PART 4

Government Analytics Using Public Servant Surveys

CHAPTER 18

Surveys of Public Servants

The Global Landscape

Ayesha Khurshid and Christian Schuster

SUMMARY

Governments around the world increasingly implement surveys of public servants to better understand—and to provide evidence to improve—public administration. As context for the subsequent chapters in *The Government Analytics Handbook* on surveys of public servants, this chapter reviews the existing landscape of governmentwide surveys of public servants. What concepts are measured in these surveys? How are these concepts measured? And what survey methodologies are used? Our review finds that while governments measure similar concepts across surveys, the precise questions asked to measure these concepts vary, as do survey methodologies—for instance, in terms of sampling approaches, survey weights, and survey modes. The chapter concludes, first, that discrepancies in survey questions for the same concepts put a premium on cross-country questionnaire harmonization, and it introduces the Global Survey of Public Servants (GSPS) as a tool to achieve harmonization. Second, the chapter concludes that methodological differences across surveys—despite similar survey objectives—underscore the need for stronger evidence to inform methodological choices in surveys of public servants. The remaining chapters in this part focus on providing such evidence.

ANALYTICS IN PRACTICE

- Surveys of public servants have been implemented by an increasing number of countries in the last two decades. They tend to measure similar concepts, focusing on a core set of employee attitudes (such as job satisfaction or engagement), on the one hand, and management practices (such as the quality of leadership), on the other.
- Despite measuring similar concepts, questionnaires across surveys of public servants are not harmonized: different governments use different measures for the same concepts.

Ayesha Khurshid is a consultant in the World Bank's Development Impact Evaluation (DIME) Department. Christian Schuster is a professor at University College London.

- Despite having similar aims, methodologies for surveys of public servants vary across countries—for instance, in terms of sampling approaches, survey weighting, survey populations, survey modes, and response rates achieved.
- Differences in survey methodology underscore the importance of robust evidence to ensure good-practice methodologies in surveying public servants, the topic of the remainder of this part.

INTRODUCTION

Understanding government and providing actionable data and evidence to public sector managers to improve the machinery of government requires microdata about government institutions (chapter 2). Surveys of public servants are one such microdata source. Many key features of the environment of public servants cannot be measured efficiently through other (administrative data) mediums. For example, how public servants are managed, their motivations, and their behaviors are all phenomena internal to an official's lived experience. Management quality is fundamentally an experienced interaction that can often only be measured by employees' or managers' reports of it. Public employees' motivations are difficult to observe outside of their own expressions of them. Thus, self-reporting through surveys becomes the primary means of measurement for many aspects of officialdom and, as detailed elsewhere in *The Government Analytics Handbook*, of the public sector production function (see chapter 1).

This section of the *Handbook* provides frontier evidence on key choices in public servant surveys—from the appropriate survey mode (chapter 19), to determining sampling sizes (chapter 21), questionnaire design (chapters 20 and 22), and the effective reporting of survey results (chapter 25). To contextualize the chapters in this section, this introductory chapter provides an overview of the state of play in public servant surveys around the world.

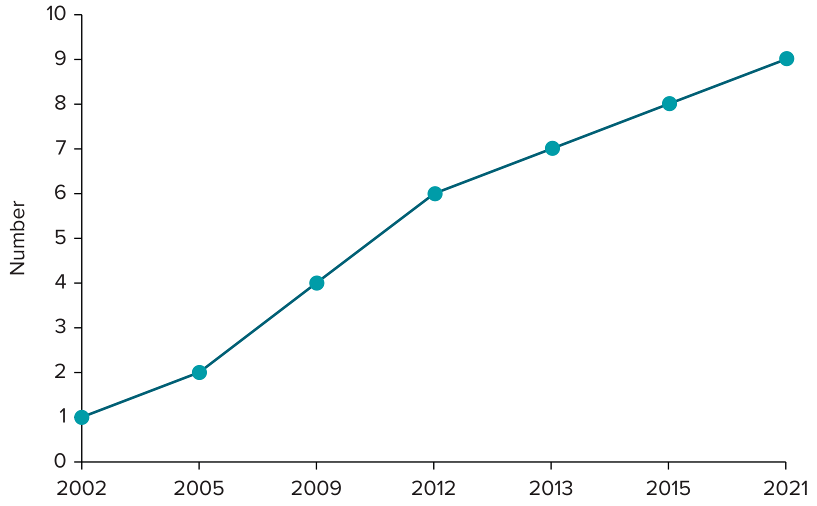
To present the state of play in this field, we review the existing landscape of regular, governmentwide employee surveys—that is, surveys that are run on a regular (annual or biannual) basis with repeated measurements (on at least three previous occasions) for a central government. We thus focus this chapter on surveys that are institutionalized as measurement and management instruments in governments. This contrasts with other reviews—in particular, Organisation for Economic Co-operation and Development (OECD 2016)—which comprise ad hoc, non-central-governmentwide surveys with varying content and methodologies.[1]

The first introductory takeaway from this review is that surveys of public servants have recently become more popular with governments. As illustrated in figure 18.1, the number of countries undertaking governmentwide employee surveys has increased continuously over the last decade, reaching nine countries in 2021. (We might, of course, underestimate the number of institutionalized surveys of public servants outside the English-speaking world, so this number is a lower bound.)

As detailed in table 18.1, all countries for which we were able to review and validate the implementation of regular surveys of public servants belong to the OECD (though some, such as Colombia, are recent OECD joiners). While most of these countries have been implementing institutionalized surveys for over a decade, countries such as New Zealand have only begun the exercise in recent years. All countries implement their surveys annually except Ireland and Canada, which implement their surveys every two years.

This chapter will provide an overview of the key features of these surveys, in part to contextualize the remainder of this section of the *Handbook*, which will provide novel empirical evidence on the design, implementation, and dissemination of public servant surveys. The chapter will first review what established surveys of public servants measure. Subsequently, it will look at survey methodologies across countries: how are surveys implemented (for instance, in terms of sampling and response rates)?

FIGURE 18.1 Countries with Regular, Governmentwide Employee Surveys, Cumulative Count, 2002–21



Source: Original figure for this publication.

TABLE 18.1 Countries with Regular, Governmentwide Employee Surveys, 2002–22

Country	Survey title	Undertaken since	Latest year	Frequency
Australia	Australian Public Service Employee Census	2012	2022	Annual
Canada	Public Service Employee Survey	2005	2020	Biannual
Colombia	Survey of the Institutional Environment and Performance in the Public Sector [Encuesta sobre ambiente y desempeño institucional nacional]	2009	2021	Annual
Ireland	Civil Service Employee Engagement Survey	2015	2020	Biannual
Korea, Rep.	Public Service Life Survey	2013	2021	Annual
New Zealand	Te Taunaki Public Service Census	2021	2021	Annual
Switzerland	Staff Survey of the Federal Administration [Enquête auprès du personnel de l'administration fédérale]	2012	2021	Annual
United Kingdom	Civil Service People Survey	2009	2021	Annual
United States	Federal Employee Viewpoint Survey	2002	2021	Annual

Source: Original table for this publication.

A REVIEW OF CONCEPTS AND MEASURES IN EXISTING SURVEYS OF PUBLIC SERVANTS

To understand the key concepts for measurement when governments undertake surveys of their employees, we summarize a review by Meyer-Sahling et al. (2021) of the concepts measured in six of the government employee surveys outlined above. This review comprises the United States' Federal Employee Viewpoint Survey, Canada's Public Service Employee Survey, the United Kingdom's Civil Service People Survey, the Australian Public Service (APS) Employee Census, Colombia's Survey of the Institutional Environment and Performance in the Public Sector, and Ireland's Civil Service Employee Engagement Survey. The focus of the review is on measurement in the last year before the COVID-19 pandemic, as the pandemic led to an exceptional focus on teleworking—rather than the implementation of the regular annual survey—in a number of countries.

Meyer-Sahling et al. (2021) frame their review within a production function of the public service (analogous to the production function presented in chapter 1 of the *Handbook*) that outlines how the productivity of public services depends on the quality and quantity of outputs relative to inputs. Inputs include staff (that is, public servants) and other resources. Inputs are converted to public sector outputs and outcomes by management practices and public or organizational policies. Whether inputs are effectively converted to outputs is moderated by exogenous factors (such as the political environment) and mediated by the attitudes and behaviors of public servants.

Surveys of public servants can be used to shed light on different components of this public service productivity chain. As detailed by Meyer-Sahling et al. (2021), surveys of public servants are particularly suitable for measuring management practices and complementary inputs, on the one hand (for example, employees' perception of the quality of leadership in their organization), and public employees' attitudes and behaviors, on the other (for example, their work motivation). These parts of the public sector production function often cannot be recorded through administrative data in a valid way. Thus, self-reporting through surveys becomes the primary measurement tool.

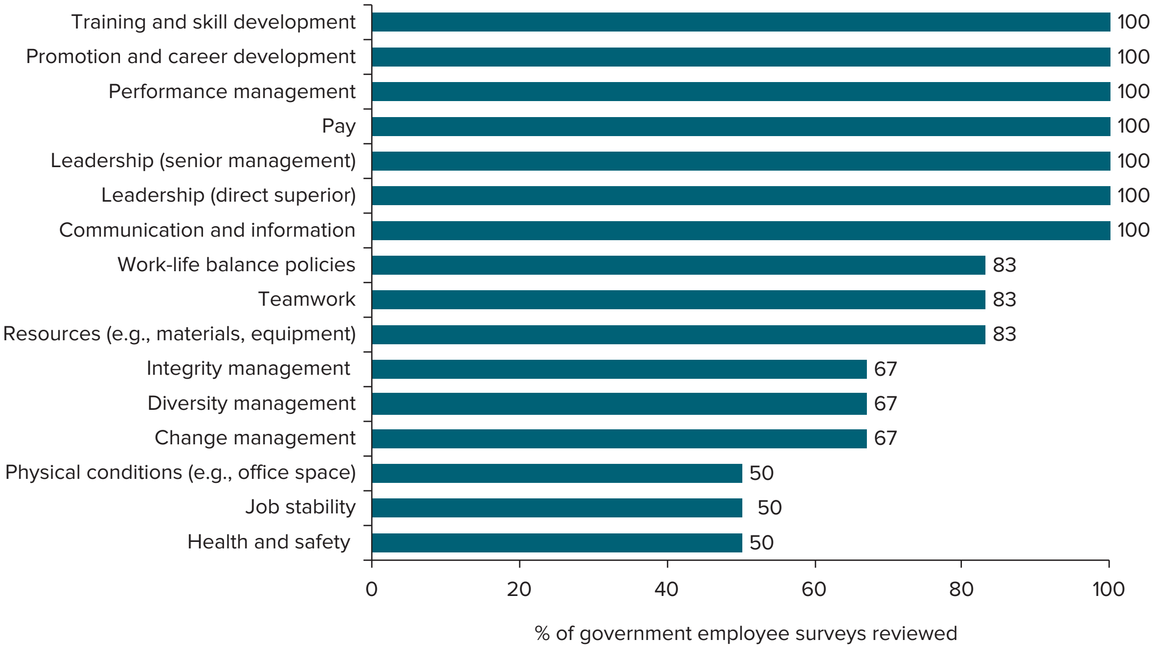
Which areas of management practice, on the one hand, and employee attitudes, on the other, do existing surveys of public servants primarily measure? By classifying topics in the six countries, seven broad areas of management practices are measured across all government employee surveys reviewed: leadership (by both the direct superior and senior management), performance management, pay, training and skills development, promotion and career development, and communication and information to employees. Three further areas—practices to foster work-life balance, teamwork, and the sufficiency of resources (for example, equipment)—were measured in all but one employee survey. As figure 18.2 shows, these 10 management areas are thus plausibly core to (almost) all government employee surveys.

Looking next at employee attitudes, the review finds that government employee surveys also measure an overlapping set of core employee attitudes and behaviors. As illustrated in figure 18.3, all reviewed government employee surveys measure the organizational commitment of public employees, their engagement with their jobs, and their perception of their workloads and work-life balance. Moreover, four additional concepts—job satisfaction, career/turnover intentions, integrity, and innovation attitudes—are measured in all but one of the government employee surveys. These six attitudes and behaviors are thus plausibly core to (almost) all government employee surveys.

Thus, governments measure similar concepts across many of their employee surveys. (Of course, governments also add idiosyncratic modules that are of particular interest to them in any given year, such as remote work during the COVID-19 pandemic.) This plausibly reflects an interest in a similar set of core management practices and employee attitudes and behaviors to improve public sector performance. At the same time, as outlined below, the exact wording of measures for the same concept frequently differs across countries (as does the precise coverage of a concept—for instance, whether pay is measured in relation to performance, satisfaction, fairness, or other pay-related factors), which is a core rationale for harmonizing this wording through the Global Survey of Public Servants (GSPS) (see below).

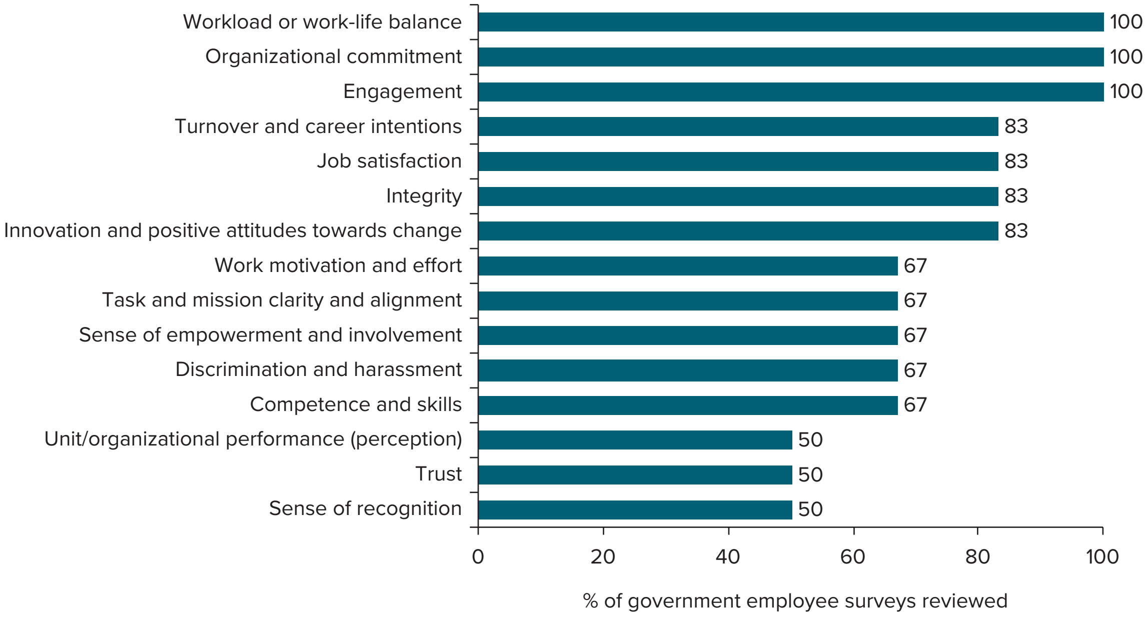
Two caveats regarding these conclusions about commonality are due. First, the review's coverage extends to OECD countries. In countries of the Global South, other concepts, such as meritocracy, politicization, and corruption, are often central to the (non)functioning of the public sector and might thus deserve greater pride of place in surveys of public servants (Meyer-Sahling et al. 2021). Second, some recent surveys have shifted toward a greater focus on directly actionable survey questions—for instance, to check for good practice in performance evaluations or onboarding procedures and showcase where basic practices are not in place (see Fukuyama et al. 2022). That most existing governmentwide employee surveys are silent on these topics suggests that focusing on more actionable survey questions is one margin for improving many existing questionnaires.

FIGURE 18.2 Management Practices Measured in Government Employee Surveys



Source: Meyer-Sahling et al. 2021.
Note: Only concepts covered in at least half of the surveys reviewed are shown.

FIGURE 18.3 Employee Attitudes and Behaviors Measured in Government Employee Surveys



Source: Meyer-Sahling et al. 2021.
Note: Only concepts covered in at least half of the surveys reviewed are shown.

METHODOLOGIES IN SURVEYS OF PUBLIC SERVANTS

Having reviewed the content of existing governmentwide employee surveys, in this section, we will review their methodologies. How are respondents sampled by governments? Are surveys conducted online, on paper, in person, or by phone? How long are public servant survey questionnaires? What response rates are achieved and how are survey weights constructed to enhance representativeness? The remaining chapters in the public servant survey section of the *Handbook* provide novel empirical evidence to enable governments and practitioners to make evidence-based choices in response to these and other methodological questions, along the decision tree in survey design, implementation, and reporting detailed in chapter 1. To contextualize these empirical and methodological chapters, the remainder of this section briefly reviews practices and methodological choices in existing governmentwide employee surveys. Table 18.2 summarizes the findings from this comparison.

Survey Mode

One of the first methodological choices in public servant surveys is the enumeration method, or survey mode. Different survey modes come with different response biases to questions and different overall response rates.

All nine government surveys reviewed in table 18.1 were implemented online, using an invitation link sent to public servants through email or shared through the administration's intranet. Additionally, to enhance accessibility (for instance, for staff with difficulty accessing or completing an online survey), Colombia, Switzerland, the UK, and a few Australian agencies offered their surveys in a paper format, while New Zealand offered its survey through paper and telephone upon request.

Field experimental evidence from the *Handbook* suggests—albeit based on data from Romania only—that these diverging survey modes do not substantially impact aggregate estimates at the national level (see chapter 19). They do, however, affect the comparability of findings across organizations, among other things (see chapter 19). Governments that offer varying survey modes should thus be careful when comparing the scores of organizations if some implement the survey primarily online while others implement it primarily on pen and paper.

SURVEY POPULATION

Across the nine surveys reviewed, the survey population generally consists of central-government civil servants, although the extent to which public sector organizations and employee contracts outside the (legally defined) civil service are covered varies—for instance, in other branches of government or frontline services.

For the UK government employee survey, all public servants from 101 agencies are eligible, excluding the Northern Ireland Civil Service, the NHS (which conducts its own survey), and frontline officials (for instance, police officers and teachers) (Cabinet Office 2020). The US survey invites all federal, nonseasonal, and permanent public servants (including all full- and part-time employees) in 82 executive branch agencies to participate (OPM 2020).

The Australian survey includes all employees from 101 agencies. While agencies set their own eligibility requirements, it generally excludes public servants on leave during the survey and those with a short tenure in the agency (Australian Public Service Commission 2021). Similarly, in Colombia, all public servants working in Bogotá with a tenure of more than six months at the central level of the executive, legislative, and judicial powers and in the headquarters of the regional autonomous corporations and public universities (200 agencies) are eligible to participate in the survey (DANE 2020).

TABLE 18.2 Methodological Choices in Public Servant Surveys

Country	Survey mode(s)	Survey population	Sampling	Response rate[a] (%)	Survey weighting[b]	Questionnaire length (number of questions)[c]
Australia	Primarily online with some agencies offering a paper-based option	All regular employees from 101 agencies with sufficient tenure in their agency	Census	77	No weights applied	112
Canada	Online	All paid employees in 90 core agencies	Census	61	Nonresponse weights applied	112
Colombia	Primarily online with a paper-based option	All employees in 200 agencies with a tenure of at least six months working in Bogotá and in the headquarters of regional autonomous corporations and public universities	Census for smaller agencies and stratified sampling for larger agencies	96	Nonresponse weights applied	65
Ireland	Online	All employees in 50 agencies in Ireland and those based abroad	Census	65	—	112
Korea, Rep.	Online	All employees from central administrative agencies and metropolitan governments	Sampled survey using multistage stratification and probability-proportional-to-size sampling	—	—	48
New Zealand	Primarily online with a paper-based and telephone option	All employees in 36 agencies and those based abroad, excluding the NZCIS and the GSCB[d]	Census	63	—	61
Switzerland	Primarily online with a paper-based option	All monthly paid employees (excluding parliamentary services and the Public Ministry of the Confederation and the courts)	Census every three years with a sampled survey in all other years	71	—	24
United Kingdom	Primarily online with a paper-based option	All employees from 101 agencies (excluding the Northern Ireland Civil Service, the National Health Service, and frontline officials)	Census	62	No weights applied	72
United States	Online	All permanently employed and nonseasonal federal employees in 82 agencies	Census every few years (2012, 2018, 2019, and 2020) with a sampled survey using stratified sampling in other years	44	Nonresponse weights applied	101

Source: Original table for this publication.

Note: The table displays "—" wherever information was unavailable to the authors.

a. Response rates are presented for the latest year for which data and/or results were available. The response rate for the Korean survey was unavailable.

b. Information about nonresponse weights in Canada, Ireland, Republic of Korea, New Zealand, and Switzerland was, unfortunately, unavailable.

c. Questionnaire lengths were reviewed for the last year before the COVID-19 pandemic.

d. New Zealand Security Intelligence Service (NZCIS); Government Communications Security Bureau (GSCB).

The Irish survey targets all public servants from 50 agencies in Ireland and those based abroad (Department of Public Expenditure and Reform 2020). Similar to Ireland, the New Zealand survey includes all public servants working in 36 public service agencies and those based overseas, apart from the New Zealand Security Intelligence Service (NZSIS) and the Government Communications Security Bureau (GSCB) (both of which conduct their own surveys) (Research New Zealand 2021). While limited information is available on the Korean survey, its target population includes all public servants from central administrative agencies and metropolitan governments (Korea Institute of Public Administration 2021).

The Canadian survey has the most flexible eligibility criteria: all indeterminate, term, seasonal, casual, and student employees in 90 core public administration agencies are eligible (excluding ministers' exempt staff, private contractors and consultants, and employees on unpaid leave) (Government of Canada 2022). Similarly, the Swiss survey population consists of all permanent staff that are paid monthly but excludes public servants working in the parliamentary services, the Public Ministry of the Confederation, and the courts (OFPER 2022).

Sampling Design

Approaches to sampling across countries vary, ranging from census to random, ad hoc, and stratified sampling. Australia, New Zealand, and the UK adopt a census approach in which all eligible public sector employees are invited to participate in the survey (Australian Public Service Commission 2021; Cabinet Office 2020; Research New Zealand 2021). Canada's Public Service Employee Survey and Ireland's Civil Service Employee Engagement Survey are also based on a census approach, albeit one with an open link offering less control over who responds (Department of Public Expenditure and Reform 2020). In Canada, public sector organizations reach out to their staff to complete the survey, but the government also makes the survey available online for anyone who decides they fit the eligibility criteria (Government of Canada 2022).

The US government Federal Employee Viewpoint Survey uses stratified randomized sampling approaches for most years but conducts a census every few years (2012, 2018, 2019, and 2020), in order to update sampling frames, with the survey link sent to all eligible respondents (OPM 2020). Similarly, Switzerland conducts a census every three years (2014, 2017, and 2020) and a sampled survey in other years (OFPER 2022). Colombia's public servant survey, in turn, uses a mixed approach: for larger organizations, a stratified sampling approach is used, while for smaller organizations (with fewer than 110 employees), a census is taken to protect anonymity. For larger organizations, the sampling frame is stratified by organization and hierarchy, and public servants are selected to participate using simple random sampling within strata (DANE 2020).

Meanwhile, the Republic of Korea adopts a sampling approach for all annual surveys. Approximately 4,000 respondents are sampled each year each using multistage stratification and probability-proportional-to-size sampling to ensure the representativeness of the sample (Korea Institute of Public Administration 2021).

As detailed later in this section of the *Handbook* (chapter 25, census approaches offer the advantage of sufficient response numbers to provide unit-level management reports based on survey results, even at more disaggregated levels. The UK government, for instance, produces over 12,000 management diagnostics or reports based on its results. At the same time, census sampling approaches are costly in terms of the opportunity cost of staff time spent on completing the survey. As detailed in chapter 20, the appropriate sampling approach thus depends on the types of inference one seeks to draw from the data. Chapter 20 offers a sampling tool to allow governments to estimate appropriate sample sizes based on the types of inference and benchmarking exercises they wish to make with the data. Interestingly, existing government approaches to sampling respondents in public servant surveys do not seem to be (explicitly) based on such a data-driven approach to sampling, suggesting that the potential to optimize sampling in surveys of public servants remains.

Response Rates and Nonresponse Weighting

Beyond their sampling approaches, surveys of public servants across governments also differ in response rates and their approaches to correcting for nonresponse bias. As detailed in table 18.2, survey response rates vary from 44 percent in the US to 96 percent in Colombia. In Colombia, the national statistical office (DANE) conducts the survey, and statistics legislation mandates that sampled respondents complete the survey. In the remaining countries, participation in the survey is voluntary, leading to relatively lower response rates.

To enhance the likelihood that the final sample is representative of the target population of public servants, Canada, Colombia, and the US apply nonresponse weights.[2] Canada uses nonresponse weights to enhance the representativeness of occupational groups in each agency (Statistics Canada 2018). To construct nonresponse weights, the US survey uses subagency identifier, supervisory status, gender, minority status, age, tenure, full- or part-time status, and location from administrative data (OPM 2020). The Colombian survey, in turn, uses nonresponse weights based on the same variables as in its sampling approach—for example, hierarchical level or the institution a respondent works for (DANE 2020).

The Australian survey checks for the representativeness of respondents across age, gender, state or territory, and classification. As survey respondents do not significantly differ from the survey population in these characteristics in the Australian case, the Australian survey does not use nonresponse weights (Australian Public Service Commission 2021). Similarly, the UK does not apply nonresponse weights to the final set of respondents.

Evidence from elsewhere in the *Handbook* suggests that the effect of nonresponse weights (constructed from demographic information) on national-level averages in particular is relatively limited, at least in the country studied in the chapter (chapter 19). This is good news for cases, like the UK, where governmentwide demographic information to construct weights is in limited supply. At the same time, some nonresponse weights are straightforward to construct for all governments—for instance, weights to correct for differential response rates in institutions of differential size. They thus merit consideration where not currently applied.

QUESTIONNAIRE LENGTH

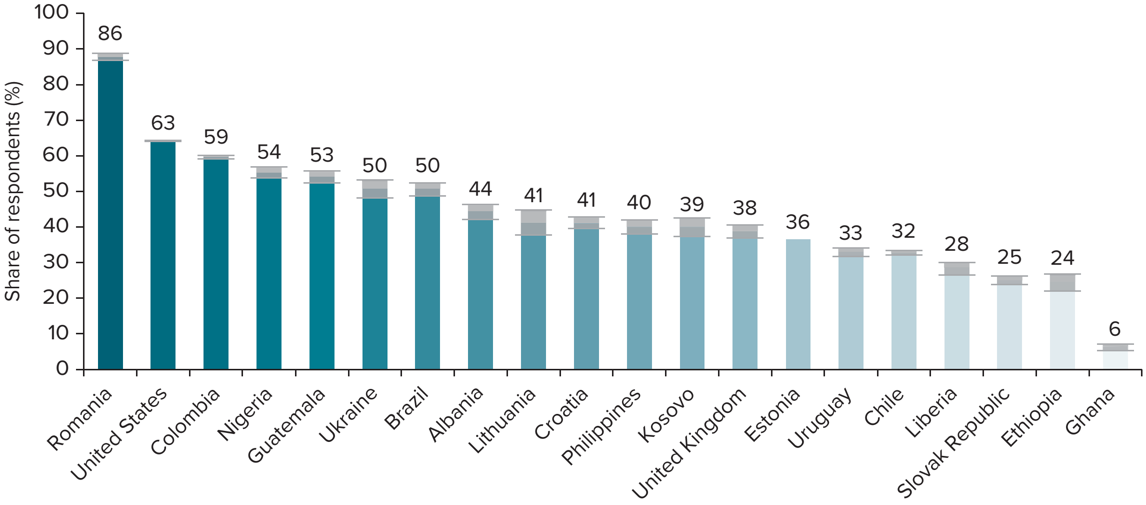
Beyond these differences in nonresponse weights, surveys of public servants also differ in questionnaire design, including length. In the last year before the COVID-19 pandemic, questionnaire lengths varied significantly.[3] Ireland and Canada implemented the longest public servant survey, with 112 questions, followed by the Australian and US surveys (100 questions each). Switzerland implemented the shortest, with 24 questions. Colombia and New Zealand (each approximately 60 questions) and Republic of Korea (48 questions) sat in between.

Longer questionnaires can generate survey fatigue, with potentially greater item nonresponse and survey dropout (Liu and Wronski 2017). For instance—though this is merely suggestive—the correlation coefficient between response rates and questionnaire length in eight of the nine countries reviewed is $r = -0.29$.[4] Question design can potentially mitigate such nonresponse. Chapter 22 of the *Handbook* assesses how to phrase questions so as to minimize item nonresponse.

THE GLOBAL SURVEY OF PUBLIC SERVANTS AS AN INSTRUMENT FOR CROSS-COUNTRY SURVEY HARMONIZATION

As this chapter has illustrated, governments often use dissimilar questions and methodologies to measure similar concepts. As a result, even though governments measure similar concepts, they cannot benchmark themselves against other governments on these concepts. This puts a premium on evidence-based, cross-country harmonization of survey questionnaires and methodologies to further the degree of consistency in measurement across surveys of public servants.

FIGURE 18.4 Share of Public Servants Satisfied with Their Pay and/or Total Benefits, Various Countries



Source: Fukuyama et al. 2022.
Note: Years of measurement vary by country. Colors denote the extent of job satisfaction, with darker shades signifying greater job satisfaction. The gray vertical bars denote 95% confidence intervals.

The GSPS was created with this objective in mind and, more broadly, to encourage the adoption of surveys of public servants by governments, good practice in public servant survey design and implementation, and the collection of cross-country and cross-institution data on public servants in governments around the world (Fukuyama et al. 2022). The aim is to increase the volume, quality, and coherence of survey data on public administration over time. The GSPS is the product of a consortium of researchers and practitioners from Stanford University, University College London (UCL), the University of Nottingham, and the World Bank.

To facilitate the harmonization of survey questions and methodologies for surveys of public servants, the GSPS presents existing questions and methods in an accessible form and provides methodological evidence on the efficacy of these questions and methods. It presents a core module of questions as a proposal for inclusion in independent surveys of public servants and publishes detailed guidance on the implementation of the core module to ease the comparison of any individual survey results with other surveys (Meyer-Sahling et al. 2021). This ensures that the data collected on public servants are comparable across independent data collection exercises.

Figure 18.4 provides an example of the type of comparison possible through the GSPS initiative, benchmarking governments on the percentage of public servants satisfied with their pay and/or total benefits. The GSPS enables governments to understand strengths and areas for development for their civil service in global comparative terms, although, as chapter 24 shows empirically, care needs to be taken when comparing responses across countries for culturally contingent concepts in particular. In figure 18.4, for instance, it is striking how differentially satisfied public servants are with their pay in countries at roughly similar levels of development, such as in the US federal government (63 percent satisfied with their pay) and the UK civil service (36 percent satisfied with their pay). This kind of comparison can help governments understand strengths and areas for development.

CONCLUSION

The number of governments implementing governmentwide surveys of public servants has increased continuously in the last two decades, though many countries have yet to implement or institutionalize the implementation of employee surveys. Our review has shown that surveys of public servants in governments are similar: they tend to measure similar concepts, focusing on a core set of employee attitudes (such as job satisfaction or engagement), on the one hand, and on management practices (such as the quality of leadership), on the other. They are thus implemented with a comparable set of measurement objectives.

At the same time, surveys across governments differ in the methodologies used and the precise measures applied to measure concepts. In terms of methodology, the review has found that surveys differ in key aspects: sampling approaches, survey weighting, survey populations, survey modes, questionnaire length, and response rates achieved. Some of these differences may stem from differences in practical or legal constraints. For instance, the civil service agency (or other entity) in charge of conducting the survey may not have a mandate for personnel management beyond the core civil service, complicating extending the survey coverage beyond the core civil service. And a central human resources management information system with demographic data about civil servants to construct survey weights may or may not be available, as detailed elsewhere in the *Handbook* (chapter 9). Some of the differences, however—for example, in sampling approaches and survey modes—are arguably due to limited methodological evidence on governmentwide surveys of public servants. The remaining chapters of this section of the *Handbook* address part of this void and can help governments make more evidence-based methodological choices in surveys of public servants. The GSPS builds on this evidence to offer governments a globally comparable set of survey questions and methodologies.

In short, the global landscape of surveys of public servants holds much promise for the future. An ever-increasing number of governments are implementing surveys, better evidence for methodological choices in surveys of public servants is becoming available, and the GSPS amplifies opportunities for global benchmarking.

NOTES

1. In line with the varying terminology used by different governments conducting such surveys, we use the terms "public servant surveys" and "government employee surveys" interchangeably.
2. Information about nonresponse weights in Ireland, Republic of Korea, New Zealand, and Switzerland was, unfortunately, unavailable.
3. The pandemic led to a number of additional pandemic and remote-work-related questions in the surveys that would ordinarily not be asked, thus reducing the generalizability of comparisons of questionnaire length during the pandemic.
4. Response rates were unavailable for the Korean survey.

REFERENCES

Australian Public Service Commission. 2021. *Australian Public Service Employee Census: Explanatory Guide 2021*. Canberra: Australian Public Service Commission, Australian Government. https://www.apsc.gov.au/initiatives-and-programs/workforce-information/aps-employee-census-2021#downloads.

Cabinet Office. 2020. *Civil Service People Survey 2020: Technical Guide*. London: Cabinet Office, United Kingdom Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/977279/Civil_Service_People_Survey_2020-_Technical_Guide.pdf.

DANE (Departamento Administrativo Nacional de Estadística). 2020. *Metodología general encuesta sobre ambiente y desempeño institucional—EDI*. DSO-EDI-MET-001, version 4. Bogotá: Dirección de Metodología y Producción Estadística (DIMPE), Departamento Administrativo Nacional de Estadística, Government of Colombia. https://www.dane.gov.co/index.php/estadisticas-por-tema/gobierno/encuesta-sobre-ambiente-y-desempeno-institucional-nacional-edi.

Department of Public Expenditure and Reform (Department of Public Expenditure, National Development Plan Delivery and Reform). 2020. *Civil Service Employee Engagement Survey*. Dublin: Department of Public Expenditure and Reform, Government of Ireland. https://www.gov.ie/en/collection/5e7009-civil-service-employee-engagement-survey/#2020.

Fukuyama, Francis, Daniel Rogger, Zahid Husnain, Katherine Bersch, Dinsha Mistree, Christian Schuster, Kim Sass Mikkelsen, Kerenssa Kay, and Jan-Hinrik Meyer-Sahling. 2022. *Global Survey of Public Servants Indicators*. https://www.globalsurveyofpublicservants.org/.

Government of Canada. 2022. "About the 2022/2023 Public Service Employee Survey." Government of Canada, October 21, 2022 (accessed March 27, 2023), https://www.canada.ca/en/treasury-board-secretariat/services/innovation/public-service-employee-survey/2022-23/about-2022-23-public-service-employee-survey.html.

Korea Institute of Public Administration. 2021. *2021 Public Service Life Survey*. Seoul: Government Data Research Center, Korea Institute of Public Administration (accessed July 1, 2022), https://www.kipa.re.kr/site/kipa/sta/selectReList.do?seSubCode=BIZ017A002.

Liu, Mingnan, and Laura Wronski. 2017. "Examining Completion Rates in Web Surveys via Over 25,000 Real-World Surveys." *Social Science Computer Review* 36 (1): 116–24. https://doi.org/10.1177/0894439317695581.

Meyer-Sahling, Jan, Dinsha Mistree, Kim Mikkelsen, Katherine Bersch, Francis Fukuyama, Kerenssa Kay, Christian Schuster, Zahid Hasnain, and Daniel Rogger. 2021. *The Global Survey of Public Servants: Approach & Conceptual Framework*. Global Survey of Public Servants. https://www.globalsurveyofpublicservants.org/about.

OECD (Organisation for Economic Co-operation and Development). 2016. *Engaging Public Employees for a High-Performing Civil Service*. OECD Public Governance Reviews. Paris: OECD Publishing. https://doi.org/10.1787/9789264267190-en.

OFPER (Office Federal du Personnel). 2022. *Aperçu des résultats de l'enquête 2021 auprès du personnel*. Bern: OFPER, Federal Council, Switzerland (accessed July 1, 2022), https://www.epa.admin.ch/epa/fr/home/themes/politique-du-personnel/enquete-aupres-du-personnel.html.

OPM (Office of Personnel Management). 2020. *2020 OPM Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: US Office of Personnel Management, United States Government. https://www.opm.gov/fevs/reports/technical-reports/technical-report/technical-report/2020/2020-technical-report.pdf.

Research New Zealand. 2021. *Technical Report: Te Taunaki Public Service Census 2021*. Report prepared for the Public Service Commission [Te Kawa Mataaho], New Zealand Government. Wellington: Research New Zealand. https://www.publicservice.govt.nz/research-and-data/te-taunaki-public-service-census-2021/.

Statistics Canada. 2018. "Public Service Employee Survey (PSES): Detailed Information for 2017." Surveys and Statistical Programs, Definitions, Data Sources and Methods, Statistics Canada. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=384108.

CHAPTER 19

Determining Survey Modes and Response Rates

Do Public Officials Respond Differently to Online and In-Person Surveys?

Xu Han, Camille Parker, Daniel Rogger, and Christian Schuster

SUMMARY

Measuring important aspects of public administration, such as the level of motivation of public servants and the quality of management they work under, requires the use of surveys. The choice of survey mode is a key design feature in such exercises and therefore a key factor in our understanding of the state. This chapter presents evidence on the impact of survey mode from an experiment undertaken in Romania that varied whether officials were administered the same survey face-to-face or online. The experiment shows that at the national level, the survey mode does not substantially impact the mean estimates. However, the mode effects have a detectable impact at the organizational level as well as across matched individual respondents. Basic organizational and demographic characteristics explain little of the variation in these effects. The results imply that survey design in public service should pay attention to survey mode, in particular in making fine-grain comparisons across lower-level units of observation.

ANALYTICS IN PRACTICE

- Most governments—and many researchers—running surveys of public officials do so online. This reduces cost, increases flexibility, and theoretically reduces biases, such as those induced by respondents' notions of socially desirable answers.
- However, online surveys tend to have lower response rates than other survey modes and a greater degree of exit before surveys are completed, leading to different samples of respondents. This raises the concern

Xu Han was a consultant at the World Bank. Camille Parker is an economist at the United States Agency for International Development. Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Christian Schuster is a professor at University College London.

that the data resulting from online surveys are not a valid representation of the population—in this case, the entire public administration.

- This chapter presents evidence from a randomized controlled trial that compares face-to-face and online survey responses. Our intention is to showcase an approach to measurement validation that can be followed by other survey teams for understanding the validity of their analyses.
- We show that the mean difference between online and face-to-face responses across all officials, which we call the *national level*, is between 0.17 and 0.35 standard deviations. Such an effect is of a similar magnitude to moving from 4.4 to 4.5 on a 1–5 scoring system (for example, "strongly disagree" to "strongly agree") on one of the aggregate variables we study. Thus, in surveys with similar mode effects, measurement mode is unlikely to make a qualitative difference to conclusions when reporting at the national level so long as such small deviations are not overanalyzed.
- At the organizational level, the modal difference across all questions is roughly consistent with the country-level average. However, several organizations exhibit a modal difference of over one standard deviation. Given the lack of objective benchmarks, we interpret sensitivity to mode as indicative of underlying measurement issues. Problems arising from sensitivity to measurement are particularly acute when ranking organizations, with mode effects having substantial impacts on the ordering of organizations. This evidence casts doubt on the validity of organization-level ranking that does not appropriately address these measurement concerns.
- At the individual level, the mode effects remain significant and substantial for most of the outcomes. We see that the survey mode effects persist across individuals matched using propensity score matching (PSM) as well across different groups, like managers and nonmanagers, although some groups appear more sensitive to survey mode than others. This evidence places a burden of proof on survey analysts to demonstrate the validity of presenting data at the unit or individual level.
- A common approach to correcting online surveys is to use survey weights. In our experiment, we find little evidence that survey weights reduce the sensitivity of results to the measurement approach.
- Identifying organizations and individuals particularly susceptible to mode effects would allow for a significant reduction in aggregate mode effects. This might be pursued through a small, face-to-face survey across organizations, upon which estimates of individual mode responses could be based.

INTRODUCTION

Measuring many aspects of public servants and their working lives is difficult. Management quality is frequently experienced rather than recorded in administrative data. Public employees' motivations are difficult to observe outside of their own expressions of their motives. Thus, self-reporting through surveys becomes the primary means of measurement for many aspects of officialdom. Externally sourced measures, perhaps from administrative data, are simply unable to record features of these important variables.

Survey design is therefore an important mediator in our understanding of the state. This part of *The Government Analytics Handbook* assesses how to determine the particular content of a survey of public servants from multiple angles. This chapter focuses on a key aspect of survey administration: whether the survey is conducted online or in person (that is, face-to-face). Though there are other modes of survey delivery, from periodic text message surveys to laboratory-in-the-field games, the debate in this context typically concerns these two forms, which will therefore be our focus (Haan et al. 2017; Heerwegh and Loosveldt 2008; Kaminska and Foulsham 2014; Tourangeau and Yan 2007).

Public servant surveys run by governments are typically carried out online, with a small or nonexistent proportion of staff allowed to use a paper form or speak to an enumerator directly (for a review of the most prominent such surveys, see chapter 18). This is done predominantly for cost reasons, but online surveys enjoy several advantages. They enable researchers to rapidly collect large amounts of data and can be quickly and flexibly deployed across a range of organizational contexts.

However, this reliance on online surveys is based on the rarely tested assumption that online surveys are able to provide valid and reliable data. This assumption may be incorrect for several reasons: online surveys often suffer from low response rates, potentially undermining the representativeness of the respondent group (Cornesse and Bošnjak 2018). Online surveys are also associated with higher levels of survey drop-off and item nonresponse than other survey modes (Daikeler, Bošnjak, and Lozar Manfreda 2020; Heerwegh and Loosveldt 2008; Peytchev 2009). The resulting higher levels of missing values may undermine the validity and reliability of the data (Baumgartner and Steenkamp 2001; Jensen, Li, and Rahman 2010; Podsakoff et al. 2003).

Face-to-face surveys can be a viable alternative to online survey data collection. Many microempirical studies, in which the individual is taken as the unit of observation, prefer to administer surveys in person. Although they consume significantly more time and resources than online surveys, face-to-face surveys tend to report significantly higher response rates and lower rates of breakoff, and they can be substantially longer without respondent exit. Talking to someone in person is a fundamentally more engaging activity than filling in a form on the screen, enabling a wider range of data to be collected from a single interview.[1] It is therefore possible that the final set of responses collected from an online survey will come from a different effective sample than would be the case in the face-to-face mode (see, for example, Couper et al. 2007).

We turn now from respondents to the answers they provide. A key feature of online surveying is that it distances the respondent from an enumerator. This potentially reduces social-desirability bias arising from a respondent's inclination to answer in a way that may be demanded by the social features of a face-to-face survey (Heerwegh 2009; Newman et al. 2002; Tourangeau and Yan 2007). An online survey is also relatively consistent in its delivery of a survey to respondents, while individual enumerators may not be.

Despite the potential reduction in social-desirability bias (Ye, Fulton, and Tourangeau 2011), online surveys may introduce other biases—for example, those derived from a lack of comprehension of the question. Where enumerators can provide clarifications, online surveys typically do not have that option, nor is it likely to be regularly used by respondents. It has also been shown that the online survey respondents engage in a larger degree of *satisficing*—that is, they more often respond "I don't know," skip questions, choose neutral response options, etc. to minimize the cognitive burden of responding (see, for example, Heerwegh and Loosveldt 2008; Krosnick and Presser 2010; see section two below for further discussion). Whereas the desire to satisfice is also present in face-to-face surveys, an experienced enumerator might probe respondents to, for example, think for a while about a question rather than saying "I don't know." Therefore, another concern is that even comparable samples of respondents may provide different responses if surveyed using different survey modes.

A series of trade-offs therefore characterizes the choice between online and face-to-face survey modes. Conceptually, there may be differences in what sample of respondents each mode attracts and how the mode affects the responses they provide. Practically, the costs and feasible lengths of the two approaches differ. While researchers and research communities typically have strong beliefs about which approach optimally resolves this tension, there is little to no rigorous empirical evidence on this subject in the field of public administration.[2]

The nature of public administration, with its hierarchical and bureaucratic communication norms, potentially implies a substantial survey mode effect. For example, written communication at work, such as filling in an online form or survey, may be regarded very differently by a public official and a private citizen. On the other hand, a 1-hour meeting to discuss public service life is similar to many of the meetings public officials have in a day. Findings from other sectors may therefore not be externally valid in a public administration setting.

