) | -0.485 [0.106]** | 0.24 [0.128]* | -0.869 [0.134]** | -0.145 [0.167] |
| Cabinet Department (0,1) | 0.724 [0.130]** | 0.915 [0.119]** | 0.449 [0.178]** | 0.583 [0.135]** |
| Whole Department (0,1) | -1.239 [0.232]** | -0.793 [0.262]** | -0.653 [0.303]** | -0.12 [0.323] |
| Workforce Skill -- Obama Administration | | | 0.519 [0.226]** | 0.627 [0.182]** |
| Constant | 0.146 [0.068]** | -0.174 [0.082]* | -0.06 [0.075] | -0.372 [0.102]** |
| R ² | | 0.35 | 0.25 | 0.54 |
| N | | 39 | 39 | 37 |

Note: ** significant at the 0.05 level; * significant at the 0.10 level in two-tailed tests. Data: *Survey on the Future of Government Service*, 2020. Dependent variable is agency performance estimate based upon aggregated responses to question: “How would you rate the overall performance of the following agencies in carrying out their missions?” (1-5). Models estimated with Ordinary Least Squares and standard errors are clustered by department.