CHAPTER 17

Government Analytics Using Data on Task and Project Completion

Imran Rasul, Daniel Rogger, Martin Williams, and Eleanor Florence Woodhouse

SUMMARY

Much government work consists of the completion of tasks, from creating major reports to undertaking training programs and procuring and building infrastructure. This chapter surveys a range of methods for measuring and analyzing task completion as a measure of the performance of government organizations, giving examples of where these methods have been implemented in practice. We discuss the strengths and limitations of each approach from the perspectives both of practice and research. While no single measure of task completion provides a holistic performance metric, when used appropriately, such measures can provide a powerful set of insights for analysts and managers alike.

ANALYTICS IN PRACTICE

- Much government activity can be conceived as discrete tasks: bounded pieces of work with definite outputs. Public sector planning is often organized around the achievement of specific thresholds; the completion of planning, strategy, or budgetary documents; or the delivery of infrastructure projects. *Task completion* is a useful conception of government activity because it allows analysts to assess public performance in a standardized way across organizations and types of activity.
- Assessing government performance based solely on the passing of legislation or the delivery of frontline services misses a substantial component of government work. Using a task completion approach pushes analysts to better encapsulate the breadth of work undertaken by public administration across government. It thus pushes analysts to engage with the full set of government tasks.

Imran Rasul is a professor in the Department of Economics, University College London. Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Martin Williams is an associate professor in the Blavatnik School of Government, University of Oxford. Eleanor Florence Woodhouse is an assistant professor in the Department of Political Science and School of Public Policy, University College London.

- A task completion approach also allows for the investigation of which units and organizations are most likely to initiate, make progress on, and complete tasks. Though not a full picture of government work—it is complementary to the analysis of process quality or sector-specific measures of quality, for example—it allows for a rigorous approach to comparisons frequently made implicitly in budgetary and management decisions.
- Collecting data across projects on determinants of progress, such as overruns, and matching them to input data, such as budget disbursements, allows for a coherent investigation of the mechanisms driving task progress across government or within specific settings.
- Attempting to assess task completion in a consistent way across government is complicated by the fact that tasks vary in nature, size, and complexity. By collecting data on these features of a task, analysts can go some way toward alleviating concerns over the variability of the tasks being considered. For example, analysis can be undertaken within particular types of task or size, and complexity can be conditioned on in any analysis. An important distinction in the existing literature is how to integrate the analysis of tasks related to the creation of physical infrastructure and tasks related to administration.

INTRODUCTION

A fundamental question for government scholars and practitioners alike is whether governments are performing their functions well. What these functions are and what performing "well" means in practice are complex issues in the public sector, given the diverse tasks undertaken and their often indeterminate nature. Despite the importance of these questions, there is little consensus as to how to define government effectiveness in a coherent way across the public service or how to measure it within a unified approach across governments' diverse task environments (Rainey 2009; Talbot 2010). Such considerations have practical importance because government entities, such as political oversight or central budget authorities, frequently have to make implicit comparisons between the relative functioning of public agencies. For example, when drawing up a budget, public sector managers must make some comparison of the likely use of funds across units and whether these funds will eventually result in the intended outputs of those units, however varied the tasks are in scope. From an analytical perspective, the more comprehensive a measure of government functioning, the greater the capacity of analytical methods to draw insights from the best-performing parts of government.

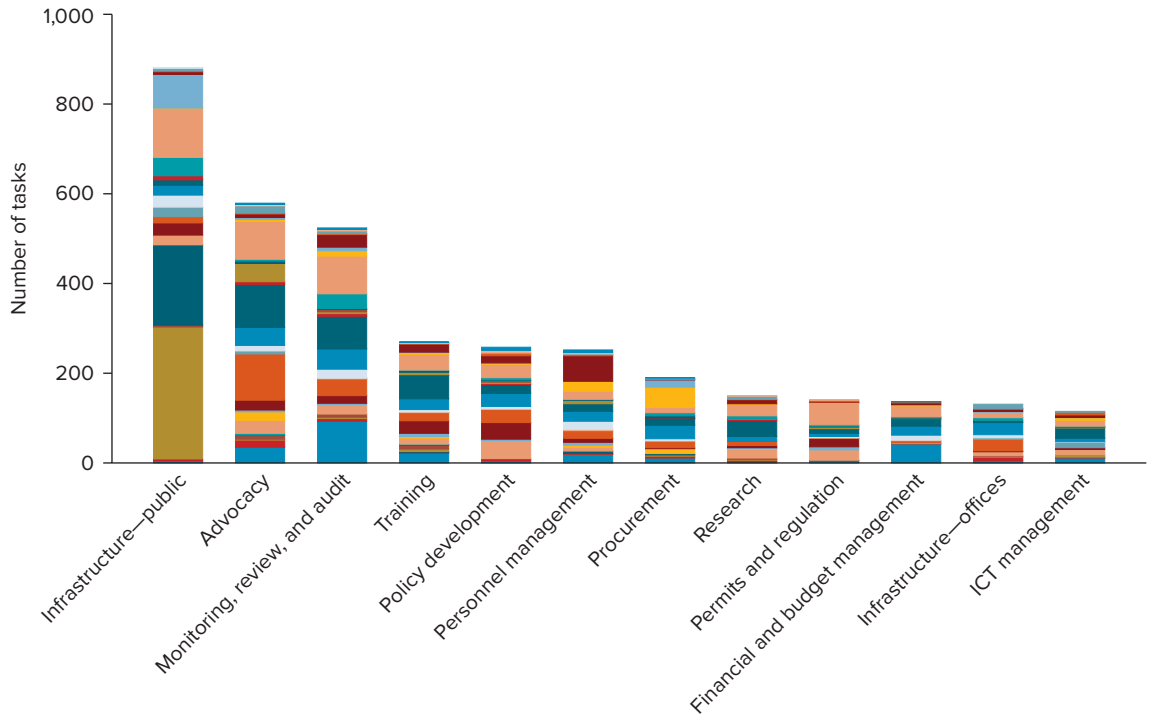
Much government activity can be conceived as discrete *tasks*: bounded pieces of work with definite outputs. Public sector planning is often organized around the achievement of specific thresholds; the completion of planning, strategy, or budgetary documents; or the delivery of projects. Government *projects* are also often conceived as bounded activities with definite outputs but frequently encompass multiple tasks within a wider conception of completion. *Task completion* (or project completion) is thus a useful conception of government activity—however that activity is conceived—because it allows analysts to assess public performance in a standardized way across organizations and types of activity. This kind of assessment contrasts with continuous regulatory monitoring, the assessment of the stability of citizens' access to frontline services, and the equity of activities related to redistribution, which are better understood as the evaluation of ongoing processes. In this chapter, we propose a way to leverage data on task completion to assess the government's effectiveness across its diverse task environments and learn from related analysis. We argue that by utilizing a unified framework for task completion, analysts can assess whether a government or government agency "does well what it is supposed to do, whether people. . . work hard and well, whether the actions and procedures of the agency and its members help achieve its mission, and, in the end, whether it actually achieves its mission" (Lee and Whitford 2009, 251; paraphrasing Rainey and Steinbauer 1999).

Task comparison is useful for three core reasons. First, task completion is a concept that can be applied across much government work, thus allowing for a broad consideration of government functioning.

We believe that by working with a task completion framework, analysts can gain a fuller and more accurate picture of the functions of government that reflects the full range of government activities—from human resource management to policy definition, infrastructure planning and implementation, service delivery, and audit and evaluation. We know little about the full distribution of tasks that public administrators undertake. As shown in figure 17.1, the few studies that do apply a task completion framework find that administrators undertake a vast range of activities—from advocacy to auditing and monitoring to planning—that go well beyond infrastructure and service delivery, the activities that are usually considered in the academic literature.