What, therefore, are public sector managers or researchers to do in collecting survey data from public servants? This question is complicated by the fact that many features of public administration, as noted above, cannot be definitively validated outside of survey data. It can be argued that the appropriate conception of management is the individual employee's specific experience of it. Thus, objective data for the purpose of benchmarking the two most common survey modes are absent for many topics. The answer to the question may also vary across topics, individuals, and settings, such that an effective answer must go beyond a simple comparison of aggregate means to understand what quantities are most affected by survey mode.

While the existing literature is an obvious foundation for our analysis, our aim in this chapter is to investigate the robustness of survey results to survey mode within a public administration setting. Given the difficulties of generating objective benchmarks for many of the topics we study, our interpretation of this robustness is used as an indicator of the validity of the underlying responses. Where feasible, we also investigate the organizational and individual determinants of mode effects, with the aim of better understanding which groups or organizations may be most impacted by differences in survey mode.

Our intention in this chapter is to showcase to survey managers and related stakeholders an approach to testing the robustness of survey responses to survey mode. We provide evidence from a single experiment to illustrate our approach, but in doing so, we provide some of the first experimental evidence on the impacts of survey mode in public administration. As such, this chapter hopes to provide frontier evidence from a single setting and a framework for investigating these issues in other surveys.

The rest of this chapter proceeds as follows. Section two outlines the existing literature on survey mode effects and how it relates to the public administration setting. A major gap in the literature on mode effects in surveys of public servants is the absence of an experimental comparison between the two modes. We address this gap through a field experiment with 6,037 public servants in 81 government institutions in Romania, in which we randomly assign each official to complete either a face-to-face or an online survey. The survey's content replicates that found in typical government employee surveys, covering both employee attitudes and management practices. By studying survey responses across the two modes with a high degree of heterogeneity in response rates, we can disentangle survey mode effects at the point of response from nonresponse bias due to the lower take-up of online compared to face-to-face surveys. Given the frontier nature of this empirical evidence, sections three to five investigate the impacts of survey mode within this data set. Section six discusses the implications of our findings for the implementation of public servant surveys and further research.

LITERATURE REVIEW

The existing literature on survey mode effects in general finds that the survey mode has significant impacts on the robustness of survey estimates across three primary dimensions: response rates, survey breakoff, and survey responses.

Response Rates

Much of the existing research on survey modes has focused on the difference in response rates between modes. In general, online surveys have been found to have significantly lower response rates compared to all other survey modes, including face-to-face (Biemer et al. 2018; Lozar Manfreda et al. 2008; Shih and Fan 2008). While not specific to public administration, a recent meta-analysis conducted by Daikeler, Bošnjak, and Lozar Manfreda (2020) summarizes the results of 114 experimental studies conducted among many different populations (students, the general public, businesses, and employees), on diverse topics (public opinion, technology, lifestyle, job, etc.), by various sponsors (academic, governmental, and commercial), both with and without participation incentives, and with varying recruitment strategies, prenotification methods, and solicitation methods. They found that in aggregate, online surveys have response rates that are 12 percentage points lower than all other survey modes.[3]

Those who do respond to online surveys tend to differ from respondents to other survey modes across several demographic characteristics, spurring concerns over the representativeness of online samples. For instance, several studies have found that online survey respondents tend to be younger and more educated than face-to-face survey respondents (Braekman et al. 2020; Couper et al. 2007; Duffy et al. 2005). A recent meta-analysis suggests that online surveys are associated with higher nonresponse biases than other survey modes (Cornesse and Bošnjak 2018). It is also worth noting that differences between respondents and nonrespondents are attributed more to the noncoverage of some population subgroups in the sample frame than to the nonresponse of people invited to participate in surveys (Couper et al. 2007). Online surveys of public servants are more likely to have a complete sample frame and, therefore, are less susceptible to nonresponse biases than online surveys of general populations.

Within public administration, there is a high level of heterogeneity in terms of response rates to existing large-scale, online public administration surveys in Organisation for Economic Co-operation and Development (OECD) countries. As shown in table 18.2 in chapter 18, while some large-scale public administration surveys, such as the survey administered in Colombia, enjoy response rates around 80 or 90 percent, others, such as the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) in the United States, have struggled to bring their response rates above 50 percent and have been experiencing a steady decline in overall response rates in the past five years.[4] Public administration surveys in non-OECD countries exhibit similarly heterogeneous response rates, ranging from 11 percent in Brazil to 47 percent in Albania. Troublingly, despite these surveys' importance in shaping public administration organizations' priorities as they relate to hiring, employee engagement, and performance management, among other topics, the question of whether declining response rates to online surveys present a threat to the overall validity of inferences about public officials drawn from the data has not been extensively studied in the public administration literature.

While response rates for country surveys tend to remain relatively consistent at the national level over time, there is a high degree of variation in survey response rates at the organizational level. For example, in the 2019 FEVS, response rates within US government organizations ranged from 86 percent to just 27 percent. While research on survey response rates in public administration is limited, the research that does exist posits several potential explanations for this variation at the organizational level. Some researchers have argued that low employee morale in certain agencies may contribute to declining response rates (de la Rocha 2015). Others, while *not* explicitly studying survey response, have found a positive relationship between voluntary behavior (such as taking a survey) and employee engagement levels (Rich, Lepine, and Crawford 2010), suggesting that organizations with higher levels of employee engagement may also experience higher response rates to employee surveys. Similarly, public employees with strong public service motivation or organizational commitment have been found to be more willing to perform extra-role tasks, including filling out surveys (Moynihan and Pandey 2010; Newell et al. 2010). Other researchers have identified links between response rates and individuals' attitudes toward the survey's sponsor institution. For instance, in a study of university students, Spitzmüller et al. (2006) find that survey nonrespondents are less likely to believe that their university values their contributions or cares about their well-being.

These differences between online respondents and nonrespondents to government surveys suggest that variation in response rates may significantly impact the degree to which online surveys provide unbiased estimates of public employees' perceptions and behaviors. In addition, the proclivity of managers and researchers to compare survey responses across organizations or other subgroups means that variation in response rates may lead to the comparison of differential subgroups of staff (Groves 2006). The self-selection issues in public administration surveys are less of a concern in the face-to-face mode because most surveys of this type record response rates close to 100 percent. For example, the Romanian face-to-face survey analyzed here collected responses from 3,316 out of 3,592 sampled individuals, yielding a response rate of 92 percent. Similar surveys in different settings give comparably high response rates: for example, Guatemala (96 percent) and Ethiopia (94 percent). Assuming successful random sampling, the almost-perfect response rate minimizes any issues arising from differences between survey respondents and nonrespondents in the face-to-face mode.

Survey Breakoff

Beyond impacting survey estimates through differential response rates, the survey mode can also impact survey estimates through different rates of breakoff. Overall, online surveys are associated with significantly higher rates of survey breakoff because they are generally less able to maintain respondents' interest and attention throughout the duration of the survey (Galesic 2006; Haan et al. 2017; Heerwegh and Loosveldt 2008; Kaminska and Foulsham 2014; Krosnick and Presser 2010; Peytchev 2009; Steinbrecher, Roßmann, and Blumenstiel 2015). This threat of breakoff can be significant: meta-analyses of the issue have found online surveys experience breakoff rates between 16 and 34 percent (Lozar Manfreda and Vehovar 2002; Musch and Reips 2000).

The ability to maintain respondents' interest throughout the survey varies depending on several survey design features, including the presence of long blocks of questions and the overall time it takes to complete the survey (Galesic 2006; Peytchev 2009; Steinbrecher, Roßmann, and Blumenstiel 2015). Many of the demographic characteristics associated with survey response are also associated with higher levels of survey breakoff, with younger, more educated respondents generally being more likely to exit an online survey before completing it (Peytchev 2009). We provide more information on this in chapter 22.

Within the public administration sector, the issue of survey breakoff has not been extensively studied, and statistics on survey breakoff in major public administration surveys, such as the FEVS, are generally not made publicly available. In the 2019 survey of the Australian Public Service, approximately 92.5 percent of respondents who began the survey completed it, for a breakoff rate of 7.5 percent (N. Borgelt, Australian Public Service Commission, pers. comm., June 24, 2020). Consistent with the survey research literature, breakoff was the highest among long blocks of matrix-style questions and questions involving a reasonably high cognitive load (such as a question asking respondents how many sick days they had taken over the last 12 months) (Peytchev 2009; Steinbrecher, Roßmann, and Blumenstiel 2015; Tourangeau, Conrad, and Couper 2013).

This evidence implies a similar concern as the above for valid inference. Comparisons of questions with higher and lower rates of breakoff may differ simply due to the subgroups that respond to them and are thus vulnerable to endogenous selection concerns. If the most self-motivated individuals are more likely to respond to motivation questions, then an analysis of these variables relative to management variables may incorrectly imply the relative importance of self-motivation over management. Once again, this issue is often minimized by a face-to-face survey interview. Such settings make the survey process more engaging to the respondent and add a possible social cost to ending the interview midstream, as this might be seen as "impolite" to the enumerator (Peytchev 2006).

Survey Responses

Finally, a substantial portion of the existing survey research literature has focused on the degree to which survey modes may impact the magnitude of survey responses directly. In general, online survey respondents tend to exhibit lower levels of motivation to answer survey questions and often pay less attention when answering questions compared to face-to-face respondents (Kaminska and Foulsham 2014; Krosnick 1991). Several studies have found that online surveys are associated with higher rates of satisficing behaviors, including selecting "I don't know" or "N/A" response options, providing less differentiation across groups of responses, and providing more neutral responses (for example, "Neither agree nor disagree" or "Neutral") than face-to-face surveys (Duffy et al. 2005; Haan et al. 2017; Heerwegh and Loosveldt 2008). Online surveys are also more likely to produce noncontingent responses (NCR), wherein there is a substantial difference between survey items that are expected to be highly correlated with each other (Heerwegh and Loosveldt 2008; Krosnick and Presser 2010). These kinds of responses imply that respondents may have simply selected answers at random or read through survey items carelessly in order to quickly complete the survey (Anduiza and Galais 2017). Taken together, these satisficing behaviors can reduce the validity and reliability of online responses (Baumgartner and Steenkamp 2001; Podsakoff et al. 2003).

At the same time, however, the existing literature suggests that online surveys may be better at eliciting candid responses to sensitive questions. Because online surveys provide respondents with a higher level of anonymity than face-to-face surveys, online survey respondents tend to be more likely to respond truthfully to questions related to socially sensitive topics (Gnambs and Kaspar 2015; Kays, Gathercoal, and Buhrow 2012; Tourangeau and Yan 2007). In the context of public administration, these findings suggest that online surveys may be particularly advantageous when measuring sensitive topics, such as ethics violations, turnover, or evaluations of organizational performance. However, the applicability of these findings to public administration has not been rigorously studied, and there is limited knowledge about the relevance of survey mode on the validity of data collected through these studies.

A SURVEY MODE EFFECTS EXPERIMENT

We address a number of these gaps in the existing literature on mode effects through a field experiment with 6,037 public servants in 81 government institutions in Romania. We randomly assigned each target respondent to complete either a face-to-face or an online survey covering several topics typical of public administration surveys: recruitment, performance appraisal, turnover, dismissal, salary, motivation, goal-setting, leadership, and ethics.[5]

How Does the Survey Mode Impact Response Rates?

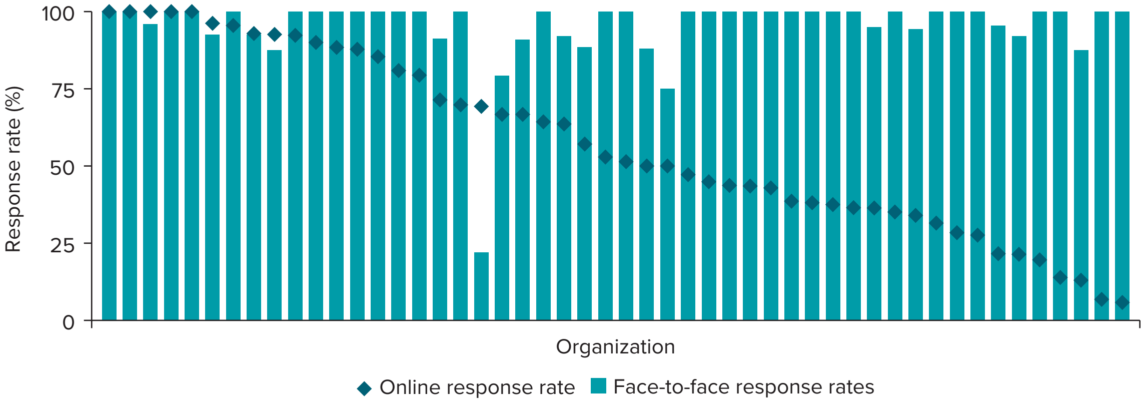
Our face-to-face survey has high response rates across most government institutions, with an average of 92.5 percent, while our online response rate—consistent with other online government employee surveys—varies widely across government institutions and ranges from a maximum value of 100 percent (5 organizations) to a minimum of 0 percent (13 organizations). For the purposes of this analysis, we remove both face-to-face and online observations from organizations who declined to participate in the online survey, as well as organizations with online response rates of less than 5 percent.[6] After this removal, the sample comprises of 4,819 public servants in 50 government institutions. Figure 19.1 presents the remaining heterogeneity in organizational response rates, with an average response rate across organizations of 86.2 percent in the face-to-face mode and 53.8 percent in the online mode.

We use heterogeneity in online response rates across organizations to disentangle survey mode effects at the point of response from nonresponse bias due to lower take-up of online surveys. By comparing questions in high online-response organizations with their face-to-face equivalents, we can abstract from selection bias. By comparing bias across the full sample, we can investigate the role of response rate in question differences.[7]

How Does the Survey Mode Affect the Distribution of Respondent Characteristics?

Table 19.1 shows the results of *t*-tests conducted between the online and face-to-face survey samples across several key demographic groups. Given that our face-to-face survey is a representative sample from staff lists and has a high average response rate, it can be seen as a reflection of the true distribution of characteristics of

FIGURE 19.1 Online and Face-to-Face Survey Response Rates, by Organization



Source: Original figure for this publication.

TABLE 19.1 Balance in Demographic Characteristics between Surveys

Variable	*N*	(1) Face-to-face sample mean [SE]	*N*	(2) Online sample mean [SE]	*T*-test difference (2)–(1)
Age	2,137	45.804 [0.191]	2,682	45.392 [0.167]	−0.412
Years worked in position	2,137	7.423 [0.136]	2,682	8.029 [0.132]	0.607***
Years worked in organization	2,137	11.565 [0.174]	2,682	10.828 [0.149]	−0.737***
Years worked in public administration	2,137	14.719 [0.175]	2,682	13.893 [0.154]	−0.826***
Employee status (1 = Civil servant)	2,137	0.873 [007]	2,682	0.91 [0.006]	0.037***
Gender (1 = Male)	2,137	0.31 [0.01]	2,682	0.26 [0.008]	−0.051***
Highest level of education attained: less than college (1 = Yes)	2,137	0.033 [0.004]	2,682	0.04 [0.004]	0.006
Highest level of education attained: undergraduate degree (1 = Yes)	2,137	0.474 [0.011]	2,682	0.433 [0.01]	−0.041***
Highest level of education attained: master's degree (1 = Yes)	2,137	0.453 [0.011]	2,682	0.481 [0.01]	0.028*
Highest level of education attained: PhD (1 = Yes)	2,137	0.035 [0.004]	2,682	0.037 [0.004]	0.001

Source: Original table for this publication.
Note: The values displayed for *t*-tests are the differences in means between the two survey modes (face-to-face and online). Significance level: * = 5 percent, ** = 1 percent, *** = 0.1 percent.

public servants. Thus, differences between the two reflect a deviation of the online survey from a representative sample.

Consistent with the existing literature, we find many statistically significant (at the 1 percent level) deviations from the population's values in the sample of online survey respondents. Most noticeably, 31 percent of face-to-face survey respondents are male, compared to only 26 percent female.[8] They are also relatively less educated, with 47.4 percent having an undergraduate degree and 48.8 percent having a Master's degree or PhD, whereas, for online respondents, these numbers stand at 43.3 percent and 51.8 percent, respectively.[9] Moreover, 87.3 percent of face-to-face respondents are civil servants (as opposed to contractors), compared to 91 percent of online respondents. We also find statistically significant differences in average tenure, but being below one year, these differences appear to be of limited magnitude.

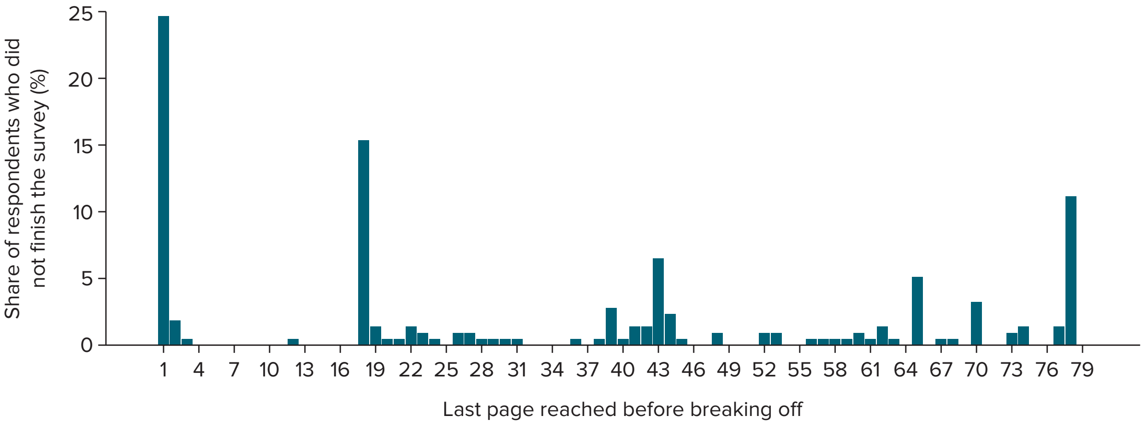
Overall, our data reflect the frequent finding that face-to-face and online samples differ along a range of margins. As many of these variables, like gender, education, and contract status, can affect survey responses, table 19.1 provides an initial rationale to look deeper into the differences between modes in the Romania survey.

How Does the Survey Mode Affect Survey Breakoff and Item Nonresponse?

Our online survey also exhibits considerably higher levels of survey breakoff than the face-to-face survey. While the breakoff rate for the face-to-face survey is almost zero, the breakoff rate for the online survey is approximately 10 percent (see figure 19.2 below, as well as figure G.2 in appendix G for the breakoff pattern by mode). While many major civil service surveys do not generally publicize their levels of survey breakoff, the evidence that does exist suggests that the breakoff rate in our survey is, generally speaking, consistent with similar public administration surveys and lower than average for surveys in general. For example, in 2019, the Australian Public Service Employee Census had a breakoff rate of 7.5 percent in its online survey (N. Borgelt, Australian Public Service Commission, pers. comm., June 24, 2020). Overall, online surveys of the general population experience an average breakoff rate of 16–34 percent (Lozar Manfreda and Vehovar 2002; Musch and Reips 2000), which suggests that civil servants are more likely to complete a survey once started.

Interestingly, as shown in figure 19.2, the largest proportion (just under a quarter of the total) of survey breakoff in the online survey occurred on the first page, suggesting that encouraging individuals to start the survey is the biggest hurdle to obtaining a complete response.[10] Survival analysis conducted on the profile of individuals who dropped out of the survey (using a Cox-Weibull hazard model) finds that demographic characteristics are poor predictors of breakoff. Only the age variable appears to have a relatively consistent impact on breakoff, with individual age, as well as average age at the organization as a whole, increasing the chances of respondents' finishing the survey (for a full summary of findings, see appendix G, table G.2).

FIGURE 19.2 Online Survey Breakoff, by Page



Source: Original figure for this publication.
Note: Minimum page = 1; maximum page = 79.

In addition to analyzing the individuals who dropped out of the survey, we also examine the profile of those who dropped out of the survey and returned to complete it later. Overall, 326 individuals dropped out of the online survey and returned to complete it later.[11] The vast majority of these individuals (80 percent) returned to the survey within one day of exiting it. However, several individuals did not return to the survey for several weeks, suggesting that subsequent reminders to complete the survey may have spurred them to revisit it.[12] There are no notable demographic differences between these individuals and the broader survey sample.

Table 19A.3 also shows that even the individuals who do not exit the online survey altogether are less likely to provide responses. The online mode of delivery is associated with all types of item nonresponse, with individuals being more likely to say "I don't know," to refuse to respond, and to skip questions. Chapter 22 discusses in greater detail the issues and determinants of item nonresponse, so here we only note that apart from larger survey nonresponse, differential demographic characteristics, and higher breakoff rate, the rate at which respondents omit particular questions should also be on the radar of researchers using online surveys, as this value is significantly larger than in equivalent face-to-face surveys.

SURVEY MODE EFFECTS ON THE VALIDITY AND RELIABILITY OF DATA

As seen above, online surveys have lower response rates, attract a nonrepresentative sample of the survey population, and suffer from survey exit more frequently than face-to-face surveys. This suggests that the *process* of responding to an online survey differs from the process of responding to a face-to-face one. But the critical question is whether any of this matters for the *measurement of outcomes* that the surveys yield. Since we undertake a randomized controlled trial that exogenously separates individual respondents into in-person and online enumeration modes, we can compare the results reached by these two methods to investigate the validity and reliability of the corresponding data. These are clearly the two most important outcomes of any change in measurement approach.

As described above, assessing which survey is best able to reflect the underlying truth is complicated by the fact that the survey mode impacts responses directly as well as through sample selection. Since we are dealing with concepts such as management and motivation that are difficult to proxy with objective data in public administration settings, our focus is on investigating the scale and determinants of any difference in the quantities the two modes yield. We interpret significant changes in question outcomes as implying vulnerability to measurement outcomes, thereby undermining the robustness of our estimates from any single approach.

Does the Survey Mode Make a Difference to Question Values?

In order to ascertain the degree to which the survey mode impacts survey estimates, we undertake an analysis with respect to the mean mode difference in survey question responses. We average the responses into three indexes: management, motivation, and ethics. In all three cases, higher index values indicate more-positive, or "desirable," traits, like exemplary leadership, job satisfaction, and aversion to bribe-taking.[13] The management index presents the average of a series of survey items related to managerial practices and performance management. The motivation index shows the average of survey items related to employees' levels of motivation and engagement in their work. Finally, the ethics index aggregates the average of survey items related to employees' perception of the prevalence of ethics violations in their organization. These dimensions reflect three of the most commonly investigated areas of public sector life in public servant surveys (see figures 18.2 and 18.3 in chapter 18).

In all instances, we compare the survey mode effects by calculating the mean response from the online survey minus the mean response from the face-to-face survey. A negative mean difference thus implies that the face-to-face survey produces higher average estimates (that is, more-positive responses) than the

TABLE 19.2 Mean Modal Difference, by Level of Analysis

	Mean	Minimum	Maximum	p25	p50	p75
(1) National level						
Management index	−0.239					
Motivation index	−0.350					
Ethics index	−0.208					
(2) Organizational level						
Management index	−0.331	−1.925	0.978	−0.617	−0.258	0.081
Motivation index	−0.308	−1.194	0.831	−0.660	−0.348	−0.039
Ethics index	−0.171	−1.430	1.099	−0.401	−0.134	0.121
(3) Individual level						
Management index	−0.242	−4.600	3.864	−1.196	−0.255	0.692
Motivation index	−0.312	−8.365	5.611	−1.247	−0.312	0.623
Ethics index	−0.196	−7.879	7.879	−0.563	0.000	0.000

Source: Original table for this publication.
Note: Panel (1) shows the full-sample differences in the means of the indexes between the online and face-to-face survey modes ($\hat{x}_{online} - \hat{x}_{f2f}$). Panel (2) calculates these differences at the level of each organization and summarizes their values for mean level $\left(\left[\frac{1}{50}\right]\left[\sum_{org=1}^{50}\left[\hat{x}_{org,online} - \hat{x}_{org,f2f}\right]\right]\right)$ and other key distribution statistics. Panel (3) shows the distribution of differences in index values between individuals matched on the following variables: organization, job tenure, organization tenure, public administration tenure, pay grade, employee status (civil servant vs. contractual staff), age, gender, and education level. Propensity score matching estimators impute the missing potential outcomes for each treated subject by using the average of the outcomes of similar subjects that receive the other treatment. Observations are matched using nearest-neighbor matching and the probability of treatment is calculated using a logit model. In the case of a tie, observations are matched with all ties with the corresponding difference averaged out.

online survey. For ease of interpretation and unless otherwise indicated, the differences are presented in terms of z-scores, so coefficients are in standard deviations.[14] Table 19.2 presents the mean survey mode effects across statistics calculated at the national, organizational, and individual levels. These three levels are discussed in turn below.

Country-Level Quantities

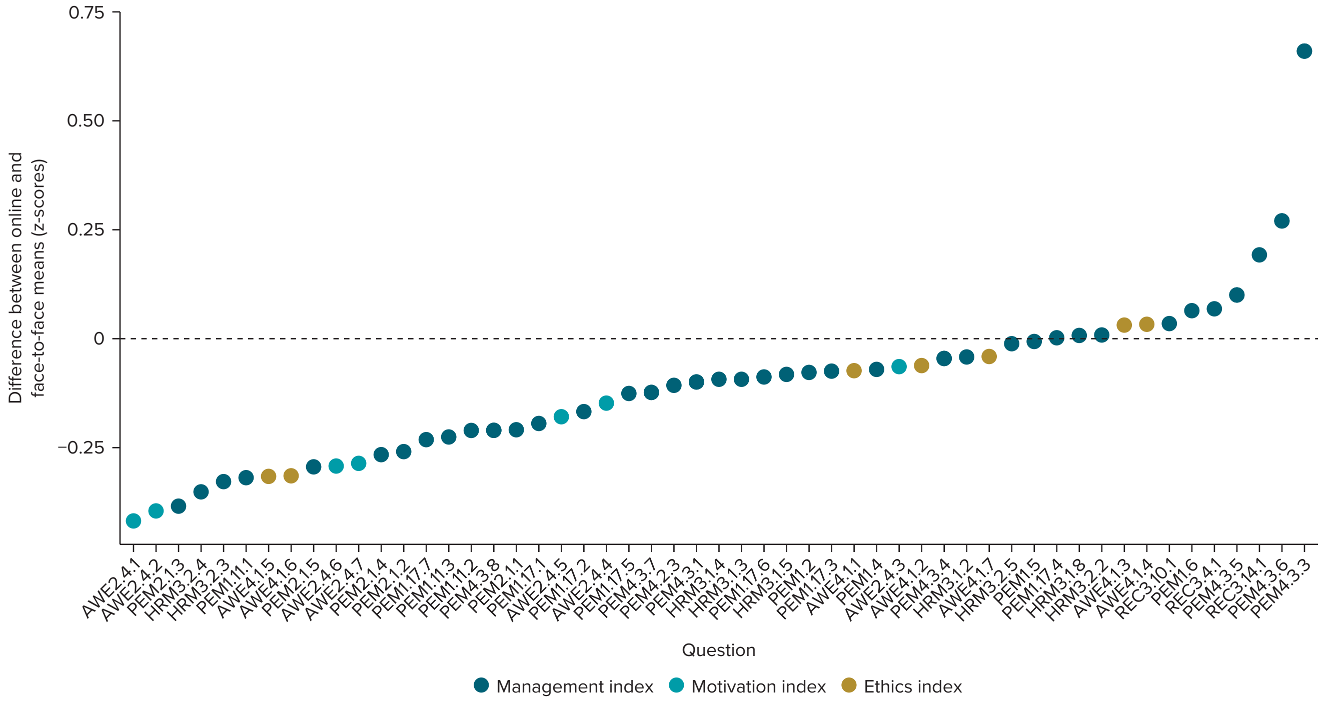
At the national level (panel 1 of table 19.2), we calculate the mean difference across all civil servants as the average score of the index in the online sample minus the average score on the same index in the face-to-face sample.

We see that the differences range from −0.208 for the ethics index, through −0.239 for the motivation index, to −0.350 for the management index. All of the average modal differences are negative, implying that the estimates produced by face-to-face surveys are, on average, higher and therefore point toward more-positive, or “desirable,” responses than those produced by online surveys.

The effect size of these differences on a 1–5 Likert scale is moving the average around 0.1 higher for the face-to-face sample than the online sample. Thus, the evidence from this experiment is that survey mode effects are small for most questions in data aggregated across all respondents. Reporting at this level seems relatively robust to the mode of data collection.

The average survey mode effects are an artifact of the survey mode effects associated with the particular questions composing a given index. Figure 19.3 presents survey mode effects by question item across all items included in the three indexes outlined above.[15] The survey mode effects vary considerably among individual question items for each index. Some items within each of the indexes are more sensitive to survey mode effects than others (Braekman et al. 2020; Gnambs and Kaspar 2015; Ye, Fulton, and Tourangeau 2011).

FIGURE 19.3 Average Item Difference, by Survey Topic



Source: Original figure for this publication.
Note: For the question text, see table G.8 in appendix G. AWE = attitudes and work environment; HRM = human resources management; PEM = performance management; REC = recruitment.

For instance, while the ethics index as a whole exhibits significant negative survey mode effects, at the item level, two items ("How frequently do employees in your institution observe unethical behavior among colleagues?" and "How frequently do employees in your institution report colleagues for not behaving ethically?") appear highly sensitive to survey mode, with mean differences of −0.40 and −0.37 standard deviations, respectively. The three other items that compose the ethics index ("How frequently do employees accept gifts or money from companies?," "How frequently do employees accept gifts or money from citizens?," and "How frequently do employees pressure other employees not to speak out against unethical behavior?") all have mean mode differences close to zero.[16]

At the national level, all of the mode effects exhibited in figure 19.3 are within relatively limited thresholds. Even for topics such as ethics, we find limited average mode effects across the population.

Organization-Level Quantities

At the organizational level (panel 2 of table 19.2), we calculate the mean difference as the average difference in online and face-to-face scores across each organization. For example, an organization's management index score as determined by the results of the face-to-face survey is subtracted from an organization's management index score as determined by the results of the online survey. These differences within organizations are then averaged to produce the mean difference in index scores. Other statistics relating to the distribution of scores across organizations are also shown.

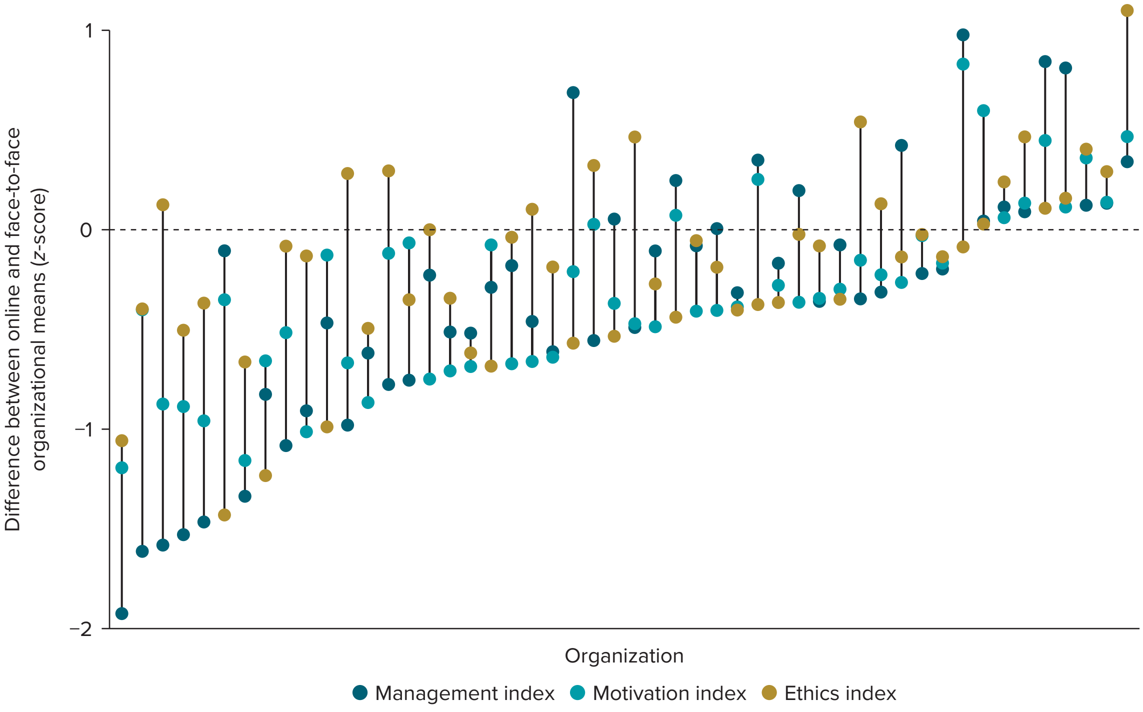
The *average* coefficients at the organizational level are not unlike those at the national level (perhaps naturally, since we are now simply producing a weighted correspondence of the national statistics). The change relative to the national level is the largest for the management index, where the mode difference increases by 38 percent. Still, the overall magnitude and direction of mean mode differences point us to the same conclusion of more-negative responses in the online mode.

However, we also see a high degree of heterogeneity in mode effects across organizations, implying that organizational characteristics may mediate respondents' experience of the survey and its mode of delivery. As shown in figure 19.4, organizations present highly varied responses to the mode of measurement. For instance, while the average mode difference across organizations for the management index is 0.331 standard deviations, seven organizations display differences above one standard deviation between the survey modes on that index. Given that the difference between organizations scoring the lowest and the highest on the management index is just above two standard deviations, this value implies a considerable impact of the survey mode on respondents *within* some organizations. Comparably large differences for some organizations are also observed for other indexes. Figure 19.4 further confirms that the survey mode effects differ across topics, as some organizations have largely different mode effects depending on the index chosen.[17]

Thus, in statistics produced at the organizational level, we start to see substantial effects of the mode of measurement, especially for a subportion of our sample. Ordinary least squares (OLS) regressions examining the relationship between the aggregate mean difference and organizational characteristics, such as organization size, gender composition, and average age, provide little evidence of the determinants of mode effects. This suggests that it is organizational characteristics typically unobservable in a public officials survey that are driving survey mode effects (for a full summary of results, see appendix G, table G.4).

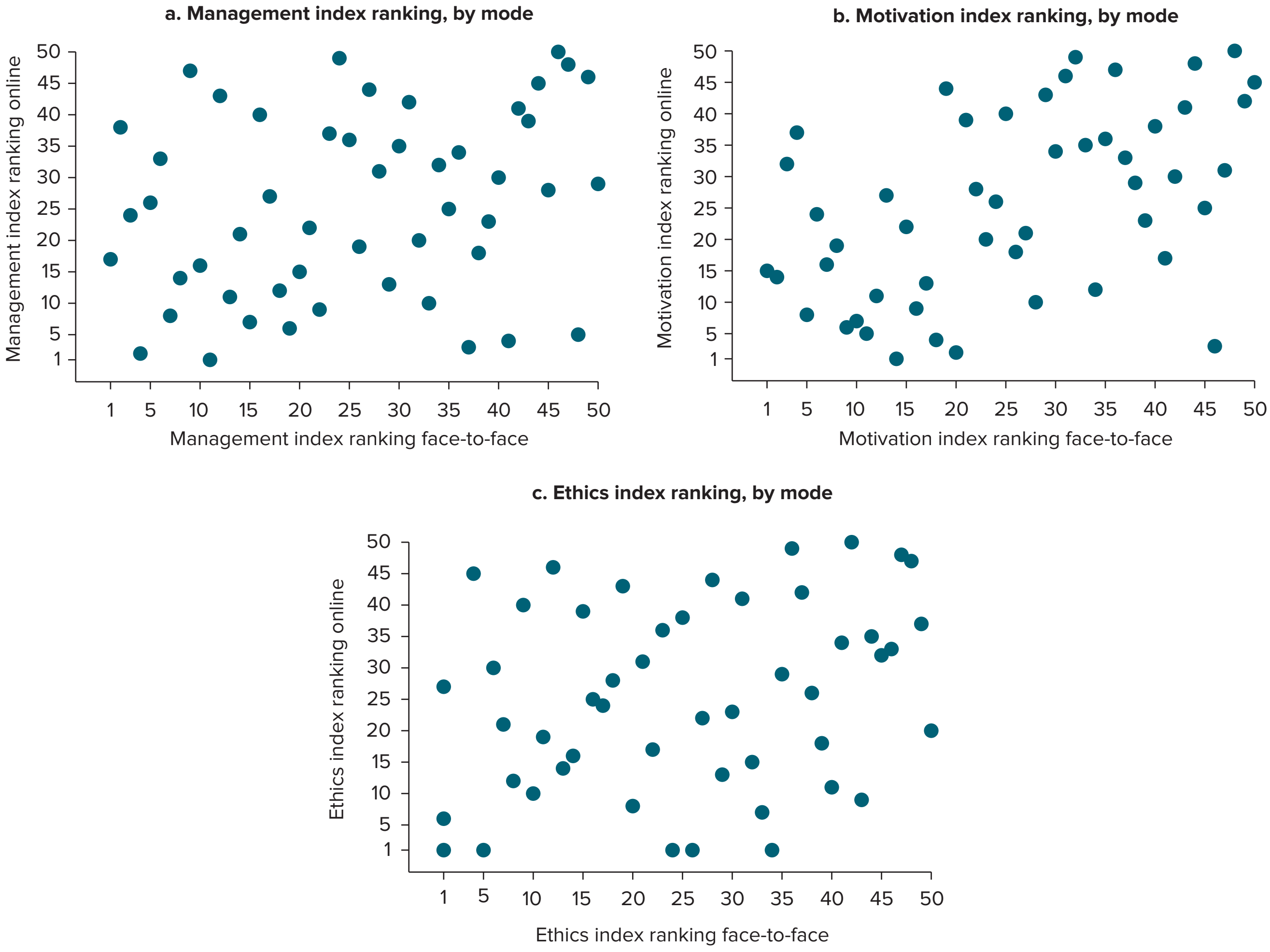
Building on the discussion in chapter 20, these heterogeneous mode effects at the organizational level are of particular concern to policy makers if they intend to present survey results as organizational rankings. Specifically, we find that the rank of a public sector organization (that is, its place on a list of organizations sorted in descending order of the value of a given index) as determined by the online survey correlates only poorly with its rank as determined by the face-to-face survey, across all three indexes.[18] Figure 19.5 plots organizations' ranks according to the face-to-face (*x* axis) and online

FIGURE 19.4 Average Modal Difference, by Organization



Source: Original figure for this publication.

FIGURE 19.5 Organization Rankings, by Index and Mode



Source: Original figure for this publication.

(y axis) surveys for the three indexes we focus on.[19] The low rank correlation between the two modes of measurement implies that such rankings are highly sensitive to measurement effects. The correlation coefficient is highest for the motivation index (coef. = 0.494, p-value = 0.00), followed by the ethics index (coef. = 0.270, p = 0.060) and the management index (coef. = 0.264, p = 0.063).

Looking at the quintile distribution of organizations across modes is even more suggestive. Out of 50 organizations included in the sample, two-thirds or more are in a different quintile when comparing face-to-face and online rankings. For the management index, 37 organizations change quintile, depending on which mode we use to rank the organizations. For the motivation index, this value is 33, and for the ethics index, it is 38 organizations.

All this suggests that benchmarking public sector organizations using employee survey results—a practice currently undertaken by several major public administration surveys—can be highly dependent on methodological choices like survey mode. These are rarely explicitly discussed in this context yet largely shape these rankings. Changes in the relative ranking of organizations may very likely be due to measurement rather than real changes in the underlying variables. As hinted at by the analyses above, this may be a concern not only regarding an organization's specific place in a ranking but also its broader position in the overall distribution of scores.

Individual-Level Quantities

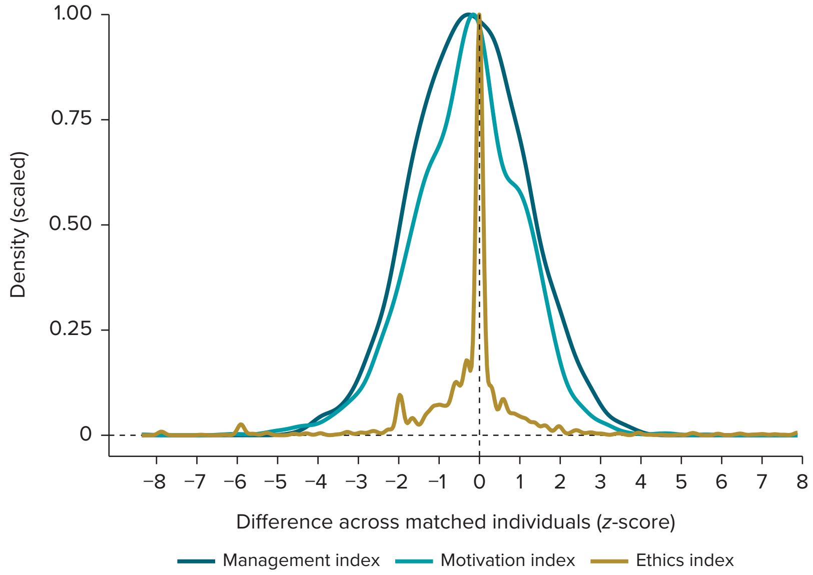
Showing the summary statistics at the individual level, as in panel 3 of table 19.2, requires matching the respondents on their observable characteristics. We use PSM to address the concern of selection bias in who chooses to respond to online surveys (Tourangeau, Conrad, and Couper 2013). PSM employs a logit model to evaluate respondents' likelihood of being in the treatment group—that is, in the online survey mode. PSM is based on the assumption that individuals with comparable observable demographic characteristics (see the note to table 19.2) should, on average, provide comparable answers. If the only meaningful difference left between matched individuals is their treatment status, then any differences in the outcomes of interest should be attributable to it. In using a PSM approach to compare survey modes, we follow earlier examples in the literature that similarly use PSM to adjust for self-selection into an online survey mode (Lee 2006; Lugtig et al. 2011). Moreover, as demonstrated in table 19.1, our experiment shows moderate signs of imbalance on key demographic items. Therefore, PSM can be seen as an additional robustness check, which ensures that these demographic imbalances between treatment arms do not taint our results.

The values shown in panel 3 of table 19.2 are calculated by taking each treated (online mode) individual and his or her index score and subtracting from it the corresponding index scores of the matched respondent(s) from the face-to-face mode. The resulting mean modal differences are comparable to their equivalents at the national and organizational levels. However, the wide distribution of survey mode effects across individuals is now clear. The minimum and maximum modal effects range between −8 and 8 standard deviations, implying that some individuals might be particularly sensitive to the nature of measurement.[20]

Figure 19.6 displays the full distribution of survey mode effects. These are conditional on the matching process we undertook to generate paired observations, though our estimates are robust to including different sets of matching variables. A large fraction (12–15 percent) of paired individuals have a mode effect of at least two standard deviations for the management and motivation indexes.

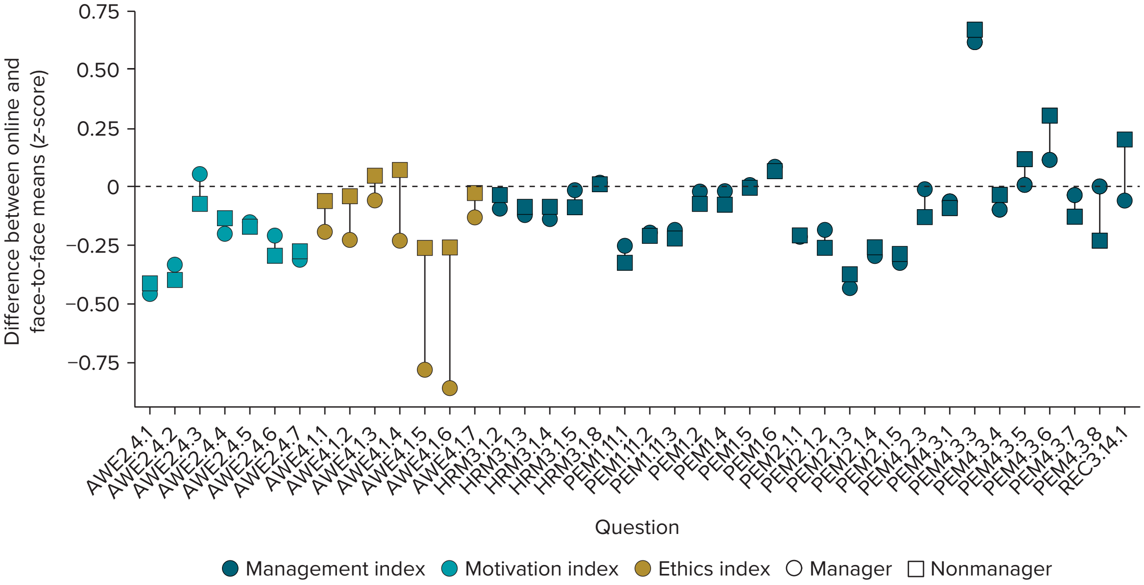
Corresponding to the finding that particular organizations are more sensitive to survey mode effects, it would seem that the distribution of sensitivity across groups of individuals is also important in

FIGURE 19.6 Distribution of Survey Mode Differences across Matched Individuals



Source: Original figure for this publication.

FIGURE 19.7 Average Item Difference, by Managerial Status



Source: Original figure for this publication. AWE = attitudes and work environment; HRM = human resources management; PEM = performance management; REC = recruitment.

understanding the wider nature of survey mode effects in survey design. We can explore how certain groups of public servants exhibit larger mode effects for certain topics. For instance, figure 19.7 shows survey mode effect differences separately for managers and nonmanagers for all individual questions included in each of our indexes (similar to figure 19.3 above). We might expect to see differences in sensitivity to the mode of survey enumeration between those two groups for multiple reasons. In a face-to-face interview with a human enumerator, managers might feel larger social pressure to keep up the good image of their work unit and therefore provide more-positive answers. Nonmanagers might feel less secure in their position, be warier of potential repercussions for answering truthfully, and, therefore, provide less-negative answers in a face-to-face setting, which is perceived as providing less anonymity. As the figure shows, the mean mode effects indeed vary between managers and nonmanagers by as much as 0.5 standard deviations. The differential sensitivity of these two groups to survey mode is particularly visible for some questions composing the ethics index, with the skew toward more-positive answers in the face-to-face mode being noticeably more pronounced for managers than for nonmanagers.

In a similar vein, we can analyze sensitivity to survey mode effects in other demographic groups. The OLS models in table 19.3 examine the relationship between the aggregate values of the three indexes, survey mode, and key individual characteristics, such as age, education level, gender, and tenure. They provide further evidence of the role of the survey mode for outcome measurement, which does not disappear after controlling for other respondent characteristics. For all three indexes, the dummy for the online mode is negative and statistically significant at 1 percent. These coefficients are also very similar in size to the coefficients in table 19.2, and they indicate that online respondents provide responses that are between 0.22 and 0.34 standard deviations more negative than face-to-face respondents.

The role of demographic controls is less consistent. Age and tenure stand out as highly significant for both the management and motivation indexes—with *older* respondents and those with *fewer* years of on-the-job experience providing more-positive answers. Table 19.3 and the further robustness checks discussed below suggest that there is little we can conclude about the independent role of measured demographic variables on our survey indexes. Across cultures, surveys, and agencies, the specific impacts of individual

TABLE 19.3 Ordinary Least Squares Results: Individual Characteristics and Mean Survey Differences

	Dependent variable		
	Management index (1)	Motivation index (2)	Ethics index (3)
Survey mode: Online	−0.244*** (0.029)	−0.341*** (0.029)	−0.222*** (0.032)
Age	0.009*** (0.002)	0.011*** (0.002)	0.002 (0.002)
Gender: Male	−0.013 (0.032)	−0.111*** (0.032)	−0.068* (0.035)
Education: Undergraduate	0.053 (0.073)	−0.097 (0.074)	−0.078 (0.083)
Education: Master's	0.045 (0.074)	−0.067 (0.075)	−0.172** (0.084)
Education: PhD	−0.112 (0.102)	−0.006 (0.103)	−0.123 (0.116)
Status: Civil servant	−0.109* (0.062)	−0.004 (0.062)	0.053 (0.067)
Pay grade	−0.020*** (0.006)	−0.006 (0.006)	−0.015** (0.007)
Managerial status: Manager	−0.470*** (0.049)	0.092* (0.049)	−0.052 (0.053)
Tenure	−0.011*** (0.003)	−0.008*** (0.003)	−0.001 (0.003)
Organizational tenure	0.009*** (0.003)	0.004 (0.003)	−0.006* (0.003)
Public administration tenure	−0.002 (0.003)	−0.003 (0.003)	−0.001 (0.003)
Constant	−0.028 (0.144)	−0.123 (0.144)	0.298* (0.158)
Observations	4,787	4,734	3,991
R^2	0.040	0.043	0.019
Adjusted R^2	0.038	0.040	0.016

Source: Original table for this publication.
Note: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

characteristics on the size of mode effects will vary. This analysis has showcased a potential route for survey analysts to investigate these issues in their own data.

Our results suggest that the high degree of uncertainty around the impact of survey modes on the responses of different organizations and employee groups is an open area for research—both as an academic concern and for the improvement of specific public service surveys. Identifying those individuals sensitive to measurement will require experimentation in the modes to which specific individuals are subject. Identifying those characteristics of public servants that predict sensitivity will make the validity of inferences about differences between individual public servants across key organizational measures significantly more robust.

THE IMPACT OF COMMON CORRECTIONS

Given the near-universal use of online surveys and the concerns that have motivated this chapter, many public servant surveys expend significant resources increasing response rates and analytical effort weighting their responses to correct for sample selection. Our experiment allows us to better understand the impacts of these efforts and their effects on the robustness of the quantities produced by analysis.

How Does the Response Rate Mediate Survey Mode Effects?

A substantial criticism of online surveys—of all types—is that they achieve generally low and varying response rates across organizations relative to face-to-face surveys. Low response rates are typically interpreted as making surveys vulnerable to systematic differences in the sample of individuals who respond and their associated responses to questions. We have seen from the Romania experiment analyzed in this chapter that online surveys do have a lower response rate overall, that it varies more dramatically than the face-to-face survey response rate across organizations, and that respondents differ from a representative sample. However, the question remains whether this leads to differential inference.

As shown in figure 19.8, survey mode effects do not appear to be significantly correlated with survey response rates. In other words, mean modal differences at the organizational level do not differ systematically between organizations with low response rates to online surveys and organizations with high response rates to online surveys (relative to face-to-face surveys with consistently high response rates). Whether response rates are particularly high or low does not seem to explain the variation we see in the robustness of online surveys to replicating the responses generated by face-to-face surveys. This suggests that aggregate responses to online surveys may be compared across organizations even when response rates between these organizations vary widely, as was the case in our survey.

These results also imply that survey mode effects are driven by selection into response and by respondents' interaction with the survey mode rather than simply differing response rates. Given that even high online response rates still exhibit large mode effects, it must be some combination of these effects that drives the wider results of this paper rather than selection alone. Thus, we cannot ultimately conclude that either mode is more accurate, but we note that respondents do seem to respond differently to different approaches to enumeration under certain conditions.

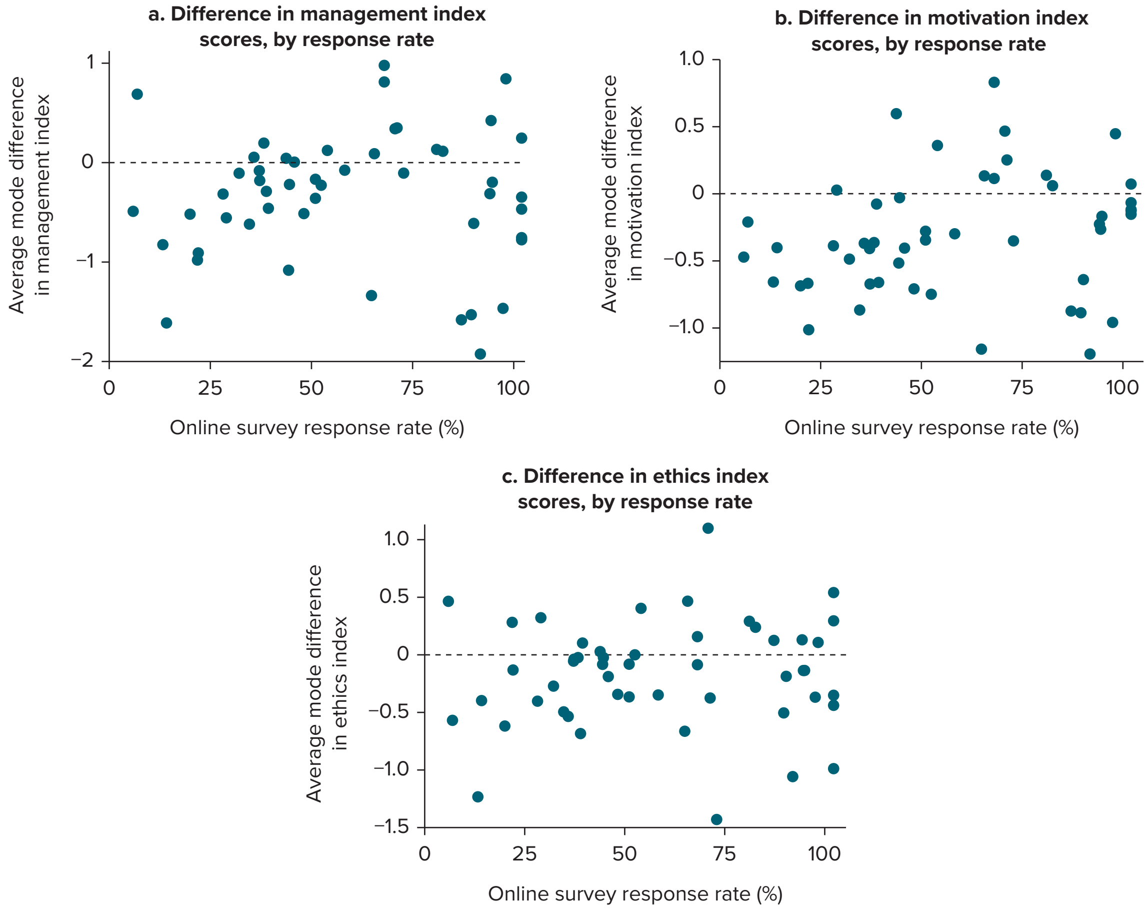
What Is the Impact of Corrections Using Survey Weights?

Many public servant surveys use weighting schemes to upweight the responses of types of officials underrepresented in the survey. To reflect these efforts, we estimate the mean modal difference, using a range of econometric weighting approaches to understand whether they impact the robustness of the corresponding estimates.

We recalculate the unweighted mean differences shown in table 19.4 using a sample weighted by a raking weight based on gender, age, and an inverse online survey response rate to adjust for differences between survey respondents and nonrespondents along these dimensions (Tourangeau, Conrad, and Couper 2013).[21] Finally, we calculate the mean modal difference using inverse probability weighting (IPW), which increases the weight of an official exactly inverse to its survey response rate. In doing so, we give responses from organizations with low response rates a larger weight.

As shown in table 19.4, the modal differences are relatively robust across the unweighted survey sample, a sample that is weighted using the raking method, and a sample that is weighted by the inverse of the organizational survey response rate. Figure 19.9 summarizes the average modal difference across all survey items at the aggregate level across three samples: one that is unweighted, one that is weighted using the raking method, and one that is weighted using IPW. The presence of mode effects is largely unchanged by either

FIGURE 19.8 Difference in Scores, by Response Rate



a. Difference in management index scores, by response rate

b. Difference in motivation index scores, by response rate

c. Difference in ethics index scores, by response rate

Source: Original figure for this publication.

weighting method. Reweighting does little to improve the robustness of the estimates and, in several cases, actually increases the magnitude of the mode effects we observe.

This suggests that the application of weights, a statistical process undertaken by many major public administration surveys, including the FEVS, may not be effective in mitigating the biases introduced by their specific measurement approaches (for a full summary of the weighting methods undertaken by major public administration surveys, see chapter 18). These results are consistent with our preceding findings that the response rate and observable characteristics of individual public servants are not key determinants of the survey mode effects we find.

DISCUSSION

Given the challenges of measuring critical aspects of public service life outside of surveys of public servants, survey design features will continue to be a critical input into our understanding of the state. Perhaps the most significant decision for a survey enumerator interviewing public officials is whether the survey should be administered in person or online. This chapter has reviewed the limited existing information on this question for the public service and presented a novel experiment that sheds light on various aspects of the choice.

This chapter has provided a framework for survey analysts to conceptualize testing survey mode effects in their own surveys, as well as benchmark evidence with which to compare their results. Experimental analysis, as in this chapter, provides a rigorous platform for better understanding the nature of the measurement of the state.

We undertake a field experiment with 6,037 public servants in 81 government institutions in Romania, in which we randomly assign each official to complete either a face-to-face or online survey. In line with predictions of the literature (Heerwegh and Loosveldt 2008; Krosnick and Presser 2010), the online survey exhibits significantly higher levels of survey nonresponse, breakoff, and item nonresponse than the corresponding face-to-face survey. This does change the sample of respondents answering each survey question, pushing the online survey away from a "representative" set of officials. Insofar as missing values impact the overall quality and usability of survey data collected, we can thus conclude that face-to-face survey modes provide higher-quality survey data with fewer missing or nonmeaningful responses. Government-run public servant surveys are almost universally online.

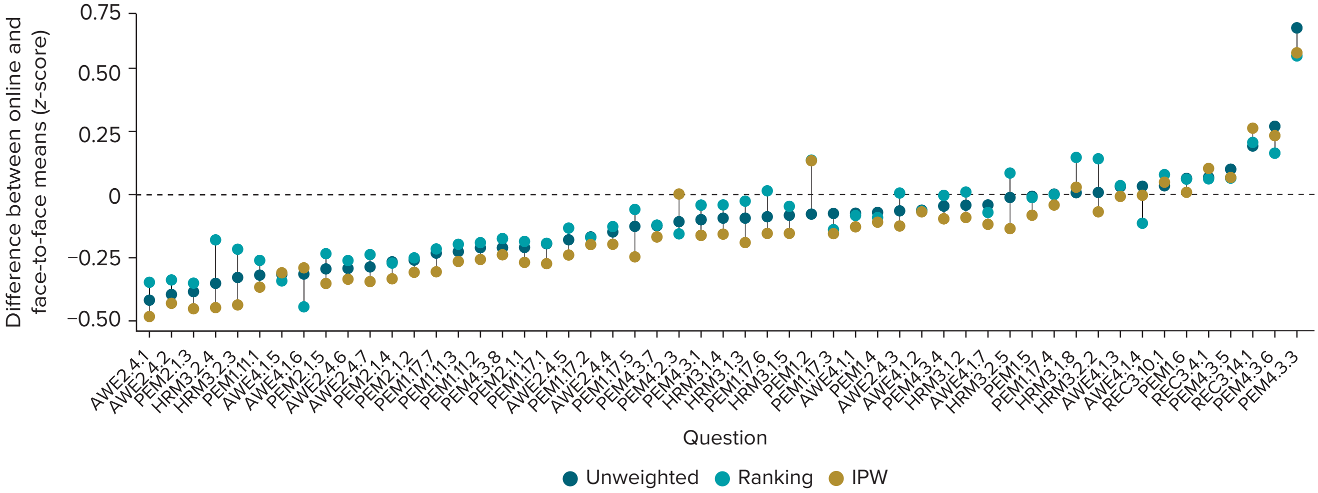
To what extent do we find evidence that the above quality concerns are leading to deviations in results from corresponding face-to-face surveys? The evidence from the experiment we analyze indicates that

TABLE 19.4 Mean Modal Differences at the National Level, by Weighting Approach

(1) Unweighted	
Management index	−0.239
Motivation index	−0.350
Ethics index	−0.208
(2) Weighted (raking)	
Management index	−0.171
Motivation index	−0.263
Ethics index	−0.278
(3) Weighted (IPW)	
Management index	−0.331
Motivation index	−0.440
Ethics index	−0.248

Source: Original table for this publication.
Note: All values reflect the mean difference in the average index values between online and face-to-face samples $(\hat{x}_{online} - \hat{x}_{f2f})$. Panel 1 shows unweighted means. Panel 2 shows the values for the sample weighted using the raking method, wherein weights are iteratively adjusted based on demographic characteristics for which the population distribution is known (in this case, age, gender, and the proportion of civil servants by employment status) until the weighted sample distribution aligns with the population distribution for those variables. Panel 3 weights the sample by inverse values of the organization-level response rate. IPW = inverse probability weights.

FIGURE 19.9 Average Item Difference, by Sample



Source: Original figure for this publication.
Note: IPW = inverse probability weighting. AWE = attitudes and work environment; HRM = human resources management; PEM = performance management; REC = recruitment.

the online treatment group provides more-negative evaluations across the topics of management, motivation, and ethics than the face-to-face group. This pattern also holds for the majority of individual survey questions, not only aggregate indexes. Though such a finding is consistent with the online mode's limiting social-desirability bias, the smallest mode effects are for the ethics index, where this bias should be the most pronounced.

Similar conclusions apply at the level of organizations and individuals. The majority of organizations record lower mean responses in surveys enumerated online. The magnitude of the difference at the national level is small, moving indexes 0.1 on a 1–5 scale. However, the difference for some organizations is above one standard deviation and is substantial enough to make rankings of organizations' scores very poorly correlated across the two survey modes. At the individual level, the magnitude of survey mode effects can be very large. Overall, we cannot make definitive statements about which survey mode is superior, but we note that measurement significantly mediates results at the organizational and individual levels. The burden of proof thus lies with survey analysts to show that results at these levels of aggregation are legitimate.

Based on a PSM analysis, we find that a number of public administrators are particularly sensitive to the survey enumeration approach. We present mode effects of considerable magnitude across matched individuals. These are not well predicted by standard observable characteristics, nor are they affected by common weighting schemes, suggesting that the survey mode is the factor responsible for the difference. Our findings hold with remarkable consistency for all three indexes. Interestingly, the mode effect is present across the whole distribution of response rates, implying that there is a limited correlation between the decision to participate and the deviation of the online survey results from a representative face-to-face survey.

These results suggest that the survey mode effects in public administration are substantial and, for some common survey conclusions to be valid, cannot be ignored. Though aggregates (say, at the national level) are least affected, the ranking of organizations, for example, can be substantially influenced by such effects. Therefore, this chapter proposes embedding an investigation of these issues into survey design generally. As particular national and service cultures mediate where mode effects are largest, corresponding survey analysts can refine their approach as each setting demands. For example, we find that certain groups of respondents and questions—like managers and ethical questions in our experiment—produce noticeably divergent results depending on the survey mode. Identifying the particular groups, questions, and circumstances that make the survey mode a more salient issue, as well as the mechanisms at work in those cases, will contribute to improving the way we measure public administration.

NOTES

We gratefully acknowledge funding from the World Bank's i2i initiative, Equitable Growth, Finance, and Institutions Chief Economist's Office, and Governance Global Practice. We are grateful to Kerenssa Kay, Maria Ruth Jones, and Ravi Somani for helpful comments. We would like to thank Lior Kohanan, Robert Lipinski, Miguel Mangunpratomo, and Sean Tu for excellent research assistance; Kerenssa Kay and Anita Sobjak for guidance and advice; and seminar participants at the World Bank for their comments. Computational reproducibility was verified by DIME Analytics.

1. Though most online surveys follow a relatively standard form, there is potential to make online surveys more engaging for the respondent. For example, the gamification of surveys or the inclusion of short clips and other multimedia extensions may enable surveys to more effectively capture respondents' attention. These have generally not been taken up or experimented with in any setting, including in public administration. One notable exception is Haan et al. (2017), who examine whether adding a video of enumerators reading online survey questions increases engagement. The study finds a null effect and concludes that the interactive component of face-to-face surveys goes beyond a video recording of the enumerators.
2. To date, the existing literature has focused on the advantages and disadvantages of online versus face-to-face surveys in the general population (Couper et al. 2007; Daikeler, Bošnjak, and Lozar Manfreda 2020; Groves and Peytcheva 2008; Heerwegh and Loosveldt 2008; Krosnick and Presser 2010; Peytchev 2009). No studies, to our knowledge, focus on this debate in the context of public administration.
3. The value was calculated as a mean difference between the ratio of the number of respondents relative to the number of invited and eligible respondents in the web mode and the equivalent ratio for the other survey mode.

4. These large country differences and declining response rates are not unlike those observed in general public opinion surveys. For example, Beullens et al. (2018) find that response rates to the European Social Survey range from well below 50 percent in countries like the United Kingdom and Germany to above 70 percent in Cyprus, Bulgaria, and Israel, all while a double-digit decline in response rates is observable in many settings.
5. Implementation of the face-to-face surveys was successful, with 99 percent of face-to-face surveys rated as having gone well or very well.
6. Our concern is that in these institutions, the relevant survey links were not adequately distributed to targeted staff. To test the robustness of this decision to our results, we also use different cutoff points for online survey response rates of 3 percent and 7 percent, and our results are qualitatively the same.
7. A remaining concern is that an organization's response rate by mode may be high for distinct reasons, and these reasons may be correlated with the variables on which we collect data. However, given the low nonresponse rates in our matched sample, there is limited scope for endogenous selection to impact our estimates (Oster 2019).
8. This is contrary to some findings in the literature that online survey respondents tend to be male (Duffy et al. 2005). This difference may be due in part to the composition of the Romanian civil service, which is predominantly female across most organizations.
9. This difference is comparable in magnitude to other surveys in the literature, which find an average difference in educational attainment of approximately 6 percentage points (Braekman et al. 2020).
10. We also see a substantial number of individuals exiting where the demographic question block begins.
11. An additional 89 revisited the survey after previously completing it. These individuals are excluded from this analysis, as it is assumed that their returning to a survey they had already taken was inadvertent.
12. Though evidence on the impact of reminders in public servant surveys is scarce, data from the 2014 FEVS shows that the number of responses is at its peak in the first week of the survey, drops dramatically in subsequent weeks, and plateaus between weeks three and six (with a slight jump in the final week). This echoes our own experience and underlies the critical importance of the survey launch.
13. The full list of questions composing each index can be found in table G.8 in appendix G.
14. The *z*-scores are calculated over the full sample of individuals used for analysis.
15. For the list of questions and their phrasing, see table G.8 in appendix G.
16. In chapter 22, we specifically focus on how the complexity and sensitivity of each question influence response patterns. For that purpose, we develop a coding framework that assesses each question in the Romania questionnaire (among others) along various margins of complexity and sensitivity, like syntax, context familiarity, privacy, and the threat of disclosure.
17. More formal tests of the difference between mode effects at the organizational level are discussed in appendix G.
18. The correlation can be expected to be even lower for individual questions, which tend to exhibit greater variation.
19. As a reminder that these graphs are not an artifact of response bias arising from extreme response rates, note again that we restrict the sample of comparison here to only those organizations with an online response rate of at least 5 percent.
20. To assess the validity of our matched estimates, in table G.5 (see appendix G), we also present results obtained if PSM controls for a different set of demographic characteristics and also for organizational fixed effects only. We find that the estimates of mean differences are qualitatively similar across various PSM approaches.
21. Iterative proportional fitting, or raking, is among the most commonly used methods for weighting survey results. The method involves choosing a set of demographic variables where the population value is known and iteratively adjusting the weight for each case until the sample distribution aligns with the population distribution for those variables (Mercer, Lau, and Kennedy 2018).

REFERENCES

Anduiza, Eva, and Carol Galais. 2017. "Answering without Reading: IMCs and Strong Satisficing in Online Surveys." *International Journal of Public Opinion Research* 29 (3): 497–519. https://doi.org/10.1093/ijpor/edw007.

Baumgartner, Hans, and Jan-Benedict E. M. Steenkamp. 2001. "Response Styles in Marketing Research: A Cross-National Investigation." *Journal of Marketing Research* 38 (2): 143–56. https://doi.org/10.1509/jmkr.38.2.143.18840.

Beullens, Koen, Geert Loosveldt, Caroline Vandenplas, and Ineke Stoop. 2018. "Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?" Survey Methods: Insights from the Field, April. https://doi.org/10.13094/SMIF-2018-00003.

Biemer, Paul P., Joe Murphy, Stephanie Zimmer, Chip Berry, Grace Deng, and Katie Lewis. 2018. "Using Bonus Monetary Incentives to Encourage Web Response in Mixed-Mode Household Surveys." *Journal of Survey Statistics and Methodology* 6 (2): 240–61. https://doi.org/10.1093/jssam/smx015.

Braekman, Elise, Rana Charafeddine, Stefaan Demarest, Sabine Drieskens, Finaba Berete, Lydia Gisle, Johan Van der Heyden, and Guido Van Hal. 2020. "Comparing Web-Based versus Face-to-Face and Paper-and-Pencil Questionnaire Data

Collected through Two Belgian Health Surveys." *International Journal of Public Health* 65 (1): 5–16. https://doi.org/10.1007/s00038-019-01327-9.

Cornesse, Carina, and Michael Bošnjak. 2018. "Is There an Association between Survey Characteristics and Representativeness? A Meta-Analysis." *Survey Research Methods* 12 (1): 1–13. https://doi.org/10.18148/srm/2018.v12i1.7205.

Couper, Mick P., Arie Kapteyn, Matthias Schonlau, and Joachim Winter. 2007. "Noncoverage and Nonresponse in an Internet Survey." *Social Science Research* 36 (1): 131–48. https://doi.org/10.1016/j.ssresearch.2005.10.002.

Daikeler, Jessica, Michael Bošnjak, and Katja Lozar Manfreda. 2020. "Web versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates." *Journal of Survey Statistics and Methodology* 8 (3): 513–39. https://doi.org/10.1093/jssam/smz008.

De la Rocha, Alexandra Mariah. 2015. "The Relationship between Employee Engagement and Survey Response Rate with Union Membership as a Moderator." Master's thesis, San José State University. https://doi.org/10.31979/etd.z4c6-uv9d.

Duffy, Bobby, Kate Smith, George Terhanian, and John Bremer. 2005. "Comparing Data from Online and Face-to-Face Surveys." *International Journal of Market Research* 47 (6): 615–39. https://doi.org/10.1177/147078530504700602.

Galesic, Mirta. 2006. "Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey." *Journal of Official Statistics* 22 (2): 313–28. https://www.proquest.com/scholarly-journals/dropouts-on-web-effects-interest-burden/docview/1266792615/se-2.

Gnambs, Timo, and Kai Kaspar. 2015. "Disclosure of Sensitive Behaviors across Self-Administered Survey Modes: A Meta-Analysis." *Behavior Research Methods* 47: 1237–59. https://doi.org/10.3758/s13428-014-0533-4.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70 (5): 646–75. https: //doi.org/10.1093/poq/nfl033.

Groves, Robert M., and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72 (2): 167–89. https://www.jstor.org/stable/25167621.

Haan, Marieke, Yfke P. Ongena, Jorre T. A. Vannieuwenhuyze, and Kees de Glopper. 2017. "Response Behavior in a Video-Web Survey: A Mode Comparison Study." *Journal of Survey Statistics and Methodology* 5 (1): 48–69. https://doi.org/10.1093/jssam/smw023.

Heerwegh, Dirk. 2009. "Mode Differences between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects." *International Journal of Public Opinion Research* 21 (1): 111–21. https://doi.org/10.1093/ijpor/edn054.

Heerwegh, Dirk, and Geert Loosveldt. 2008. "Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality." *Public Opinion Quarterly* 72 (5): 836–46. https://doi.org/10.1093/poq/nfn045.

Jensen, Nathan M., Quan Li, and Aminur Rahman. 2010. "Understanding Corruption and Firm Responses in Cross-National Firm-Level Surveys." *Journal of International Business Studies* 41 (9): 1481–504. https://doi.org/10.1057/jibs.2010.8.

Kaminska, Olena, and Tom Foulsham. 2014. "Real-World Eye-Tracking in Face-to-Face and Web Modes." *Journal of Survey Statistics and Methodology* 2 (3): 343–59. https://doi.org/10.1093/jssam/smu010.

Kays, Kristina, Kathleen Gathercoal, and William Buhrow. 2012. "Does Survey Format Influence Self-Disclosure on Sensitive Question Items?" *Computers in Human Behavior* 28 (1): 251–56. https://doi.org/10.1016/j.chb.2011.09.007.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5 (3): 213–36. https://doi.org/10.1002/acp.2350050305.

Krosnick, Jon A., and Stanley Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research*, edited by Peter V. Marsden and James D. Wright, 2nd ed., 263–314. Bingley: Emerald.

Lee, Sunghee. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22 (2): 329–49. https://www.researchgate.net/publication/259497319PropensityScoreAdjustmentasaWeightingSchemeforVolunteerPanelWebSurveys.

Lozar Manfreda, Katja, Michael Bošnjak, Jernej Berzelak, Iris Haas, and Vasja Vehovar. 2008. "Web Surveys versus Other Survey Modes: A Meta-Analysis Comparing Response Rates." *International Journal of Market Research* 50 (1): 79–104. https://doi.org/10.1177/147078530805000107.

Lozar Manfreda, Katja, and Vasja Vehovar. 2002. "Survey Design Features Influencing Response Rates in Web Surveys." Paper delivered at the International Conference on Improving Surveys, Copenhagen, August 25–28, 2002. http://www.websm.org/uploadi/editor/LozarVehovar2001Surveydesign.pdf.

Lugtig, Peter, Gerty J. L. M. Lensvelt-Mulders, Remco Frerichs, and Assyn Greven. 2011. "Estimating Nonresponse Bias and Mode Effects in a Mixed-Mode Survey." *International Journal of Market Research* 53 (5): 669–86. https://doi.org/10.2501/IJMR-53-5-669-686.

Mercer, Andrew, Arnold Lau, and Courtney Kennedy. 2018. "How Different Weighting Methods Work." In *For Weighting Online Opt-In Samples, What Matters Most?*, 7–14. Washington, DC: Pew Research Center. https://www.pewresearch.org/methods/2018/01/26/how-different-weighting-methods-work.

Moynihan, Donald P., and Sanjay K. Pandey. 2010. "The Big Question for Performance Management: Why Do Managers Use Performance Information?" *Journal of Public Administration Research and Theory* 20 (4): 849–66. https://doi.org/10.1093/jopart/muq004.

Musch, Jochen, and Ulf-Dietrich Reips. 2000. "A Brief History of Web Experimenting." In *Psychological Experiments on the Internet*, edited by Michael H. Birnbaum, 61–87. Cambridge, MA: Academic Press. https: //doi.org/10.1016/B978-0-12-099980-4.X5000-X.

Newell, Carol E., Kimberly P. Whittam, Zannette A. Urielle, and Yeuh-Chun (Anita) Kang. 2010. *Non-Response on U.S. Navy Quick Polls*. NPRST-TN-10-3. Millington, TN: Navy Personnel Research, Studies, and Technology, Bureau of Naval Personnel. https://apps.dtic.mil/sti/citations/ADA516853.

Newman, Jessica Clark, Don C. Des Jarlais, Charles F. Turner, Jay Gribble, Phillip Cooley, and Denise Paone. 2002. "The Differential Effects of Face-to-Face and Computer Interview Modes." *American Journal of Public Health* 92 (2): 294–97. https://doi.org/10.2105/ajph.92.2.294.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business and Economic Statistics* 37 (2): 187–204. https://doi.org/10.1080/07350015.2016.1227711.

Peytchev, Andy. 2006. "A Framework for Survey Breakoffs." Paper presented at the 61st Annual Conference of the American Association for Public Opinion Research, Montréal, May 18–21, 2006. In JSM Proceedings, Survey Research Methods Section, 4205–12. Alexandria, VA: American Statistical Association. http://www.asasrms.org/Proceedings/y2006/Files/JSM2006-000094.pdf.

Peytchev, Andy. 2009. "Survey Breakoff." *The Public Opinion Quarterly* 73 (1): 74–97. https://www.jstor.org/stable/25548063.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies." *Journal of Applied Psychology* 88 (5): 879–903. https://doi.org/10.1037/0021-9010.88.5.879.

Rich, Bruce Louis, Jeffrey A. Lepine, and Eean R. Crawford. 2010. "Job Engagement: Antecedents and Effects on Job Performance." *Academy of Management Journal* 53 (3): 617–35. https://doi.org/10.5465/amj.2010.51468988.

Shih, Tse-Hua, and Xitao Fan. 2008. "Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis." *Field Methods* 20 (3). https://doi.org/10.1177/1525822X08317085.

Spitzmüller, Christiane, Dana M. Glenn, Christopher D. Barr, Steven G. Rogelberg, and Patrick Daniel. 2006. "'If You Treat Me Right, I Reciprocate': Examining the Role of Exchange in Organizational Survey Response." *Journal of Organizational Behavior* 27 (1): 19–35. https://doi.org/10.1002/job.363.

Steinbrecher, Markus, Joss Roßmann, and Jan Eric Blumenstiel. 2015. "Why Do Respondents Break Off Web Surveys and Does It Matter? Results from Four Follow-Up Surveys." *International Journal of Public Opinion Research* 27 (2): 289–302. https://doi.org/10.1093/ijpor/edu025.

Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. *The Science of Web Surveys.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199747047.001.0001.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859–83. https://doi.org/10.1037/0033-2909.133.5.859.

Ye, Cong, Jenna Fulton, and Roger Tourangeau. 2011. "More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice." *Public Opinion Quarterly* 75 (2): 349–65. https://doi.org/10.1093/poq/nfr009.

CHAPTER 20

Determining Sample Sizes

How Many Public Officials Should Be Surveyed?

Robert Lipinski, Daniel Rogger, Christian Schuster, and Annabelle Wittels

SUMMARY

Determining the sample size of a public administration survey often entails a trade-off between the benefits of increasing the precision of survey estimates and the high costs of surveying a larger number of civil servants. Survey administrators ultimately have to decide on the sample size based on the types of inference they want the survey to yield. This chapter aims to quantify the sample sizes necessary to make a range of inferences that are commonly drawn from public administration surveys. It does so by employing Monte Carlo simulations and past survey results from Chile, Liberia, Romania, and the United States. The analyses show that civil service–wide estimates can be reliably derived using sample sizes considerably smaller than the ones currently used by these surveys. By contrast, comparison across demographic groups—gender and managerial status—and ranking individual public administration organizations both require large sample sizes, often substantially larger than those available to survey administrators. These results suggest that not all types of inference and comparison can be drawn from surveys of civil servants, which, instead, may need to be complemented by other research tools, like interviews or anthropological research. This chapter is also linked to an online toolkit that allows practitioners to estimate the optimal sample size for a survey given the types of inference expected to be drawn from it. Together, the chapter and the toolkit allow practitioners involved in survey design for the civil service to understand the trade-offs involved in sampling and what types of comparison can be reliably drawn from the data.

ANALYTICS IN PRACTICE

- Sample size is one of the key factors affecting survey quality. An accurately selected sample of adequate size is indispensable to making survey results reliable and actionable. Choosing the number of respondents is, therefore, a crucial decision faced by any survey designer. This chapter details what factors should be considered to make an optimal choice in the context of sampling for civil servant surveys.

Robert Lipinski is a consultant and Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Christian Schuster is a professor at University College London. Annabelle Wittels is an independent researcher.

- Efficient survey sampling strategies need to balance the precision of estimates against the costs of expanding the sample size. Sampling more people tends to improve the accuracy of survey results and the number of comparisons that can be reliably drawn from the responses. However, for logistical and financial reasons, it is not always possible to survey everyone. Thus, the benefits of increasing the sample size and the costs of running a survey need to be balanced against each other.
- The required survey sample size crucially depends on the types of comparison a researcher plans to make based on the data. Obtaining precise civil service–wide aggregates requires a considerably smaller sample than drawing comparisons between demographic groups of civil servants or institutions within public administration. Survey designers have to decide in advance what inferences they need to draw from their surveys and adjust the sample size accordingly.
- Civil servant surveys often oversample for the purpose of determining civil service–wide aggregate measures. On the basis of past civil servant surveys, we conclude that most common civil servant survey measures, like job satisfaction, work motivation, and merit-based recruitment, could be accurately estimated at the level of the civil service as a whole by surveying 50–70 percent of the current sample.
- Comparisons of survey responses between different demographic groups (such as male vs. female or manager vs. nonmanager) require sample sizes equivalent to or larger than those currently used. Decreasing current sample sizes would likely lead to incorrect comparisons between demographic groups—due to nonrepresentative samples—or prevent them altogether—due to insufficient responses from each group of interest to enable comparison. Although this topic is not covered here, the present analysis indicates that comparisons between more than two demographic groups, like civil servants of different education levels or ethnic backgrounds, would require sample sizes larger than the ones currently prevalent.
- Precise ranking of institutions within the civil service according to survey measures, like job satisfaction or motivation, requires larger sample sizes than currently prevalent. Given the standard sample sizes and the variation in estimates, survey questions are unlikely to determine an exact ranking of institutions within public administration. Institutions might not be sufficiently large for such comparisons, or samples of respondents drawn from them would need to become considerably larger than is currently the case. Rather than an exact ranking position, the quintile position of an institution (for example, if it is in the top 20 percent of institutions on a given measure) can be more reliably determined.

INTRODUCTION

The usefulness of surveys as a research tool is determined by multiple factors, but one of the most crucial is sample size. The number of people who provide responses to a survey determines the confidence one can have in its results and the types of inference and comparison one can draw from it. In general, the more people are surveyed, the more reliable and actionable the results of a survey. To take the simplest example, a survey of 1,000 people in, say, a ministry of education is more likely to yield the true value of the quantity of interest, like the level of job satisfaction, than a survey of 10 people. It would also be more likely to allow for the comparison of job satisfaction levels between men and women, managers and nonmanagers, or different departments within the ministry.

However, surveying as many people as possible is not always a useful guideline for survey designers, especially in the context of public administration surveys. For one, many surveys in this context are administered face-to-face. This may be due to technical reasons (for example, low access to the internet) or methodological considerations (for example, face-to-face surveys tend to decrease item nonresponse; see chapter 19). Moreover, each additional person surveyed, regardless of the mode of survey delivery, increases the direct and indirect costs associated with running a survey. The direct costs of survey administration are particularly pronounced in face-to-face surveys, in which travel time and enumerator staff costs increase for each extra person surveyed. Even in online surveys—in which survey administration costs are often fixed—indirect survey costs can be significant. For instance, completing surveys takes time. Each minute taken away from the workday of a public sector employee incurs a cost to the public purse. Half an hour of the time of the average public sector employee in the United States costs the taxpayer US$19.81.[1] If the number of US civil servants surveyed in the annual Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) were reduced by 10 percent, the opportunity cost of the survey would be reduced by US$3 million. These costs cannot be eliminated, but they can be reduced by limiting survey time, which might limit the scope of inferences drawn from the survey, and—the focus of this chapter—by optimizing the number of people surveyed.

The goal of a public sector survey should be to sample efficiently in order to save resources on one survey and free up resources for other work tasks or for more frequent, targeted surveying, which can improve the quality and breadth of data available for decision-making. For instance, the United Kingdom's Office for National Statistics (ONS) publishes "experimental statistics."[2] The ONS collects data on the UK labor market every three months but provides model estimates for single months and weeks. Their accuracy is repeatedly assessed to establish whether surveying on a three-month basis provides statistics that are accurate enough to make decisions about the performance of the labor market in a single month, or even in a single week. More frequent surveying of civil servants could be supported by creating surveys that sample a smaller pool of people and are thus quicker and less costly to administer. Slashing sample sizes, however, entails considerable risk: if sample sizes are too small, the error bounds around estimates become too large to reliably assess progress on key performance targets or to compare different groups of civil servants or individual organizations within public administration.

What, then, are the appropriate sample sizes for civil service surveys? And are existing approaches to civil service survey sampling efficient? To assess these conundrums, this chapter conducts Monte Carlo simulations with civil service survey data from Chile, Liberia, Romania, and the United States. Our results suggest that appropriate sample sizes depend on the inferences governments wish to make from the data. To estimate averages for countries or large organizations within public administration, sample sizes could often be reduced. This holds all the more for survey measures—such as measures of work motivation or job satisfaction—that vary only to a limited extent (cf. chapter 21). Where detailed comparisons among public sector organizations—ranking the organizations by the mean values of survey question responses—or groups of public servants—for example, by gender or managerial status—are sought, sample sizes are typically too small. This holds in particular for those survey measures with limited variation and high skew, such as work motivation, which require high levels of precision to enable comparisons that detect statistically significant differences between groups of public servants or organizations within public administration. Our chapter thus concludes that a detailed elaboration of the desired uses of the survey results should precede the determination of sample sizes. It also offers an online sampling toolkit for survey designers to estimate appropriate sample sizes depending on the intended uses of the survey data.[3]

SAMPLING BEST PRACTICES AND THE CIVIL SERVICE SURVEY CONTEXT

Several governments regularly survey their employees, yet approaches to sampling vary. For instance, in Australia and the United Kingdom, all public sector employees are invited to take the survey (a census approach), whereas other countries employ a mix of random, ad hoc sampling, and census approaches.[4] For example, the FEVS uses stratified random sampling approaches in most years but conducts a census every few years (2012, 2018, and 2019) to update the sampling frames. Canada's Public Service Employee Survey recruits public sector organizations to reach out to their staff to complete the survey and also makes the survey available online for anyone who decides they fit the eligibility criteria. In Colombia, the annual national public employee survey (the Survey of the Institutional Environment and Performance in the Public Sector) uses a mixed approach: for larger organizations, a stratified sampling approach is used, while for smaller organizations, a census is taken. This is similar to the approach that the United States uses during noncensus years. Countries that have run surveys for several years have the advantage of looking back at historical data to assess what future sample sizes would be adequate, given the distributions and variations of the indicators they use. However, in many countries, surveys are not yet routinized and survey questions or approaches have changed, so there is a dearth of data to make informed decisions. This chapter addresses this problem by illustrating how countries can determine what sample size is adequate for their needs.

Determining adequate sample sizes ideally requires information on the following factors:

- **The size and proportion of the units of comparison.** The ideal approach to sampling entails drawing up so-called sampling frames, which list all relevant persons to be surveyed. In countries that lack routinized surveys of the public sector, a common obstacle to efficient sampling for public sector surveys is that complete and up-to-date records of public sector staff are not centralized, not fully digitized, or generally contain gaps (Bertelli et al. 2020). The creation and maintenance of complete sampling frames is a first step toward improving the efficiency of sampling.
- **The types of comparison—between countries, organizations, subunits, key personnel groups, previous years, or industry benchmarks.** It is also important to consider what types of comparison governments want to make using survey results. In most cases, public sector organizations desire to provide feedback to the managers of organizational subunits. In these cases, sampling should be stratified at the subunit level to increase the chances of an adequate sample size at the subunit level. However, this is often not possible because staff lists at the subunit level are incomplete or not centralized. In such a case, a minimum number of observations per subunit should be used as a target. Another consideration is whether sampling approaches are adequate for the types of comparison that governments desire to make. For example, are organizations to be benchmarked against industry (public sector) averages? Should their performance be compared with the previous year? Are comparisons required between key employee groups, such as managers and nonmanagers? It might be the case that some comparisons are not possible in certain contexts. For example, if all subunits are composed of only a few civil servants, ranking them by average survey responses might not be possible even if all of them were surveyed. Therefore, the desired comparisons should account for all the external limitations present.
- **The distributions of key variables (for example, mean and variance).** Which sample sizes allow comparisons to be meaningful depends on the distribution of these indicators (and also, but to a lesser extent, the number of comparisons that are planned). If distributions are narrow (for example, for measures such as motivation; see chapter 21), then fewer respondents are needed to arrive at the true value of aggregate-level statistics, like the mean or median. However, such distributions make it difficult to discern differences between different groups or units within public administration.
- **The desired degree of precision for the estimates.** Pinpointing the exact value of the quantity of interest is almost never possible when sampling from a larger population. However, the sampling strategy depends on how wide of uncertainty survey designers are willing to tolerate. If the representativeness of the sample is maintained, having more respondents tends to mean a more precise estimate. However, survey designers have to decide what degree of precision is acceptable. For example, if a mean estimate within ±0.1 points of the true value on a 1–5 Likert scale is sufficient, then it would be unnecessary to increase the sample size, and therefore the costs of running a survey, in order to narrow the precision even further.

Advice on sampling for surveys outside the public sector is available. Since Cochran (1977), conventions for how to sample have been well known. Textbooks, such as SAGE Publishing's "little green book" (Kalton 1983), an encyclopedia of common research methods, typically suggest the following approach to determining sample sizes for survey research: using simple random sampling, first determine the degree of precision that is required from the estimates, add a design factor—a multiplier that inflates the sample size—if you use clustered sampling approaches, and adjust the sample for the expected level of nonresponse.

While this approach is sensible in many instances, simplification carries several dangers. As Fowler (2009) cautions, the size of the population from which a sample is drawn has little effect on the precision of the estimates, all sample size requirements need to be decided on a case-by-case basis, and increasing the sample size does not necessarily reduce the error of estimates.