Figure 17.1 displays the frequency of the most prevalent tasks undertaken by Ghanaian public officials in their daily duties. Infrastructure provision for the public (rather than upgrading government facilities) is the most common activity, partly motivating our particular attention to it in this chapter. However, the figure indicates the broad diversity of tasks undertaken by the public service. The distinct colors in each bar of the histogram indicate different organizations undertaking that type of task. Thus, it is clear that each type of task is undertaken by many different organizations. Second, a task completion framework pushes analysts to think carefully about the characteristics of each of the tasks assessed. We define projects above as collections of tasks; an obvious question is how to apply boundaries to tasks or projects uniformly across government. There is very limited research on the characteristics of the tasks undertaken by public administrators and how to assess whether they are being undertaken adequately. The task completion framework pushes analysts to think in detail about the activities that administrators engage in and how successfully they do so. That is to say, they must think not only about whether a bridge is completed but what the full conception of the bridge project is, whether the bridge was of a complex design that was hard to implement, whether the quality of the implementation of the bridge is adequate, whether it was completed within a reasonable time frame given the complexity of the project, and so on.

FIGURE 17.1 Task Types across Organizations



Source: Rasul, Rogger, and Williams 2021.
Note: The task type classification refers to the primary classification for each output. Each color in a column represents an organization implementing tasks of that type, but the same color across columns may represent multiple organizations. Figures represent all 30 organizations with coded task data. ICT = information and communication technology.

Third, by creating comparators from across government, an integrated measurement approach yields analytical benefits that more than make up for the losses from abstraction for many types of analysis. Analysts can investigate the determinants of successful task completion from a large sample, with varying management environments, buffeted by differential shocks, and on which a greater range of statistical methods can be effectively applied. The task completion framework has the advantage of capturing a wide range of activities and being comparable across departments. Thus, the framework can pool task types to allow analysts to draw conclusions about government effectiveness more broadly, rather than, for example, allowing only for inferences about a specific type of task (for example, the delivery of a specific service, such as passport processing times, versus a range of government tasks that are indicative of administrative effectiveness more broadly). By leveraging these data—on the nature, size, and complexity of tasks—analyses can be undertaken within particular types of tasks—for example, distinguishing between the completion of physical and nonphysical outputs. Taking the analysis a step further, having measures of task completion that are comparable across teams or organizations can enable researchers to identify the determinants of task completion—although this entails its own methodological challenges and is beyond the scope of this measurement-focused chapter.

Much of the existing literature that seeks to describe how well governments perform their functions focuses on upstream steps in the public sector production process (such as budgetary inputs or processes), which are useful for management (but not so relevant to the public) (Andrews et al. 2005; Lewis 2007; Nistotskaya and Cingolani 2016; Rauch and Evans 2000). Or it focuses on final outcomes (such as goods provided or services delivered), which are relevant to the public (but not always so useful for management) (Ammons 2014; Boyne 2003; Carter, Klein, and Day 1992; Hefetz and Warner 2012). The completion of tasks and projects falls between these two approaches: it is useful for management and relevant to the public. The task completion framework helps analysts address the gap between inputs and final outcomes in terms of how they measure government performance. It gives analysts a way to engage with the full distribution of government tasks and to assess the characteristics of the tasks themselves. The task approach outlined in this chapter is closely aligned with the discussion in chapter 15 of *The Government Analytics Handbook*. There, the relevant task is the processing of an administrative case. Clearly, there are other types of tasks in government, and this chapter aims to present a framework that can encapsulate them all. Given the scale of case processing in government, however, chapter 15 presents a discussion specific to that type of task. Related arguments can be made for chapter 12 on procurement, chapter 14 on customs, and chapter 29 on indicators of service delivery. Across these chapters, the *Handbook* provides discussions of the specific analytical opportunities afforded by different types of government activity. These chapters contain some common elements, such as discussions of some form of complexity in relation to the task under focus. This chapter showcases the considerations required for an integrated approach across task types.

Through the task completion framework, we aim to encourage practitioners and scholars alike to conceive of government activity more broadly and to leverage widely available data sources, such as government progress reports or independent public expenditure reviews, to do so. As well as being widely available, these kinds of objective data are highly valuable because they usually cover a wide range of different task types performed by numerous different agencies and departments.

This chapter continues as follows. First, we conceptualize government work as task completion. Second, we use tasks related to the creation of physical infrastructure to illustrate the task completion framework. Third, we show how the framework applies to other types of tasks. Fourth, we explore how to measure task characteristics (considering the complexity of tasks and their ex ante and ex post clarity). Fifth, we discuss key challenges in integrating these measures with each other and into management practice. Finally, we conclude.

CONCEPTUALIZING GOVERNMENT WORK AS THE COMPLETION OF TASKS

Much of the literature in public administration has focused on how to measure government effectiveness by relying on the tasks of single agencies (Brown and Coulter 1983; Ho and Cho 2017; Lu 2016), on a set of

agencies undertaking the same task (Fenizia 2022), or on a broad conception of the central government as a single entity (Lee and Whitford 2009). These approaches limit analysis to a single conception of government effectiveness, which in the case of a single agency or sector, can be precisely defined. However, almost by definition, this limits analysis to a subset of government work and thus raises concerns over what such analysis tells us about government performance as a whole or how performance in one area of government affects other areas.

In addition to measuring government effectiveness on the basis of a partial vision of government, many studies that have sought to investigate government effectiveness have relied on perception-based measures of effectiveness, based on the evaluations of either government employees or external stakeholders and experts (Poister and Streib 1999; Thomas, Poister, and Ertas 2009; Walker et al. 2018). Such measures are frequently available only at an aggregate or even country level because of how distant these individuals are from actual government tasks, and they frequently assess not the outputs of those tasks directly but some perception of "general effectiveness."[1]

Objective measures of government functioning have frequently been eschewed because of obstacles related to data availability, their purported inability to capture the complexity of government work, or conflicting understandings of what effectiveness means. However, many government agencies produce their own reports on the progress they have made across the full distribution of their work. Similarly, agencies often have administrative data on the totality of their activities that provide quantities related to the complexity of task completion that can be repurposed for analytics. These data are collected for management and reporting purposes as part of the daily duties of agency staff. These reports frequently contain characteristics of the tasks undertaken and progress indicators outlining how far tasks have progressed. These reports can be the basis of an integrated analysis of government functioning.

For tasks related to physical infrastructure and administration, analysts can use quantities from these reports, or similar primary data collection, to conceptualize government work in a unified task completion framework. The following discussion of the strengths and limitations or challenges of such an approach focuses on a small set of research papers that have applied a task completion framework to the assessment of government functioning. It thus aims to illustrate the utility of the task completion framework rather than being in any way comprehensive. Where relevant, we provide a number of examples of how public officials have taken a similar approach.

We rely on two simple definitions throughout the chapter. First, a *task* is the bounded activity for which a given organization, team, or individual in the government is responsible. Second, an *output* is the final product a government organization, team, or individual delivers to society. An output is the result of a successful task. In government performance assessment, outputs are defined as "the goods or services produced by government agencies (e.g., teaching hours delivered, welfare benefits assessed and paid)."[2] An example of a government task might be developing a draft competition policy or organizing a stakeholder meeting to validate the draft competition policy (Rasul, Rogger, and Williams 2021, appendix). The corresponding outputs would be the draft competition policy itself and the holding of the stakeholder meeting. These tasks are usually repeated and are completed within varying time frames, depending on the complexity and urgency of the activity at hand.

More granular guidance on how to define a task is challenged by the fact that the appropriate conception of a task will vary by the focus of the analysis. However, to illustrate common conceptions, some examples from the analyses that will be discussed in this chapter include the design, drilling, and development of a water well (including all taps linked to a single source of water); the design, construction, and finishing of a school; the renovation of a neighborhood sewage system; a full maintenance review and associated activities, such as resurfacing, to bring a road up to a functioning state as determined by local standards; the development of a new public health curriculum for primary school students; and the updating of a human resources management information system with current personnel characteristics for all health-related agencies.

By conceiving all government activity as consisting of tasks with intended outputs, analysts can construct a standardized measure of government performance and can gather multiple tasks together to assess government performance across teams within an organization, across organizations, and over time.

Government performance can be defined as the frequency with which particular government actors are able to produce outputs from corresponding tasks. We now turn to considerations in the definition of a task or project and an output in the case of physical and nonphysical outputs.

Physical Outputs

We first consider a task completion framework as it pertains to the accomplishment of physical infrastructure, or, more precisely, tasks relating to the production of physical outputs. In lower-middle-income countries in particular, the noncompletion of infrastructure projects is a widespread and costly phenomenon, with recent estimates suggesting that over one-third of the infrastructure projects started in these countries are not completed (Rasul and Rogger 2018; Williams 2017).