The following illustrates Fowler's first point: although the population of the United States is 16 times larger than that of Romania, one would not need a sample size 16 times larger to make estimates about the public sector in the United States versus in Romania. Rather, the dispersion of scores matters. If civil servants in the United States answer more similarly to one another than those in Romania, it is possible that, despite the Romanian civil service being considerably smaller, a larger sample size would be needed for Romania than for the United States.

With regard to Fowler's second point, rules of thumb can be useful. For example, one rule that is often used is that one should have at least 30 observations in each subgroup in order to calculate nonparametric statistics, such as the chi-square statistic. However, without knowledge about the underlying distribution of metrics and likely error rates, rules of thumb can result in highly unsatisfactory sample sizes. What sample size is satisfactory is thus an empirical question. For instance, while many survey companies routinely use a target precision of ±3 percentage points, one should ask how this compares to the dispersion of the underlying scores and whether it provides for meaningful differences. For instance, if one organization differs by 0.05 standard deviations from another in a given year, can this be considered a meaningful difference? If so, then the sample size should be large enough to detect such differences. If not, then the sample size should be revised to capture a difference that is meaningful to the question at hand.

Finally, error caused by insufficiently large sample sizes needs to be understood as a part of the total survey error. The total survey error refers to a compound measure of error. It includes, but is not limited to, error created by sampling; it includes error deriving from the choice of scale, the survey mode, and interview techniques. For instance, if more resources are deployed to sample more people, this might come at the expense of pretesting survey scales or training enumerators, which can inflate the variance of survey answers and thereby make estimates more imprecise.

What is more, algorithmic approaches to gauge sample size can lead to misleading conclusions when survey design and analysis approaches are more complex. For instance, one needs to assess whether clustered or stratified approaches were used.

THEORETICAL BACKGROUND

Surveys can either be targeted at collecting information from the entire population or universe of interest (a census approach) or at collecting information from a fraction of the population—a sample. Typically, surveys are used to estimate means, medians, and modes for certain responses for the entire target sample and subgroups of interest. The desire is that these estimates are an accurate or accurate-enough representation of the measures of the population. This might be impossible because of errors introduced by sampling and such things as the interview process or the coding of data. Bjarkefur et al. (2021) provides more in-depth information on how to address issues related to nonsampling error. Sampling bias can occur because of issues related to who was targeted by the survey recruitment, self-selection into survey participation, and nonresponse bias (on this topic, see chapter 22). Finally, error can be introduced by sampling variance—the fact that measurements vary and that the sample technique and size need to be adequate for the underlying dispersion of responses targeted for estimation (on this topic, see chapter 21).

Why sample size matters can be demonstrated by looking at how the standard error of two group means is calculated. Typically, inferences from surveys will pertain to comparisons between groups of observations (for example, between two agencies, between managers and nonmanagers, etc.). The standard error of the estimate of the difference in mean scores between the two groups is the square root of the sum of their individual squared standard errors:

$$SE(\mu_1 - \mu_2) = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{SD_1^2}{n_1}\frac{SD_2^2}{n_2}}. \qquad (20.1)$$

The standard error is mechanically smaller, the larger the respective sample sizes of each of the groups in the comparison are (n_1 and n_2). At the same time, it is positively correlated with the values of standard deviation.

METHODOLOGY

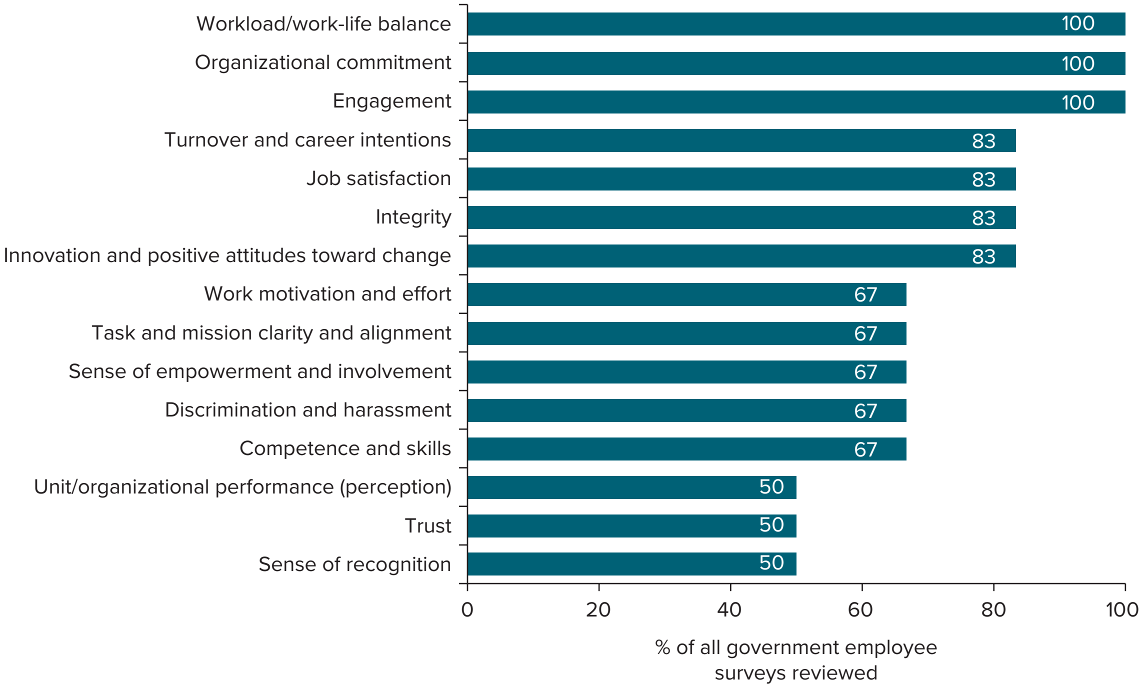
In this chapter, we illustrate what sampling error can be expected based on the variance observed in typical measures used in civil servant surveys and, consequently, what types of inference can be reliably drawn from them.

Since the approach taken in this chapter is to provide sampling guidelines for survey practitioners by extrapolating from existing civil service survey data and practice, we base the analyses upon the wealth of information provided by surveys of civil servants conducted in recent years by the World Bank, the Global Survey of Public Servants (GSPS) academic consortium, and national governments. Together, they allow us to present a wide range of statistical tests and a breadth of examples. The following surveys are included:

- A survey of civil servants in **Chile**, which takes a census approach, targeting all employees in a sample of 65 central government institutions. (The survey was part of the GSPS consortium's effort to collect more data on public administrations around the world.)
- A survey of civil servants in **Liberia**, which uses random sampling, stratified by institution. (The survey was conducted by the World Bank.)
- A survey of civil servants in **Romania**, which follows a stratified sampling approach, by which respondents are sampled in each department of a sample of organizations. (The survey was part of the GSPS consortium's effort to collect more data on public administrations around the world.)
- The Federal Employee Viewpoint Survey (FEVS), an annual survey administered by the **United States** Office of Personnel Management (OPM)—a federal agency—which first launched in 2002 under the name Federal Human Capital Survey. The FEVS aims to recruit a sample representative of the different types of US federal agencies. In 2012, 2018, and 2019, the FEVS took a census approach. In other years, the FEVS has used stratified random sampling, whereby the sample is stratified by work units within organizations. Work units smaller than 10 employees are merged. All senior executives are targeted by the survey, while lower-rank individuals are subject to random sampling within their strata. A target sample size for each organization is calculated. When this target rate amounts to 75 percent or more of an organization's entire staff, a census approach, whereby all employees are targeted, is employed instead. The FEVS has served as an important benchmark for multiple surveys of public administrators around the world.

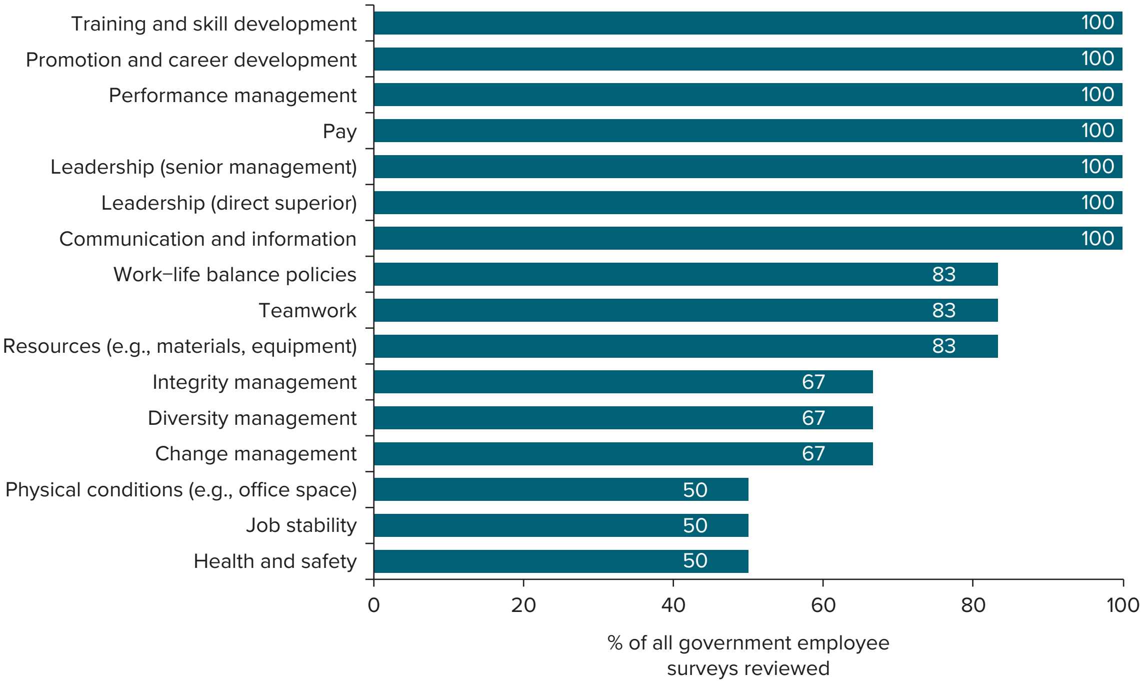
The selected surveys cover four continents and divergent socioeconomic contexts, as well as different sampling approaches and a range of widely used survey questions and indicators. Our analyses focus on a set of questions about job satisfaction, work motivation, performance review (evaluation), and merit-based recruitment. The chosen measures reflect some of the most commonly used indicators in surveys of public servants around the world, as a review by the GSPS has indicated (see figures 20.1 and 20.2). Most measures

FIGURE 20.1 Most Commonly Used Survey Measures of Attitudes and Behaviors across Civil Servant Surveys



Source: Meyer-Sahling et al. 2021.

FIGURE 20.2 Most Commonly Used Survey Measures of Management Practices across Civil Servant Surveys



Source: Meyer-Sahling et al. 2021.

are indicators, which are averages across several questions. We highlight where single questions, rather than indexes, are used for analysis.

As table 20.1 summarizes, the surveys selected for analysis in this chapter were all conducted between 2017 and 2019. Most surveys were conducted online using a structured format with closed-ended questions. The Liberia survey used a semi-structured format, akin to that used by the World Management Survey.[5] Trained enumerators asked open-ended questions and then selected a precoded answer option based on the responses that participants provided.

The Romania survey used two approaches: online and face-to-face. As chapter 19 on survey mode effects shows in more detail, surveys conducted via face-to-face enumeration tend to have higher response rates. For simplicity, in the simulations underpinning this chapter, we assume that these response rates remain the same.[6]

All surveys identify organizations within the sample. For each survey, the means for institutions were calculated in order to compare their performance. The number of organizations ranges from 30 to 65 per survey.

To foster comparability in our sampling simulations, we select survey questions that are similar, to the extent possible, across surveys. The exact wording can be found in table 20.2. To foster the generalizability of our findings to other surveys, the selected survey questions cover a range of core and frequently asked-about topics in civil service surveys—such as work motivation, job satisfaction, performance management, leadership, and the quality of management practices.

The distributions of each of the included variables in each of the countries and public sector organizations are visualized in figure 20.3.

Monte Carlo Simulations

We show, based on these data, what sample sizes might be needed to draw the most common types of inference—defining country-level aggregates, comparing key demographic groups of civil servants (male vs. female and manager vs. nonmanager), and ranking organizations within public administration. Our hope is that these examples will help practitioners find examples that are similar to their own cases. This will provide

TABLE 20.1 Characteristics of Surveys Included for Simulations

Country	Sampling strategy	Year	Key indicators	Key comparisons made	Mode	Sample size	No. of orgs.	Response rate
Chile	Simple random	2019	Motivation, leadership, performance, recruitment practices	Organization; unit	Online	23,636	65	44%
Liberia	Stratified random	2017	Management, recruitment practices	Organization; unit	Face-to-face	2,790	33	48%
Romania	Cluster random	2019	Motivation, leadership, performance, recruitment practices	Organization; unit	Face-to-face Online	2,721 3,721	30	92% 24%
United States	Cluster stratified random	2019	Engagement, satisfaction	Organization; previous years	Online	615,395	45	43%

Source: Original table for this publication.

TABLE 20.2 Overview of Survey Questions for All Items Included in the Simulations, by Survey

Survey	Indicator	Question	Original scale
Chile	Satisfaction question	I am satisfied with my job.	1 (strongly disagree) to 5 (strongly agree)
	Motivation question	I do my best to do my job, regardless of the difficulties.	1 (strongly disagree) to 5 (strongly agree)
	Performance review question	Did you have the opportunity to discuss the results of your last individual performance appraisal with your line manager?	0–1 dummy
	Merit-based recruitment question	Thinking about how you got your first job in the public sector—which of the following evaluations did you have to go through? (Written examination.)	0–1 dummy
	Motivation index	I am willing to start my workday earlier or stay after my hours of work to finish a pending job.	1 (strongly disagree) to 5 (strongly agree)
		I perform extra tasks at work, even if they are not really required.	1 (strongly disagree) to 5 (strongly agree)
		I put my best effort to perform my work, regardless of difficulties.	1 (strongly disagree) to 5 (strongly agree)
	Leadership index	My supervisor leads by setting a good example.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor says things that make employees proud to be part of this institution.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor communicates clear ethical standards to subordinates.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor personally cares about me.	1 (strongly disagree) to 5 (strongly agree)
	Performance	My superior evaluates my performance in a just manner.	1 (strongly disagree) to 5 (strongly agree)
		The feedback that I receive about my work helps me to improve my performance.	1 (strongly disagree) to 5 (strongly agree)
		If I put more effort in my work, I will obtain a better evaluation of my performance.	1 (strongly disagree) to 5 (strongly agree)
		A positive evaluation of my performance could lead to an increase in my salary.	1 (strongly disagree) to 5 (strongly agree)
		A positive evaluation of my performance could help me in obtaining a promotion.	1 (strongly disagree) to 5 (strongly agree)
		A negative evaluation of my performance could be a reason for termination.	1 (strongly disagree) to 5 (strongly agree)
Liberia	Satisfaction question	To what extent would you say you are satisfied with your experience of the civil service?	1 (very dissatisfied) to 4 (very satisfied)
	Motivation question	How motivated are you to work as a civil servant today?	0 (not motivated at all) to 10 (extremely motivated)
	Management index	Does your unit have clearly defined targets?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		How are targets and performance measures communicated to staff in your unit?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		When arriving at work every day, do staff in the unit know what their individual roles and responsibilities are in achieving the unit's goals?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		Does your unit track its performance to deliver services?	0–1 dummy
		How does your unit track its performance to deliver services?	Five descriptive answers progressively aligned from least to most positive description of the practices in question

(continues on next page)

TABLE 20.2 Overview of Survey Questions for All Items Included in the Simulations, by Survey *(continued)*

Survey	Indicator	Question	Original scale
Liberia *(continued)*	Management index *(continued)*	How much discretion do staff in your unit have when carrying out their assignments?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		Can most of the staff in your unit make substantive contributions to the policy formulation and implementation process?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		Is your unit's workload evenly distributed across its staff, or do some groups consistently shoulder a greater burden than others?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		Consider about the projects that your unit has worked on. Do the managers try to use the right staff for the right job?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		Does your unit try to adjust how it does its work based on the needs of the unit's clients/stakeholders who benefit from the work?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		How flexible is your unit in responding to new and improved work practices and reforms?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		How do problems in your unit get exposed and fixed?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		Consider if you and your colleagues agreed to an Action Plan at one of your meetings. What would happen if the plan was not being implemented or failed to meet the set deadlines?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		In your opinion, do the management of your unit think about attracting talented people to your unit and then do their best to keep them?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		If two senior-level staff joined your unit five (5) years ago and one performed better at their work than the other, would he/she be promoted through the service faster?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
Romania	Satisfaction question	Overall, I am satisfied with my job.	1 (strongly disagree) to 5 (strongly agree)
	Motivation question	I put forth my best effort to get my job done regardless of any difficulties.	1 (strongly disagree) to 5 (strongly agree)
	Performance review question	Has your superior discussed the results of your last performance evaluation with you after filling in your performance evaluation report?	0–1 dummy
	Merit-based recruitment question	Have you ever participated in a recruitment competition in the public administration?	0–1 dummy
	Motivation index	I am willing to do extra work for my job that isn't really expected of me.	1 (strongly disagree) to 5 (strongly agree)
		I put forth my best effort to get my job done regardless of any difficulties.	1 (strongly disagree) to 5 (strongly agree)
		I stay at work until the job is done.	1 (strongly disagree) to 5 (strongly agree)

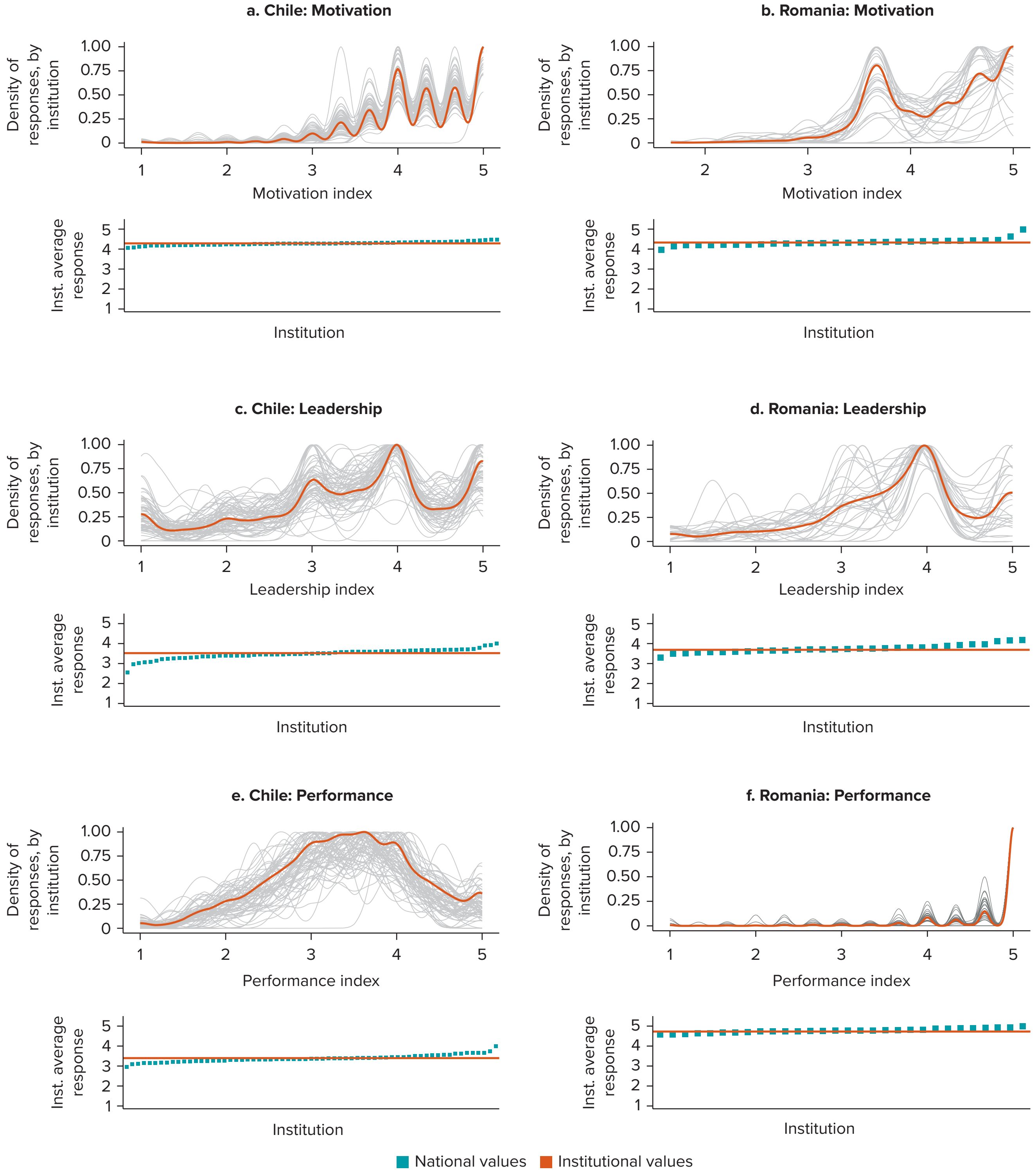
(continues on next page)

TABLE 20.2 Overview of Survey Questions for All Items Included in the Simulations, by Survey *(continued)*

Survey	Indicator	Question	Original scale
Romania *(continued)*	Leadership index	How frequently does your direct superior undertake the following actions? (Leads by setting a good example.)	1 (never) to 5 (always)
		How frequently does your direct superior undertake the following actions? (Says things that make employees proud to be part of this institution.)	1 (never) to 5 (always)
		How frequently does your direct superior undertake the following actions? (Communicates clear ethical standards to subordinates.)	1 (never) to 5 (always)
		How frequently does your direct superior undertake the following actions? (Personally cares about me.)	1 (never) to 5 (always)
	Performance	My performance indicators measure well the extent to which I contribute to my institution's success.	1 (strongly disagree) to 5 (strongly agree)
		My superior has enough information about my work performance to evaluate me.	1 (strongly disagree) to 5 (strongly agree)
		My superior evaluates my performance fairly.	1 (strongly disagree) to 5 (strongly agree)
United States	Satisfaction question	Considering everything, how satisfied are you with your job?	1 (strongly disagree) to 5 (strongly agree)
	Motivation question	I am willing to do extra work for my job that isn't really expected of me.	1 (strongly disagree) to 5 (strongly agree)
	Engagement index	In my organization, senior leaders generate high levels of motivation and commitment in the workforce.	1 (strongly disagree) to 5 (strongly agree)
		My organization's senior leaders maintain high standards of honesty and integrity.	1 (strongly disagree) to 5 (strongly agree)
		Managers communicate the goals of the organization.	1 (strongly disagree) to 5 (strongly agree)
		I have a high level of respect for my organization's senior leaders.	1 (strongly disagree) to 5 (strongly agree)
		Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor?	1 (strongly disagree) to 5 (strongly agree)
		Supervisors in my work unit support employee development.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor listens to what I have to say.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor treats me with respect.	1 (strongly disagree) to 5 (strongly agree)
		I have trust and confidence in my supervisor.	1 (strongly disagree) to 5 (strongly agree)
		Overall, how good a job do you feel is being done by your immediate supervisor?	1 (strongly disagree) to 5 (strongly agree)
		I feel encouraged to come up with new and better ways of doing things.	1 (strongly disagree) to 5 (strongly agree)
		My work gives me a feeling of personal accomplishment.	1 (strongly disagree) to 5 (strongly agree)
		I know what is expected of me on the job.	1 (strongly disagree) to 5 (strongly agree)
		My talents are used well in the workplace.	1 (strongly disagree) to 5 (strongly agree)
		I know how my work relates to the agency's goals.	1 (strongly disagree) to 5 (strongly agree)

Source: Original table for this publication.

FIGURE 20.3 Full-Sample Distribution of Key Indicators: National and Institutional Values across Surveys



a. Chile: Motivation

b. Romania: Motivation

c. Chile: Leadership

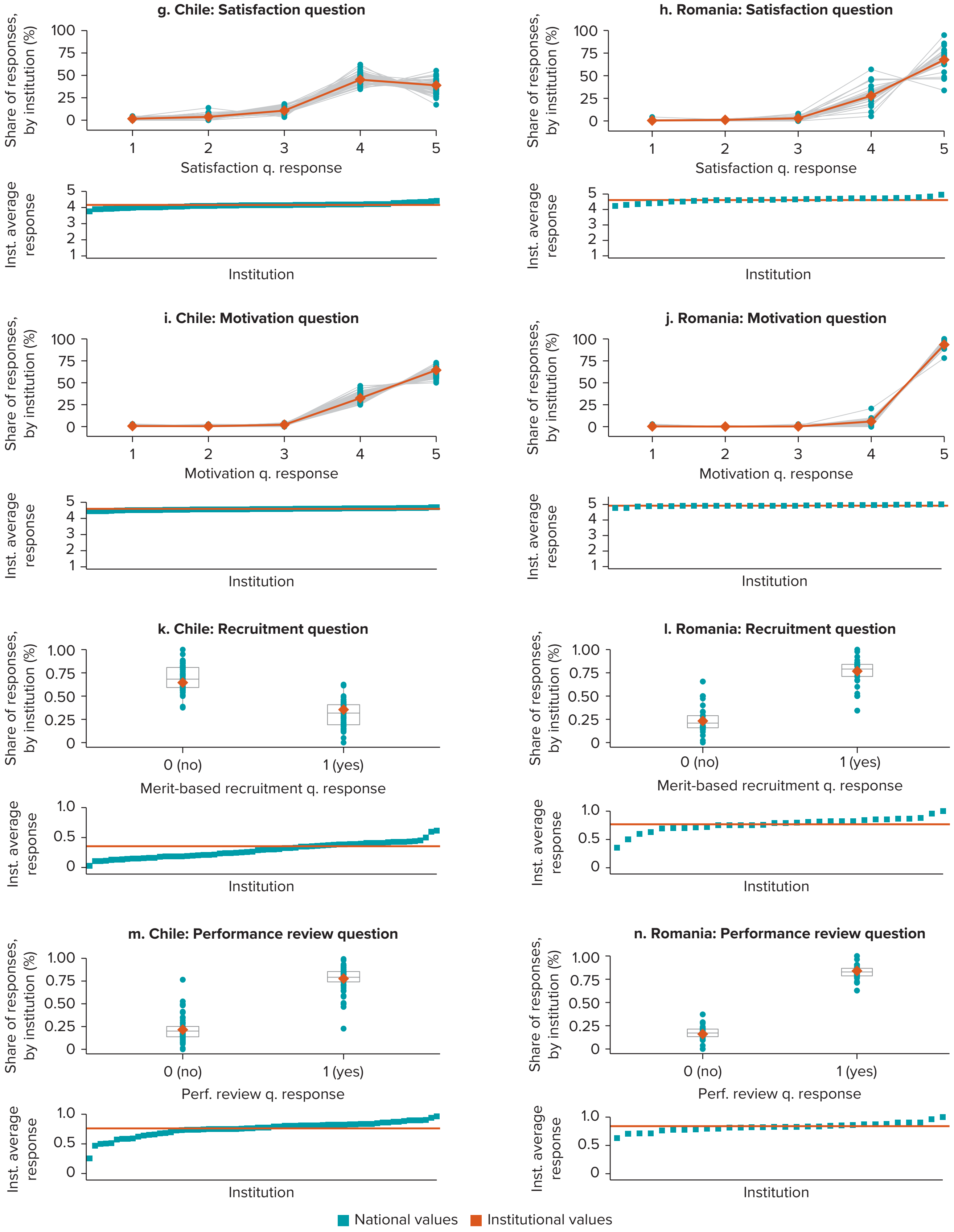
d. Romania: Leadership

e. Chile: Performance

f. Romania: Performance

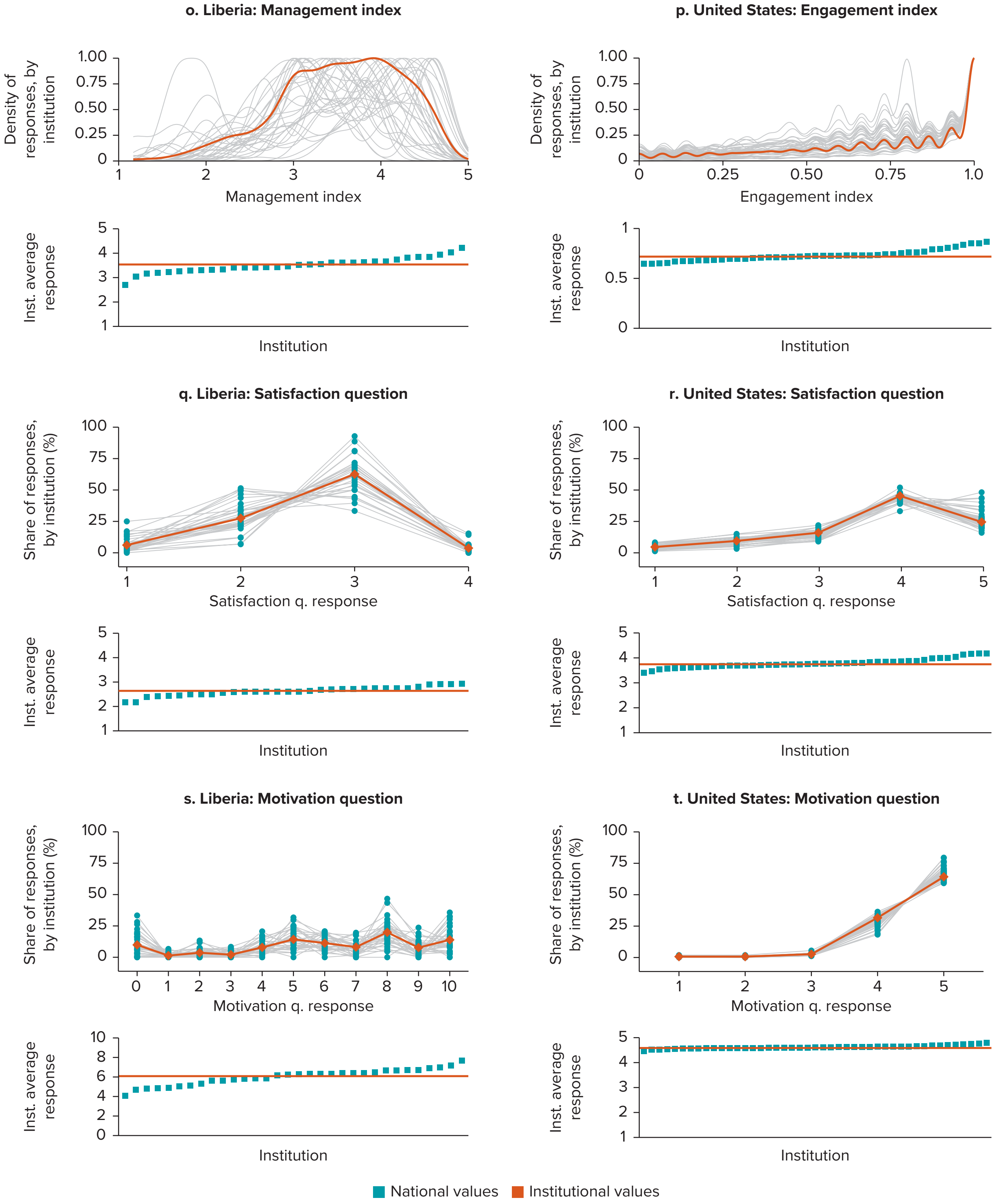
(continues on next page)

FIGURE 20.3 Full-Sample Distribution of Key Indicators: National and Institutional Values across Surveys *(continued)*



(continues on next page)

FIGURE 20.3 Full-Sample Distribution of Key Indicators: National and Institutional Values across Surveys *(continued)*



Source: Original figure for this publication.
Note: This figure shows the distribution of all key indicators analyzed in this chapter for the full sample of each of the surveys. Each subfigure refers to one indicator and survey and is divided into two panels. The top panel shows the distribution of responses, and the bottom panel shows ordered institution-level averages for a given indicator and survey. Red lines and points refer to aggregate values at the level of individual institutions within the civil service, whereas the blue ones refer to national-level values. Inst. = institution; perf. = performance; q. = question.

guidance on which sample sizes are more likely to yield satisfactory results—a goal that is further supported by the online sampling tool published alongside this chapter.

To do so, we use Monte Carlo simulation procedures to estimate:

- Sample means, standard deviations, and confidence intervals, to illustrate how sampling affects the statistical precision of estimates, and
- Differences in means between organizations and two groups of public servants that are often compared (managers and nonmanagers), to illustrate how sampling affects the possibility of statistically significant benchmarking between public sector institutions and groups of public servants—which is one primary use, in practice, of civil service survey data.

A random seed is set, from which a defined number of individual response IDs is randomly drawn, following the sampling strategy of the survey in question. This is repeated 1,000 times for each sampling proportion. All statistics presented here average across the number of simulations, providing an estimate for the average conclusions one would draw, given a certain number of individuals sampled, if the survey had been repeated 1,000 times. As a robustness check, we repeat each run of 1,000 simulations with a total of three different random seeds and record whether results deviate by more than 0.005 points on the answer scales. The results reported here have passed this robustness check.[7]

The results of the simulations are compared to means, standard deviations, confidence intervals, differences in means between manager and nonmanagers, and organizational rankings derived from the original surveys. In other words, we accept the statistics derived from these original surveys as the true sample statistics. We do not make statements of how these original means compare to “true” population means. We simply assume that the original sample sizes provide the best feasible estimates of underlying truths.

This approach has the advantage of not making assumptions about the population distribution beyond the information available to us. However, it is possible that the original sample sizes also over- or underestimated the true population parameters. If this is the case, results that indicate bias should be interpreted as lying even further away from the truth than when the original sample sizes were employed.

We evaluate the adequacy of the sample sizes using the following metrics:

- **The proportion of cases that fall within 95 percent of the confidence interval of the estimated means derived from the original samples.** Note that this metric is the inverse of what is typically used in statistics textbooks for the following reason: in our simulations, we sample smaller fractions of the original sample and see how well they perform in terms of recovering the original estimates. Mechanically, the confidence intervals for the estimates derived from samples with a small N will be larger than those derived from samples with a larger N. This means that it is more likely that a small sample includes the original mean, as it is wider. We instead want to know whether the estimated means of our new, smaller samples are close enough to the original mean (that is, within its confidence interval of 95 percent). For simplicity, we refer to estimates that fall within the 95 percent confidence interval of the original samples as estimates that have successfully been recovered.
- **The proportion of cases in which we find a significant difference between group means although there is none in the original data (type I error) and in which no significant difference is found although a difference between groups exists in the original data (type II error).** For the metrics presented here, we do not distinguish between the types of error that occur; we simply report the rate at which an error is made.
- **The proportion of cases in which an organization’s rank based on one of the metrics shifts into another performance quintile.** We use the proportion of shifts for ease of interpretation. For a more granular measure, we also calculate the Kendall’s tau rank correlation coefficient.[8]

The first metric illustrates the likelihood, given a sample size, that the means obtained are meaningfully different from those obtained from the original target sample size. The second metric illustrates the risk of drawing misleading inferences about differences in organizational subgroups. For smaller sample sizes, the risk increases that one might wrongly conclude—for instance—that managers rate organizational characteristics differently than nonmanagers, when they do not, or conclude the opposite, when they indeed think differently. The third metric illustrates the extent to which the robustness of organizational rankings is affected by reductions in sample size. One frequent use of civil service surveys—and employee engagement surveys in the private sector (for example, Harter et al. 2020)—is the benchmarking of organizations and units—be that the benchmarking of different public sector organizations, units within public sector organizations, or organizations across the public and private sector, or benchmarking with other countries.

Benchmarking is often deemed crucial to understanding strengths and weaknesses by showcasing how well a unit or organization performs in comparison to other, similar organizations or units. Given the limited variation and skew of many variables typically included in civil service surveys (see chapter 21)—and, as a result, the small differences between organizations—the individual ranks of organizations are likely to be highly sensitive to sample composition changes. We thus instead assess whether changes in sample size can move an organization into an entirely different tranche of organizations in benchmarking. For instance, if a unit changes from ranking in the bottom 20 percent of performers to the midrange, this can have serious consequences for how problematic or nonproblematic its performance is perceived to be. We thus focus on quintile changes due to sample composition changes.

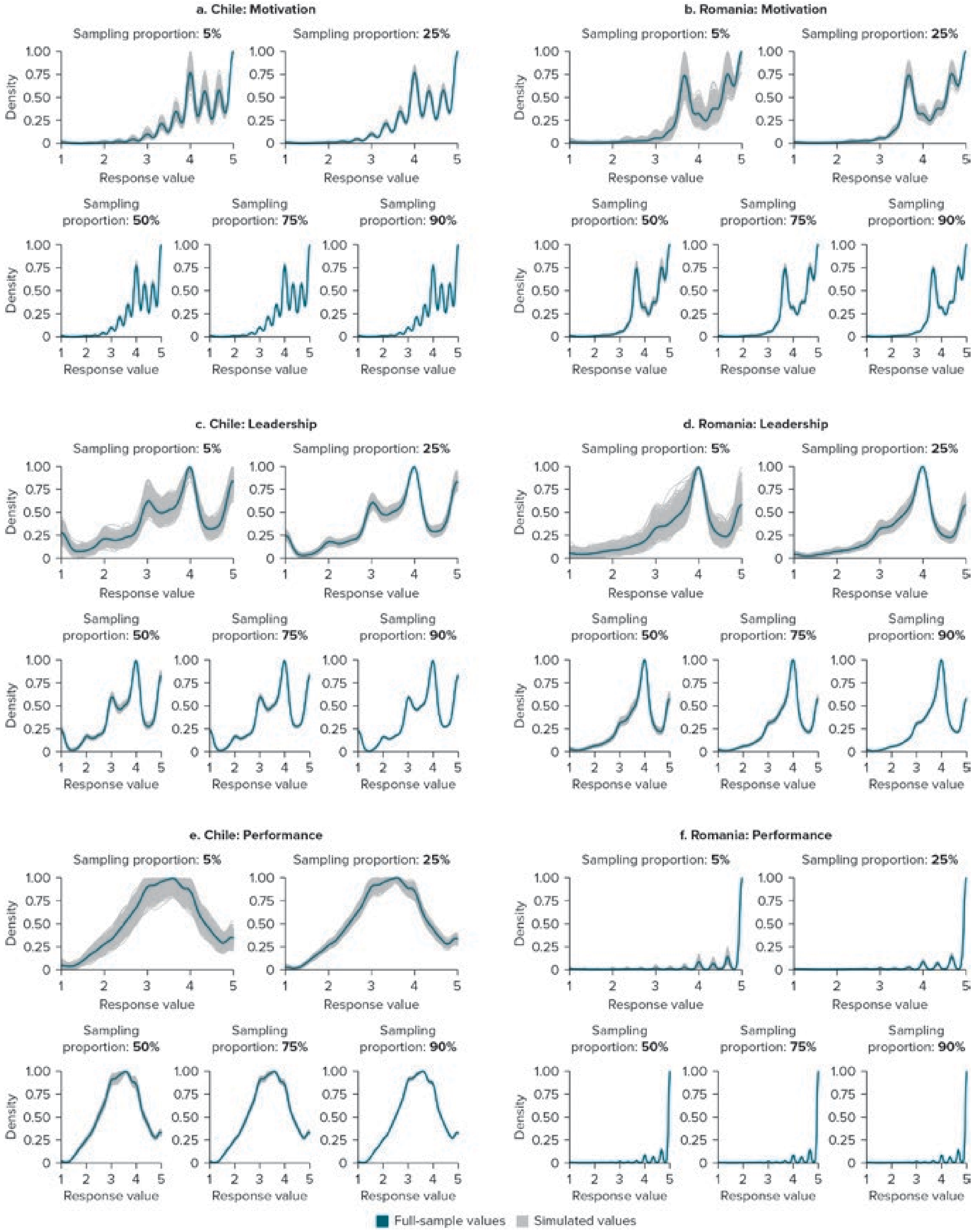
EMPIRICAL FINDINGS ON SAMPLE SIZE REQUIREMENTS FOR PUBLIC ADMINISTRATION SURVEYS

Figures 20.4 and 20.5 present our results graphically. Figure 20.4 showcases how the distribution of results would change with distinct sample sizes relative to what was collected in the case of each survey. Figure 20.5 showcases how the accuracy of the results—benchmarked against the statistics derived from the full sample—varies with the proportion of sample that is used.

The results of our simulations against our three metrics underscore that appropriate sample sizes are largely a result of the intended use of the survey results. Assessing, first, statistical precision—our first metric—we find, for most metrics across all four surveys, 50–60 percent of the original sample size suffices to estimate means that fall within the 95 percent confidence interval of the original mean. In other words, if the objective of a civil service survey is to recover reasonably precise statistical estimates about civil servants at the country level, all four surveys currently oversample respondents. While single random surveys with a considerably smaller sample size can lead to substantial over- and underestimates of means, on average, differences are small. They range between 0.002 and 0.13 points on a five-point scale, or, expressed differently, 4 percent and 15 percent of the original standard deviation. This can be considered a very small difference. The extent of these deviations varies somewhat across questions and country. Most countries score very similarly on measures of motivation and job satisfaction. For such measures, smaller sample sizes suffice when the goal is to calculate simple country averages. As detailed in chapter 21, questions on management practices, by contrast, offer more variation. For instance, for countries like Chile, where there is considerable variation across organizations in terms of whether and how they conduct performance reviews, larger sample sizes are required to assess these indicators adequately.

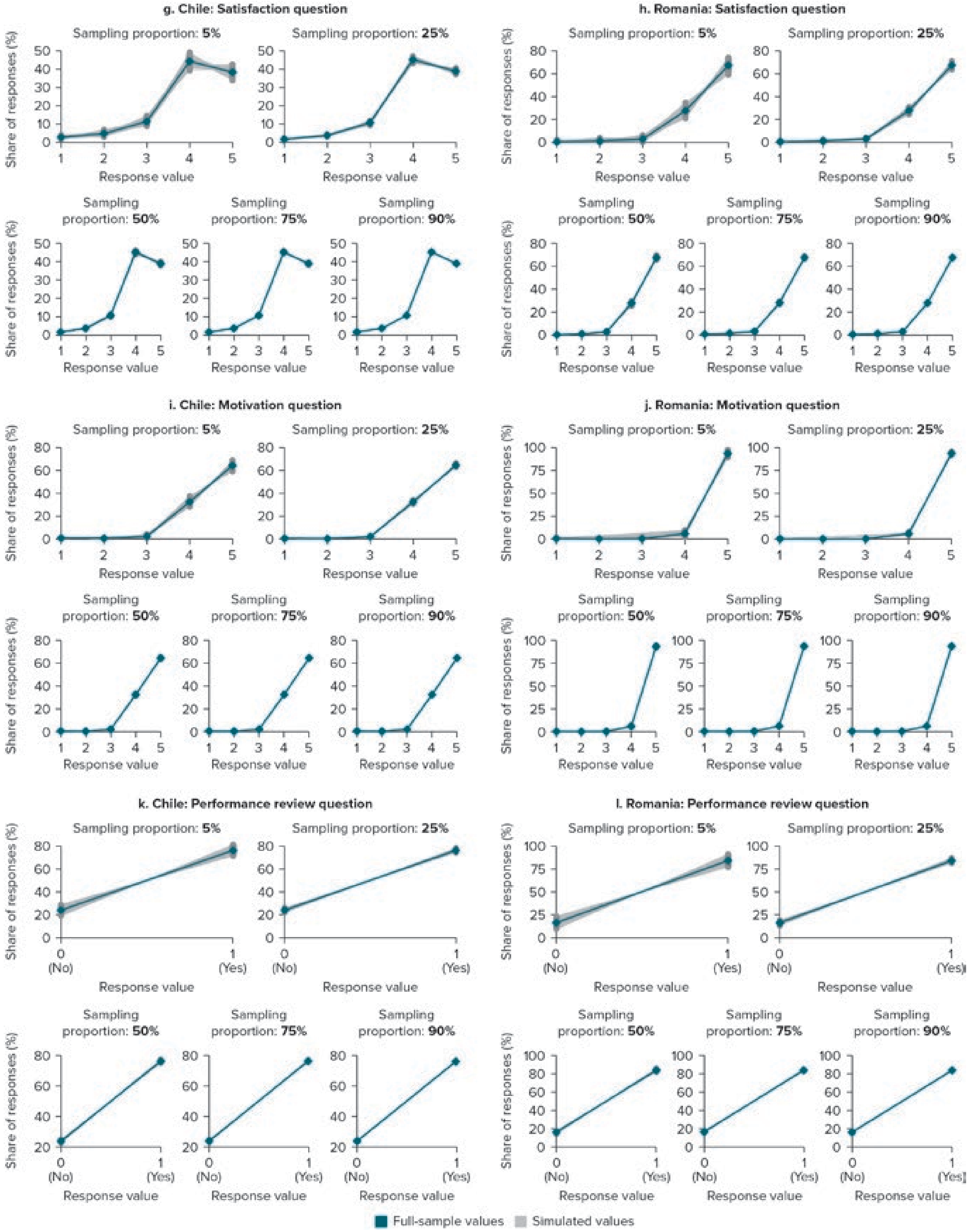
While our first metric suggests that countries oversample, our second and third metrics yield different conclusions. Consider, first, the results on the benchmarking of organizations. We find, as expected given the limited variation in many civil service survey indicators, that individual ranks are highly susceptible to changes in sample composition. In particular, if fewer than 80–90 percent of civil servants are sampled, conclusions about how institutions rank on key measures change significantly. For Romania, for instance, even when only 10 percent fewer civil servants are sampled, 50 percent of institutions change rank. At 90 percent sampled, most institutions get shuffled by one rank (there are 30 organizations in total in the sample). When only 60 percent are sampled, this increases to two to three ranks, and when only 40 percent are sampled, to

FIGURE 20.4 Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions



(continues on next page)

FIGURE 20.4 Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions *(continued)*



g. Chile: Satisfaction question

h. Romania: Satisfaction question

i. Chile: Motivation question

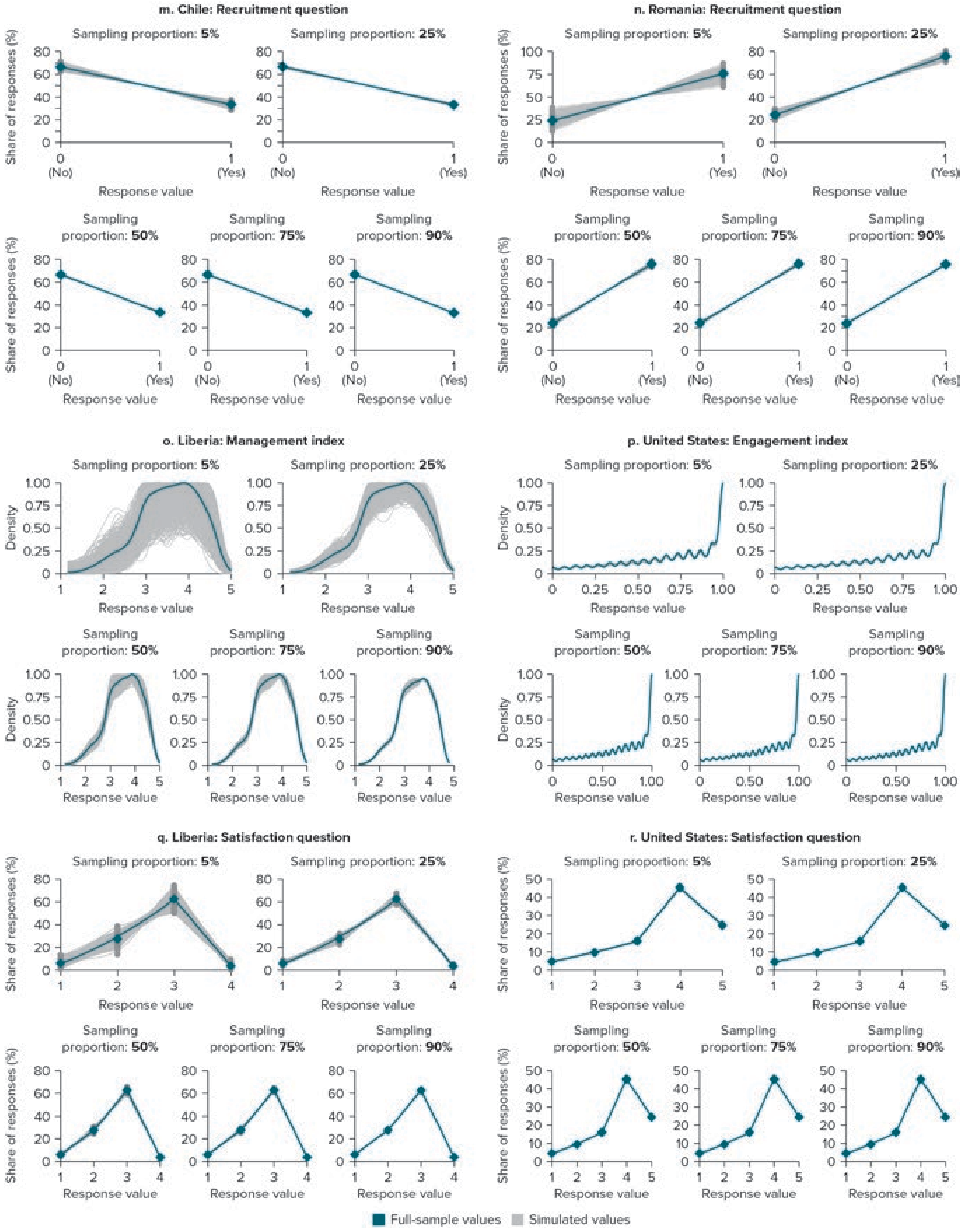
j. Romania: Motivation question

k. Chile: Performance review question

l. Romania: Performance review question

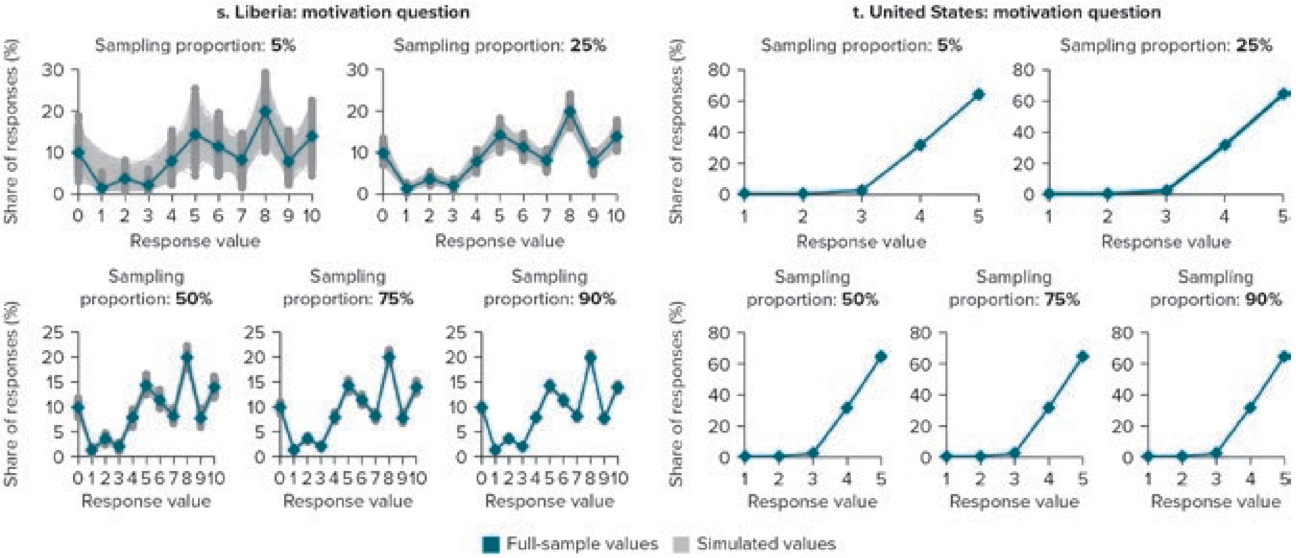
(continues on next page)

FIGURE 20.4 Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions *(continued)*



(continues on next page)

FIGURE 20.4 Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions *(continued)*



Source: Original figure for this publication.
Note: This figure shows the distribution of all key indicators analyzed in this chapter across each of 1,000 simulations at different sampling proportions. The sampling proportion is specified in percentage terms on top of each line plot. Therefore, each line plot shows 1,000 simulated distributions of responses to a given question, which were obtained when a given percentage of respondents were randomly sampled from the original full-sample distribution. Gray lines and points refer to national-level distributions obtained from each simulation, whereas the blue ones refer to full-sample values obtained in the actual survey.

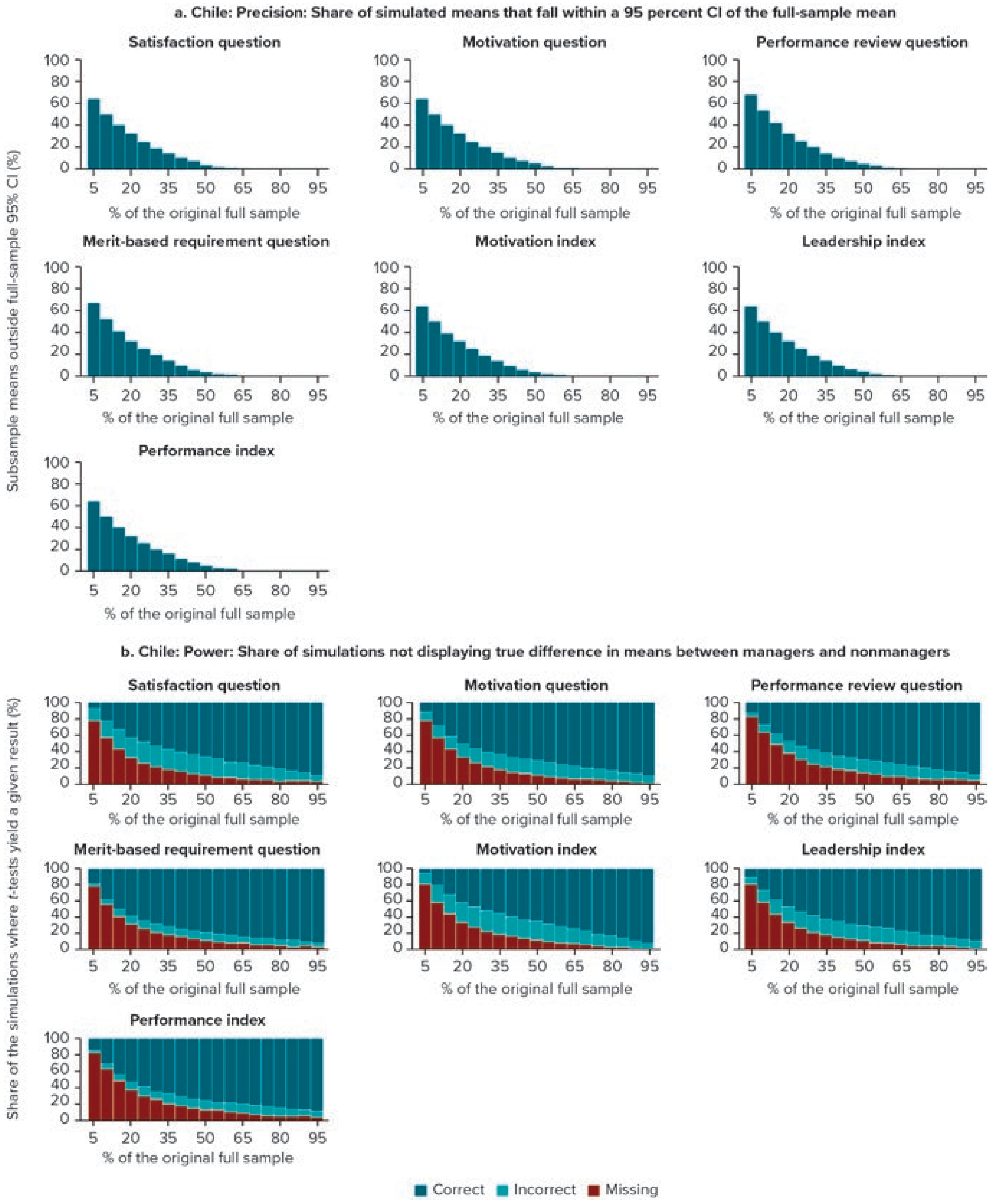
three to four ranks (see appendix H for the variation in institution-level values across simulations). We can also express this in terms of the Kendall's tau rank correlation coefficient, which indicates how well rankings obtained from the original data set correlate with those of the smaller samples. A rank correlation of one indicates a perfect match, and one of zero that no ranks matched. A correlation of 0.8 or more is considered desirable. This is only attainable when 80 percent or more of the original sample is surveyed for most measures. For measures with more condensed variances (motivation), sampling 60 percent or more of the original sample can achieve a similar result.

Looking at absolute shifts, however, might allow variability to appear disproportional. Often, governments, watchdogs, and international organizations group institutions into high and low performers. If we group institutions into quintiles, even at 80 percent sampled, 20–30 percent of them shift into another quintile. In other words, when 20 percent fewer civil servants are sampled, 20–30 percent of the institutions can end up being erroneously placed into the bottom 20 percent instead of the middle 20–40 percent of performers.

Another common type of analysis conducted on data derived from civil servant surveys is subgroup analysis. Statistics are typically broken down by characteristics such as job level, gender, or minority status. In our simulation example, we illustrate what the sample size requirements would be if one were to compare statistics for managers and nonmanagers. For simplicity, we report the rate of total errors committed in tests of independence. Across surveys, we find that error rates are high as soon as anything less than the original sample size is sampled. This is the case because initial differences on most indicators are very small. For example, in the original Chile survey, managers' and nonmanagers' assessments of leadership, motivation, and performance differ by less than 0.1 standard deviations (SD) for leadership and performance indicators and by about 0.2 SD for motivation. Differences in the original surveys conducted in Romania (0.1 SD), Liberia (0.2 SD), and the United States (0.1–0.2 SD) are similarly small.

These small differences imply that the proportion of each of these subgroups needs to be rather large to be able to capture differences between the groups. Error rates for most indicators remain at around 20 percent with reduced sample sizes—considerably higher than the widely accepted 5–10 percent—until 90 percent or more of the original sample is recovered. For any sample sizes smaller than 50–60 percent of the original, indicators with an initially high variance, such as leadership in Chile, motivation in Romania,

FIGURE 20.5 Precision and Power of Simulated Distributions across Surveys

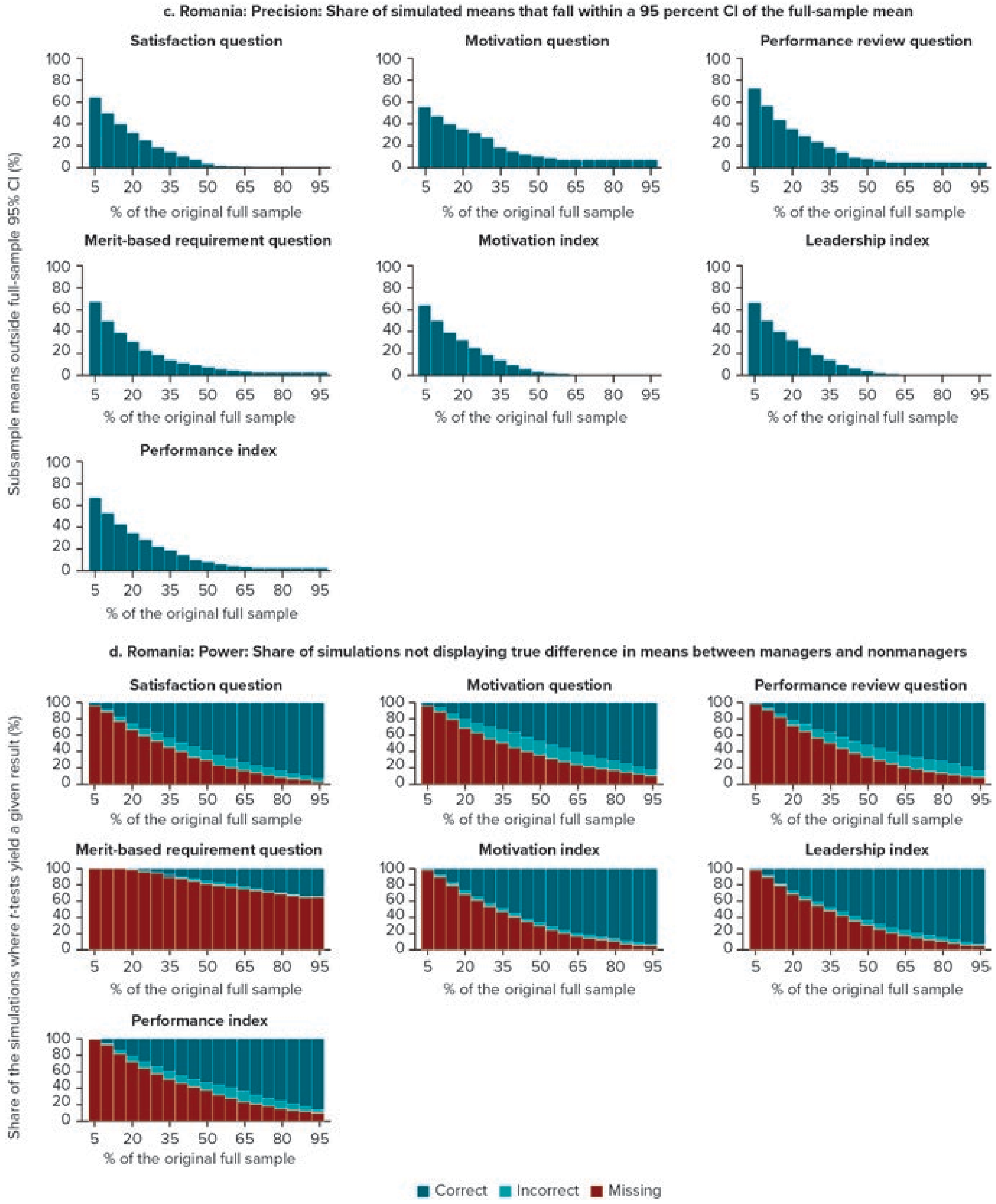


a. Chile: Precision: Share of simulated means that fall within a 95 percent CI of the full-sample mean

b. Chile: Power: Share of simulations not displaying true difference in means between managers and nonmanagers

(continues on next page)

FIGURE 20.5 Precision and Power of Simulated Distributions across Surveys *(continued)*

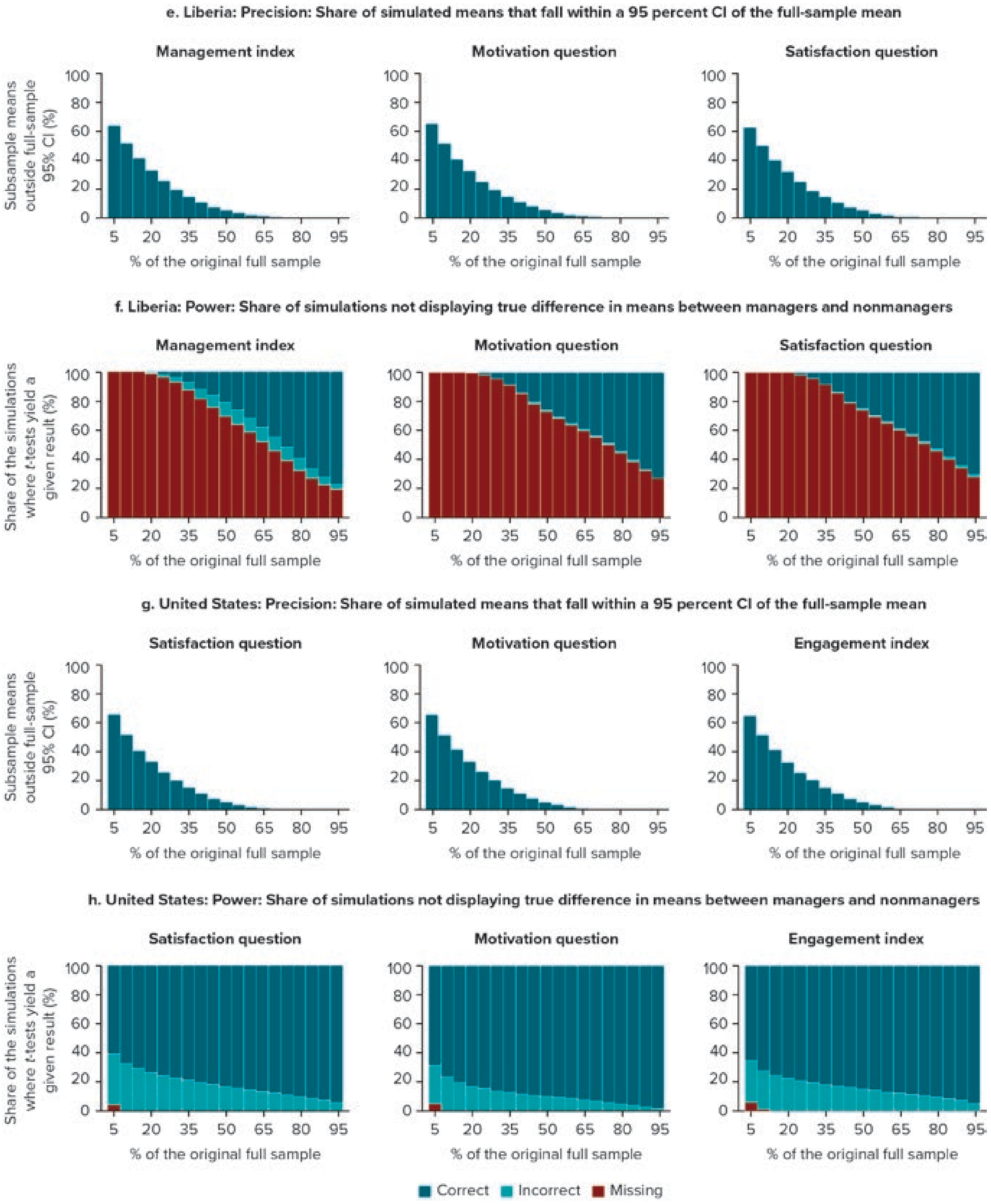


c. Romania: Precision: Share of simulated means that fall within a 95 percent CI of the full-sample mean

d. Romania: Power: Share of simulations not displaying true difference in means between managers and nonmanagers

(continues on next page)

FIGURE 20.5 Precision and Power of Simulated Distributions across Surveys *(continued)*



Source: Original figure for this publication.
Note: CI = confidence interval; perf. = performance; q. = question.

or management practices in Liberia, have very high error rates—most of the time, estimates fall far away from the true statistics.

The second challenge that conductors of civil service surveys should expect is that as sample sizes are reduced, it becomes more likely that statistics cannot be computed at all. For instance, simulations indicate that if one takes a rather conservative threshold of a minimum five observations per cell required to conduct comparisons, and if only 60–70 percent of the original sample is surveyed, in 20–30 percent of the cases, the statistic cannot be computed. The rate of failure quickly increases to 60–80 percent for questions that have a high rate of nonresponse (for example, the recruitment question in Romania).

DISCUSSION AND CONCLUSION: IMPLICATIONS FOR CIVIL SURVEY SAMPLING

In sampling respondents, civil service survey designers face a trade-off between the costs of additional survey responses and the benefits of more precise survey estimates with greater sample sizes. What, then, are the appropriate sample sizes in civil service surveys? To assess this conundrum, this chapter has conducted Monte Carlo simulations with civil service survey data from the United States, Chile, Liberia, and Romania. Our results suggest that appropriate sample sizes depend, most of all, on the inferences governments wish to make from the data. Conclusions differ depending on which indicators are chosen and which comparisons are made. Assessing sample size requirements on a case-by-case basis, depending on government needs and survey topics, thus remains paramount.

With that said, some common patterns across civil service surveys do exist that can inform future sampling decisions. For one, on attitudinal measures—such as work motivation or job satisfaction—smaller sample sizes might be sufficient if the objective is relatively precise means (though no benchmarking). To estimate averages for countries or larger organizations, sample sizes could often be reduced. Where there are differences in practice that vary substantially by institution, however, the required sample sizes for the country increase.

At the same time, where detailed comparisons among public sector organizations—or individual rankings—are sought, sample sizes are typically too small, not least because many survey measures do not offer large variation between organizations and thus require high levels of precision to enable comparison.

However, in such instances, practitioners should first assess whether the magnitude of historical differences is likely to be sufficiently meaningful to increase sample sizes to obtain statistically significant differences. For instance, does it merit changing organizational strategies if nonmanagers are 0.05 standard deviations less satisfied? Or would the gap need to be closer to one full standard deviation (which suggests a sizeable gap) to be substantively meaningful? If the answer is the latter, then increasing sample sizes to obtain statistically significant differences on the former would not be meaningful.

This chapter thus concludes that a determination of the use for survey results should precede the determination of sample sizes. Once that discussion has been had, practitioners can turn to an online toolkit to estimate appropriate sample sizes depending on the intended uses of the survey data. We recommend that practitioners look for countries, survey measures, and comparisons or benchmarking similar to their own use case in the online tool for guidance on which sample sizes are likely required in their own surveys.

NOTES

1. Our calculation is based on data from the CBO (2017).
2. More information about the experimental statistics program is available on the website of the ONS at https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/guidetoexperimentalstatistics.

3. Interested readers can access the toolkit at https://encuesta-col.shinyapps.io/sampling_tool/.
4. In this chapter, we assess sample sizes from the perspective of the types of inference that can be drawn from civil service survey data. For political reasons, governments may, of course, choose to undertake a census irrespective of whether this is necessary from a statistical perspective, to give every public employee the opportunity for *voice*—that is, the opportunity to give their feedback on matters of concern in the survey.
5. More information about the World Management Survey can be found on its website, https://worldmanagementsurvey.org/.
6. In practice, this would mean that if one were to sample again in the same country, using the same survey mode, response rates would look the same as for the last survey that was conducted.
7. Random seeds are used to enable replicable research. However, no computer-generated seed is truly random. Further, even if the starting seed is random, it is possible—although, by definition, very unlikely—that the random draws started from this seed end up being a very rare combination, leading to results not reflective of what most random draws would yield. Therefore, it is advisable to rerun all simulations with different seeds.
8. Kendall's tau is defined as: $\tau = \frac{2\ \left(n_{concordant} - n_{discordant}\right)}{n\ \left(n-1\right)}$.

REFERENCES

Bertelli, Anthony M., Mai Hassan, Dan Honig, Daniel Rogger, and Martin J. Williams. 2020. "An Agenda for the Study of Public Administration in Developing Countries." *Governance: An International Journal of Policy, Administration, and Institutions* 33 (4): 735–48. https://doi.org/10.1111/gove.12520.

Bjarkefur, Kristoffer, Luiza Cardoso de Andrade, Benjamin Daniels, and Maria Ruth Jones. 2021. *Development Research in Practice: The DIME Analytics Data Handbook.* Washington, DC: World Bank. http://hdl.handle.net/10986/35594.

CBO (Congressional Budget Office). 2017. *Comparing the Compensation of Federal and Private-Sector Employees, 2011 to 2015.* Washington, DC: CBO, Congress of the United States. https://www.cbo.gov/publication/52637.

Cochran, William G. 1977. *Sampling Techniques.* 3rd ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Fowler, Floyd. 2009. *Survey Research Methods.* 4th ed. Applied Social Research Methods. Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9781452230184.

Harter, James K., Frank L. Schmidt, Sangeeta Agrawal, Anthony Blue, Stephanie K. Plowman, Patrick Josh, and Jim Asplund. 2020. *The Relationship between Engagement at Work and Organizational Outcomes: 2020 Q12 Meta-Analysis.* 10th ed. Washington, DC: Gallup. https://www.gallup.com/workplace/321725/gallup-q12-meta-analysisreport.aspx.

Kalton, Graham. 1983. *Introduction to Survey Sampling.* Quantitative Applications in the Social Sciences. Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9781412984683.

Meyer-Sahling, Jan, Dinsha Mistree, Kim Mikkelsen, Katherine Bersch, Francis Fukuyama, Kerenssa Kay, Christian Schuster, Zahid Hasnain, and Daniel Rogger. 2021. *The Global Survey of Public Servants: Approach and Conceptual Framework.* Global Survey of Public Servants. Last updated May 2021. https://www.globalsurveyofpublicservants.org/about.

CHAPTER 21

Designing Survey Questionnaires

Which Survey Measures Vary and for Whom?

Robert Lipinski, Daniel Rogger, Christian Schuster, and Annabelle Wittels

SUMMARY

Many aspects of public administration, such as employee satisfaction and engagement, are best measured using surveys of public servants. However, evaluating the extent to which survey measures are able to effectively capture underlying variation in these attributes can be challenging given the lack of objective benchmarks. At a minimum, such measures should provide a degree of discriminating variation across respondents to be useful. This chapter assesses variation in a set of typical indicators derived from data sets of public service surveys from administrations in Africa, Asia, Europe, and North and South America. It provides an overview of the most commonly used measures in public servant surveys and presents the variances and distributions of these measures. The chapter thus provides benchmarks against which analysts can compare their own surveys and an investigation of the determinants of variation in this field. Standard deviations of the measures we study range between 0.72 and 1.24 on a five-point scale. The determinants of variation are mediated by the focus of the variable, with country fixed effects the largest predictors for motivation and job satisfaction, and institutional structure key for leadership and goal clarity.

ANALYTICS IN PRACTICE

- Effective questionnaire design and efficient sampling strategies both rely on an understanding of the performance of relevant survey measures. This chapter presents the variation in common measures used in public servant surveys from settings across the world (see table 21.2 later in the chapter for a full listing

Robert Lipinski is a consultant and Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Annabelle Wittels is an independent researcher. Christian Schuster is a professor at University College London.

of summary statistics). Such statistics provide individual survey analysts a means of comparative analysis with their own results.

- In the sample of surveys we assess, measures related to personal characteristics, such as motivation, do not vary as substantially as those relating to institutional variables, such as leadership. The design of questions about motivation and satisfaction may therefore need to be reconceptualized to better discriminate between degrees of these categories with the highest response likelihood.
- Commonly used measures of quantities of interest in public administration show significant negative skew across most contexts and for most measures. This indicates that responses are located mostly on the more-positive (higher) end of the answer scale. This might reflect response bias or an underlying lack of latent variation. It also indicates a need for redesigning standard measures in public service surveys to better discriminate between values at the top end of indexes.
- Where analysts would like to test for positive-response (or social-desirability) bias, they can include scale items specifically designed to capture such bias and apply regression or weight adjustments to averages, include alternative methods of capturing more nuanced levels of variation, or apply transformation techniques to address extreme skew before data analysis.
- The determinants of the variation we observe are mediated by the nature of the variable. Demographics generally explain a very small proportion of variation across measures (less than 2 percent). Country fixed effects account for the highest degree of variation for measures of motivation and job satisfaction. Institutional divisions (unit and subunit identifiers) explain a greater proportion of the variation in measures related to the quality of leadership and the clarity of a respondent's goals and tasks. Thus, survey measures associated with organizational features of the public service are more likely to exhibit variation than those that probe aspects influenced by servicewide or national cultures.
- Since many countries use different survey approaches and questionnaires, it is difficult to establish to what extent these differences are artificially created by differences in measurement as opposed to differences in environment, institutions, and management practices. This underlines the necessity to further standardize a core of public servants' surveys in order to make cross-country comparisons meaningful and informative for public administration reform.

INTRODUCTION

Surveys of attitudes, perceptions, and reported behaviors often try to assess two factors: the average or most common value of a concept of interest for respondents (or respondents in a particular subgroup) and variation in those responses. For example, one might want to find out the level of job satisfaction for public servants in an agency and the variation in satisfaction across that agency, across agencies, or across public servants of different managerial ranks.

Large-scale surveys are of particular relevance where a feature of the population being surveyed varies substantially. If satisfaction, or any other variable of interest, were known to be the same everywhere, analysts would only be required to carefully measure a single instance of the phenomenon. This would then be sufficient to know the value of the variable in the population at large. A practical example of this in the public service is the de jure nature of laws and regulations. Recording a single instance of a universal law is sufficient to understand its nature everywhere.

By contrast, once a phenomenon can vary across individuals, units, departments, agencies, time, and so on, surveys provide a tool to measure the underlying variation. Mapping this variation allows analysts to understand the average of the variable, its spread, where the feature takes extreme or unusual values, and so on. Again taking satisfaction as an example, a central public service agency may look for agencies that have the lowest levels of staff satisfaction, those where satisfaction is falling fastest, or those in which there are the largest disparities. Or, taking the universal-law example, analysts may be interested in how the law is de facto implemented across agencies, which may differ significantly. Surveys provide a tool for mapping the corresponding variation.[1]

Features of public administration increase the likelihood that there will be variation in key elements of the work environment. Unlike the private sector, there are no market forces driving work units to specific standards. The diverse range of activities undertaken by the public sector and the myriad outputs it produces imply potentially very different approaches to production. The challenges to measuring many aspects of public administration—externalities created by both tasks and public outputs, for example—compound the challenges to creating a common approach to management.

Not all phenomena of interest in the public service vary. It is conceivable that in some settings, public officials are universally oriented toward public service, or the opposite, such that even the best measures will exhibit no variation. The tension at the heart of measurement in public administration, where so few benchmark measures exist of a wide range of phenomena of interest, is how to identify which elements of the public service truly do vary, and for whom.

We proxy this underlying variation through observed variation in survey measures. Yet it is variation in the underlying phenomenon that we are interested in, rather than a proxy measured by a survey. We want measures that reflect true levels of satisfaction and allow us to discriminate between levels that are meaningfully different. If variation in survey measures solely reflects biases induced by the way questions are formulated or measurement error, it does not provide valid information upon which to base decision-making.[2] The validity of the variation in surveys of public administration is thus of key concern to understanding the public service.

This chapter aims to investigate the validity of measures from public servant surveys. The challenge all such exercises face is that many important concepts in public administration—such as satisfaction, motivation, and quality of management—are inherently internal phenomena. Assessing different measures against objective benchmarks, such as measures of satisfaction against turnover data, presents many issues of comparison. Alternative indicators of the validity of survey data are that conceptually related items should covary and that the same measurement should attain comparable variation across measurement in time and across survey contexts. This chapter assesses validity by comparing common measures across settings. By benchmarking which measures consistently vary across settings, we identify those measures that consistently provide differentiating variation.

The assumption of this approach to assessing the validity of variation is that the biases and measurement error that may drive variation in one setting are distinct from those in another. Thus, where we observe a measure providing discriminating variation across settings, we can infer that it is providing valid data. The disadvantage of such an approach is revealed when this assumption fails and measurement is affected by common bias across settings. For our approach to be valid, we also require that there be variation in the underlying phenomena across settings.

The payoffs for undertaking a valid assessment of variation in public servant surveys are substantial. Knowing what type and shape distributions of measurements of concepts of interest take is important for picking appropriate survey designs. Sampling strategies (see more on this in chapter 20) and question design (see chapters 22 and 23) require an understanding of underlying variances to be fit for purpose. The validity of our measurements of public administration, and of the corresponding survey designs, is at the core of our ability to understand the functioning of the state.

This chapter's perspective is that despite concerns about comparing measures of public service across time and space, such assessments act as rare and therefore valuable benchmarks to a single survey's results. Knowing that a specific measure has limited variation in many other settings allows analysts to take a more informed perspective on the use of that measure in their own contexts.

The importance of variation and its relation to validity and reliability has been investigated widely in volumes on statistics and survey sciences (for example, Brandler et al. 2007; Fink and Litwin 1995;

Grosh and Glewwe 2000; Wright and Marsden 2010). However, there is little systematic evidence available on patterns of variation of the measurements typically used to assess concepts central to the analysis of public administration. A review of common source bias in civil servant surveys conducted by George and Pandey (2017) makes evident that public administration as a field suffers from a reliance on surveys to measure both independent and dependent variables.[3] This approach inflates correlations between different variables and can make it difficult to distinguish between effects driven by individual-level error, individual-level differences, and generalizable relationships between different factors.[4] Most other work relates to scale validation. (For example, for public service motivation, see Kim 2009; Mikkelsen, Schuster, and Meyer-Sahling 2021; Perry 1996; for job satisfaction, see Cantarelli, Belardinelli, and Bellé 2016; for policy alienation, see van Engen 2017; for public leadership, see Tummers and Knies 2016 and chapter 24 on invariance.)

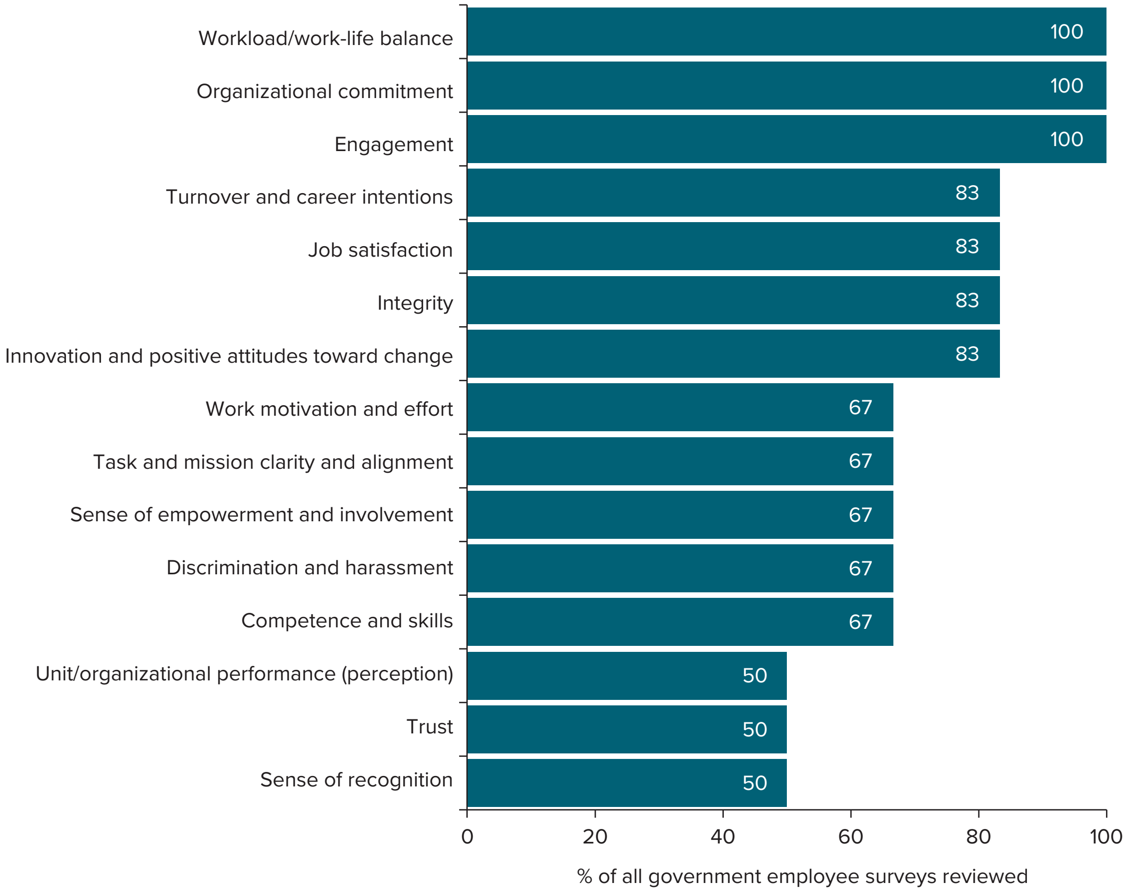
This chapter, therefore, assesses variation in a set of typical indicators derived from data sets of public service surveys to add to the existing literature and provide survey designers and analysts with benchmarks against which to assess their own efforts. The data span administrations in Africa, Asia, Europe, and North and South America. The chapter provides an overview of the most commonly used measures in public servant surveys and presents the variances and distributions of these measures. It then describes the extent to which observed variance can be explained by demographic and institutional characteristics typical of the analysis undertaken by analysts of public servant surveys.

KEY CONCEPTS IN PUBLIC SERVANT SURVEYS AND THEIR MEASUREMENT

What are the phenomena that surveys of public servants typically seek to measure? Meyer-Sahling et al. (2021) undertake a review of major surveys of public servants. They find that government employee surveys almost universally ask questions about workload or work-life balance and organizational commitment and engagement, and they tend to ask about career intentions, job satisfaction, integrity, and attitudes toward organizational change. Figure 21.1 provides a breakdown of the proportion of surveys that attempt to measure each of these phenomena. The figure indicates that though the precise set of measures used in government employee surveys varies, the core concepts themselves overlap significantly. Why have these specific topics become the major areas of investigation in surveys of public servants? Some measures that are central to existing government employee surveys, such as engagement, do not emerge clearly from reviews of academic models of public service governance (Meyer-Sahling et al. 2021). This is, in large part, due to the fact that "models" of public service governance do not engage in depth with organizational psychology. Such considerations are critical to the actual management of the public service. Management practices, such as work-life balance policies or leadership to generate enthusiasm for the mission of a public sector organization, are important predictors of the attitudes and behaviors of public servants (see, for example, Esteve and Schuster 2019) and feature prominently in government employee surveys, yet models of public service governance are (with some exceptions) silent about them.[5] Major surveys of public servants thus reflect the priorities of those who manage them, typically central agencies of public service management.

For this chapter, we assess the topics within surveys for which we both have access to the underlying microdata and which contain required identifiers (such as organization). We focus on those measures dealt with in a comparable way across these surveys. As we describe in more detail later, these topics are job satisfaction, pay satisfaction, motivation, assessments of leadership's trustworthiness and tendency to motivate, a measure of performance management related to promotion, and the clarity respondents have over goals and tasks. As can be seen from figure 21.1, these overlap closely with many of the standard topics in surveys of public servants. In this section, we review the existing evidence on the quality of measurement of these topics in surveys of public servants and their relationship to the effective functioning of public administration.

FIGURE 21.1 Topics Measured in Government Employee Surveys



Source: Meyer-Sahling et al. 2021.
Note: Meyer-Sahling et al. (2021) review the Federal Employee Viewpoint Survey (FEVS) in the United States, Canada's Public Service Employee Survey, the United Kingdom's Civil Service People Survey, the Australian Public Service Employee Census, Colombia's Survey of the Institutional Environment and Performance in the Public Sector, and Ireland's Civil Service Employee Engagement Survey. Questionnaires were reviewed for the last survey year prior to the COVID-19 pandemic. Only concepts covered in at least half of the surveys are shown.

Measuring Job Satisfaction

A review conducted by Cantarelli, Belardinelli, and Bellé (2016) finds that about a quarter of studies in public administration use a single item to measure job satisfaction. Three-quarters use an index based on several question items. They all use Likert-type response scales. Some measure overall feeling ("How satisfied are you?") while others measure specific aspects of job satisfaction, such as pay, career prospects, and work-life balance. Several surveys, such as the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS), use a mixture.

This diversity in measurement approach can partly be explained by the lack of a coherent theory as to how job satisfaction factors into public sector performance. Myriad authors have postulated links between satisfaction and public service performance, as well as corresponding interpretations of satisfaction (Hameduddin and Engbers 2022; Mehra and Joshi 2010; Potts and Kastelle 2010).[6]

Cantarelli, Belardinelli, and Bellé's (2016) meta-analysis reports significant correlations between the various measures of job satisfaction used in the literature and organizational commitment, feelings of inclusion, justice, goal clarity, turnover intentions, leadership perceptions, and positive perceptions of promotion systems and monetary incentives. Job satisfaction measures have also been shown to be strongly correlated with measures of employee mental health and burnout (Faragher, Cass, and Cooper 2013; Scanlan and Still 2019). These within-survey correlations are taken as indicators of the validity of measures of satisfaction.

Returning to the discussion on objective benchmarks of satisfaction, studies typically use turnover as an independent measure of satisfaction. However, longitudinal studies suggest that the relationship between job satisfaction and staff turnover might be temporal or spurious (Cramer 1996; Sousa-Poza and Sousa-Poza 2007). It is simply unclear to what extent turnover is a good marker of organizational productivity and good management in the context of a noncompetitive job market. Staff might simply stay because they have no attractive exit options.

Measuring Motivation

Motivation is most commonly measured via a scale developed by Perry (1996) or its adaptation by Kim (2009). Perry's scale consists of 24 items and has six dimensions: attraction to public policy making, commitment to the public interest, civic duty, social justice, self-sacrifice, and compassion. Each of the dimensions is measured by four questions with a Likert-type response scale. The measure developed by Kim consists of 12 items and has four dimensions: attraction to policy making (APM), commitment to the public interest (CPI), compassion (COM), and self-sacrifice (SS).

Mikkelsen, Schuster, and Meyer-Sahling (2021) have found that Kim's scale can be used to compare relationships between motivation and other variables across cultural contexts but that means cannot be meaningfully compared across country contexts. In terms of concept validity, a large body of correlation-based studies shows robust correlations between measures of public service motivation and effort (Camilleri and Van Der Heijden 2007; Moynihan and Pandey 2007; Naff and Crum 1999). However, as the vast majority of these studies rely on correlations between self-assessments completed in a survey, they all suffer from the threat of common source bias (Favero and Bullock 2015; George and Pandey 2017; Meier and O'Toole 2010): the risk that correlations are an artifact of individuals' providing multiple ratings about themselves at the same point in time, without any external validation.