We focus on task completion measures developed from coding administrative data that are at least somewhat comparable across organizations and can be implemented at scale, rather than on performance audits of specific programs (for example, by national audit offices or international financial institutions' internal performance reports) or on the evaluation of performance against key performance indicators (for example, in leadership performance contracts or through central target-setting mechanisms). Many governments or government agencies have infrastructure-project-tracking databases (either electronic or in paper-based files). These records may be for implementation management, for budgeting and fiduciary reasons, or for audit and evaluation. These databases keep records of how far physical projects have been implemented relative to their planned scope.

For example, in Nigeria, Rasul and Rogger (2018) use independent engineering assessments of thousands of projects from across the government implemented by the Nigerian public service to assess the functioning of government agencies. They complement this with a management survey in the agencies responsible for the projects and examine how management practices matter for the completion rates of projects. The analysis exploits a specific period in the Nigerian public service when "the activities of public bureaucracies were subject to detailed and independent scrutiny" (2) and a special office was set up to track the quality of the project implementation of a broad subset of government activities. This was due to an effort by the presidency to independently verify the status of many of the public infrastructure projects funded by the proceeds of debt relief and implemented by agencies across the federal government. The records of this tracking initiative allowed the authors to quantify both the extent of project implementation and the assessment of the quality of the public goods provided.

A second application of the task completion framework to an empirical setting examining physical outputs is provided by Williams (2017), who collects, digitizes, and codes district annual progress reports in Ghana. These reports, which are written annually by each district's bureaucracy and submitted to the central government, include a table listing basic information about projects that were ongoing or active during the calendar year. Such reports are widely produced but not frequently available in a digital format or used for government analytics. The potential of these data for useful insights into government performance is great. Williams uses the reports on physical projects to examine the determinants of noncompletion, presenting evidence that corruption and clientelism are not to blame but rather a dynamic collective action process among political actors facing commitment problems in contexts of limited resources.

Similarly, Bancalari (2022) uses district administrative data on sewerage projects in Peru to explore the social costs of unfinished projects. She uses a combination of mortality statistics, viability studies, annual budget reports on sewerage projects (which allow her to identify unfinished and completed projects), spatial topography data, and population data in order to provide evidence that infant mortality and under-five mortality increase with increases in unfinished sewerage projects. She also finds that mayors who are better connected to the national parliament are able to complete more projects.

Beyond using administrative data, analysts have also undertaken primary fieldwork to explore the completion of physical projects. For example, Olken (2007) uses various surveys on villages, households, individuals, and the assessments of engineering experts to investigate the level of corruption involved in building roads in Indonesia. Olken is able to produce a measure of corruption in terms of missing expenditures by

calculating discrepancies between official project costs and an independent engineer's estimate of costs defined by the survey responses. Primary field activity also allows analysts to undertake randomized controlled trials of potential policies to improve government functioning. In the case of Olken (2007), randomized audits of villages are used to estimate the effect of top-down monitoring on the quality of government outputs: in this case, the building of roads. Such a research design and measure are highly valuable and capture a very important feature of government activity, although they come at a high cost in terms of the resources needed to capture these government tasks.

Other papers have studied the maintenance rather than the construction of physical outputs. In these cases, task completion is the effective continuation of physical outputs. Once again using primary fieldwork to collect required data, Khwaja (2009) uses survey team site visits and household surveys to measure the maintenance of infrastructure projects in rural communities in northern Pakistan (Baltistan) as a form of task completion. Maintenance here is measured through surveys of expert engineers who assess the maintenance of infrastructure projects in terms of their physical state (that is, how they compare to their initial condition), their functional state (that is, the percentage of the initial project purpose satisfied), and their maintenance-work state (that is, the percentage of required maintenance that needs to be carried out). Khwaja (2009) uses these data to examine whether project design can improve collective success in maintaining local infrastructure. The paper presents within-community evidence that project design makes a difference to maintenance levels: "designing projects that face fewer appropriation risks through better leadership and lower complexity, eliciting greater local information through the involvement of community members in project decisions, investing in simpler and existing projects, ensuring a more equitable distribution of project returns, and emulating NGOs can substantially improve project performance even in communities with low social capital" (Khwaja 2009, 913).

We have seen several examples of "government analytics" that seek to measure the completion rate of tasks related to the provision (or maintenance) of physical outputs. From Nigerian federally approved social sector projects, such as providing dams, boreholes, and roads, to Indonesian road building, analysts have defined measures of task completion based on physical outputs. The analysis has used administrative data, existing household surveys, and primary fieldwork (sometimes in combination with one another) to generate insights into the determinants of government functioning.

These papers measure task completion in a series of different ways that all aim to capture the underlying phenomenon of what share of the intended outputs are completed. But there are important commonalities to their approaches. First, the definition of a task or project is determined by a common, or consensus, engineering judgment that crosses institutional boundaries. Thus, though a ministry of urban development may bundle the creation of multiple water distribution points, the building of a health center, and road repaving into a single "slum upgrading" project, the analysts discussed above split these groupings into individual components that would be recognizable across settings, and thus across government. A water distribution point will be conceived as a discrete task whether it is a component of a project in an agriculture, education, health, or water infrastructure project. The wider point is that an external conception of what makes up a discrete activity, such as the common engineering conception of a water distribution point, provides discipline on the boundaries of what is conceived as a single task for any analytical exercise.

Second, within these conceptions of projects, an externally valid notion of completion and progress can be applied. For example, the threshold for a water distribution point is that it produces a sufficient flow of water over a sustained period for it to be considered "completed." Williams (2017) uses the engineering assessments included in administrative data to categorize projects into bins of "complete" (for values such as "complete" or "installed and in use") or "incomplete" (for values such as "ongoing" or "lintel level"). Rasul and Rogger (2018) use engineering documents specific to each project to define a percentage scale of completion for each project allowing for a more granular measure of task progress, mapping them along a 0–1 continuum. Thus, highly varied project designs are mapped into a common scale of progress by consideration of the underlying production function for that class of infrastructure. What constitutes a halfway point in the development of a water distribution point and a dam will differ, but both can be feasibly assessed as having a halfway point.

Third, notions of scale or complexity can be determined from project documentation, providing a basis for improving the credibility of comparisons across tasks. As will be discussed in section three, there is little consensus about how to proxy such complexity across tasks. The literature on complexity in project

management and engineering emphasizes the multiple dimensions of complexity (Remington and Pollack 2007). This can be seen as a strength, in that a common framework for coding complexity can be flexibly adapted to the particular environment or analytical question. In the above examples, planned (rather than expended) budget is frequently used as one way to proxy scale and complexity. The challenge is that the planned budget may already be determined by features related to task completion. For example, the history of task completion at an agency may influence contemporary budget allocations.

For this reason, physical infrastructure tasks can be conceptualized and judged by external conceptions and scales that discipline the analysis. A strength of these measurement options is that they offer a relatively clear, unambiguous measure of task completion. Fundamentally, generating a sensible binary completion value requires understanding how progress maps onto public benefit (for example, an 80 percent finished water distribution point is of zero public value). With this basic knowledge across project types, task completion indicators can be computed for the full range of physical outputs produced by government.[3]

However, this type of task completion framework measurement also comes with limitations. It is easier to measure completion than quality with these types of measures. Quality is typically multifaceted, such that it is more demanding to collect and harmonize into an indicator that can be applied across project types. In Rasul and Rogger (2018), assessors evaluate the quality of infrastructure projects on a coarse scale related to broad indicators that implementation is of "satisfactory" quality relative to professional engineering norms. Analysis can then be defined by whether tasks are, first, completed, and second, completed to a satisfactory level of quality. Administrative progress reports vary in their information content but tend to assume quality and focus on the technical fulfillment of different stages in the completion process.

One way to gain information on quality is to undertake independent audits or checks, though these tend to be highly resource intensive relative to the use of administrative data. For example, Olken (2007, 203) relies on a team of engineers and surveyors to assess the quality of road infrastructure, who "after the projects were completed, dug core samples in each road to estimate the quantity of materials used, surveyed local suppliers to estimate prices, and interviewed villagers to determine the wages paid on the project." From these data, Olken constructs an independent estimate of the quality of each road project.