Causal studies are limited because intrinsic motivation cannot be manipulated directly. However, several experiments have shown that when extrinsic, nonfinancial rewards are increased, the performance of public sector employees improves (Ashraf et al. 2020; Bellé 2014, 2015; Bellé and Cantarelli 2015).

Measuring Leadership

Leadership in public administration has typically been measured with scales based in or borrowed from management science and psychology (Tummers and Knies 2016). Scales commonly measure specific types of leadership. Two types studied extensively are transformational and transactional leadership (for example, Hameduddin and Engbers 2022; Kroll and Vogel 2014; Pandey et al. 2016; Vigoda-Gadot 2007). More recently, there has been a movement to develop scales that are particular to the leadership challenges faced by public sector managers, such as working with a large and diverse network of stakeholders and responding to the demands of political principals, all while remaining accountable to a broad public (Tummers and Knies 2016; Vogel, Reuber, and Vogel 2020).

In terms of the validity of leadership scales, Hameduddin and Engbers (2022) show in a meta-analysis of studies on public service motivation that there is considerable evidence that assessments of leadership predict motivation levels in staff. The relationship seems to be consistent across country contexts. Problematically, as with motivation, the majority of studies measure both motivation and leadership with a single survey. They are thus subject to common source bias, and there is a risk that relationships are spurious (see George and Pandey 2017). Evidence derived through other methods increases confidence, however, that this is not the case. For example, a field experiment conducted by Bellé (2014) finds that transformational leadership interventions can increase motivation as measured by output quantity. Other examples are 360-degree assessments of leadership, whereby assessments are collected from managers themselves and staff (Vogel and Kroll 2019), and varying whether questions are asked at the organizational or individual level (see chapter 23). Both approaches have found significant relationships between assessments of leadership and motivation.

Measuring Performance Management

In the academic literature, the measurement of public performance is often equated with the management approaches in place (Bouckaert 2021). A lot of measurement on public sector productivity now is concerned with integrating the measurement of inputs with administrative data on outputs (for recent reviews and discussions, see Heinrich 2002; Somani 2021). Questions about performance management in public service surveys typically ask staff to report what management approaches are used in their organization. Survey items include questions on the frequency and adequacy of performance reviews; goal setting and goal clarity; and the recognition of good performance, promotion, and financial incentives. However, little consensus exists on the appropriate approach to the measurement of these concepts in public service surveys.

To the knowledge of the authors, no comprehensive reviews of the validity or reliability of such question items and scales have been published. However, government agencies in charge of running public service surveys report routinely undertaking reviews of the relevance and internal validity of such measures.[7]

Measuring Goal and Task Clarity

The concept of *goal clarity* was first developed under the umbrella of organizational psychology. Latham and Locke (1991) define it as a spectrum varying from "vague ('work on this task') to specific ('try for a score of 62 correct on this task within the next 30 minutes')." Greater goal clarity is theorized to improve performance because resources can be targeted at goals and there is less waste. It is also theorized to have a motivational role because it becomes clear to individuals what they need to do in order to do well (Latham and Locke 1991). Goal clarity and its opposite—goal ambiguity—have been regarded as particularly important in the public sector because the mission statements of government organizations are often vast, and outcomes are hard to measure (Jung 2014). Jung (2012) distinguishes between clarity relating to "target, time limit, and external evaluation."

Goal clarity is typically measured via self-ratings with questions such as "The work I do is meaningful to me," "I understand my agency's mission," and "I understand how I contribute to my agency's mission" (see, for example, Hoek, Groeneveld, and Kuipers 2018). Survey research and laboratory studies have suggested that such self-ratings are positively associated with perceived performance (Hoek, Groeneveld, and Kuipers 2018) and performance as measured by quantity and quality of output (Anderson and Stritch 2016).

Rasul, Rogger, and Williams (2021) demonstrate with observational data that expert ratings of *task clarity* ("How precise, specific, and measurable is what the division actually achieved?" and "How precise, specific, and measurable is the target?") are strongly associated with differences in tasks completed by public sector workers. Importantly, they find that for tasks rated as high in ambiguity, greater control over workers backfires (a reduction of 14 percentage points in completion rates in response to a standard deviation increase in their corresponding measure of management), while giving workers greater autonomy over how they manage their work increases task completion rates (a corresponding increase of 21 percentage points).

In sum, the public administration literature has established the relevance of the latent concepts we focus on in this chapter to key concerns of management: performance, motivation, and turnover. However, most studies have focused on correlations—and, to a lesser extent, causal relationships—without providing an overview of the expected distribution of these variables. It remains unclear to what extent practitioners and scholars should expect substantial variation from these variables, how they are typically distributed across civil servant populations, and to what extent such distributions should be expected to be uniform across employee groups. Answers to these questions are crucial for picking appropriate research designs—including questionnaire design, sampling strategies, and analytic approaches—and for spurring the improvement of existing survey measures.

SELECTION OF SURVEYS AND MEASURES

Assessing variation requires access to micro-level public service survey data across countries. To maximize microdata coverage, we combine data collected in public service surveys conducted by the Global Survey of Public Servants (GSPS) consortium (which was cofounded by two of the authors) with micro-level public service survey data published by the US federal government. To undertake the required analysis, we also require surveys that can identify the public sector organization (unit) of respondents, can identify the department (subunit) within the organization within which the sampled public administrator works, and that measure the topics most commonly covered in public administration surveys.

This process leads us to focus our analysis on 10 surveys from across Africa (Ethiopia, Ghana, and Liberia), Asia (China and the Philippines), Europe (Romania), North America (the FEVS in the United States), and South America (Chile, Colombia, and Guatemala). All surveys except the FEVS were undertaken by members of the GSPS. They were conducted between 2014 and 2020 and include a mix of online and face-to-face efforts. Each survey featured in this analysis targets core administrative entities—ministries, all nationwide (or federal) agencies, and, where applicable, local governments.

Although we select surveys to maximize comparability, we are not able to measure all concepts consistently across all settings. We therefore focus our analyses on concepts that can be compared across the majority of surveys and have been identified as concepts of interest for the public administration literature, as described above. The resulting set of questions pertains to job satisfaction, pay satisfaction, motivation, assessments of leadership's trustworthiness and tendency to motivate, a measure of performance management related to promotion, and the clarity respondents have over goals and tasks. Comparison with the topics in figure 21.1 indicates that the topics we focus on are key elements of major surveys of public servants.

Table I.1 in appendix I provides further details on the survey questions used from each of the surveys, their original and transformed scales, and the extent of missing observations in the underlying data. Across surveys, question items related to job and pay satisfaction, leadership, performance incentives, and goal and task clarity are nearly identical except for some small adjustments implemented in response to testing in the local context.[8]

All the measures are based on single items. Though this deviates from some common practices, such as the use of Perry scales to measure public service motivation, it reduces the dimensionality of comparison in our setting, where few surveys used similar indexes. Most survey outcomes use a response scale ranging from 1 to 5, where 1 is the most negative and 5 is the most positive response.[9] For surveys where this was not the case, we rescale outcomes to fit on a 1–5 scale. Where a midpoint is missing, scores are split and rounded up to the next full point on a 1–5 scale.

The resulting data set combines multiple surveys of public servants in as coherent a way as possible given the differences in the underlying questions. Given the paucity of published public servant survey data, this provides a relatively unique opportunity to understand the spread of responses to the typical measures contained in such surveys.

To augment the analysis of variance, we consolidate a set of explanatory variables from the surveys that are frequently used for subgroup analysis in reporting on public servant surveys. One of the two most common ways public administration statistics are investigated is by demographic categories. Breaking down statistics by employee demographics can be valid in its function to provide accountability to different interest groups (for example, ethnic minorities and women in the workforce).

It is unclear to what extent demographic characteristics have explanatory value. Parola et al. (2019) find in a meta-analysis that age and gender are significantly related to different levels of public sector motivation. However, the confidence intervals are large and span zero for gender in many specifications. The analysis does not control for other individual and job characteristics, such as time in the job and job type. Cantarelli, Belardinelli, and Bellé (2016) find no significant association between gender and age and leadership assessments. The literature on correlates between demographic variables and other measurements, such as job satisfaction, is largely lacking or based on studies with ad hoc and very small samples. For our analysis, we

TABLE 21.1 Surveys Used in the Analysis

Survey country	Year	Unit	*N*	Subunit	*N*
Chile	2019	Organization	31	Subunit within the government organization	417
China	2015	City administration	4	Subunit defined by nature of sector and associated task	28
Colombia	2020	Central government/local government	84	Ministry within central government/local government organization	488
Ethiopia	2016	Central government/local government	53	Ministry within central government/sectoral office within local government	198
Ghana	2018	Central government organization	40	Subunit within the government organization	196
Guatemala	2019	Organization	15	Region	176
Liberia	2016	Central government organization	30	Subunit within the government organization	104
Philippines	2014	Central government agency	6	Locality within central agency	18
Romania	2019	Central government/regional government/local government	13	Ministry within central government/sectoral office within regional and local government	54
United States	2019	Agency	30	Level one units (one level below agency)	222

Source: Original table for this publication.

thus refer to the following set of variables as *demographics*: gender, age, tenure in public service, and managerial level. Where demographic characteristics are missing, we impute the median response for continuous and ordinal variables and the mode for categorical variables.[10]

The second type of explanatory variable typically used in studies of public administrations is institutional markers (for example, local or regional governments, organizations, agencies, and teams). Governments naturally want to understand how different government organizations and subunits compare in order to develop targeted strategies to improve performance. Once again, whether institutional divisions are strong predictors of variation in public service surveys is unclear. In their review of studies on motivation and leadership, Hameduddin and Engbers (2022) find no significant differences in the relationship between the two variables by the level of government. Table 21.1 provides details of the hierarchical level we use in each country to approximate organization (unit) and department (subunit).

WHICH PUBLIC SERVICE SURVEY MEASURES VARY?

Table 21.2 presents descriptive statistics for the surveys we assess, and figure 21.2 presents corresponding standardized distributions of the variables across surveys. In general, pay satisfaction is low while motivation and, to some extent, job satisfaction is high, in line with theories of the public service that see pay satisfaction as a limited component of public sector motivation. Assessments of leadership and meritocratic promotion receive some of the lowest scores across countries.

There is a degree of variation in all measures and in all countries. As shown in table 21.2, standard deviations in the aggregate panel (the means of the statistics in the rest of the table) range between 0.72 and 1.24 on a five-point scale. As a benchmark, if responses are uniformly distributed over a five-point scale, the standard deviation is 1.15.

A number of features stand out. First, there is a distinct negative skew to the variables, with modal responses of 4 or 5. This interpretation is summarized by the motivation scales' highly negative skew (−2.44 in the aggregate panel), indicating that most people report high levels of motivation.[11] Assessments of task and goal clarity and job satisfaction also show considerable—albeit more positive—skew, followed by those

TABLE 21.2 Descriptive Statistics for Surveys of Public Servants

Country	Variable	Mean	Median	SD	Skew	Shannon's entropy	*N*
Aggregate	Job satisfaction	3.88	4.29	0.92	–1.16	1.05	7
	Pay satisfaction	2.8	2.88	1.12	0.21	1.21	8
	Motivation	4.42	4.67	0.72	–2.44	0.85	6
	Leadership trust	3.84	4.2	1.13	–1.01	1.27	5
	Leadership motivates	3.66	4	1.12	–0.88	1.36	5
	Meritocratic promotion	3.54	3.75	1.24	–0.83	1.3	8
	Goal clarity	4.01	4.25	0.89	–1.31	1.13	8
	Task clarity	4.15	4.38	0.8	–1.5	1.02	8
Chile	Job satisfaction	4.16	4	0.87	–1.23	1.15	10,926
	Pay satisfaction	2.82	3	1.23	0.11	1.53	11,082
	Motivation	4.6	5	0.61	–1.99	0.78	10,955
	Leadership trust	3.75	4	1.18	–0.85	1.43	10,605
	Leadership motivates	3.52	4	1.22	–0.57	1.51	10,675
	Meritocratic promotion	2.7	3	1.4	0.24	1.58	9,303
	Goal clarity	4.42	5	0.75	–1.66	0.97	10,973
	Task clarity	4.46	5	0.73	–1.67	0.94	10,978
China	Job satisfaction	3.85	4	0.68	–1	0.96	2,477
	Pay satisfaction	—	—	—	—	—	—
	Motivation	—	—	—	—	—	—
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	3.62	4	0.93	–0.64	1.3	2,473
	Goal clarity	—	—	—	—	—	—
	Task clarity	—	—	—	—	—	—
Colombia	Job satisfaction	4.43	5	0.76	–1.76	0.96	9,693
	Pay satisfaction	—	—	—	—	—	—
	Motivation	4.57	5	0.59	–1.7	0.77	9,710
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	—	—	—	—	—	—
	Goal clarity	—	—	—	—	—	—
	Task clarity	4.27	4	0.84	–1.51	1.08	17,595
Ethiopia	Job satisfaction	3.04	4	1.31	–0.3	1.2	1,117
	Pay satisfaction	2.12	2	1.17	0.81	1.13	1,125
	Motivation	—	—	—	—	—	—
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	2.91	3	1.54	–0.01	1.56	1,121
	Goal clarity	3.13	3	0.85	0.03	1.25	368
	Task clarity	2.93	3	0.81	0.41	1.17	368

(continues on next page)

TABLE 21.2 Descriptive Statistics for Surveys of Public Servants *(continued)*

Country	Variable	Mean	Median	SD	Skew	Shannon's entropy	*N*
Ghana	Job satisfaction	—	—	—	—	—	—
	Pay satisfaction	1.33	1	0.87	2.78	0.65	2,632
	Motivation	4.49	5	0.81	–2.18	0.93	1,103
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	4.25	5	1.09	–1.61	1.15	1,384
	Meritocratic promotion	4.6	5	1.05	–2.73	0.66	1,276
	Goal clarity	4.32	5	0.95	–1.37	1.12	1,503
	Task clarity	4.44	5	0.82	–1.51	1.01	1,510
Guatemala	Job satisfaction	—	—	—	—	—	—
	Pay satisfaction	3.18	4	1.07	–0.3	1.2	1,138
	Motivation	—	—	—	—	—	—
	Leadership trust	3.59	4	1.05	–1.11	1.26	579
	Leadership motivates	3.47	4	1.06	–0.93	1.32	585
	Meritocratic promotion	3.02	3	1.22	–0.18	1.54	574
	Goal clarity	4.08	4	1.01	–0.9	1.27	748
	Task clarity	4.28	5	0.88	–1.01	1.13	747
Liberia	Job satisfaction	3.31	4	1.09	–0.74	0.94	2,651
	Pay satisfaction	2.33	2	1.13	0.61	1.09	2,670
	Motivation	3.33	3	1.31	–0.43	1.55	2,687
	Leadership trust	3.89	5	1.36	–0.77	1.13	839
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	4.39	5	1.03	–1.91	0.94	486
	Goal clarity	3.84	4	1.04	–0.71	1.35	1,407
	Task clarity	3.82	4	1	–0.4	1.34	1,410
Philippines	Job satisfaction	—	—	—	—	—	—
	Pay satisfaction	2.9	3	1.14	–0.08	1.42	1,768
	Motivation	—	—	—	—	—	—
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	—	—	—	—	—	—
	Goal clarity	3.78	4	0.93	–1.04	1.2	1,766
	Task clarity	—	—	—	—	—	—
Romania	Job satisfaction	4.62	5	0.65	–2.18	0.79	2,716
	Pay satisfaction	4.09	4	1.17	–1.5	1.2	2,690
	Motivation	4.92	5	0.35	–6.29	0.27	2,726
	Leadership trust	4.01	4	0.88	–1.26	1.16	1,624
	Leadership motivates	3.88	4	0.96	–1.02	1.26	1,667
	Meritocratic promotion	3.82	4	1.5	–0.96	1.33	612
	Goal clarity	4.83	5	0.52	–3.97	0.48	2,707
	Task clarity	4.88	5	0.45	–4.94	0.38	2,723

(continues on next page)

TABLE 21.2 Descriptive Statistics for Surveys of Public Servants *(continued)*

Country	Variable	Mean	Median	SD	Skew	Shannon's entropy	*N*
United States	Job satisfaction	3.75	4	1.07	–0.88	1.36	573,255
	Pay satisfaction	3.59	4	1.15	–0.73	1.43	572,853
	Motivation	4.58	5	0.65	–2.04	0.82	601,274
	Leadership trust	3.96	4	1.18	–1.08	1.36	582,758
	Leadership motivates	3.17	3	1.25	–0.29	1.55	565,650
	Meritocratic promotion	3.25	3	1.23	–0.43	1.52	558,198
	Goal clarity	3.64	4	1.1	–0.84	1.39	569,466
	Task clarity	4.13	4	0.87	–1.38	1.12	598,601

Source: Original table for this publication.
Note: The table shows the mean, median, standard deviation (SD), skew, and Shannon's entropy for each variable in each survey data set we analyze. Skew indicates the extent to which observed values divert from the balance of observations on each side of the scale characteristic of normal distributions. Shannon's entropy is a measure of variation for categorical variables. It describes the log likelihood of a category's being observed. If the measure equals zero, then all observations are of the same category. If the measure equals one, then the number of observations per category is equal (or near equal). For the aggregate panel, the numbers presented are a simple average of those presented for individual surveys in the rest of the table. — = not measured.

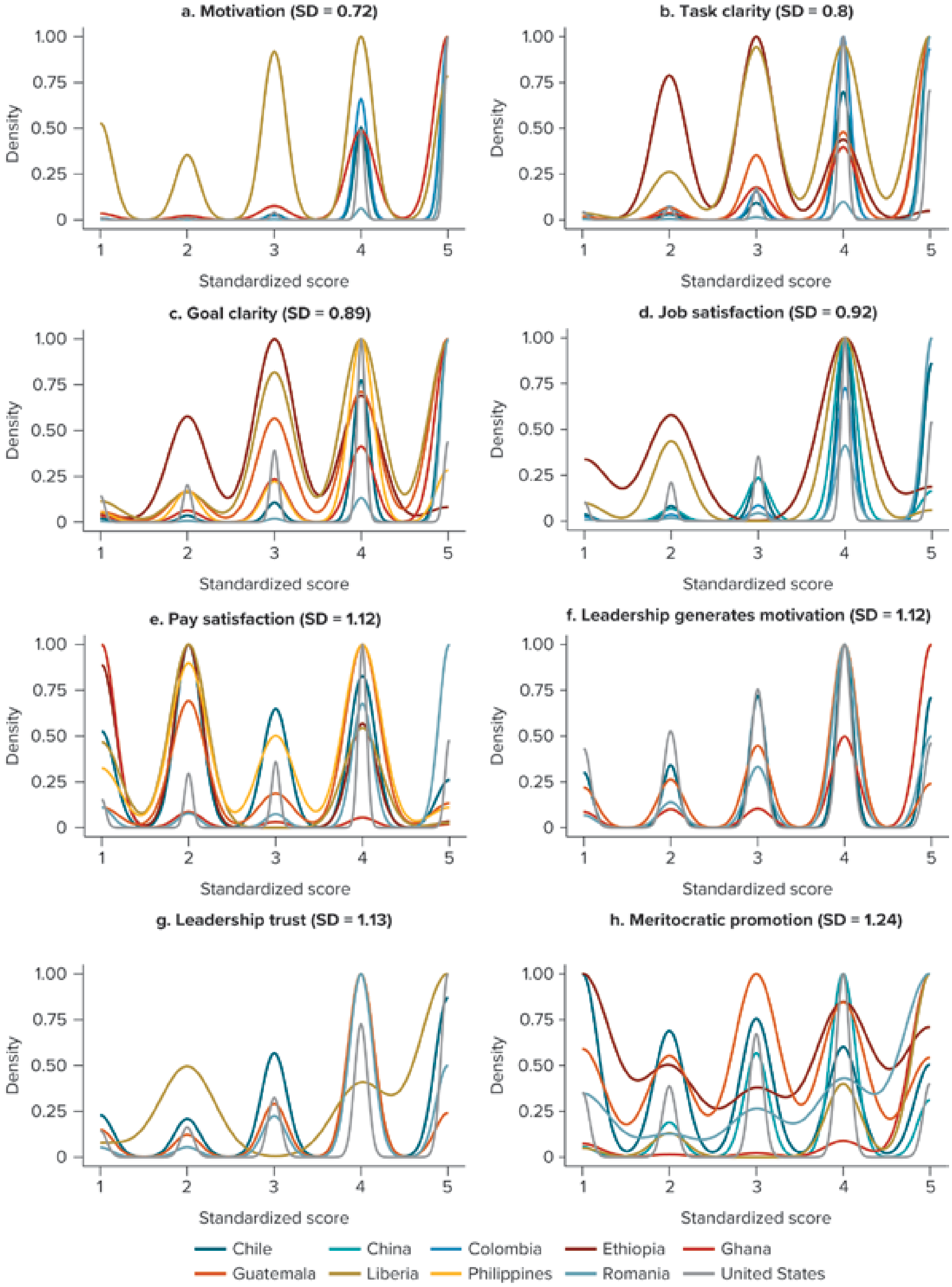
on leadership and promotion. Thus, current practice in measure design compacts a significant proportion of the variation into a minority of the response categories. This limits discriminating variation and the effectiveness of the measures.

Second, we find that responses vary more for measures that assess institutional characteristics, such as leadership, and the extent to which respondents perceive promotion to be linked to performance, relative to questions about individual characteristics. As the aggregate panel of table 21.2 shows, measures related to these topics all have standard deviations above one, while variation in measures of job satisfaction and motivation are lower. As can be seen in figure 21.2, responses about the respondent's characteristics tend to be rated toward the top of the scale for most individuals. This pattern tends to hold in all of the countries we assess.

A different way to see the relative variation in variables related to organizational features is presented in figure 21.3. The figure presents the standard deviation of individual topics across our surveys. The first three topics relate to measures of the self (how motivated a person is), the second group to interactions between the individual and the organization (the extent to which an individual understands the organization's goals), and the third group to perceptions of organizational characteristics (whether leadership is trusted). Figure 21.3 makes clear that across surveys and countries, we see a surprisingly large degree of commonality in which topics vary more than others and in the extent of variation for a single topic. This implies that there are commonalities in the nature of survey responses across settings. Comparing across topic groups, we see that concepts related to a general assessment of the organization exhibit greater variation than those more focused on the self.

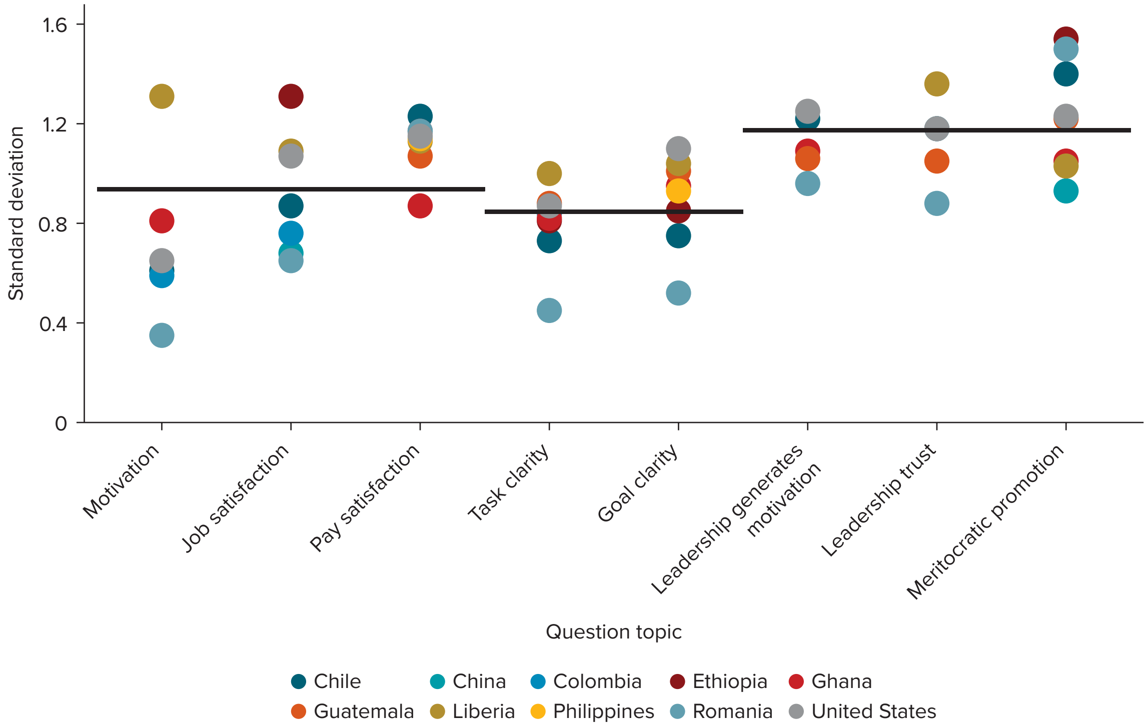
Since the data are rather skewed, relying on standard deviation as a summary statistic has its drawbacks. Patterns gleaned from looking at the standard deviation and skew of all measures are reflected in Shannon's entropy index. The index equals zero when all observations assume one value and increases as observations tend to assume different values in equal proportions. Using Shannon's entropy index, the variation in the aggregate panel ranges from 0.85 to 1.36. As a benchmark, if responses are uniformly distributed over a five-point scale, Shannon's entropy index equals one. Similar to values of standard deviations, the highest diversity according to Shannon's entropy index lies in leadership and performance questions, while motivation has the lowest index value (0.85), meaning the least diversity. When looking at measures within countries, rankings of diversity hold up in most cases where comparisons can be made, suggesting that patterns of diversity in responses apply across country contexts.

FIGURE 21.2 Distributions of Standardized Scores



Source: Original figure for this publication.
Note: The panels (corresponding to distinct topics) are ordered in ascending order of the standard deviation (SD) of the measure across surveys. The standard deviation shown above each panel is an average of those in the underlying surveys.

FIGURE 21.3 Average Response Variance across Surveys and Question Topics



Source: Original figure for this publication.
Note: Horizontal lines illustrate the average standard deviation (*y*-axis variable) across a given group of questions.

Fourth, the results in table 21.2 suggest that sample size is not a central mediating factor in the extent of variation. The degree of variation in surveys with a thousand or so respondents is not dissimilar to those with tens or hundreds of thousands. One interpretation of this fact is that the underlying distributions across public service entities are relatively stable and are not simply artifacts of measurement error. This could be taken as validation of our approach.

FOR WHOM DO PUBLIC SERVICE SURVEY MEASURES VARY?

To what extent do the survey measures documented in the last section vary substantively by country, organization, unit within the organization, and demographics? To investigate this question, we fit fixed-effect models to our focal measures and compare the explanatory power of these features.[12] For all surveys, we exclude subunits that have fewer than two respondents (2.2 percent of all respondents with nonmissing values of unit and subunit variables).

The results of a series of analysis of variance (ANOVA) exercises are displayed in table 21.3. Here, the dependent variable is the measure of the topic of interest, and each row of a panel is a distinct regression including variables outlined in the "Models" column. Demographic models include the variables listed above—that is, gender and tenure in public service, as well as age (for all except the United States) and managerial status (for Chile, Colombia, Ghana, Guatemala, and the United States). "Country," "unit," and "subunit" indicate the inclusion of fixed effects at the corresponding level, with unit and subunit defined along the lines outlined in table 21.1.

A broad assessment of the measures we include, which are typical factors drawn on in reports on public service surveys, indicates that they all explain a significant portion of the variation we seek to explore. Of the 24 *F*-tests we undertake, all are significant at the 5 percent level.

TABLE 21.3 Compare Models: ANOVAs, Nested

Variable	Model	Residual df	RSS	df	Sum of squares	*F*-stat	Pr	R^2	Adjusted R^2
Job satisfaction	Demographics	616,400	694,485					0.01	0.01
	Demographics + country	616,394	685,684	6	8,801.12	1,354	0.00	0.02	0.02
	Demographics + country + unit	616,091	675,248	303	10,436.06	32	0.00	0.04	0.04
	Demographics + country + unit + subunit	614,047	665,135	2,044	10,113.33	4.57	0.00	0.05	0.05
	Nested RE model							0.24	
Pay satisfaction	Demographics	609,406	826,499					0.01	0.01
	Demographics + country	609,399	795,521	7	30,977.73	3,586	0.00	0.05	0.05
	Demographics + country + unit	609,130	782,511	269	13,010.01	39.2	0.00	0.07	0.06
	Demographics + country + unit + subunit	607,351	749,507	1,779	33,004.34	15.03	0.00	0.10	0.10
	Nested RE model							0.40	
Motivation	Demographics	642,380	277,121					0.01	0.01
	Demographics + country	642,375	272,429	5	4,692.82	2,243	0.00	0.02	0.02
	Demographics + country + unit	642,093	270,640	282	1,788.93	15.16	0.00	0.03	0.03
	Demographics + country + unit + subunit	640,062	267,870	2,031	2,769.50	3.26	0.00	0.04	0.04
	Nested RE model							0.44	
Leader: Trust	Demographics	608,171	839,113					0.01	0.01
	Demographics + country	608,167	838,490	4	622.87	115.30	0.00	0.01	0.01
	Demographics + country + unit	608,033	830,463	134	8,027.20	44.36	0.00	0.02	0.02
	Demographics + country + unit + subunit	607,044	819,823	989	10,640.20	7.97	0.00	0.03	0.03
	Nested RE model							0.05	
Leader: Motivation	Demographics	591,842	928,523					0.01	0.01
	Demographics + country	591,838	924,733	4	3,789.80	638.46	0.00	0.01	0.01
	Demographics + country + unit	591,679	89,700	159	27,730.55	117.5	0.00	0.04	0.04
	Demographics + country + unit + subunit	590,411	876,150	1,268	20,852.75	11.08	0.00	0.06	0.06
	Nested RE model							0.14	
Meritocratic promotion	Demographics	585,446	868,379					0.02	0.02
	Demographics + country	585,439	861,128	7	7,250.71	733.26	0.00	0.03	0.03
	Demographics + country + unit	585,171	842,951	268	18,177.21	48.01	0.00	0.05	0.05
	Demographics + country + unit + subunit	583,544	824,321	1,627	18,629.89	8.11	0.00	0.07	0.07
	Nested RE model							0.27	
Goal clarity	Demographics	601,764	723,096					0.01	0.01
	Demographics + country	601,757	712,487	7	10,608.59	1,327	0.00	0.03	0.03
	Demographics + country + unit	601,488	698,168	269	14,319.39	46.60	0.00	0.05	0.04
	Demographics + country + unit + subunit	599,824	685,129	1,664	13,039.43	6.86	0.00	0.06	0.06
	Nested RE model							0.23	
Task clarity	Demographics	649,539	486,185					0.01	0.01
	Demographics + country	649,532	482,436	7	3,748.90	735.4	0.00	0.02	0.02
	Demographics + country + unit	649,155	477,083	377	5,353.09	19.50	0.00	0.03	0.03
	Demographics + country + unit + subunit	646,343	470,732	2,812	6,351.41	3.10	0.00	0.04	0.04
	Nested RE model							0.34	

Source: Original table for this publication.

Note: The first four lines for each variable summarize test statistics for analyses of variance and how the model fit compares to the next more complex model. The first row refers to a model that only includes demographic predictor variables. These include the respondent's gender and tenure in public service, as well as age (present in all surveys except for the United States) and managerial status (for Chile, Colombia, Ghana, Guatemala, and the United States). Individual missing values for age and tenure are imputed using the median and mean values, respectively. Missing values for the gender and managerial status variables are assigned to the "missing" category. Rows two through four progressively add country, unit, and subunit level dummies to the model. The *F*-test for each model indicates whether it has a better fit than the simpler model specified above. Models with lower residual sums of squares (RSS) and a higher (adjusted) *R*-squared explain a larger proportion of the variance. The last, fifth, line for each variable reports the model fit for a nested model that nests subunits into units and units into countries. If the *R*-squared of the nested model is larger than the values in the lines above it, the nested model is a better fit. ANOVAs = analyses of variance; df = degrees of freedom; Pr = probability associated with the *F*-statistic; RE = residual error.

Our analysis begins with an assessment of the extent to which basic demographic characteristics of respondents are predictive of their answers. Demographics explain between 0 and 2 percent of the variation across measures, with no clear pattern across different measures. Of the demographic variables, managerial position tends to explain the largest portion of variation, followed by age, gender, and tenure. Thus, public service measures vary most for managers compared with nonmanagers in the data sets we study.

Our ANOVA results suggest that the determinants of variation we observe are mediated by the nature of the variable. Country effects are significant throughout the analysis, but these may pick up both national commonalities in responses as well as differences in survey wording, enumeration, and so on. They are particularly important for respondents' assessments of their own characteristics, such as motivation, job satisfaction, and pay satisfaction. Thus, though more intimate features of self-identity vary the least, they are the most likely to be predicted by demographic features or national boundaries.

In rows three and four of each panel in table 21.3, we add measures of institutional structures indicating the unit and subunit the respondent works in. Focusing on the sum of squares each set of variables explains, we see that relative to country fixed effects, the institutional features explain a small proportion of the variance in job satisfaction, pay satisfaction, and motivation, in comparison to their much more significant role in assessments of leadership and organizational features (such as the extent to which promotions are generally meritocratic and how individual respondents understand organizational goals and tasks and their relationships to them). Institutional variables therefore appear to have more predictive power for those variables more closely aligned to hierarchy.

Intuitively, institutional structures are more predictive of those features of public service life generated by those structures. This implies that elements of public service defined most fully within the individual respondent, such as motivation, are in fact relatively stable across institutional settings. The core motivation of public servants seems relatively robust to their office, while perceptions of the quality of leadership are highly dependent on the unit and subunit in which an official works.

We perform a series of robustness exercises. Since three countries use different scales for three measures, we perform a robustness check whereby we rerun the main models excluding these countries. The results are presented in table I.2 in appendix I. We also rerun all analyses on data without imputation, using a listwise deletion instead. The results are presented in table I.3 in appendix I. Regression diagnostics indicate that none of the variables of interest has normally distributed error terms (see table I.4 in appendix I). Therefore, we rerun all models with the outcome variables transformed using Box-Cox transformations (see figure I.1 and table I.3 in appendix I for details). The robustness checks broadly support our core results.

Finally, we also fit mixed models in row five of each panel of table 21.3. Fixed-effect models do not account for the nested structure of data—public administrators who are located within subunits are nested in units that belong to organizations.[13] The mixed models have fixed effects for demographics, country, and unit and random effects for subunits nested within these. We do not fit random slopes as our main set of predictor variables is categorical and we have no clear hypotheses of interactions between predictor variables. Taking into account the nested structure of the explanatory variables does not significantly alter the interpretation.

DISCUSSION

There is little systematic evidence available on variation in the measurements typically used to assess the nature of public administration. In this chapter, we have provided descriptive statistics for, and assessed variation in, a range of the most common indicators of public administration. We have done so based on a unique data set of public service surveys conducted in 10 countries in Africa, Asia, Europe, and North and South America. The statistics presented in table 21.2 provide benchmarks for other analysts to use in assessing variation in their own surveys of public servants. They answer the question "Which public service survey measures vary?" The analysis in table 21.3 provides evidence of which features of public administration are predictive of these measures, and thus answers "For whom do public service survey measures vary?"

Our results point to less variation in measures related to personal characteristics, such as motivation, than in institutional variables, such as assessments of leadership. Personal characteristics are predicted more strongly by demographics and country fixed effects than institutional features, which are more strongly predicted by the units and subunits in which respondents work. The most substantial variation in surveys of public servants is in organizational characteristics, and these are determined by the office a respondent works in.

Our findings may reflect both the design of questions common in public servant surveys as well as a skew in the latent features of public service on which we have focused. For example, it may be that motivation is very high across all the public service entities we study, and our measures accurately reflect this. However, the negative skew we observe may be an indication that survey questions could be better designed and analyzed to explore the variation at the top of affected measures. Given the ambition of this chapter to inform the design of public servant surveys, we conclude with a discussion of avenues for responding to this finding.

Developing More-Discriminatory Measures

Validity of Scales and Skew

The first question raised by the compressed variance and extreme skew in most of these measures is whether these are artifacts of the survey measures employed, or whether they capture the realities of public administrations. As the introduction of this chapter summarized, for some measures, particularly motivation and leadership scales, an extant body of research on their validation reinforces our findings as reflections of reality. However, this does not preclude the possibility that current measures do not adequately capture distributions of concepts in real populations.

Observed patterns of skew could be driven by several factors related to measurement: social-desirability bias (see Kim and Kim [2016] for a discussion related to public service motivation), cognitive biases related to the choice of reference category, and extreme response bias (Tourangeau 2003). Public administrators may feel pressured to indicate high levels of motivation, for instance, in case their responses are ever disclosed (even if such disclosure never occurs in practice). Alternatively, there may be no desirability bias at play, but skew and kurtosis may simply be driven by cognitive biases. For instance, the *medium fallacy* is a common psychological bias that makes people believe they are better than the average person (which, statistically, cannot be true for everyone). Extreme response bias may also explain some of the observed patterns. It has been shown that some individuals have a greater tendency to pick extreme points on scales than others (Hibbing et al. 2019). One approach to these concerns is to tweak questions so that their scales have a greater range of options to discriminate between higher values of response. Another is to provide anchors to which respondents can relate their experiences.

Analysis Strategies and Skew

If measures are valid, the second concern raised by our observations of extreme skew in the data pertains to analysis methods. How can an analyst approach highly skewed data? There are several strategies that can be pursued to help address them.

The first is to include other questions in surveys that allow analysts to quantify the potential drivers of skew. For example, surveys could include social-desirability scales, which could then be used in regression analysis to (partially) control for bias introduced via this avenue.

A second strategy is to reweight data points by using transformations such as the log or Box-Cox transformations, as used in this chapter. Such an approach can "smooth out" the distribution of a skewed variable, conditional on a reinterpretation of the corresponding results.

Another strategy is to approach skewed responses differently than other points in the data. Several sophisticated strategies have also been developed to deal with extreme response bias. For example, item response tree (IRTree) models adjust for extreme responses by modeling a two-stage decision-making process. The multidimensional nominal response model (MNRM) recodes extreme responses as a separate dimension and includes them as dummies in regressions. Partial credit models use random effects to control for biases introduced by extreme responses (see Falk and Ju [2020] for a recent evaluation of their comparative performance).[14]

As this chapter has illustrated, there is a danger that the error terms of skewed variables are not normally distributed (and are potentially also heteroskedastic). Analysts can employ regression diagnostics, as used in table I.4 in appendix I, to assess the nature of their data more thoroughly. In response to the nonnormality of measures, they might consider employing bootstrapping methods in their analysis (see Afifi et al. 2007).

Building the Evidence Base Further

The specific culture of public service will determine the challenges to survey measurement that analysts will face. Though international comparisons are useful, particularly given the commonalities we have observed across surveys in this chapter, generating evidence on survey design is best done at the survey level. The analysis undertaken in this chapter could be repeated for multiple rounds of the same survey or for distinct departments or geographical regions covered by a survey. Such work builds a picture of which measures of public administration provide discriminating variation and which do not.

It has been difficult to assess the predictive validity of measures standard in public servant surveys. Assessments of discriminant validity are more common, but they could be expanded to address the theoretical overlap and imprecision of many concepts utilized in public administration research (see chapter 24 on discriminant validity for a recent evaluation). One key problem is that the vast majority of research in public administration, and the validation relating to the measurement used, is reliant on surveys (see Strauss and Smith [2009] for a discussion of developments in the philosophy of science on construct validity). Using the same methodology to test a measure can severely inflate its construct validity. Future research thus faces a pressing need to link survey and self-reported data to other ways of measuring the same concepts, such as administrative and behavioral data (for example, turnover, sick leave, performance ratings, output efficiency, and career progression). None of these measures is superior on its own to survey measurement. However, using Campbell and Fiske's (1959) multitrait-multimethod matrix methodology, the robustness of validity assessments of key concepts in public administration research can be improved: if the measured concepts are universal, they should manifest in different contexts and be detectable with a variety of methods. Their quantities should not change substantially as a function of method. Where adequate quantitative data are missing, qualitative methods could help to assess the validity of survey measures (see chapter 4 for a discussion of the problems with monolithic approaches to methodology).

Where experimentation is feasible, analysts may build evidence as to what is driving the (skewed) variation in responses. Cognitive biases could be addressed by using randomized controlled trials to systematically evaluate which features of a survey might cause greater skew in response. By combining this evidence with objective measures, where available, analysts can answer the questions posed in the title of this chapter with increasingly granular detail for their survey(s) of interest.

NOTES

1. Given that so many features of public administration may vary across units of observation, and the challenge of measuring these features, the use of surveys seems a natural response. An alternative approach would be to use administrative data to measure variation—for example, by using the extent to which officials leave a department (turnover data) as a measure of

satisfaction. But such a measure is very crude, only measuring satisfaction once it is at its lowest level and an official leaves the department, and has a range of other issues. Survey variation helps us understand the extent to which respondents perceive or experience things differently or similarly across the full distribution of values by asking the party of interest directly.

2. A survey measure is valid when it appears to measure the concept of interest (*face validity*) and covers relevant dimensions of the concept of interest (*content validity*), as well as to the extent that the measurement correlates with those that theoretically should be correlated (*criterion validity*) and captures variation not already captured by other variables (*discriminant validity*) (see chapter 24).
3. This reliance is very much based in the difficulty of accessing alternative data sources and the latent nature of most concepts of interest.
4. Guajardo (1996) looks into variation in demographic variables used as a proxy for diversity and representation in the public sector but restricts attention to studies with this common source bias.
5. This is not to say, of course, that there is not significant scholarship on these organizational psychology concepts in the public service (Esteve and Schuster 2019). Dozens of studies have, for instance, focused on leadership in the public sector. These organizational psychology concepts have, however, not been aggregated into a separate model of civil service governance, akin to Weberianism or new public management.
6. Measures of pay satisfaction are intimately linked to job satisfaction. They are commonly measured as a part of job satisfaction or separately, via a single item.
7. As evidenced by private correspondence between the authors and the Australian, Canadian, Irish, UK, and US governments, which forms the basis of the more in-depth case studies of measurement featured in chapters 25 and 26.
8. For example, in Ethiopia, the question on pay satisfaction was phrased "To what extent would you say you are satisfied with your salary?"; in Liberia, it was "How satisfied are you with your total income?"; and in Ghana, it was "My salary is very satisfactory."
9. The exception is work motivation, which is measured the same for all but three surveys. In Ethiopia, Liberia, and the Philippines, civil servants were asked to compare their motivation today to when they started. In Ethiopia and the Philippines, they were provided with an answer scale ranging from 0 to 100, while in Ethiopia, a 0–10 scale was used. In Ghana, civil servants were asked to rate the extent to which they "would feel an obligation to take time from my personal schedule to generate ideas/solutions for the organization if it is needed."
10. All surveys have data available on the respondent's gender and tenure in public service. The age variable is missing for the United States, and the managerial level is missing for China, Ethiopia, Liberia, the Philippines, and Romania.
11. Regression diagnostics indicate that none of the variables of interest has normally distributed error terms (see appendix I).
12. Our core approach uses *F*-tests to test for statistical significance, but we also run Wald tests using robust standard errors, and results do not differ.
13. Note that the unit-inside-organization classifier is not homogeneous across countries. For instance, in some cases, units are ministries, while in others, they are local governments, while subunits may refer to teams inside ministries or regional offices of ministries, for instance.
14. All such models can easily be implemented in standard statistical software; for example, in R, using packages such as *mirt* and *eRm*.

REFERENCES

Afifi, Abdelmonem A., Jenny B. Kotlerman, Susan L. Ettner, and Marie Cowan. 2007. "Methods for Improving Regression Analysis for Skewed Continuous or Counted Responses." *Annual Review of Public Health* 28: 95–111. https://doi.org/10.1146/annurev.publhealth.28.082206.094100.

Anderson, Derrick M., and Justin M. Stritch. 2016. "Goal Clarity, Task Significance, and Performance: Evidence from a Laboratory Experiment." *Journal of Public Administration Research and Theory* 26 (2): 211–25. https://doi.org/10.1093/jopart/muv019.

Ashraf, Nava, Oriana Bandiera, Edward Davenport, and Scott Lee. 2020. "Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services." *American Economic Review* 110 (5): 1355–94.