Some conceptions of quality go as far as the citizen experience of the good or service or how durable or well managed it is. Rasul and Rogger (2018) also include assessments of citizen satisfaction with the project overall as determined by civil society assessors, but such data are almost never available in administrative records and have to be collected independently.

There are also issues pertaining to the reliability and interpretation of task completion that are worth highlighting. First, doubts may be raised when the progress reports that act as the foundation for task completion assessments are provided by the same public organizations that undertake the projects themselves (see the discussion in chapter 4). For this reason, they may not constitute reliable measures of progress, or at least may be perceived as unreliable. The problem is whether organizations can be considered reliable in their assessments of their own work. Measures of task progress sourced from administrative data must thus be used with care and, ideally, validated against a separate (independent) measure of progress. A good example of this comes from Rasul, Rogger, and Williams (2021), who match a subsample of tasks from government-produced progress reports to task audits conducted by external auditors in a separate process.[4] Such validation exercises can be very helpful in providing evidence that the measures produced by government organizations on their own performance are credible, thus salvaging an important source of data that might otherwise be deemed unusable.

Additionally, *noncompletion* can mean different things depending on how the timeline of infrastructure procurement, construction, and operation is organized. This is especially clear in the case described by Bancalari (2022), where it is hard to establish whether the effect uncovered is an effect of noncompletion or delays and cost overruns in delivery.[5] It can be hard to distinguish noncompletion (a project will remain unfinished) from delays (a project will be completed but is running over schedule). Here, the point in time when one decides to measure completion and the initial time frame set for a given task become important and can affect how one interprets task noncompletion.

Finally, a separate issue pertains to whether tasks are completed as planned, not simply whether they are completed. The existing literature from management studies has mostly focused on overruns, delays, and

over-estimated benefits rather than on noncompletion per se (Bertelli, Mele, and Whitford 2020; Post 2014). This body of literature tends to focus on the service and goods delivery side of government rather than on the full range of government activities. However, it is an important complement to the task completion framework precisely because it focuses on whether the tasks governments undertake are being completed *and* are being completed in the time frame and up to the standard that they were planned for. For example, a vast body of literature emphasizes the value-for-money or cost calculations of infrastructure projects rather than the efficiency or effectiveness of the processes via which they are delivered (for example, Engel, Fischer, and Galetovic 2013). Scholars such as Flyvbjerg (2009, 344) have argued that the "worst" infrastructure gets built because "ex ante estimates of costs and benefits are often very different from actual ex post costs and benefits. For large infrastructure projects the consequences are cost overruns, benefit shortfalls, and the systematic underestimation of risks."

Nonphysical Outputs

Now we turn to the task completion framework as it applies to the production of nonphysical outputs. Examples of nonphysical outputs are auditing activities, identifying localities where infrastructure is required, raising awareness about a given social benefit scheme, or planning for management meetings. These types of task, in short, involve government activities that pertain to the less visible side of government: not delivery in the form of physical goods or services but the planning, monitoring, information sharing, reviewing, and organizational tasks of government.

Rasul, Rogger, and Williams (2021) use administrative data on the roughly 3,600 tasks that civil servants undertook in the Ghanaian civil service in 2015. The data on these tasks are extracted from quarterly progress reports and represent the full spectrum of government activities. As can be seen from figure 17.1, a large proportion of these tasks are related to nonphysical outputs. For each type of task, in relation to both physical and nonphysical outputs, the researchers identify a scheme by which to judge task completion by allocating a threshold of progress to represent completion for each task type.

Rasul, Rogger, and Williams (2021) also collect data on the management practices under which these tasks are undertaken via in-person surveys with managers covering six dimensions of management: roles, flexibility, incentives, monitoring, staffing, and targets. Together, the task and management data allow for an assessment of how public sector management impacts task completion, allowing for the comparison of the effect of management practices on the same tasks across different organizations. Their data demonstrate, first, that there is substantial variation in task completion across types of task and across civil service organizations. Second, there is also substantial variation in the types of management practice that public servants are subject to across organizations, and the nature of management correlates significantly with task completion rates.

Integrating the analysis of tasks related to both physical and nonphysical outputs allows for a broad assessment of government functioning, encompassing the many interactions between tasks of different natures. Such a holistic approach also enables the assessment of tasks with different underlying characteristics, which has long been identified as a core determinant of government performance.

Rasul, Rogger, and Williams (2021) are interested in exploring whether different management techniques are differentially effective, depending on the clarity of the task in project documents. They build on the literature arguing that where settings involve intensive multitasking, coordination, or instability, management techniques using monitoring and incentive systems are likely to backfire. The question, as they put it, harking back to the Friedrich vs. Finer debate (Finer 1941; Friedrich 1940), is "to what extent should [civil servants] be managed with the carrot and the stick, and to what extent should they be empowered with the discretion associated with other professions?" (Rasul, Rogger, and Williams 2021, 262). Their central finding is that there are "positive conditional associations between task completion and organizational practices related to autonomy and discretion, but negative conditional associations with management practices related to incentives and monitoring" (Rasul, Rogger, and Williams 2021, 274).[6] The authors distinguish between government tasks with high and low ex ante and ex post clarity. Incentives and monitoring-intensive management approaches are hypothesized (and found) to be more effective when ex ante task clarity is high

(and ex post task clarity is low), whereas autonomy and discretion-intensive management approaches are relatively more effective when ex ante task clarity is low (and ex post task clarity is high).

The main contribution of Rasul, Rogger, and Williams (2021) to the discussion of this chapter is providing a holistic, output-based organizational performance metric. However, their approach also takes a holistic account of the multifarious nature of management practices in government and showcases the value of combining such data. The authors "conceptualize management in public organizations as a portfolio of practices that correspond to different aspects of management, each of which may be implemented more or less well. Bureaucracies may differ in their intended management styles, that is, what bundle of management practices they are aiming to implement, and may also differ in how well they are executing these practices" (262). That is, there is a combination of both intent and implementation when it comes to management practices that may affect the effectiveness of an organization. The task completion framework, with its focus on both the breadth of activities that government bodies undertake and on the detail of the characteristics of government tasks, represents an important stepping stone toward a more holistic and realistic understanding of government work and effectiveness.

A separate body of literature that brings together tasks and projects of distinct types into a single analytical framework is the literature on donor projects. For example, using data on the development projects of international development organizations (IDOs)—specifically, eight agencies—including project outcome ratings of holistic project performance, Honig (2019) investigates the success of IDO projects according to internal administrative evaluations. The success ratings are undertaken by IDO administrators, who employ a consistent underlying construct across different IDOs, with an OECD-wide standard in place. These ratings are combined with a host of other variables capturing various features of the projects (for example, their start and end dates, whether there was an IDO office presence in situ, what the sector of the project is, etc.).

Honig (2019, 172, 196) uses "variation in recipient-country environments as a source of exogenous variation in the net effects of tight principal control" to find that "less politically constrained IDOs see systematically lower performance declines in more unpredictable contexts than do their more-constrained peers." That is to say that monitoring comes with costs in terms of reducing the ability of agents to adapt, particularly in less predictable environments.

Similarly, Denizer, Kaufmann, and Kraay (2013, 288) leverage a data set of over 6,000 World Bank projects (over 130 developing countries) to "simultaneously investigate the relative importance of country-level 'macro' factors and project-level 'micro' factors in driving project level outcomes." The authors leverage Implementation Status Results Reports completed by task team leaders at the World Bank, which report on the status of the projects, as well as Implementation Completion Reports, which include a "subjective assessment of the degree to which the project was successful in meeting its development objective" (290), plus more detailed ex post evaluations of about 25 percent of projects, in order to assess project outcomes. They find that roughly 80 percent of the variation in project outputs occurs across projects within countries, rather than between countries, and that a large set of project-level variables influence aid project outputs.

A related but separate body of literature considers nonphysical task completion by frontline delivery agents. For example, using the case of the Department of Health in Pakistan, Khan (2021) undertakes an experiment in which he randomly emphasizes the department's public health mission to community health workers, provides performance-linked financial incentives, or does both. He measures task completion through a combination of internal administrative data on service delivery and outputs, gathered as part of routine monitoring processes, and household surveys of beneficiaries. Mansoor, Genicot, and Mansuri (2021), instead, use the case of the agriculture extension department in Punjab, Pakistan, to measure both objective task completion and supervisors' subjective perception of performance. They measure this through a combination of household surveys and data from a mobile phone tracking app that frontline providers use to guide and record their work.