Bellé, Nicola. 2014. "Leading to Make a Difference: A Field Experiment on the Performance Effects of Transformational Leadership, Perceived Social Impact, and Public Service Motivation." *Journal of Public Administration Research and Theory* 24 (1): 109–36. https://doi.org/10.1093/jopart/mut033.

Bellé, Nicola. 2015. "Performance-Related Pay and the Crowding Out of Motivation in the Public Sector: A Randomized Field Experiment." *Public Administration Review* 75 (2): 230–41. https://doi.org/10.1111/puar.12313.

Bellé, Nicola, and Paola Cantarelli. 2015. "Monetary Incentives, Motivation, and Job Effort in the Public Sector: An Experimental Study with Italian Government Executives." *Review of Public Personnel Administration* 35 (2): 99–123. https://doi.org/10.1177/0734371X13520460.

Bouckaert, Geert. 2021. "Public Performance: Some Reflections and Lessons Learned." In *The Public Productivity and Performance Handbook,* 3rd ed., edited by Marc Holzer and Andrew Ballard, 68–73. New York: Routledge.

Brandler, Sondra, Camille P. Roman, Gerald J. Miller, and Kaifeng Yang. 2007. *Handbook of Research Methods in Public Administration.* Boca Raton, FL: CRC Press.

Camilleri, Emanuel, and Beatrice I. J. M. Van Der Heijden. 2007. "Organizational Commitment, Public Service Motivation, and Performance within the Public Sector." *Public Performance and Management Review* 31 (2): 241–74. https://doi.org/10.2753/PMR1530-9576310205.

Campbell, Donald T., and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56 (2). https://doi.org/10.1037/h0046016.

Cantarelli, Paola, Paolo Belardinelli, and Nicola Bellé. 2016. "A Meta-analysis of Job Satisfaction Correlates in the Public Administration Literature." *Review of Public Personnel Administration* 36 (2): 115–44. https://doi.org/10.1177/0734371X15578534.

Cramer, Duncan. 1996. "Job Satisfaction and Organizational Continuance Commitment: A Two-Wave Panel Study." *Journal of Organizational Behavior* 17 (4): 389–400. https://www.jstor.org/stable/2488549.

Esteve, Marc, and Christian Schuster. 2019. *Motivating Public Employees.* Elements in Public and Nonprofit Administration. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781108559720.

Falk, Carl F., and Unhee Ju. 2020. "Estimation of Response Styles Using the Multidimensional Nominal Response Model: A Tutorial and Comparison with Sum Scores." *Frontiers in Psychology* 11: 72. https://doi.org/10.3389/fpsyg.2020.00072.

Faragher, E. Brian, Monica Cass, and Cary L. Cooper. 2013. "The Relationship between Job Satisfaction and Health: A Meta-analysis." In *From Stress to Wellbeing: The Theory and Research on Occupational Stress and Wellbeing,* edited by Cary L. Cooper, 254–71. London: Palgrave.

Favero, Nathan, and Justin B. Bullock. 2015. "How (Not) to Solve the Problem: An Evaluation of Scholarly Responses to Common Source Bias." *Journal of Public Administration Research and Theory* 25 (1): 285–308. https://doi.org/10.1093/jopart/muu020.

Fink, Arlene, and Mark S. Litwin. 1995. *How to Measure Survey Reliability and Validity.* Thousand Oaks, CA: SAGE Publications. https://doi.org/10.4135/9781483348957.

George, Bert, and Sanjay K. Pandey. 2017. "We Know the Yin—–But Where Is the Yang? Toward a Balanced Approach on Common Source Bias in Public Administration Scholarship." *Review of Public Personnel Administration* 37 (2): 245–70. https://doi.org/10.1177/0734371X17698189.

Grosh, Margaret, and Paul Glewwe, eds. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study.* 3 vols. Washington, DC: World Bank. https://www.worldbank.org/en/programs/lsms/publication/designing-household-survey-questionnaires-for-developing-countries.

Guajardo, Salomon A. 1996. "Representative Bureaucracy: An Estimation of the Reliability and Validity of the Nachmias-Rosenbloom MV Index." *Public Administration Review* 56 (5): 467– 77. https://doi.org/10.2307/977046.

Hameduddin, Taha, and Trent Engbers. 2022. "Leadership and Public Service Motivation: A Systematic Synthesis." *International Public Management Journal* 25 (1): 86–119. https://doi.org/10.1080/10967494.2021.1884150.

Heinrich, Carolyn J. 2002. "Outcomes-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness." *Public Administration Review* 62 (6): 712–25. https://doi.org/10.1111/1540-6210.00253.

Hibbing, Matthew V., Matthew Cawvey, Raman Deol, Andrew J. Bloeser, and Jeffery J. Mondak. 2019. "The Relationship between Personality and Response Patterns on Public Opinion Surveys: The Big Five, Extreme Response Style, and Acquiescence Response Style." *International Journal of Public Opinion Research* 31 (1): 161–77. https://doi.org/10.1093/ijpor/edx005.

Hoek, Marieke van der, Sandra Groeneveld, and Ben Kuipers. 2018. "Goal Setting in Teams: Goal Clarity and Team Performance in the Public Sector." *Review of Public Personnel Administration* 38 (4): 472–93. https://doi.org/10.1177/0734371X16682815.

Jung, Chan Su. 2012. "Developing and Validating New Concepts and Measures of Program Goal Ambiguity in the US Federal Government." *Administration and Society* 44 (6): 675–701. https://doi.org/10.1177/0095399711413730.

Jung, Chan Su. 2014. "Why Are Goals Important in the Public Sector? Exploring the Benefits of Goal Clarity for Reducing Turnover Intention." *Journal of Public Administration Research and Theory* 24 (1): 209–34. https://doi.org/10.1093/jopart/mus058.

Kim, Sangmook. 2009. "Revising Perry's Measurement Scale of Public Service Motivation." *American Review of Public Administration* 39 (2): 149–63. https://doi.org/10.1177/0275074008317681.

Kim, Seung Hyun, and Sangmook Kim. 2016. "Social Desirability Bias in Measuring Public Service Motivation." *International Public Management Journal* 19 (3): 293–319. https://doi.org/10.1080/10967494.2015.1021497.

Kroll, Alexander, and Dominik Vogel. 2014. "The PSM–Leadership Fit: A Model of Performance Information Use." *Public Administration* 92 (4): 974–91. https://doi.org/10.1111/padm.12014.

Latham, Gary P., and Edwin A. Locke. 1991. "Self-Regulation through Goal Setting." *Organizational Behavior and Human Decision Processes* 50 (2): 212–47. https://doi.org/10.1016/0749-5978(91)90021-K.

Mehra, Kavita, and Kirti Joshi. 2010. "The Enabling Role of the Public Sector in Innovation: A Case Study of Drug Development in India." *Innovation* 12 (2): 227–37. https://doi.org/10.5172/impp.12.2.227.

Meier, Kenneth J., and Laurence J. O'Toole. 2010. "Organizational Performance: Measurement Theory and an Application: Or, Common Source Bias, the Achilles Heel of Public Management Research." Paper presented at the Annual Meeting of the American Political Science Association, September 1–5, 2010, Washington, DC.

Meyer-Sahling, Jan, Dinsha Mistree, Kim Mikkelsen, Katherine Bersch, Francis Fukuyama, Kerenssa Kay, Christian Schuster, Zahid Hasnain, and Daniel Rogger. 2021. *The Global Survey of Public Servants: Approach and Conceptual Framework.* Global Survey of Public Servants. Last updated May 2021. https://www.globalsurveyofpublicservants.org/about.

Mikkelsen, Kim Sass, Christian Schuster, and Jan-Hinrik Meyer-Sahling. 2021. "A Cross-Cultural Basis for Public Service? Public Service Motivation Measurement Invariance in an Original Survey of 23,000 Public Servants in Ten Countries and Four World Regions." *International Public Management Journal* 24 (6): 739–61. https://doi.org/10.1080/10967494.2020.1809580.

Moynihan, Donald P., and Sanjay K. Pandey. 2007. "The Role of Organizations in Fostering Public Service Motivation." *Public Administration Review* 67 (1): 40–53. https://doi.org/10.1111/j.1540-6210.2006.00695.x.

Naff, Katherine C., and John Crum. 1999. "Working for America: Does Public Service Motivation Make a Difference?" *Review of Public Personnel Administration* 19 (4): 5–16. https://doi.org/10.1177/0734371X9901900402.

Pandey, Sanjay K., Randall S. Davis, Sheela Pandey, and Shuyang Peng. 2016. "Transformational Leadership and the Use of Normative Public Values: Can Employees Be Inspired to Serve Larger Public Purposes?" *Public Administration* 94 (1): 204–22. https://doi.org/10.1111/padm.12214.

Parola, Heather R., Michael B. Harari, David E. L. Herst, and Palina Prysmakova. 2019. "Demographic Determinants of Public Service Motivation: A Meta-analysis of PSM-Age and -Gender Relationships." *Public Management Review* 21 (10): 1397–1419. https://doi.org/10.1080/14719037.2018.1550108.

Perry, James L. 1996. "Measuring Public Service Motivation: An Assessment of Construct Reliability and Validity." *Journal of Public Administration Research and Theory* 6 (1): 5–22. https://doi.org/10.1093/oxfordjournals.jpart.a024303.

Potts, Jason, and Tim Kastelle. 2010. "Public Sector Innovation Research: What's Next?" *Innovation* 12 (2): 122–37. https://doi.org/10.5172/impp.12.2.122.

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.

Scanlan, Justin Newton, and Megan Still. 2019. "Relationships between Burnout, Turnover Intention, Job Satisfaction, Job Demands and Job Resources for Mental Health Personnel in an Australian Mental Health Service." *BMC Health Services Research* 19 (1): 1–11. https://doi.org/10.1186/s12913-018-3841-z.

Somani, Ravi. 2021. *Public-Sector Productivity (Part 1): Why Is It Important and How Can We Measure It?* Equitable Growth, Finance and Institutions Insight. Washington, DC: World Bank. http://hdl.handle.net/10986/35165.

Sousa-Poza, Alfonso, and Andrés A. Sousa-Poza. 2007. "The Effect of Job Satisfaction on Labor Turnover by Gender: An Analysis for Switzerland." *Journal of Socio-Economics* 36 (6): 895–913. https://doi.org/10.1016/j.socec.2007.01.022.

Strauss, Milton E., and Gregory T. Smith. 2009. "Construct Validity: Advances in Theory and Methodology." *Annual Review of Clinical Psychology* 5: 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639.

Tourangeau, Roger. 2003. "Cognitive Aspects of Survey Measurement and Mismeasurement." *International Journal of Public Opinion Research* 15 (1): 3–7. https://doi.org/10.1093/ijpor/15.1.3.

Tummers, Lars, and Eva Knies. 2016. "Measuring Public Leadership: Developing Scales for Four Key Public Leadership Roles." *Public Administration* 94 (2): 433–51. https://doi.org/10.1111/padm.12224.

van Engen, Nadine A. M. 2017. "A Short Measure of General Policy Alienation: Scale Development Using a 10-Step Procedure." *Public Administration* 95 (2): 512–26. https://doi.org/10.1111/padm.12318.

Vigoda-Gadot, Eran. 2007. "Leadership Style, Organizational Politics, and Employees' Performance: An Empirical Examination of Two Competing Models." *Personnel Review* 36 (5): 661–83. https://doi.org/10.1108/00483480710773981.

Vogel, Dominik, and Alexander Kroll. 2019. "Agreeing to Disagree? Explaining Self–Other Disagreement on Leadership Behaviour." *Public Management Review* 21 (12): 1867–92. https://doi.org/10.1080/14719037.2019.1577910.

Vogel, Dominik, Artur Reuber, and Rick Vogel. 2020. "Developing a Short Scale to Assess Public Leadership." *Public Administration* 98 (4): 958–73. https://doi.org/10.1111/padm.12665.

Wright, James D., and Peter V. Marsden. 2010. "Survey Research and Social Science: History, Current Practice, and Future Prospects." In *Handbook of Survey Research,* 2nd ed., edited by Peter V. Marsden and James D. Wright, 3–26. Bingley, UK: Emerald.

CHAPTER 22

Designing Survey Questionnaires

To What Types of Survey Questions Do Public Servants Not Respond?

Robert Lipinski, Daniel Rogger, and Christian Schuster

SUMMARY

Surveys of public servants differ sharply in the extent of item nonresponse: respondents' skipping or refusing to respond to questions. Item nonresponse can affect the legitimacy and quality of public servant survey data. Survey results may be biased, for instance, if those least satisfied with their jobs are also most prone to skipping survey questions. Understanding why public servants respond to some survey questions but not others is thus important. This chapter offers a conceptual framework and empirical evidence to further this understanding. Drawing on the existing literature on survey nonresponse, the chapter theorizes that public servants are less likely to respond to questions that are complex (because they are unable to) or sensitive (because they are unwilling to). This argument is assessed using a newly developed coding framework for survey question complexity and sensitivity, which is applied to public service surveys in Guatemala, Romania, and the United States. The results imply that one indicator of complexity—the unfamiliarity of respondents with the subject question—to be the most robust predictor of item nonresponse across countries. By contrast, other indicators in the framework or machine-coded algorithms of textual complexity do not predict item nonresponse. The findings point to the importance of avoiding questions that require public servants to speculate about topics with which they are less familiar.

Robert Lipinski is a consultant and Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Christian Schuster is a professor at University College London.

ANALYTICS IN PRACTICE

- Surveys of public servants typically rely on voluntary responses from public servants. For this reason, they may suffer not only from unit nonresponse—that is, public servants' not responding to surveys at all—but also item nonresponse—that is, public servants' not responding to particular survey questions.
- Assessments of three public servant surveys spanning three continents imply that item nonresponse is a significant concern in the public sector. In some survey modules, nonresponse can be as high as 30 percent.
- Public servants are typically more educated than the average survey respondent, and their daily duties are closely aligned with the task of filling in a questionnaire. As such, the determinants of nonresponse in surveys of public servants may be distinct from those identified in the existing literature.
- This chapter presents a coding framework that allows survey analysts to measure the complexity and sensitivity of different questions in a public service questionnaire. Such assessments provide an important exercise in assessing survey quality.
- The analysis finds one indicator of complexity—the unfamiliarity of respondents with the subject question—to be the most robust predictor of item nonresponse across countries. Surveys of public servants should carefully consider the need for questions that require public servants to speculate about topics they are less familiar with, as they are associated with greater item nonresponse.
- In contrast, no other margin of complexity or sensitivity is a particularly acute source of nonresponse. At least in terms of missing data, the current analysis implies that public officials can handle many aspects of complex and sensitive topics.
- The manual coding approach is compared to common machine-coded assessments of complexity and find that a manually coded assessment of unfamiliarity outperforms machine-coded variables.

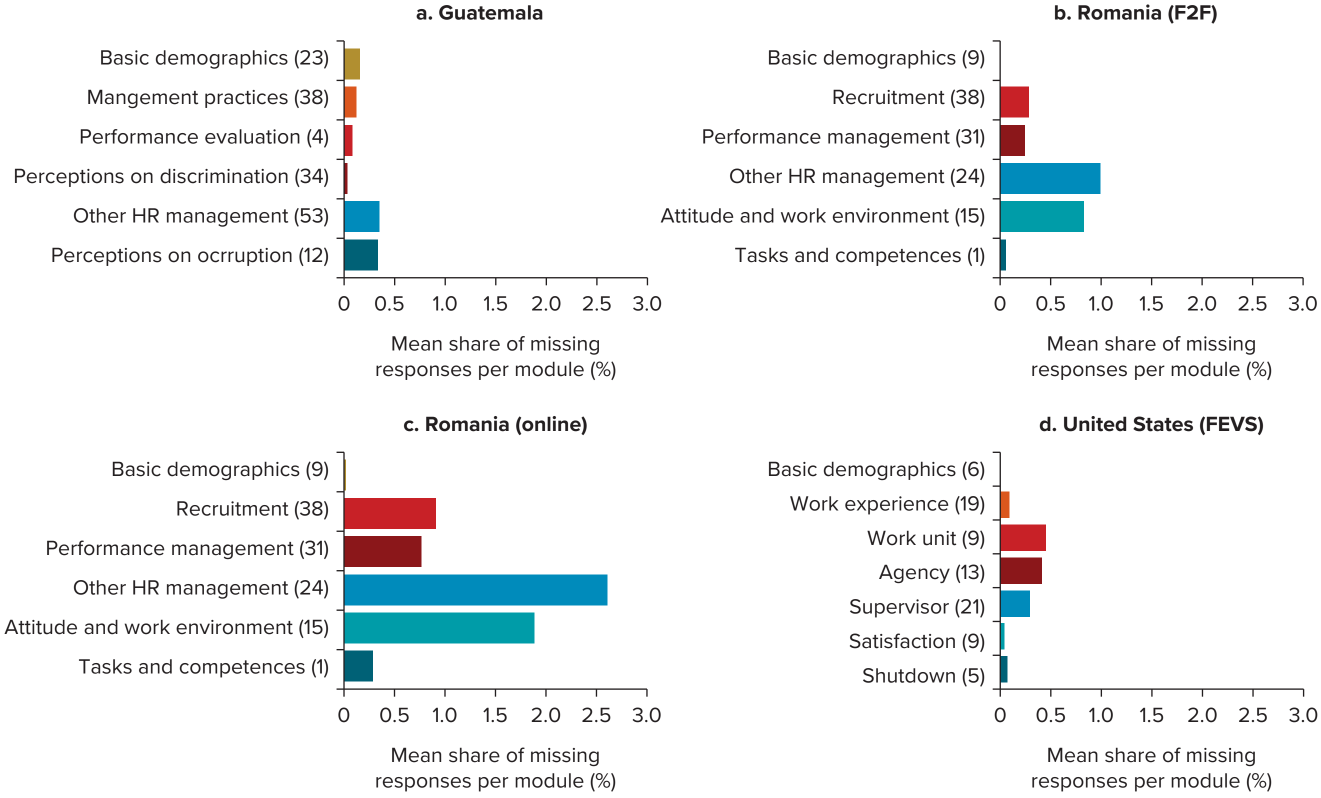
INTRODUCTION

Surveys of public servants typically rely on voluntary responses from public servants. For this reason, they may suffer not only from unit nonresponse—that is, public servants' not responding to surveys at all or dropping out of the survey (see chapter 19)—but also item nonresponse: public servants' not responding to particular survey questions. They may, for instance, skip survey questions in online surveys, refuse to answer questions in face-to-face surveys, or simply indicate "I don't know" in response to questions.

Item nonresponse is a challenge for both the quality and legitimacy of public service survey data. Item nonresponse may undermine the quality of public service survey data because having fewer responses enhances the variance of items. From a legitimacy perspective, high item nonresponse undermines potential uses of the data, as skeptics can critique the inferences drawn from items with high nonresponse as not representative of the survey population. If nonrespondents differ in a systematic way from respondents, questions can produce biased point estimates (Haziza and Kuromi 2007). This is not inconceivable: survey results may be biased, for instance, if those least satisfied with their jobs or those with reason to hide their behavior are also most prone to skipping survey questions.

Understanding what types of questions public servants tend to respond to and what types of questions prompt item nonresponse is thus important for survey designers. It provides a basis for designing questions that reduce item nonresponse and thus for enhancing public service survey quality and legitimacy. This is

FIGURE 22.1 Share of Missing Responses, by Survey Module



Source: Original figure for this publication.
Note: The labels on the *y* axis in each graph contain numbers in parentheses indicating the number of questions in each module. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face; HR = human resources.

important not least because surveying public servants about certain topics—such as satisfaction, motivation, or assessments of leadership—is often the only means to obtain data on these topics. Given the absence of other data sources to measure them, improved questionnaire design is the only alternative for valid data collection.

To date, public service surveys have varied in the extent to which their questions yield nonresponse. As illustrated in figure 22.1—which draws on data from public service surveys in Guatemala, Romania, and the United States (and which will be used throughout this chapter)—item nonresponse varies across survey modules from almost 0 percent to almost 30 percent in some settings (and up to 60 percent for certain individual questions). These figures imply that for certain topics, nonresponse is a substantive concern in public service surveys. The variation observed across questions also implies that question characteristics determine the likelihood that a question will be answered.

Why do public servants respond to some survey questions but not to others? This chapter offers a conceptual framework and empirical evidence to better understand this question. Conceptually, we build on the survey methodology literature, which has broadly argued for two causes of item nonresponse: question complexity and question sensitivity. Question complexity leads to item nonresponse when respondents are unable to answer a question, even if they are willing. This is due to an excessive cognitive burden on one or more steps in the mental process of answering a question: (1) comprehension of the question, (2) information retrieval from memory, (3) information integration, and (4) translation to the correct response option (Tourangeau 1984; Tourangeau and Rasinski 1988). As detailed below, this burden might arise because a question is formulated using complicated or vague language, because a question asks for information that is not readily accessible in the respondent's memory, because a question asks for a simultaneous evaluation of several factors, making it more difficult to render a judgment, or because a respondent's judgment does not correspond to the available answer categories. This burden might also be larger for certain groups of respondents—for example, the elderly.

Question sensitivity, by contrast, leads to item nonresponse when respondents are not willing to answer a question, even if they are able to. A sensitive question might infringe on respondents' privacy or make them reluctant to answer due to a fear of social or legal repercussions should the answer become known to third parties (Tourangeau, Rips, and Rasinski 2000; Tourangeau and Smith 1996).

While the survey methodology literature on question complexity and sensitivity is substantial, it is typically based on assessments of citizen or household surveys. It is unclear whether its findings are applicable to surveys of public servants. Public servants typically respond to employee surveys as part of their work duties, thus potentially enhancing their willingness to invest cognitive effort into question understanding. Moreover, public servants are usually relatively educated and accustomed to bureaucratic language, which is often highly technical and more complex than the language used in regular conversations.[1] Therefore, public officials should find it easier to interpret complex syntax and vague terms, and their education should enable them to integrate varied information and perform required calculations or information retrieval from memory more easily. At the same time, questions in public employee surveys often ask for more complex inferences than household surveys—for instance, about employees' perception of the organization or senior management practices. These diverging characteristics of public officials, the environment in which they respond to surveys, and the content of surveys put a premium on empirically assessing item nonresponse in public employee surveys, rather than simply extrapolating findings about item nonresponse from household surveys.

This chapter does this by analyzing missing response patterns in three public administration surveys—the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) and two World Bank surveys of public officials in Guatemala and Romania.[2] The analysis is based on the creation of a coding framework to assess different elements of question complexity and sensitivity, the application of this framework to code each of the questions in the aforementioned surveys in terms of complexity and sensitivity, and, finally, regressions to assess which of the elements of complexity and sensitivity predict item nonresponse.

We find, contrary to literature findings in other contexts, that public officials do not appear to shy away from answering questions that are longer, that are characterized by more complex syntax, or that require more cognitive effort to answer. We also find only limited evidence that question sensitivity is associated with greater item nonresponse. By contrast, we find robust evidence that one subindicator of complexity—the unfamiliarity of topics in questions—is associated with item nonresponse across all countries. Public officials prefer to not answer questions about topics outside of their immediate experience—for instance, about practices in their units and organization at large—a feature we term *unfamiliarity*. In sum, it appears that relatively highly educated public officials do not struggle with terminologically complex questions but are more unwilling or unable to answer questions about their broader working environment or to integrate different aspects of the functioning of their organization into one response option.

Given the manual nature of the approach to coding the complexity and sensitivity of survey questions, one natural criticism is that machine-coded measures may perform as effectively in determining problem questions but at a lower cost. We therefore perform a comparison of the core results to the predictive ability of machine-coded measures. We find that the unfamiliarity index continues to be the most effective approach to identifying questions that suffer from nonresponse in public servant surveys.

The chapter is organized as follows. Section 2 presents an overview of past work on survey complexity and sensitivity. Section 3 shows how the coding framework was constructed and how it relates to the past research, as well as the present research design. Section 4 details the results, which are followed by a discussion in section 5. The final section concludes and outlines avenues for future research.

UNDERSTANDING ITEM NONRESPONSE: LESSONS FROM THE SURVEY METHODOLOGY LITERATURE

In essence, the survey methodology literature posits two broad underlying causes of item nonresponse: respondents are either unable to answer survey questions (due to different dimensions of question complexity) or are unwilling to answer survey questions (due to different dimensions of question sensitivity) (Rässler and Riphahn 2006). We follow the literature in assessing these two central causes of potential item nonresponse. To build the coding framework, we discuss the literature on complexity and then on sensitivity.

Complexity

Questions assessing the same underlying concept can be expressed in more or less complex ways. "What is your age?" is an extremely common survey question. It is also a question that virtually everyone can understand and answer. "How many orbital periods have passed on the third planet from the sun since your hour of birth?" asks for the same information but could leave respondents confused about what the question is actually asking for. Although this example is needlessly complicated, some survey questions are longer and more convoluted or otherwise hinder respondents who are willing to respond from providing answers. The literature typically refers to this quality as *question complexity* (Knäuper et al. 1997; Yan and Tourangeau 2008).

Complexity is a multidimensional concept. While its definition is contested, it can perhaps best be conceptualized as a set of hurdles that respondents can encounter on their mental pathway from the moment they are presented with a question to providing an answer (Tourangeau and Rasinski 1988). Or as Knäuper et al. (1997, 181) phrase it, "Question answering involves a series of cognitive tasks that respondents have to resolve to provide high-quality data." These tasks may be objectively more or less difficult, but the effort they require may also depend on respondents' characteristics. To go back to the example used at the beginning of this section, the more complex version of the age question would likely pose relatively less trouble to a native English-speaking astrophysicist than to someone for whom English is a second language and who has never learned about physics.

The literature on cognitive psychology commonly refers to four steps in the question-answering process, as outlined by Tourangeau (1984), Tourangeau and Rasinski (1988), and Tourangeau, Rips, and Rasinski (2000). These steps are as follows: (1) question comprehension, (2) the retrieval of necessary information from memory, (3) the integration of the retrieved information into a judgment or estimate, and (4) the translation of the judgment into an appropriate response. Depending on the type and format of a question, these steps might vary in length and cognitive difficulty. For example, for a question about age, information is easily retrieved from memory but might require some mapping process if the response is not numerical but rather matched to predefined age bands.

In the first step, respondents have to comprehend the language used in a question and its intent (Holbrook, Cho, and Johnson 2006). Faaß, Kaczmirek, and Lenzner (2008, 2) write that "comprehending a question involves two processes which cannot be separated: decoding semantic meaning and inferring pragmatic meaning." Therefore, a question with more elaborate syntax and sentence construction, as well as technical or unfamiliar words, requires more cognitive effort to be understood by respondents (Knäuper et al. 1997)—an effort they may or may not be able or willing to perform.

It is less obvious whether questions that are longer have a positive or negative impact on comprehension. On the one hand, a question might be longer because it explains its purpose and content in more detail, thus reducing the cognitive effort required on the part of respondents. On the other hand, a long question may simply be convoluted, touch on too many topics, or be difficult to remember in full when providing the answer, thus increasing difficulties for respondents (Holbrook, Cho, and Johnson 2006; Knäuper et al. 1997).

Other features of a question, like the number of propositions and logical operators (for example, *or* and *not*), dense nouns (accompanied by many adjectives or adverbs), or left-embedded syntax, can interact with the above to complicate even relatively short words and sentences (Faaß, Kaczmirek, and Lenzner 2008). Cognitive difficulties in comprehension might also depend on individual working memory capacity. Research by Just and Carpenter (1992) shows that working memory is a key element of both information storage and the computations necessary for language comprehension.

Once respondents have comprehended what information is required, they have to search their memories to retrieve it. This task is more difficult when the required information refers to the more distant past (Krosnick 1991). It is clear that recalling what one had for breakfast this morning, for example, is easier than recalling the same information from a week ago. In psychology, this is the well-known phenomenon of *attitude* (or *information*) *accessibility* (Fazio 1986). More-accessible attitudes are retrieved from the memory more easily and quickly, or, in other words, with lower cognitive effort. The more recently an individual has thought about a particular matter, the more accessible this and related considerations are when answering a survey (Zaller 1992). Zaller (1992) terms the predominant use of easily retrievable information the "accessibility axiom."

Apart from the temporal reference frame, attitudes that refer to direct, more recent, or recurrent experiences tend to be more accessible (Berger and Mitchell 1989; Fazio 1989; Fazio and Roskos-Ewoldsen 2005). Memories of events that were emotional, unique, or drawn out are more likely to be accessible from memory, possibly biasing survey responses in favor of such events (Tourangeau 1984). Finally, it is less burdensome to retrieve information related to one item or topic rather than two or more, and surveys should therefore avoid what are called *double-barreled* questions (Krosnick 1991).

The information retrieved then needs to be integrated into a judgment. Depending on the question, the difficulty of this process can range from null to very high. Information about one's gender or age and other factual questions about oneself require little integration. By contrast, in other cases, the format in which questions are asked can shape the difficulty of integration. Consider the following example of three different question formats to measure the role of personal connections in public sector recruitment:

1) Were personal connections (friends and family in the institution) important to get your first public sector job?
1 - Yes; 2 - No; 3 - Don't know

2) How important were personal connections (friends and family in the institution) to getting your first public sector job?
1 - Very unimportant; 2 - Somewhat unimportant; 3 - Neither important nor unimportant; 4 - Somewhat important; 5 - Very important; 6 - Don't know

3) Please rank the following criteria in order of the importance they had for obtaining your first public sector job:
1 - Personal connections (friends and family in the institution); 2 - Political connections; 3 - Educational background; 4 - Previous work experience; 5 - Work-related skills

The first version of the question only requires respondents to make a binary choice about the importance of personal connections. The second version requires a more fine-grain evaluation—not only about whether personal connections were important but also how important. In the third version, respondents have not only to judge the importance of personal connections but also of four other considerations and to evaluate them against each other. Clearly, this last approach requires the greatest cognitive effort from respondents.

Much work in psychology has been conducted to determine how people formulate judgments from available information. According to Anderson's (1971) information integration theory, when people formulate a judgment, they gather all available pieces of information, assigning value and weight to each of them, before summing them up to form a final judgment. Another view, developed mainly in the work of Tversky and Kahneman, is that people tend to use a range of heuristic methods to arrive at judgments, like using only readily available instances and examples, using resemblance to a prototype, or anchoring based on initial information (Tourangeau 1984; Tversky and Kahneman 1974). A combination of these views has been adopted by Zaller (1992) in his "response axiom," which argues that individuals answer survey questions by averaging different considerations, but only those that are immediately salient or accessible to them.

The final stage of question answering is mapping the answer onto the available response options (Tourangeau and Rasinski 1988). Holbrook, Cho, and Johnson (2006) mention two possible difficulties at this stage. One is the problem of mental multitasking, which occurs because respondents have to simultaneously remember the question and the answer options and to map their formed judgments onto them. This might be an issue, particularly for individuals who have problems with remembering information—for example, the elderly. It might also be overly taxing if response options are descriptive rather than articulated on a frequency or Likert-like scale, or if they contain vague words and complex phrases. Second, response formats that are hard to understand or that have an ambiguous set of possible responses might compound mapping difficulties. Whereas multitasking as an obstacle depends mainly on the respondent, problems with the response format are usually due to faulty questionnaire design. To ease the process of translating a formed judgment into a response, it is particularly important to ensure that the set of responses to each question is both exhaustive and mutually exclusive (Krosnick and Presser 2009).

Across each of these stages, survey question complexity can have multiple effects. Some are less consequential for survey data quality—such as longer response times (Faaß, Kaczmirek, and Lenzner 2008; Yan and Tourangeau 2008) or respondents' asking the interviewer for clarification (Holbrook, Cho, and Johnson 2006). Some effects of question complexity, however, are more consequential. In particular, complexity can invite *acquiescence bias* or *satisficing*, in which respondents tend to agree with a complex statement, regardless of their true position, in order to avoid cognitive overload (Knäuper et al. 1997; Krosnick 1991; Lenski and Leggett 1960). Apart from agreeing with a statement, respondents might ease the cognitive burden by selecting the first available response option, choosing randomly, skipping the question, or selecting the "I don't know" option. This last option is an example of strong satisficing because it requires no cognitive effort whatsoever.[3]

In short, the survey methodology literature suggests that complex survey questions heighten the cognitive effort required along the mental process of answering a question and may thus lead to satisficing, including item nonresponse. The empirical literature that complements the theoretical considerations outlined here finds supporting evidence that each of these answering stages can increase nonresponse. For example, Knäuper et al. (1997) find that respondents answer "I don't know" more often to questions that, among other things, contain ambiguous terms or require retrospective or quantity reports. Including these more complex question characteristics raised item nonresponse in their study by between 0.5 and 7.7 percentage points (and, as expected, more so for individuals with lower cognitive ability). This is substantial, considering that in most subgroups, the total share of "I don't know" responses stayed well below 10 percent.

Sensitivity

Irrespective of how complicated a question is, the extent to which it requests personally sensitive information may also impact nonresponse. "How many bribes have you accepted in the last month?" has simple syntax, uses precise terms, and has a clearly defined, short, and direct reference frame. It is not a complex question. However, the question is sensitive—it asks about behavior that is typically both morally wrong and illegal—which is a second source of concern for survey designers.

Unlike complex questions, when people are asked sensitive questions, they usually know the correct or true response but are unwilling to provide it. Or, in other words, "data quality does not only depend on the accurate recall of facts but also depends on the degree of peoples' self-disclosure" (Gnambs and Kaspar 2015, 1238). Sensitivity is unavoidable in some surveys. In fact, the whole purpose of a survey might be to elicit information that cannot be obtained from other data sources because people conceal it and avoid discussing it in public (Lensvelt-Mulders 2008). Typical topics of concern include drug use, sexuality, and gambling. In the context of public administration, the issue of sensitivity may arise with topics such as corruption and integrity, discrimination inside the public service, or the sexual harassment of employees.

The most commonly used classification of sources of sensitivity was developed by Roger Tourangeau, along with several coauthors (Tourangeau, Rips, and Rasinski 2000; Tourangeau and Yan 2007). According to them, sensitivity derives from three primary sources—a question may touch on a taboo subject, a truthful answer may violate social norms, or a truthful answer may lead to negative formal consequences.[4]

In the first instance, respondents might feel that the topic of a question is not supposed to be discussed in public but rather kept private. In other words, it is considered a *taboo* subject (Tourangeau and Yan 2007). This may be a concern for various topics, from sexual orientation to salary level. McNeeley (2012) describes how talking about such topics may lead to distress and uneasiness for respondents (and, in some cases, enumerators as well) and, for demographic items, also lead to the threat of identification.[5] Unlike social-desirability bias and other sources of sensitivity discussed below, these topics are not problematic because revealing the requested information could lead to some type of sanctions. Instead, these topics are perceived as sensitive regardless of a respondent's true position (Krumpal 2013; Tourangeau and Yan 2007) because it is not common to discuss them in public or with strangers, like an enumerator. Therefore, these questions often lead to item nonresponse rather than misreporting (Höglinger, Jann, and Diekmann 2016) because respondents simply do not want to discuss the topics at all.

Second, but arguably most commonly, the wariness to truthfully answer sensitive questions is explained with reference to *social-desirability bias*. This refers to an inner desire to conform to established social norms in a given circle, be it a workplace, a family, or society at large. Admitting that one has committed an action that violates a common norm, either by doing something considered "wrong" (for example, taking a bribe) or failing to do "good" (for example, not helping a colleague in need), is undesirable (Tourangeau and Yan 2007) because, if someone found out, the violator could be frowned upon, criticized, or shunned. The impact social-desirability bias has on responses further depends on the specific social norms respondents identify with and how concerned they are about not violating them. For example, Kim and Kim (2016) find that national culture significantly moderates the degree and pattern of social-desirability bias in public service motivation surveys.

Apart from this *extrinsic* threat, answering sensitive questions also poses *intrinsic* threats to the self-image of respondents (Lensvelt-Mulders 2008, 462). Touching on sensitive topics may raise feelings of guilt, embarrassment, or shame in respondents for having done (or for failing to do) something, or they may be stressful to discuss for respondents in general. Therefore, to avoid negative consequences from others as well as one's own conscience, respondents may prefer not to answer a sensitive question or to answer it in a socially "expected" way. In face-to-face surveys, even respondents who believe in the full confidentiality of their responses may want to create a positive image of themselves or earn social approval from the enumerator (Krumpal 2013) and thus may succumb to social-desirability bias.

Psychologists have long debated the precise causes of social-desirability bias. Paulhus (1984) suggests it has two parts. One is impression management—that is, a desire to present oneself in a positive light in front of others to avoid negative feedback from them. Another is self-deception, which means holding favorably biased views about oneself while honestly believing them to be true.

Third, a related but distinct source of sensitivity comes from questions that ask about actions that are formally (rather than socially or informally) prohibited. For example, hiring one's family members and friends might be a widespread and socially accepted practice. However, if nepotism is formally prohibited, then admitting it in a survey might lead to legal sanctions, like a fine, a disciplinary note, or being fired. Or, as Tourangeau and Yan (2007, 859) note, "possessing cocaine is not just socially undesirable; it is illegal, and people may misreport in a drug survey to avoid legal consequences rather than merely to avoid creating an unfavorable impression."

Informal and formal sources of sensitivity might also interact with each other. Research by Galletly and Pinkerton (2006) suggests that there is an interaction between social stigma and formal sanctions in the case of HIV disclosure laws in US states. The introduction of legal sanctions for some actions may add to the already existing social stigma around them. Alternatively, the threat of social disapproval may be a more undesirable consequence than a formal sanction that is small or unlikely to follow. Likewise, if social and legal norms are not perfectly matched, admitting to an illegal but socially acceptable practice might be less difficult for respondents. For example, if the law prohibits hiring one's family members and friends but society generally accepts this practice, then admitting to some degree of nepotism might come more easily to a survey respondent than if this practice were socially unacceptable.

As with question complexity, item nonresponse is one of several possible behavioral responses to sensitivity (Krumpal 2013; Lensvelt-Mulders 2008; McNeeley 2012; Tourangeau and Smith 1996; Tourangeau and Yan 2007). Respondents, when aware of survey topics, may decline to participate altogether (McNeeley 2012). Moreover, respondents may believe that not answering sensitive questions is "revealing" in itself (Tourangeau and Yan 2007, 877). Instead, respondents may choose simply to answer in an expected way that is certain not to result in any negative consequences (Bradburn et al. 1978; Krumpal 2013; McNeeley 2012). For example, refusing to answer a question about bribe-taking might seem suspicious in itself, so bribe-takers may avoid any suspicion or feeling of shame by simply saying that they have never taken bribes rather than refusing to answer.

In sum, item nonresponse may increase as a result of increased question sensitivity—though, compared with complexity, this effect may be diluted by respondents who answer sensitive items in a socially desirable way rather than not answering at all (Sakshaug, Yan, and Tourangeau 2010). Tourangeau and Yan (2007), for example, report that item nonresponse in the National Survey of Family Growth (NSFG) Cycle 6 female questionnaire tends to rise by fewer than 3 percentage points when comparing questions with very low sensitivity (for example, education [0.04 percent nonresponse rate] and age [0.39 percent]) with high-sensitivity items (for example, the number of times the respondent had sex in the past four weeks [1.37 percent] and their number of sexual partners [3.05 percent]). Only the income question has more noticeable nonresponse, at 8.15 percent. And whereas experimental methods that aim to reduce question sensitivity, such as the unmatched count technique, do significantly affect the mean estimates obtained, they have a far smaller effect on item nonresponse (Coutts and Jann 2011), suggesting that biasing rather than avoiding an answer is a more prevalent response for people presented with sensitive questions.[6] Comparing the effects of unit (although not item) nonresponse and measurement error in reports of voting behavior, Tourangeau, Groves, and Redline (2010) suggest that the latter is around two times larger and can elevate the reported prevalence of voting from the true value of 47.6 percent to 69.4 percent.

METHODOLOGY

Case Selection

We evaluate question complexity and sensitivity and their relationship to item nonresponse in three 2019 governmentwide public administration surveys in Guatemala, Romania, and the United States.

The surveys in Guatemala and Romania were nationally representative surveys of public officials conducted by the World Bank in 2019 and 2020. The survey in Guatemala was a face-to-face survey conducted from November to December 2019. It covered 14 central government and four decentralized institutions. A sample of 205 respondents was selected from each institution (of which one-quarter were supervisors and three-quarters were subordinates). In total, 3,465 public officials provided answers, resulting in a response rate of 96 percent (World Bank 2020a). All respondents were surveyed in person by trained enumerators.

The survey in Romania used a mixed-mode delivery, with a randomly chosen set of officials answering the survey online and another set answering it in face-to-face (F2F) interviews with enumerators. The face-to-face questionnaire was longer than the online one, and, therefore, only the questions overlapping between the two versions are used in the analyses below. The Romanian data were collected from June 2019 to January 2020 across 81 institutions that agreed to participate (out of 103 invited). The targeted sample of respondents was drawn from the institutional census of employees. In total, 2,721 public officials answered the online questionnaire (for a response rate of 24 percent), and 3,316 answered the face-to-face one (for a response rate of 92 percent; for details see World Bank [2020b]).

Responding to a survey online may increase respondents' sense of comfort and privacy, thus reducing the perceived threat posed by sensitive questions (McNeeley 2012). On the other hand, online surveys lack an enumerator, who can clarify complex questions or encourage respondents to answer (De Leeuw 1992). We thus estimate the effects of complexity and sensitivity for Romania separately for online and face-to-face respondents.

The FEVS has been fielded by the US OPM biannually since 2004 and annually since 2010 (see chapter 26). It covers all types of employees across federal government departments and agencies that choose to participate. It is delivered in an online, self-administered form. In the latest available iteration, from 2019, which is used here, it was conducted as "a census administration that included all eligible employees from 36 departments and large agencies as well as 47 small and independent agencies" (OPM 2019). In total, over 615,000 government employees responded to the survey, for a response rate of 42.6 percent.

The case selection enables us to understand item nonresponse in public service surveys of countries from across diverse cultures, regions, and levels of development and education. Findings about item nonresponse that travel across all three contexts are plausibly generalizable to other surveys of public administrators.

Coding Framework

Understanding item nonresponse—and whether different dimensions of complexity and sensitivity shape item nonresponse in public service surveys—requires measuring complexity and sensitivity consistently across and within surveys. To do so, coding framework is developed that allows us to assign a numerical value reflecting the degree of complexity and sensitivity of every survey question. The approach builds on the existing literature summarized above and resembles research by Bais et al. (2019), who similarly integrate several aspects of complexity and sensitivity into a manual coding framework.[7]

The complexity and sensitivity indexes comprise several subdimensions, as described in tables 22.1 and 22.2. The complexity index is composed of 10 subdimensions, which are conceptually based on the four-stage mental process of answering a question (see Tourangeau and Rasinski 1988; Tourangeau, Rips, and Rasinski 2000), and synthesizes the measures proposed by, among others, Belson (1981); Holbrook, Cho, and Johnson (2006); and Knäuper et al. (1997). The subdimensions include the complexity of the syntax, the number of subquestions, the presence of a reference frame, and the unfamiliarity of the subject.

The sensitivity index is constructed using four subdimensions suggested by the literature: invasion of privacy, the social-emotional threat of disclosure, the threat of formal sanctions (Tourangeau, Rips, and Rasinski 2000; Tourangeau and Yan 2007), and the interaction between informal and formal sanctions (see, for example, Galletly and Pinkerton 2006).

We evaluate each question in the three surveys studied along the dimensions outlined in tables 22.1 and 22.2 and score it a value of 0, 1, or 2. The value of 0 is given to questions that do not present a particular subdimension of complexity or sensitivity at all. The value of 1 refers to cases in which questions potentially create problems for respondents in a given subdimension, whereas 2 is used for cases where such problems are clearly substantive. The full coding framework is presented in appendix J.

Three research assistants applied the coding framework to assess the complexity and sensitivity of each of the questions in the three surveys. (For examples of this process, see box 22.1.) Each research assistant first coded the values independently, and then their scores were compared. In 86.8 percent of cases, all coders working on a given question agreed on the score. In the instances where they were not in agreement, differences were discussed and resolved with a view to maximizing consistency in coding across survey questions. The values of both indexes are calculated as arithmetic means of the scores across their respective subdimensions.