Analogous to the physical outputs case, then, to apply a task completion framework to tasks related to nonphysical outputs, we require common definitions of tasks that cross institutional boundaries, externally valid notions of completion and progress, and notions of scale or complexity. Such external standards for what completion and quality look like across institutions are rare, but they do exist in some fields, such as health care (see the example of the joint health inspection checklist in Bedoya, Das, and

Dolinger [forthcoming]). Creating an analogous approach to these issues for tasks related to nonphysical outputs ensures comparability with tasks related to physical outputs. However, they are also valid pillars for analysis even within the set of tasks related to nonphysical outputs only.

For many tasks related to nonphysical outputs, there are, in fact, natural conceptions of task and output. For example, a curriculum development project is only complete once the curriculum is signed off on by all stakeholders, and an infrastructure monitoring program is only complete when a census of the relevant infrastructure has been completed. Similarly, such an approach can be developed for measures of progress. The curriculum development will typically be broken down into substantive stages in planning documents, and each of these stages can be assigned a proportion of progress. In the infrastructure monitoring case, a simple proportion of infrastructure projects assessed, perhaps weighted by scale or distance measures, seems fitting. Not all cases will be so clear-cut. To identify a consensus definition of task by task type that could apply across institutional boundaries, Rasul, Rogger, and Williams (2021) employ public servants at a central analytics office (in the Ghanaian case, this was the Management Services Department) to agree on relevant definitions using data from across government. As will be seen below, this team also defines measures of complexity relevant across the full set of tasks, including (as mentioned above) clarity of design. Decisions as to how to define task completion will be influenced by, but then very much influence, the approach to data collection. Table 17.1 summarizes the approaches analysts have taken to measuring task completion for physical and nonphysical outputs.

While we have focused our discussion mainly on research-oriented examples of measuring task completion, there are also examples of government organizations' use of task completion measures for tasks related to physical and nonphysical outputs—with varying degrees of formality. For example, the United Kingdom Infrastructure and Projects Authority conducts in-depth annual monitoring of all large-scale projects across UK government departments—235 as of 2022—and publishes an annual report with a red/amber/green project outlook rating (IPA 2022). At the other end of the formality and resource-intensiveness spectrum, in their engagement with the government of Ghana in 2015–16 in the course of conducting fieldwork, Rasul, Rogger, and Williams (2021) found that Ghana's Environmental Protection Agency tallied the percentage of outputs completed by each unit in their quarterly and annual reports for internal monitoring purposes. In between these two examples, the Uganda Ministry of Finance and the International Growth Centre (IGC) have partnered to apply Rasul, Rogger, and Williams's (2021) coding methodology (supplemented with qualitative interviews) to monitor the implementation progress of 153 priority policy actions across government and examine the determinants of their completion (Kaddu, Aguilera, and Carson n.d.). And of course, as argued above, many if not most government organizations do some form of task or output completion measurement in the course of their own routine reporting—despite most not taking the next step of using these data for formal analytical purposes.

TABLE 17.1 Selected Measures of Task Completion

Task type	Potential data sources and measurement methods	Selected examples
Physical tasks	• Site visits by expert teams • Site visits by survey teams • Compilation from other secondary sources (for example, media or project reports) • Administrative data from periodic reports	Olken (2007); Rasul and Rogger (2018) Khwaja (2009) Flyvbjerg, Skamris Holm, and Buhl (2002); Williams (2017) Bancalari (2022)
Nonphysical tasks	• Surveys of beneficiaries or citizens • Tracking app used by frontline personnel • Administrative data from periodic reports • Administrative data from internal management monitoring sources • International donor project evaluation reports	Khan (2021) Mansoor, Genicot, and Mansuri (2021) Rasul, Rogger, and Williams (2021) Mansoor, Genicot, and Mansuri (2021), Khan (2021) Denizer, Kaufmann, and Kraay (2013), Honig (2019)

Source: Original table for this publication.

Applying the task completion framework to nonphysical outputs comes with its challenges and limitations, building on those noted above for physical outputs. The issues pertaining to assessing the quality of the implementation of tasks related to nonphysical outputs are twofold. First, establishing how to assess quality is not straightforward, and second, the nature of a task can render the difficulty of assessing quality differentially complex. For instance, if the task one is measuring is the completion of a bridge, one first has to establish the criteria that dictate whether it can be considered a high- or low-quality bridge, whereas if one is also considering nonphysical outputs, such as the development of an education strategy, then one faces a potentially even greater challenge in defining what "high-quality" means for such a project (see Bertelli et al. [2021] for a discussion of this).

There are certain types of task, in short, for which establishing objective benchmarks is more difficult than for others. It does not seem like too much of a leap, for example, to hypothesize that the nonphysical tasks we have considered in this section might frequently be more complex to benchmark in terms of quality than the physical outputs we described earlier.

This difficulty creates discontinuity in measurement quality across physical and nonphysical goods, which, in turn, raises the issue of the potential endogeneity of task and output selection. That is to say, out of the universe of possible government tasks, the types of tasks we are best able to measure may be correlated with particular outputs. This could provide us with a distorted image of the types of tasks that are conducive to producing certain outputs.

MEASURING TASK CHARACTERISTICS

As we outline in the introduction to this chapter, a task completion framework is helpful to analysts in two main senses. First, it pushes analysts to better encapsulate the breadth of work undertaken by public administration across government. Second, it encourages them to think carefully about the characteristics of the tasks themselves. In this section, we will focus on the latter feature of a task completion framework: how to measure task characteristics.

There are, naturally, a plethora of government task characteristics on which one could focus. Here, we will focus on several of the most relevant characteristics from the perspective of implementation. We concentrate on implementation because it has been the focus of the literature on task completion and because it is of direct relevance to the work of practitioners, the intended audience of this chapter.

We start by considering task complexity. When examining government outputs and their relationship to phenomena such as management practices, government turnover, or risk environment, it is often important to understand their relationship with project or task complexity (Prendergast 2002). This is because the complexity of the task will frequently be strongly correlated with variables such as time to completion, total cost, the likelihood of delays, and customer satisfaction, which might be of interest to scholars or practitioners interested in task completion. Table 17.2 summarizes how the analysts described in this paper have attempted to implement measurement of complexity, as well as how authors have measured two further important features of government tasks to which we will turn next, visibility and clarity.

Rasul and Rogger (2018, 12), in their study of public services in the Nigerian civil service, create complexity indicators that capture "the number of inputs and methods needed for the project, the ease with which the relevant labour and capital inputs can be obtained, ambiguities in design and project implementation, and the overall difficulty in managing the project." They are thus able to condition on the complexity of projects along these margins when exploring the relationship between managerial practices and project completion rates. However, such an approach does not account for the fact that worse-performing agencies may be assigned easier (less complex) tasks in a dynamic process over time. So in background work for the study, Rasul and Rogger assess the extent to which there was sorting of projects across agencies by their level of complexity, a task only feasible with appropriate measures. They do not find any evidence of such sorting.

TABLE 17.2 Selected Measures of Task Characteristics

Task or project characteristic	Potential data sources and measurement methods	Selected examples
Complexity	• Expert data coding from site visits • Semi-expert data coding from administrative reports • International donor project evaluation reports	Khwaja (2009); Rasul and Rogger (2018); Rasul, Rogger, and Williams (2021); Denizer, Kaufmann, and Kraay (2013)
Visibility	• Project-level data from infrastructure database assembled from governmental and financial sources	Woodhouse (2022)
Clarity (ex ante and ex post)	• Semi-expert data coding from administrative reports	Rasul, Rogger, and Williams (2021)

Source: Original table for this publication.

Khwaja (2009, 915), instead, captures project complexity by creating an index that measures whether "the project has greater cash (for outside labor and materials) versus noncash (local labor and materials) maintenance requirements, . . . the community has had little experience with such a project, and . . . the project requires greater skilled labor or spare parts relative to unskilled labor for project maintenance." In this way, he is able to distinguish group-specific features—such as social capital—from features of task design—such as degree of complexity—in order to better understand their relative importance to one another.

Denizer, Kaufmann, and Kraay (2013) also consider complexity in their study of how micro (project-level) or macro (country-level) factors are correlated with aid project performance, albeit as a secondary focus. Using three proxies for project complexity (the extent to which a project spans multiple sectors, a project's novelty, and the size of the project), they find "only some evidence that larger—and so possibly more complex—projects are less likely to be successful. On the other hand, greater dispersion of a project across sectors is in fact significantly associated with better project outcomes, and whether a project is a 'repeater' project or not does not seem to matter much for outcomes" (Denizer, Kaufmann, and Kraay 2013, 302).

Given, then, that the issue of accounting for complexity is widespread and often relies upon assessments that are not anchored to an external concept or measure of what complexity is, what are some of the ways that analysts can validate their measures of complexity? Rasul, Rogger, and Williams (2021), in their construction of a measure of the complexity of the tasks being undertaken by Ghanaian civil servants, ensure that coding is undertaken by two independent coders because the variables they measure require coders to make judgment calls about the information reported by government agencies. They also implement reconciliation by managers in cases where there are differences between coders. Discussion between coders and managers about how they see different categories or levels of complexity can be a good way to iron out differences in the measurement of complexity.

Another way to ensure consistency in measuring complexity can be to randomly reinsert particular tasks into the set of tasks being assessed by the coders to check whether they award the same complexity score to identical tasks. This is something that Rasul and Rogger (2018) do in their construction of a measure of task complexity completed by the Nigerian civil service. Rasul and Rogger (2018) also assess the similarity of scores between their two coders and leverage the passing of time to get one of the coders to recode a subsample of projects from scratch (without prompting) to assess the consistency of coding in an additional way.

In a similar spirit, audits of coding can be an effective way to validate a measure of complexity, albeit a costly one. For example, Rasul, Rogger, and Williams (2021, 265) use an auditing technique to check the validity of their measure of task completion; they "matched a subsample of 14% of tasks from progress reports to task audits conducted by external auditors through a separate exercise." Although this technique was applied to task completion, a similar method could easily be used to validate a complexity measure in many contexts; if there are data available on the technical complexity of a task (for example, from engineers or other field specialists), such assessments could be used to check a subsample of the analyst's own evaluations of complexity. Rasul and Rogger (2018), for example, work with a pair of Nigerian engineers to get them to assess the complexity of government tasks according to five dimensions.

Another salient feature of government tasks is how easy it is to define a given task and to evaluate whether and when it has been completed. This feature is related to, but conceptually separate from, the *complexity* of the task. Rasul, Rogger, and Williams (2021) call this feature ex ante and expost *task clarity*. According to their definition, bureaucratic tasks are "ex ante clear when the task can be defined in such a way as to create little uncertainty about what is required to complete the task, and are ex post clear when a report of the actual action undertaken leaves little uncertainty about whether the task was effectively completed" (Rasul, Rogger, and Williams 2021, 260).

Task clarity is an important characteristic to consider, especially in relation to management practices, because the types of management strategy that one wishes to implement may be heavily influenced by the types of task that they govern. Indeed, Rasul, Rogger, and Williams (2021, 260) hypothesize, and find evidence, that "top-down control strategies of incentives and monitoring should be relatively more effective when tasks are easy to define ex ante because it is easier to specify what should be done and construct an appropriate monitoring scheme." On the other hand, they also theorize (and, again, find evidence) that "empowering staff with autonomy and discretion should be relatively more effective when tasks are unclear ex ante, as well as when the actual achievement of the task is clear ex post" (260).

The clarity of task definition is thus also important to take into consideration when exploring questions pertaining to the management of public administration. The degree to which a task is easy to describe and evaluate has a significant bearing on the types of management strategy that make sense to employ when undertaking that task. Task clarity can also impact a number of other features of government work, such as the level of political and citizen support it enjoys—with simpler, more visible projects tending to garner more interest from politicians and support from citizens (Mani and Mukand 2007; Woodhouse 2022)—or the degree to which a task is subject to measurement or performance-pay mechanisms.

Task clarity is important to measure for its potential interactions with the concepts of effort substitution and gaming (Kelman and Friedman 2009). If performance measures are applied only to those tasks that are ex ante and ex post clear, such tasks may be prioritized to the detriment of others because they are subject to measurement or because bureaucrats seek to "game" the system by focusing their attention on improving statistics relating to their performance but not their actual performance. As we have seen in the work of Honig (2019) and Khan (2021), it is especially in complex, multidimensional task environments where granting autonomy or discretion to bureaucrats can have beneficial results. In short, thinking about the nature of the task at hand and its interaction with features such as the management practices being adopted and individual behavioral responses on the part of public servants and politicians is highly important if one wants to get to the bottom of "what works" in government.

DISCUSSION: KEY CHALLENGES

The previous sections have reviewed the scattered and relatively young literature on the systematic measurement of task and project completion in government organizations. The measurement methods and data sources identified hold great promise for practitioners and researchers but also present a number of conceptual and practical challenges. While we have discussed some of these above in relation to specific papers or measurement methods, in this section, we briefly highlight some cross-cutting issues for measurement and analysis as well as for integration into management practice and decision-making.

The first challenge is determining what a *task* is. At the beginning of this chapter, we defined outputs as the final products delivered by government organizations to society and tasks as the intermediate steps taken by individuals or teams within government to produce those outputs. We characterized both as discrete, bounded, and clearly linked to each other. While this is conceptually useful and can serve as a guide for measurement, it is also a profound simplification of the messy, interlinked, and uncertain reality of work inside most government organizations. Indeed, the research insights produced by several of the studies we have discussed emphasize that the ambiguity, complexity, and interconnection of tasks and bureaucratic actions

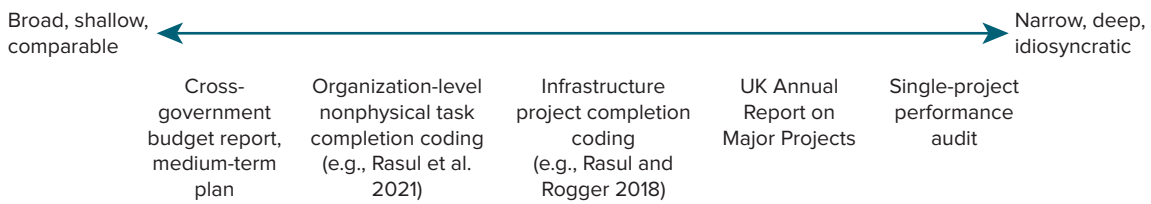
often mean that simplistic management efforts do not produce their anticipated effects. Analysts interested in measuring task completion must thus strike a difficult balance between identifying distinct tasks, projects, and outputs in order to measure their completion and simultaneously calibrating their analysis and inference to capture the nuances of the effective performance of these tasks.

A second and related challenge is drawing appropriate inferences from measures of task completion, which, in itself, is just a descriptive fact of the level of task performance. On its own, measuring task completion does not diagnose the causes of task (in)completion, predict future levels of performance, pinpoint needs for improvement, or measure the performance of the individual personnel responsible for a task (since factors outside their control may also matter). It does, however, provide a foundation upon which to conduct further analysis along these lines. Indeed, for most of the studies cited above, the measurement of task completion simply provides a dependent variable for analysis of a diverse range of potential factors and mechanisms. This chapter has focused mainly on the measurement of this dependent variable; linking it to causes and consequences requires additional analysis, which will differ in its aims and methods depending on an analyst's purposes.

A third challenge relates to integrating the measurement of task completion into practice and management—that is, taking action based on it. One main challenge relates to the well-known potential for gaming and distorting effort across multiple tasks (Dixit 2002; Propper and Wilson 2003), exemplified by "Goodhart's Law": "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (Goodhart 1984). In other words, it may well be possible to accurately measure task completion in government organizations, but using these measures for the purpose of management—particularly if it involves benefits or consequences for the actors involved—risks undermining the validity of the measures and their linkage to bureaucratic performance. See the discussion in chapter 4. While some strategies can be put in place to mitigate such effects (for example, data quality audits or measuring multiple dimensions of bureaucratic performance), these are nearly always imperfect. Analysts should thus seek to innovate in measuring task completion as a means of improving understanding while being cautious and selective in how they use it to guide management actions.

A final consideration in deciding what tasks to measure and how is the trade-off between prioritizing breadth and comparability, on the one hand, and specificity and depth, on the other. Figure 17.2 illustrates this trade-off. In general, task completion measures that are widely applicable across the whole of government will naturally tend to be less specific to (and hence less informative of) the performance of any given unit or task. An example of this might be the type of data contained in a government's annual report, budget execution report, or multiyear plan, which usually cover the whole of government activity but do so at a relatively shallow level. At the other extreme, researchers or practitioners can gather a great deal of information about the completion of a specific task, as a performance audit might do. This gives a very informative picture of the completion of that particular task but permits little comparison across tasks or units. In between, one can locate the various measurement options we have discussed in this chapter. For example, Rasul and Rogger's (2018) project completion data set focuses on physical infrastructure projects, which are likely to be more comparable to each other and across organizations than Rasul, Rogger, and Williams's (2021) data set of both physical and nonphysical outputs—but at the cost of less comprehensive coverage of government activity. The optimal place on this spectrum for any given measure of task completion naturally depends on

FIGURE 17.2 A Spectrum of Task Completion Measures, with Selected Examples



Source: Original figure for this publication.

the analytical purpose for which it is being created. From the standpoint of advancing measurement, the aim is to find ways to surmount this trade-off by increasing both the comparability and the rigor of task completion measures.

CONCLUSION

We conclude by returning to the question with which we opened: how do we know if governments are performing their functions well? In this chapter, we have sought to describe and demonstrate how to apply the task completion framework in order to answer precisely this question. The framework conceives government activity in such a way as to allow analysts to assess public performance in a standardized manner across organizations and types of activity. As such, it gives us a fuller and more accurate picture of government work, forces us to think more carefully about the characteristics of the tasks that different agencies perform, and facilitates comparison of performance on a large sample that spans many types of organizations.

We have applied the framework to different categories of tasks in order to illustrate both its strengths and its limitations. In the case of tasks related to physical outputs, we have shown how data such as engineering assessments, annual progress reports, and budget reports can be merged with other data, such as management or user surveys, to provide a hitherto-inaccessible vision of the extent of project implementation and the quality of the work undertaken.

Much of this work relies, at least partly, on data that already exist but have to be digitized or rendered usable in some other way. The existence of objective, external benchmarks—produced, for example, by experts such as infrastructure engineers—means that the development of projects of many different types can be mapped onto a comparable continuum. The strength of the evaluation of physical outputs is that analysts can produce a meaningful measure of completion that gives the user some sense of how task completion maps onto public benefit. However, the weakness of the approach, as applied to physical outputs, is that the quality of task completion is often overlooked because it rests upon more complex, multifaceted assessments that are difficult to harmonize into a single indicator. Moreover, the reliability of such measures may be called into question where completion rates are reported by the same organizations that undertake the tasks themselves (although this can be counteracted to some degree if external audits of task reports are available to validate the measure).

In the case of nonphysical outputs (such as auditing, planning, or awareness-raising activities), we have demonstrated how data may come from existing sources, such as progress reports, that need to be digitized or processed to be used for analysis. The strength of extending task completion assessments to nonphysical outputs is that this provides a much richer and fuller picture of the activities that governments engage in and allows for meaningful comparisons across departments. However, the task completion framework as applied to nonphysical outputs also suffers from the same potential misreporting concern associated with physical outputs and comes with additional challenges in terms of how to measure the quality of the tasks being completed. The challenges of measuring quality are distinct from those for physical outputs, in that quality is not necessarily overlooked but is more difficult to define. For example, how do you assess the quality of a health strategy objectively and in such a way that it is comparable with, for example, education strategies or fiscal strategies?

The task completion framework, in short, moves us in the right direction when it comes to measuring the performance of governments in a way that takes into account the full breadth of government activity. However, there is much room for improvement when it comes to the measurement of the quality of the provision of both physical and nonphysical outputs. For physical outputs, expert benchmarks are often taken at face value without critical engagement with what the index or evaluation actually captures; whereas, for nonphysical outputs, benchmarks are often nonexistent, with no way to anchor quality assessments that makes them comparable across organizations. This is where we see the frontier in terms of the measurement of government performance; we need to expand the application of the task

completion framework and complement this with greater attention to how technical benchmarks are used in the measurement of physical outputs and the development of workable benchmarks for the measurement of nonphysical outputs.

NOTES

The authors gratefully acknowledge funding from the World Bank's i2i initiative, Knowledge Change Program, and Governance Global Practice. We are grateful to Galileu Kim and Robert Lipinski for helpful comments.

1. See, for instance, the World Bank's World Governance Indicators, available at https://info.wordlbank.org/governance/wgi, and the Millennium Challenge Corporation scorecards—for example, on the website of the Millennium Challenge Coordinating Unit for Sierra Leone, http://www.mccu-sl.gov.sl/scorecards.html.
2. *Outputs* are not to be confused with *outcomes*, or "the impacts on social, economic, or other indicators arising from the delivery of outputs (e.g., student learning, social equity)." *OECD Glossary of Statistical Terms*, s.vv. "output," "outcome" (Paris: OECD Publishing, 2022), http://stats.oecd.org/glossary.
3. Such indicators do not rely upon subjective citizen-survey responses, which are limited by their reliance on human judgment and prey to multiple biases and recall issues (Golden 1992), both from the researcher designing the study and the experts or citizen respondents evaluating the government.
4. No evidence was found that completion levels differed significantly across auditors and agencies, with 94 percent of completion rates being corroborated across coding groups (Rasul, Rogger, and Williams 2021, 265).
5. The measure of unfinished projects is a "combination of projects still underway (on time or delays) and abandoned (temporarily or indefinitely) in a given district" (Bancalari 2022, 10).
6. It is important to note that their findings are relative to one another—that is, "organizations appear to be overbalancing their management practice portfolios toward top-down control measures at the expense of entrusting and empowering the professionalism of their staff" (Rasul, Rogger, and Williams 2021, 261).

REFERENCES

Ammons, David N. 2014. *Municipal Benchmarks: Assessing Local Performance and Establishing Community Standards*. 3rd ed. London: Routledge.

Andrews, Rhys, George A. Boyne, Kenneth J. Meier, Laurence J. O'Toole Jr., and Richard M. Walker. 2005. "Representative Bureaucracy, Organizational Strategy, and Public Service Performance: An Empirical Analysis of English Local Government." *Journal of Public Administration Research and Theory* 15 (4): 489–504. https://doi.org/10.1093/jopart/mui032.

Bancalari, Antonella. 2022. "Can White Elephants Kill? Unintended Consequences of Infrastructure Development in Peru." IFS Working Paper 202227, Institute for Fiscal Studies, London. https://ifs.org.uk/publications/can-white-elephants-kill-unintended-consequences-infrastructure-development.

Bedoya, Guadalupe, Jishnu Das, and Amy Dolinger. Forthcoming. "Randomized Regulation: The Impact of Minimum Quality Standards on Health Markets." Working paper, World Bank, Washington, DC.

Bertelli, Anthony Michael, Eleanor Florence Woodhouse, Michele Castiglioni, and Paolo Belardinelli. 2021. *Partnership Communities.* Cambridge Elements: Public and Nonprofit Administration. Cambridge: Cambridge University Press.

Bertelli, Anthony Michael, Valentina Mele, and Andrew B. Whitford. 2020. "When New Public Management Fails: Infrastructure Public-Private Partnerships and Political Constraints in Developing and Transitional Economies." *Governance: An International Journal of Policy, Administration, and Institutions* 33 (3): 477–93. https://doi.org/10.1111/gove.12428.

Boyne, George A. 2003. "What Is Public Service Improvement?" *Public Administration* 81 (2): 211–27. https://doi.org/10.1111/1467-9299.00343.

Brown, Karin, and Philip B. Coulter. 1983. "Subjective and Objective Measures of Police Service Delivery." *Public Administration Review* 43 (1): 50–58. https://doi.org/10.2307/975299.

Carter, Neil, Rudolf Klein, and Patricia Day. 1992. *How Organisations Measure Success: The Use of Performance Indicators in Government*. London: Routledge.

Denizer, Cevdet, Daniel Kaufmann, and Aart Kraay. 2013. "Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance." *Journal of Development Economics* 105: 288–302. https://doi.org/10.1016/j.jdeveco.2013.06.003.

Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review." *Journal of Human Resources* 37 (4): 696–727. https://doi.org/10.2307/3069614.
Engel, Eduardo, Ronald Fischer, and Alexander Galetovic. 2013. "The Basic Public Finance of Public-Private Partnerships." *Journal of the European Economic Association* 11 (1): 83–111. https://www.jstor.org/stable/23355049.
Fenizia, Alessandra. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84. https://doi.org/10.3982/ECTA19 244.
Finer, Herman. 1941. "Administrative Responsibility in Democratic Government." *Public Administration Review* 1 (4): 335–50. https://doi.org/10.2307/972907.
Flyvbjerg, Bent. 2009. "Survival of the Unfittest: Why the Worst Infrastructure Gets Built—And What We Can Do about It." *Oxford Review of Economic Policy* 25 (3): 344–67. https://doi.org/10.1093/oxrep/grp024.
Flyvbjerg, Bent, Mette Skamris Holm, and Soren Buhl. 2002. "Underestimating Costs in Public Works Projects: Error or Lie?" *Journal of the American Planning Association* 68 (3): 279–95. https://doi.org/10.1080/01944360208976273.
Friedrich, Carl J. 1940. "Public Policy and the Nature of Administrative Responsibility." In *Public Policy: A Yearbook of the Graduate School of Public Administration, Harvard University* 1: 1–20.
Golden, Brian R. 1992. "The Past Is the Past—Or Is It? The Use of Retrospective Accounts as Indicators of Past Strategy." *Academy of Management Journal* 35 (4): 848–60. https://doi.org/10.2307/256318.
Goodhart, Charles A. E. 1984. "Problems of Monetary Management: The UK Experience." In *Monetary Theory and Practice: The UK Experience*, 91–121. London: Red Globe Press. https://doi.org/10.1007/978-1-349-17295-5.
Hefetz, Amir, and Mildred E. Warner. 2012. "Contracting or Public Delivery? The Importance of Service, Market, and Management Characteristics." *Journal of Public Administration Research and Theory* 22 (2): 289–317. https://doi.org/10.1093/jopart/mur006.
Ho, Alfred Tat-Kei, and Wonhyuk Cho. 2017. "Government Communication Effectiveness and Satisfaction with Police Performance: A Large-Scale Survey Study." *Public Administration Review* 77 (2): 228–39. https://doi.org/10.1111/puar.12563.
Honig, Dan. 2019. "When Reporting Undermines Performance: The Costs of Politically Constrained Organizational Autonomy in Foreign Aid Implementation." *International Organization* 73 (1): 171–201. https://doi.org/10.1017/S002081831800036X.
IPA (Infrastructure and Projects Authority). 2022. *Annual Report on Major Projects 2021–22.* Reporting to Cabinet Office and HM Treasury, United Kingdom Government. London: IPA. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachmentdata/file/1092181/IPAAR2022.pdf.
Kaddu, M., J. Aguilera, and L. Carson. n.d. *Challenges to Policy Implementation in Uganda (Review of Policy Implementation in Uganda)*. London: International Growth Centre, London School of Economics and Political Science.
Kelman, Steven, and John N. Friedman. 2009. "Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service." *Journal of Public Administration Research and Theory* 19 (4): 917–46. https://doi.org/10.1093/jopart/mun028.
Khan, Muhammad Yasir. 2021. "Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan." Working paper delivered at 25th Annual Conference of the Society for Institutional Organizational Economics, June 24–26, 2021 (accessed February 8, 2023). https://y-khan.github.io/yasirkhan.org/muhammadyasirkhanjmp.pdf.
Khwaja, Asim Ijaz. 2009. "Can Good Projects Succeed in Bad Communities?" *Journal of Public Economics* 93 (7–8): 899–916. https://doi.org/10.1016/j.jpubeco.2009.02.010.
Lee, Soo-Young, and Andrew B. Whitford. 2009. "Government Effectiveness in Comparative Perspective." *Journal of Comparative Policy Analysis* 11 (2): 249–81. https://doi.org/10.1080/13876980902888111.
Lewis, David E. 2007. "Testing Pendleton's Premise: Do Political Appointees Make Worse Bureaucrats?" *The Journal of Politics* 69 (4): 1073–88. https://doi.org/10.1111/j.1468-2508.2007.00608.x.
Lu, Jiahuan. 2016. "The Performance of Performance-Based Contracting in Human Services: A Quasi-Experiment." *Journal of Public Administration Research and Theory* 26 (2): 277–93. https://doi.org/10.1093/jopart/muv002.
Mani, Anandi, and Sharun Mukand. 2007. "Democracy, Visibility and Public Good Provision." *Journal of Development Economics* 83 (2): 506–29. https://doi.org/10.1016/j.jdeveco.2005.06.008.
Mansoor, Zahra, Garance Genicot, and Ghazala Mansuri. 2021. "Rules versus Discretion: Experimental Evidence on Incentives for Agriculture Extension Staff." Unpublished manuscript.
Nistotskaya, Marina, and Luciana Cingolani. 2016. "Bureaucratic Structure, Regulatory Quality, and Entrepreneurship in a Comparative Perspective: Cross-Sectional and Panel Data Evidence." *Journal of Public Administration Research and Theory* 26 (3): 519–34. https://doi.org/10.1093/jopart/muv026.
Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–49. https://doi.org/10.1086/517935.
Poister, Theodore H., and Gregory Streib. 1999. "Performance Measurement in Municipal Government: Assessing the State of the Practice." *Public Administration Review* 59 (4): 325–35. https://doi.org/10.2307/3110115.
Post, Alison E. 2014. *Foreign and Domestic Investment in Argentina: The Politics of Privatized Infrastructure.* Cambridge, UK: Cambridge University Press.

Prendergast, Canice. 2002. "The Tenuous Trade-Off between Risk and Incentives." *Journal of Political Economy* 110 (5): 1071–102. https://doi.org/10.1086/341874.

Propper, Carol, and Deborah Wilson. 2003. "The Use and Usefulness of Performance Measures in the Public Sector." *Oxford Review of Economic Policy* 19 (2): 250–67. https://doi.org/10.1093/oxrep/19.2.250.

Rainey, Hal G. 2009. *Understanding and Managing Public Organizations.* 4th ed. New York: John Wiley & Sons.

Rainey, Hal G., and Paula Steinbauer. 1999. "Galloping Elephants: Developing Elements of a Theory of Effective Government Organizations." *Journal of Public Administration Research and Theory* 9 (1): 1–32. https://doi.org/10.1093/oxfordjournals.jpart.a024401.

Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal* 128 (608): 413–46. https://doi.org/10.1111/ecoj.12418.

Rasul, Imran, Daniel Rogger, and Martin J. Williams. 2021. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory* 31 (2): 259–77. https://doi.org/10.1093/jopart/muaa034.

Rauch, James E., and Peter B. Evans. 2000. "Bureaucratic Structure and Bureaucratic Performance in Less Developed Countries." *Journal of Public Economics* 75 (1): 49–71. https://doi.org/10.1016/S0047-2727(99)00044-4.

Remington, Kaye, and Julien Pollack. 2007. *Tools for Complex Projects.* Aldershot, UK: Gower.

Talbot, Colin. 2010. *Theories of Performance: Organizational and Service Improvement in the Public Domain.* Oxford: Oxford University Press.

Thomas, John Clayton, Theodore H. Poister, and Nevbahar Ertas. 2009. "Customer, Partner, Principal: Local Government Perspectives on State Agency Performance in Georgia." *Journal of Public Administration Research and Theory* 20 (4): 779–99. https://doi.org/10.1093/jopart/mup024.

Walker, Richard M., M. Jin Lee, Oliver James, and Samuel M. Y. Ho. 2018. "Analyzing the Complexity of Performance Information Use: Experiments with Stakeholders to Disaggregate Dimensions of Performance, Data Sources, and Data Types." *Public Administration Review* 78 (6): 852–63. https://doi.org/10.1111/puar.12920.

Williams, Martin J. 2017. "The Political Economy of Unfinished Development Projects: Corruption, Clientelism, or Collective Choice?" *American Political Science Review* 111 (4): 705–23. https://doi.org/10.1017/S0003055417000351.

Woodhouse, Eleanor Florence. 2022. "The Distributive Politics of Privately Financed Infrastructure Agreements." Unpublished manuscript.