TABLE 22.1 Complexity Coding Framework

Subdimension	Description	Guatemala	Romania	United States (FEVS)	Aggregate
Comprehension					
Complex syntax	This component assesses the length of a question, which is measured by the number of characters (*n*), and the complexity of the syntax or the grammatical arrangements of words and phrases, which is determined by the sentence structure. The term *simple syntax* indicates simple sentence(s) with three parts of speech, *moderately difficult syntax* indicates simple sentence(s) with more than three parts of speech, and *complicated syntax* indicates complex or complex-compound sentences.	0.85 (0.65)	1.09 (0.61)	1.2 (0.53)	1 (0.63)
Vagueness	This component assesses the extent to which the language used in a question is vague, unclear, imprecise, ambiguous (Edwards et al. 1997), or open to interpretation. Common terms such as "good" are predetermined in a list of vague words.	0.34 (0.48)	0.64 (0.5)	0.54 (0.55)	0.48 (0.52)
Reference category	This component assesses the extent to which the necessary frame(s) of reference are available in a question so that respondents understand the question in the way intended.	0.17 (0.47)	0.26 (0.51)	0.16 (0.48)	0.2 (0.48)
Number of questions	This component measures the number of subquestions embedded in the question block to which a question belongs. A subquestion must only ask for one issue, so a compound subquestion is not counted as one subquestion.	0.3 (0.55)	0.11 (0.33)	0.21 (0.46)	0.22 (0.48)
Information retrieval					
Unfamiliarity	This component assesses the extent to which respondents are knowledgeable on the subject of a question. The coding presumes that respondents are more familiar with subjects they have a closer knowledge of (for instance, their own experience versus their perceptions of the experiences of other employees in the organization).	0.93 (0.79)	0.34 (0.53)	0.35 (0.51)	0.62 (0.72)
Recalling	This component assesses the extent to which respondents are required to remember information based on the question's level of specificity and time frame of interest (past/present).	0.91 (0.4)	1 (0.44)	1.04 (0.4)	0.96 (0.42)
Information integration					
Computational intensity	This component assesses the extent to which basic arithmetic computations (addition, subtraction, multiplication, and division) are required to reach an answer.	0.07 (0.29)	0.07 (0.26)	0.02 (0.16)	0.06 (0.26)
Scope of information	This component assesses the extent to which answers are derived from information beyond the personal experience of respondents.	0.38 (0.52)	0.46 (0.55)	0.32 (0.47)	0.39 (0.52)
Translation to answer					
Category mismatch	This component assesses the extent to which the available answer options match the true answer to the question.	0.06 (0.28)	0.09 (0.41)	0.04 (0.25)	0.06 (0.32)
Number of responses	This component assesses the extent to which respondents are required to pick more than one answer to the question.	0.06 (0.28)	0.01 (0.09)	0 (0)	0.03 (0.2)

Source: Original table for this publication.
Note: The final four columns show the mean and standard deviation (in parentheses) of scores for each subdimension and survey. FEVS = Federal Employee Viewpoint Survey.

TABLE 22.2 Sensitivity Coding Framework

Subdimension	Description	Guatemala	Romania	United States (FEVS)	Aggregate
Privacy					
Invasion of privacy	This subindicator measures the extent to which respondents are asked to discuss taboo or private topics that may be inappropriate in everyday conversation. Questions related to a respondent's income or religion may fall into this category.	0.08 (0.27)	0.31 (0.6)	0.04 (0.19)	0.14 (0.41)
Informal sensitivity					
Social-emotional threat of disclosure	This subindicator measures the degree to which respondents may be concerned with the social or emotional consequences of a truthful answer, should the information become known to a third party. In the case of informal sensitivities, this type of question is only considered sensitive if the respondent's truthful answer departs from socially desirable behaviors or social norms.	0.55 (0.51)	0.7 (0.65)	0.79 (0.56)	0.65 (0.58)
Formal sensitivity					
Threat of formal sanctions	This subindicator measures the degree to which respondents may be concerned with the legal and/or formal consequences of a truthful answer, should the information become known to a third party. This type of question is only sensitive if the respondent's truthful answer departs from legal behaviors defined by formal institutions and legal regulations.	0.26 (0.6)	0.15 (0.46)	0.15 (0.5)	0.2 (0.54)
Interaction					
Relationship between informal and formal sensitivity	This subindicator measures the likelihood that a behavior or attitude may cause a threat of both social-emotional disclosure and formal sanctions. This type of question is logically more sensitive than ones that violate one type of institution while conforming to another. A behavior may be frowned upon in one's social circle—for example, reporting colleagues taking bribes might be considered "snitching"—but it may also be a legal obligation. In such instances, asking about it should be less sensitive compared to a situation where both informal and formal norms were violated. Galletly and Pinkerton (2006) suggest such an interaction between social stigma and formal sanctions (in the case of HIV disclosure laws).	0.2 (0.41)	0.2 (0.48)	0.13 (0.49)	0.19 (0.45)

Source: Original table for this publication.
Note: The final four columns show the mean and standard deviation (in parentheses) of scores for each subdimension and survey. FEVS = Federal Employee Viewpoint Survey.

Analysis

To investigate nonresponse in a public administration setting, we assess the impact of the complexity and sensitivity measures outlined above on responsiveness in the three surveys under study. The regressions take the respondent-question as the unit of observation, meaning that each row corresponds to a particular respondent's answer to a given question. We define item nonresponse as an "I don't know" answer, a refusal to answer, or skipping the question.

We control for individual-level characteristics that might affect nonresponse, including age and education (both of which are correlated with respondents' cognitive abilities to deal with complexity; Holbrook, Cho, and Johnson [2006]; Yan and Tourangeau [2008]), gender (which can shape item nonresponse for

BOX 22.1 Applying the Coding Framework: Illustrative Examples from Romania

Complexity—information retrieval (recalling): "Which of the following factors were important for getting your current job in the public administration?" This question pertains to the past, asks for a specific level of information, and requires the respondent to consider the importance of many factors: academic qualifications, job-specific skills, knowing someone with political links, having personal connections, and so on. Therefore, this question is coded as 2.

Complexity—information integration (computational intensity): "How many years have you been in your current institution?" This question requires respondents to calculate their length of service in their current institution by subtracting their starting year from the current year. This is not a complicated calculation, but it still is not likely to be performed often and may require some mental effort if respondents joined a long time ago, are confused about whether periods like initial internships should be included, and so on. Therefore, this question is coded as 1.

Sensitivity—threat of formal sanctions: "How frequently do employees in your institution undertake the following actions? Accepting gifts or money from citizens." This question is coded as 2 because respondents may feel that there are social consequences for disclosing this information or even formal ones if they did not inform relevant authorities about the bribe-taking behavior.

sensitive questions—for instance, on harassment), tenure in the organization, and managerial status (more-experienced workers and managers might have more work-related knowledge and a different cost-benefit calculus when deciding whether to answer a survey), as well as job satisfaction as a proxy measure for willingness to respond (with more-satisfied respondents potentially more willing to respond to employee surveys or, alternatively, dissatisfied workers more eager to respond to report reasons for their dissatisfaction).[8]

We further control for the overall response rate in the government organization or agency to which a respondent belongs. A lower response rate might reflect unobservable characteristics of the organization or its employees that shape item nonresponse. We also control for the position of a question within a questionnaire (coded as integer variables starting from one). This is to take into account the fact that respondents might skip more questions or become less willing to cognitively engage with questions as the survey progresses and fatigue or dullness sets in (Krosnick 1991). Our data thus take a "long" format, with each row corresponding to a particular respondent's answer to a question, accompanied by the respondent's individual characteristics and the variables pertaining to his or her organization; the question's complexity, sensitivity, and position in the questionnaire; and, finally, whether the respondent answered a given question (1) or not (0). In general, it is found that men tend to have lower item nonresponse and that nonmanagers and less-satisfied employees skip questions more often, although the pattern doesn't hold in all settings and regression specifications. Questions appearing later in the questionnaire are also omitted more often, as hypothesized.

We first look at simple correlations between the key variables of interest and then go on to regress the item nonresponse variable on the indexes of complexity and sensitivity, as well as their various subdimensions in ordinary least squares (OLS) regressions. In order to account for the possible correlation of residual errors in the data set, we use multiway clustering on the individual and question levels, which allows us to correctly estimate standard errors and corresponding significance levels.

RESULTS

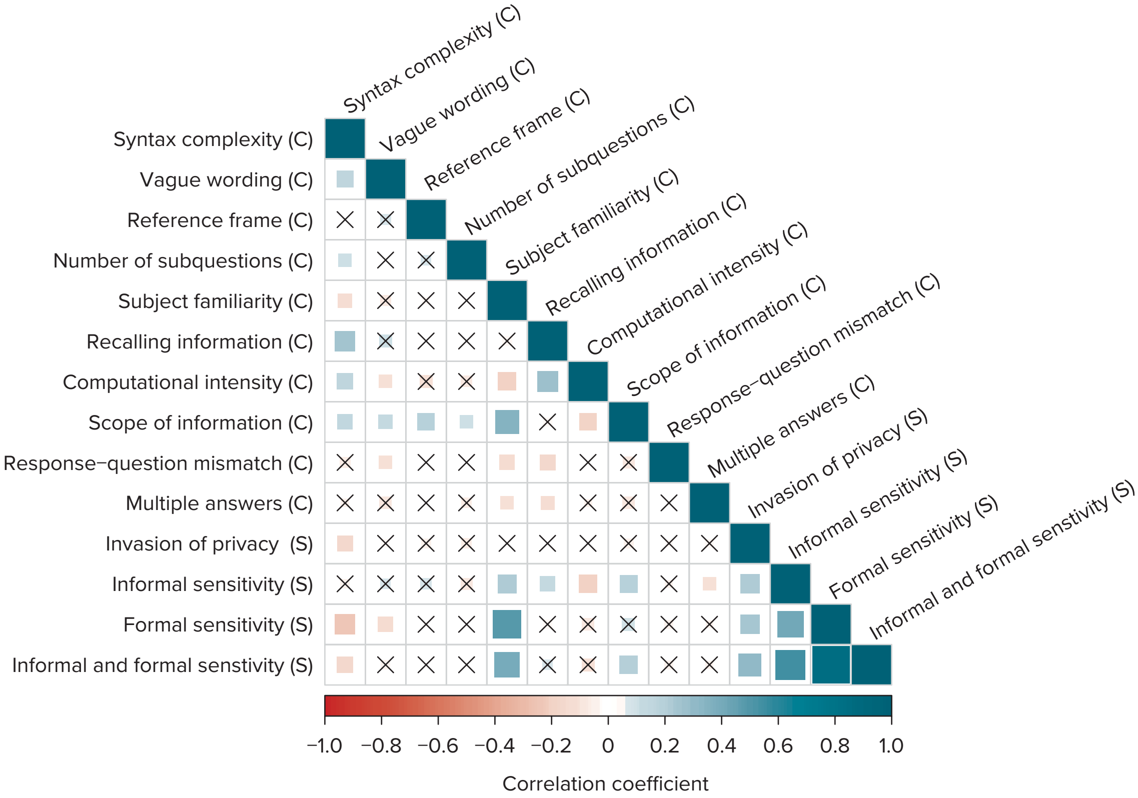
Descriptively, our independent variables vary across the components of complexity and sensitivity we code. As detailed in tables 22.1 and 22.2, among the components of complexity, complexity of syntax and unfamiliarity are the variables with the highest variance. On the other end of the scale, components related to translation to answer seem the least variable. Among the sensitivity components, the social-emotional threat of disclosure records both the highest mean score and the greatest variation. Invasion of privacy scores the lowest in mean and standard deviation.

To ensure the coding framework meaningfully captures distinct subdimensions or components of complexity and sensitivity, we assess correlations between different components or subdimensions of complexity and sensitivity. Figure 22.2 shows that for complexity, most of the correlations are not significant, suggesting, as theorized, that different components relate to different mental processes and aspects of a question. Where there is some conceptual overlap, however, we do see significant correlations, such as between syntax complexity and vague wording or between the scope of information and subject unfamiliarity.

In the case of sensitivity, all correlations are significant and strong. This is conceptually plausible. Informal and formal sensitivity most often occur simultaneously, while questions about illegal or socially disapproved behaviors are plausibly also often too private or embarrassing to discuss in public. In sum, the observed correlations yield a degree of credibility to the coding framework and its application.

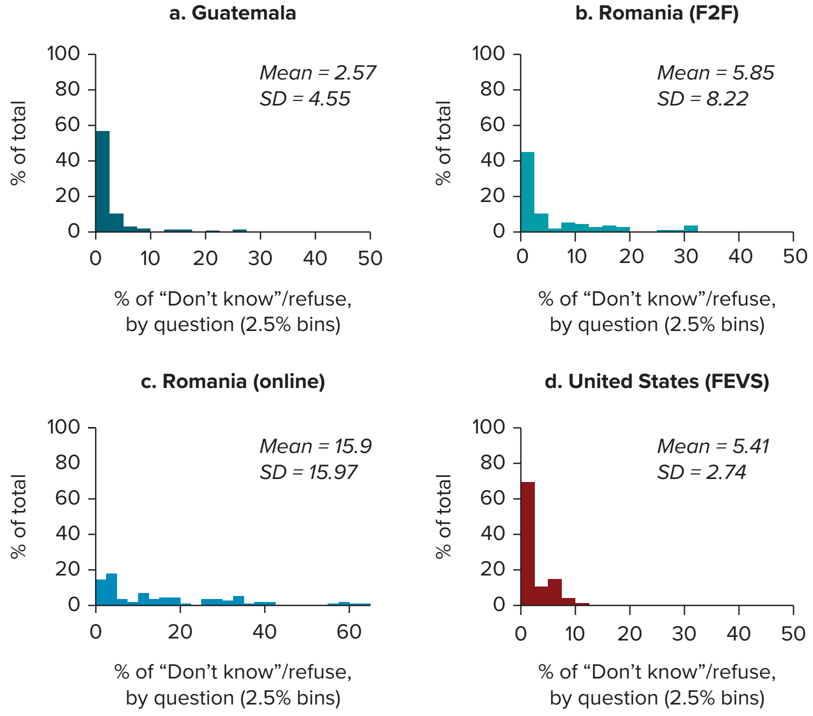
Next, as presented below, we can observe that item nonresponse is a challenge across the three surveys, though to a varying extent. As illustrated in figure 22.3, in the FEVS online survey, questions have an average item nonresponse of 2.4 percent. This number increases to 2.6 percent in the face-to-face public service

FIGURE 22.2 Correlation between Subdimensions of Complexity and Sensitivity



Source: Original figure for this publication.
Note: Correlations are obtained by pooling questions across all three surveys. Crosses mark correlations that are insignificant at the 5 percent level. C = complexity; S = sensitivity.

FIGURE 22.3 Share of Missing Responses



Source: Original figure for this publication.
Note: FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face; SD = standard deviation.

survey in Guatemala and 5.9 percent in the face-to-face public service survey in Romania. In the online version of the survey in Romania, in turn, average item nonresponse increases to 15.9 percent.[9] To what extent do complexity and sensitivity predict item nonresponse?

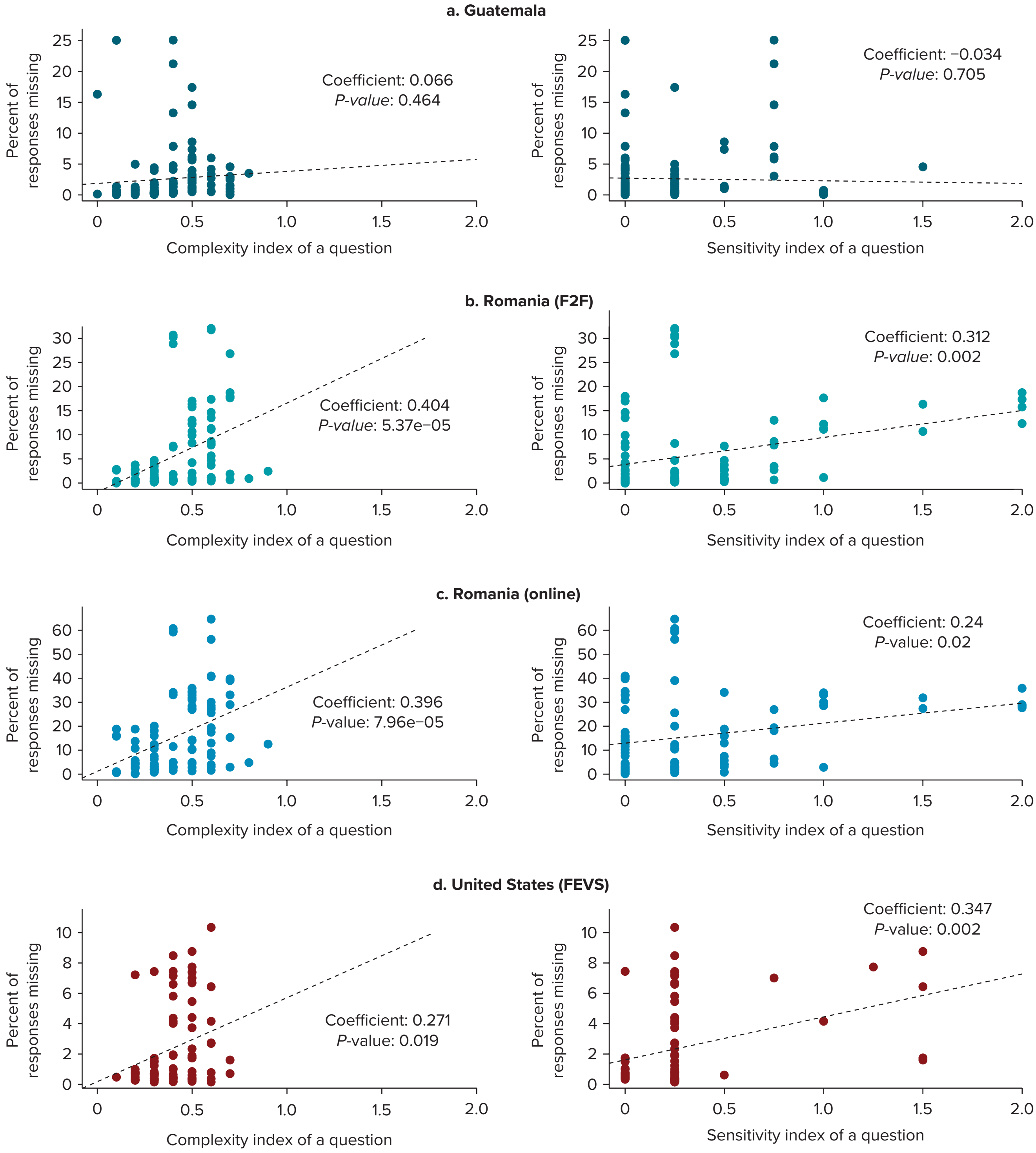
Looking first at correlations, we find that there is a positive association between complexity and item nonresponse. Figure 22.4 presents the correlations separately for each survey. Correlation coefficients range from 0.066 to 0.404 and are significant at the 5 percent level, except for Guatemala. We observe a similar pattern, although slightly weaker in Romania, for correlation between item nonresponse and sensitivity.

Table 22.3 presents regressions of item nonresponse on standardized sensitivity and complexity (both separately and jointly) with and without the aforementioned set of controls. We observe evidence from the United States and Romania that both complexity and sensitivity increase the probability of survey nonresponse in surveys of public officials. The results in Guatemala are not significant at the 10 percent level. The effect sizes are relatively small, with a standard deviation increase in the indexes having a 1 percentage point increase in nonresponse in the United States. In Romania, a one standard deviation increase in complexity is associated with an at most 6 percentage point increase in nonresponse, depending on the specification and mode of enumeration.

On average, the indexes of complexity and sensitivity thus predict item nonresponse in some but not all cases. Of course, however, it could be that our indexes—which simply average out different potentially relevant subcomponents of complexity and sensitivity—are not appropriately aggregated. The various subcomponents of complexity and sensitivity may not, as theorized, measure a single underlying dimension. To assess this, exploratory factor analysis (EFA) is performed across all 14 subdimensions pooled together. Indeed, instead of finding that two factors are sufficient to describe the data (as would be expected if the subdimensions measured only two dimensions: complexity and sensitivity), we find that at least four factors are needed to properly describe the data in each survey.[10]

The results of the EFA with four factors are presented in table 22.4. The results suggest that across countries, sensitivity subdimensions load onto a single factor (first factor). While the scores for the second

FIGURE 22.4 Relationship between Complexity and Sensitivity Indexes and the Share of Missing Responses



a. Guatemala

b. Romania (F2F)

c. Romania (online)

d. United States (FEVS)

Source: Original figure for this publication.
Note: FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.

TABLE 22.3 The Impacts of Complexity and Sensitivity on Item Nonresponse

	OLS estimates					
	(1)	(2)	(3)	(4)	(5)	(6)
Guatemala						
Sensitivity	−0.002 (0.004)		−0.002 (0.004)	0.002 (0.004)		−0.002 (0.004)
Complexity		0.003 (0.006)	0.003 (0.006)		0.003 (0.005)	0.002 (0.006)
Romania (F2F)						
Sensitivity	0.025*** (0.004)		0.020*** (0.004)	0.020*** (0.006)		0.015** (0.006)
Complexity		0.035*** (0.009)	0.030*** (0.009)		0.033*** (0.009)	0.031*** (0.009)
Romania (online)						
Sensitivity	0.039*** (0.007)		0.030*** (0.009)	0.027** (0.010)		0.017 (0.010)
Complexity		0.062*** (0.015)	0.056*** (0.015)		0.060*** (0.015)	0.057*** (0.015)
United States (FEVS)						
Sensitivity	0.009** (0.004)		0.008* (0.004)	0.009** (0.003)		0.008* (0.004)
Complexity		0.007* (0.003)	0.004 (0.003)		0.008* (0.003)	0.005 (0.003)
Controls	No	No	No	Yes	Yes	Yes

Source: Original table for this publication.
Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered “I don’t know,” refused to answer, or skipped a particular question. Sensitivity and complexity scores are calculated as *z*-scores estimated across questions in a given survey. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. The controls are described in detail in the analysis subsection of the methodology section. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.
Significance level: * = 10 percent, ** = 5 percent, *** = 1 percent.

TABLE 22.4 Exploratory Factor Analysis

	First factor	Second factor	Third factor	Fourth factor
Guatemala				
Complexity				
Comprehension: complex syntax	–0.304	0.219		0.476
Comprehension: vagueness				0.508
Comprehension: reference category				
Comprehension: number of questions				0.348
Information retrieval: unfamiliarity	0.349	0.798	–0.355	–0.333
Information retrieval: recalling		0.445	0.263	
Information integration: computational intensity	–0.237		0.952	
Information integration: scope of information		0.329	–0.215	0.416
Translation to answer: categories mismatch	0.290	–0.342		
Translation to answer: number of responses		–0.281		

(continues on next page)

TABLE 22.4 Exploratory Factor Analysis *(continued)*

	First factor	Second factor	Third factor	Fourth factor
Sensitivity				
Invasion of privacy	0.259			–0.325
Social-emotional threat	0.484			
Formal threat of sanctions	0.776	0.241		–0.404
Informal-formal threat interaction	0.964			
Romania				
Complexity				
Comprehension: complex syntax			0.247	0.566
Comprehension: vagueness				–0.236
Comprehension: reference category		0.37		–0.447
Comprehension: number of questions				
Information retrieval: unfamiliarity		0.715		
Information retrieval: recalling			0.967	0.206
Information integration: computational intensity		–0.235		0.376
Information integration: scope of information		0.989		
Translation to answer: categories mismatch				
Translation to answer: number of responses			–0.252	
Sensitivity				
Invasion of privacy	0.532			
Social-emotional threat	0.65			
Formal threat of sanctions	0.806	0.314		
Informal-formal threat interaction	0.938	0.321		
United States (FEVS)				
Complexity				
Comprehension: complex syntax			0.982	
Comprehension: vagueness				–0.378
Comprehension: reference category		–0.356		–0.24
Comprehension: number of questions	0.403			
Information retrieval: unfamiliarity	0.207	0.544		
Information retrieval: recalling				
Information integration: computational intensity				0.597
Information integration: scope of information	0.23	0.795		
Translation to answer: category mismatch				
Translation to answer: number of responses				
Sensitivity				
Invasion of privacy				0.547
Social-emotional threat	0.489	0.438		–0.259
Formal threat of sanctions	0.989			
Informal-formal threat interaction	0.909			

Source: Original table for this publication.
Note: Only loadings with absolute values higher than 0.2 are shown. FEVS = Federal Employee Viewpoint Survey.

factor exhibit more variation, two subdimensions consistently score highly across countries: *unfamiliarity* and *scope of information*. Both these factors measure whether a question asks about the personal or at least proximate experiences of a respondent rather than the broader working environment (for example, the behavior of employees in the organization as a whole). Both thus relate closely to the unfamiliarity (of a topic). The remaining two factors vary, in terms of significant subdimensions, across countries and thus do not offer a clear conceptual interpretation.

We next assess whether the four factors from the EFA models—and, in particular, the sensitivity factor (first factor) and the unfamiliarity factor (second factor)—predict item nonresponse (table 22.5). We find that the first factor (sensitivity) does not predict item nonresponse. By contrast, the second factor (unfamiliarity) does predict item nonresponse in two of the three countries (Romania and the United States) (the third and fourth factors do not display clear patterns).[11]

As the EFA pointed to the sensitivity index as meaningfully reflecting the empirical structure of the subdimensions, while the complexity index consists of unfamiliarity and other complexity items, we next regress item nonresponse on unfamiliarity, sensitivity, and complexity without unfamiliarity (table 22.6). We find that unfamiliarity significantly predicts item nonresponse in Romania and the United States. It is also associated with greater item nonresponse in Guatemala, though this relationship is not significant at the standard significance levels.

The coefficients on the unfamiliarity index are larger than those on the basic indexes in table 22.3. A standard deviation increase in the unfamiliarity index (implying that the questions are *less* familiar) increases nonresponse by 3 percentage points in the United States and by almost 20 percentage points in the online survey in Romania. Relative to the baseline levels of nonresponse of 2.4 percent in the US FEVS and 5.9 percent and 15.9 percent in Romania's face-to-face and online surveys, respectively, these are large effects. By contrast, within this framework, the sensitivity index and complexity without unfamiliarity do not have significant effects. The evidence we present points to unfamiliarity, in the sense we have coded it, as the key driver of nonresponse.

TABLE 22.5 Factor Analysis Regression

	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
First factor	0.005 (0.005)	0.006 (0.007)	0.0001 (0.010)	0.006 (0.004)
Second factor	–0.002 (0.005)	0.048*** (0.007)	0.095*** (0.013)	0.018*** (0.003)
Third factor	–0.002 (0.003)	–0.009* (0.003)	–0.011 (0.006)	0.003 (0.003)
Fourth factor	0.011* (0.005)	0.016* (0.008)	0.029 (0.015)	–0.002 (0.002)
Controls	Yes	Yes	Yes	Yes
N	378,472	181,614	161,793	667,425
Adjusted R^2	0.005	0.057	0.094	0.015

Source: Original table for this publication.
Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered "I don't know," refused to answer, or skipped a particular question. Factor scores are obtained from exploratory factor analysis models with four factors, as presented in table 22.4. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is the response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.
Significance level: * = 10 percent, ** = 5 percent, *** = 1 percent.

TABLE 22.6 Impact of Sensitivity, Complexity, and Unfamiliarity on Nonresponse Rate

	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
Sensitivity	–0.005 (0.005)	0.004 (0.007)	–0.005 (0.011)	0.004 (0.004)
Complexity (without unfamiliarity subdimensions)	–0.004 (0.005)	–0.011 (0.008)	–0.027 (0.015)	–0.002 (0.003)
Unfamiliarity	0.007 (0.007)	0.097*** (0.017)	0.196*** (0.028)	0.030*** (0.007)
Controls	Yes	Yes	Yes	Yes
N	378,472	181,614	161,793	667,425
Adjusted R^2	0.001	0.061	0.100	0.012

Source: Original table for this publication.
Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered “I don’t know,” refused to answer, or skipped a particular question. Sensitivity and complexity scores are calculated as *z*-scores estimated across questions in a given survey. Unfamiliarity is calculated as a mean value of the “information retrieval: unfamiliarity” and “information integration: scope of information” subdimensions. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is the response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face. Significance level: * = 10 percent, ** = 5 percent, *** = 1 percent.

The Performance of Manual versus Machine-Coded Complexity

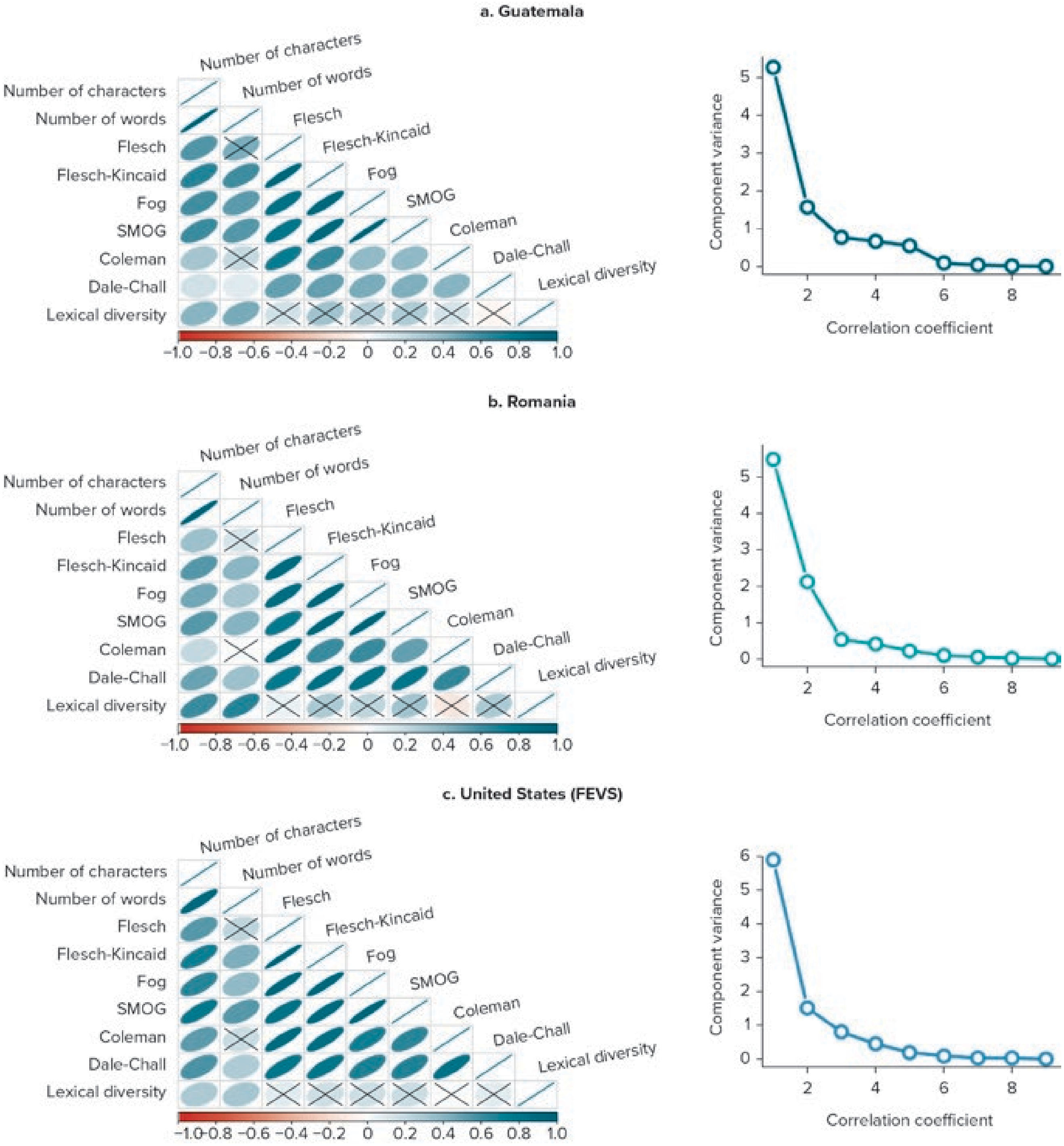
A potential criticism of our approach is that automated measures of complexity may be as effective a means of identifying potential problem questions—or more—while requiring far less investment. We therefore turn to assessing the relative effectiveness of the approach with respect to machine-coded complexity.

Machine coding has the disadvantage that it must rely on relatively basic indicators of syntax. Computer-based complexity indicators are usually based on mathematical formulas that score the complexity (or, as it is described more commonly, the *readability*) of a text based on purely textual features, like the number of characters, syllables, words, and sentences. Some measures also check the text against predefined lists of words regarded as easy or difficult. As there are dozens of such indexes, with no agreement on which one is optimal, we select nine that are commonly used and calculate their values for each survey question. Correlations among their final scores can be seen in the first column of figure 22.5. The correlation between models (shown by the intensity of shading) varies, though is understandably relatively high across the comparisons made. Given a very high degree of correlation, instead of using all nine scores in a regression, a principal component analysis is performed across them and extract the first principal component (which explains between 59 and 66 percent of the overall variance—see the second column in figure 22.5) to serve as a predictor in the regressions.

Tables 22.7 and 22.8 evaluate the predictive power of machine-driven complexity scores. When not controlling for the manually coded indexes (sensitivity, unfamiliarity, and complexity without unfamiliarity), we find some evidence for an effect of machine-coded complexity, with significant positive effects in the United States only.

Once we condition on the indexes of complexity (excluding complexity of syntax to avoid multicollinearity with the machine-coded measure) and sensitivity, we no longer find any evidence that the machine-coded complexity measure is predictive of greater item nonresponse. Table 22.8 presents the full regressions. The measure of how unfamiliar questions are is a significant and positive predictor of item nonresponse for the United States and both modes of the Romania survey. In Guatemala, the coefficient on unfamiliar is positive,

FIGURE 22.5 Relationship between Machine-Coded Complexity Scores: Correlograms and Scree Plots from Principal Component Analysis



a. Guatemala

b. Romania

c. United States (FEVS)

Source: Original figure for this publication.
Note: Crosses mark correlations that are insignificant at the 5 percent level. FEVS = Federal Employee Viewpoint Survey; SMOG = simple measure of gobbledygook.

TABLE 22.7 Impact of Machine-Coded Complexity on Nonresponse Rate

	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
Machine-coded complexity	−0.008 (0.006)	0.009 (0.010)	0.022 (0.015)	0.010*** (0.003)
Controls	Yes	Yes	Yes	Yes
N	378,472	181,614	161,793	667,425
Adjusted R^2	0.002	0.014	0.027	0.007

Source: Original table for this publication.
Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered "I don't know," refused to answer, or skipped a particular question. Machine-coded complexity is calculated as the first principal component across nine different machine-coded complexity scores (number of characters, number of words, Flesch's Reading Ease Score, Flesch-Kincaid Readability Score, Gunning's Fog Index, SMOG Index, Coleman's Readability Formula, Dale-Chall Readability Formula, and lexical diversity), as described in detail in appendix J. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is the response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.
Significance level: * = 10 percent, ** = 5 percent, *** = 1 percent.

TABLE 22.8 Full Model: Impact of Sensitivity, Complexity, Unfamiliarity, and Machine-Coded Complexity

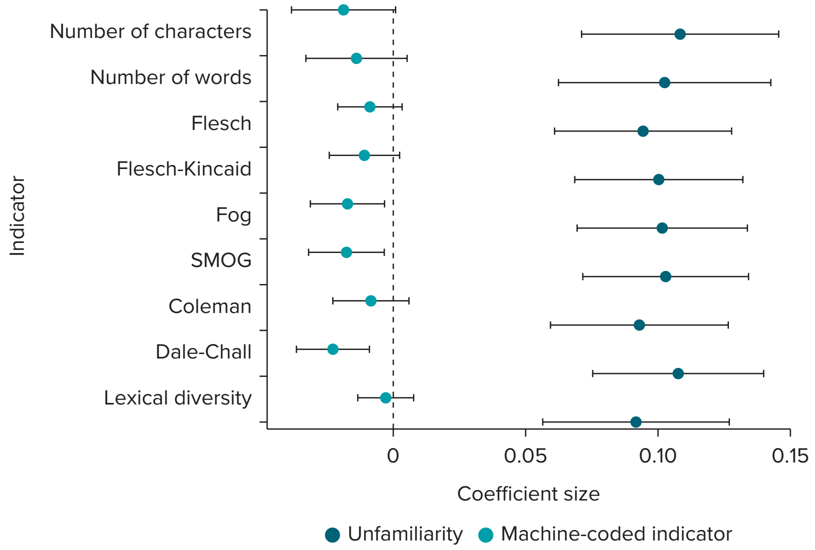
	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
Sensitivity	−0.002 (0.007)	0.002 (0.006)	−0.008 (0.010)	0.003 (0.003)
Complexity	−0.003 (0.005)	−0.011 (0.008)	−0.026 (0.015)	−0.002 (0.003)
Unfamiliarity	0.010 (0.006)	0.109*** (0.016)	0.217*** (0.026)	0.028*** (0.008)
Machine-coded complexity	−0.009 (0.009)	−0.017* (0.008)	−0.029 (0.015)	0.003 (0.003)
Controls	Yes	Yes	Yes	Yes
N	378,472	181,614	161,793	667,425
Adjusted R^2	0.003	0.064	0.103	0.013

Source: Original table for this publication.
Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered "I don't know," refused to answer, or skipped a particular question. Sensitivity and complexity scores are calculated as *z*-scores estimated across questions in a given survey. Unfamiliarity is calculated as a mean value of the "information retrieval: unfamiliarity" and "information integration: scope of information" subdimensions. Machine-coded complexity is calculated as the first principal component across nine different machine-coded complexity scores (number of characters, number of words, Flesch's Reading Ease Score, Flesch-Kincaid Readability Score, Gunning's Fog Index, SMOG Index, Coleman's Readability Formula, Dale-Chall Readability Formula, and lexical diversity), as described in detail in appendix J. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is the response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.
Significance level: * = 10 percent, ** = 5 percent, *** = 1 percent.

although smaller in size, and just misses the 10 percent threshold of statistical significance. The coefficients vary in size from 0.010 in Guatemala to 0.217 in the online mode of the Romania survey. The coefficients for sensitivity and the restricted measure of hand-coded complexity are insignificant and small across all the models.

To illustrate the relative predictive power of the framework relative to machine-coded methods, figure 22.6 also presents the coefficient sizes of each individual measure of machine-coded readability

FIGURE 22.6 Machine-Coded Complexity Indicators, Romania (F2F)



Source: Original figure for this publication.
Note: Values show the size of the coefficients of machine-coded complexity indicators when they are entered individually into a regression model with the standard dependent and control variables. The size of the coefficient for unfamiliarity entered into the same regression is shown for comparison. Error bars indicate 95 percent confidence intervals. F2F = face-to-face; SMOG = simple measure of gobbledygook.

and the unfamiliarity index (see appendix J to get details on each of the individual readability scores presented). For ease of presentation, we present coefficients from the Romanian face-to-face survey only, but the patterns are similar throughout. Measuring lack of familiarity clearly has a far greater predictive ability than any of the syntax-based, machine-coded measures.

CONCLUSION

The loss of precision and potential biases introduced by item nonresponse can hinder valid inference from surveys of public servants. Why do public servants respond to some questions but not others? The importance of this question stems from the proliferation of such surveys and their use for management reforms in government. Yet, to our knowledge, prior studies have not assessed item nonresponse in surveys of public officials.

This chapter contributes to addressing this gap. Building on the survey methodology literature, we design a unique coding framework to coherently assess the roles of question complexity and sensitivity in nonresponse in surveys of public servants. We apply this framework to governmentwide surveys of public officials in Guatemala, Romania, and the United States. As in the existing literature, we find that complexity matters for item nonresponse. Contrary to much prior work on item nonresponse, however, public servants do not seem to shy away from questions that are complex due to, for instance, syntax, computational intensity, or the number of response options (see, for example, Knäuper et al. 1997). As we argued in the introduction, this may be because public officials tend to be more educated and more accustomed to complex technical language in their day-to-day bureaucratic work. As such, they may be better able to cope with these dimensions of complexity. We find that asking public officials about issues with which they have lower familiarity is the feature of question design that is most robustly associated with item nonresponse. Questions that ask for assessments of public sector organizations as a whole or departments within them, for instance, lead to greater item nonresponse than questions about public servants themselves. By contrast, the findings provide little evidence that public officials shy away from answering sensitive questions. This does not, however, imply that responses to sensitive questions are not biased.

The implication for survey designers is clear: asking about topics public officials are less familiar with—such as their organizations or departments, rather than their immediate work environment—is associated with greater item nonresponse, with concomitant concerns about greater variance and potential biases in estimates. Where data aim to assess practices in larger units or institutions, it would thus be preferable, from a nonresponse perspective, to ask respondents about their individual-level experiences with organizational practices and aggregate these.

We have also compared the predictive ability of the findings to models that include machine-coded measures of complexity. The findings underscore the importance of manual assessments by survey designers to assess question complexity. While machine-coded estimates have some predictive power, this was eclipsed by the manual coding approach, once it was added to the models. Algorithms themselves appear to be an imprecise guide when assessing question complexity in surveys of public servants.

Future research could, in the first place, use the coding framework to understand whether our findings travel beyond Guatemala, Romania, and the United States. The diverse case contexts give us confidence that the findings are generalizable. Probing generalizability should not only extend to testing different country settings but also different survey administrators. The noticeably lower item nonresponse in the FEVS compared to the World Bank–administered surveys may reflect differential levels of trust in the survey administrator itself, for instance. The framework could equally be employed in employee surveys in private sector companies. One worthwhile area of investigation is to understand whether the findings are unique to public officials or would apply similarly to (educated) private sector administrators in a workplace survey.

Survey designers in the public service can utilize the coding framework to adjust survey questions in terms of their complexity and sensitivity. They can randomly roll out survey variations with different levels of these concepts—in particular, unfamiliarity—and assess experimentally whether this leads to improvements in item response rates in their setting.

The limitations of the findings should be kept in mind. In the first place, we only assess item nonresponse. Other threats to validity—such as overall survey nonresponse or response bias—may be of equal or greater concern. Sensitivity, for instance, was not robustly associated with greater item nonresponse across all of the surveys but may well lead to significant response bias.

Moreover, the inferences are necessarily limited by the number of surveys (three countries) and the types of questions included. In particular, the surveys contained relatively few highly sensitive questions (see figure 22.4), which might partially explain the null results obtained. It is possible that more discernible patterns in item nonresponse could be observed in surveys focused more squarely on sensitive topics—say, for instance, a corruption survey.

Overall, we present an analytically coherent approach to assessing survey item nonresponse that highlights a particular aspect of complexity—unfamiliarity—as the fundamental driver of nonresponse.

NOTES

The authors would like to thank Lior Kohanan, Miguel Mangunpratomo, and Sean Tu for excellent research assistance, Kerenssa Kay for guidance and advice, and seminar participants at the World Bank for their comments.

1. According to the Worldwide Bureaucracy Indicators published by the World Bank, the share of publicly paid employees with tertiary education across the world is 54.2 percent, whereas in the private sector, it is around half of that: 26.9 percent (average over 2010–18).
2. More information on the surveys used and the reason for their selection is presented in the methodology section of this chapter.
3. Apart from the degree of complexity, which is a stable feature of a question, the likelihood of engaging in satisficing also depends on respondents' characteristics that might increase or decrease their cognitive capacity (for example, age, education, or tiredness) and on their willingness to answer the question. Contextual variables, like the pace at which the interviewer conducts an interview or time pressure (for example, having only a 20-minute slot to take a survey), can also impact the degree to which respondents are willing and able to engage in high cognitive effort (Fazio and Roskos-Ewoldsen 2005; Lessler, Tourangeau, and Salter 1989).

4. For a more in-depth discussion, see, for example, Paulhus (2002).
5. This might be a particular concern in restricted-sample settings, like the ones in which public administration surveys are usually conducted. This is because "individuals who complete surveys may worry that their unique responses to demographic questions could allow researchers to identify them, especially if they are part of a known sample, such as a survey conducted within one's workplace" (McNeeley 2012, 4380).
6. Some other methods, like the randomized response technique, actually lead to higher item nonresponse, but due to the convoluted instructions they entail rather than the nature of the question itself.
7. Bais et al. (2019) do not disaggregate sensitivity to the same extent we do and, more importantly, do not assess whether their measures of complexity and sensitivity predict item nonresponse.
8. The exclusion of this variable does not change the substantive conclusions.
9. This is consistent with prior work that associates online surveys with higher nonresponse than face-to-face surveys (Heerwegh and Loosveldt 2008).
10. The decision was made based on the *p*-values of the EFA models. In Guatemala, the model was significant at the 5 percent level only when five factors were used, but for cross-country consistency, we employ a four-factor model throughout the analyses.
11. In the following regression tables, we present results only with the standard set of controls. The results, however, hold in unconditional regressions as well.

REFERENCES

Anderson, N. 1971. "Integration Theory and Attitude Change." *Psychological Review* 78 (3): 171–206.

Bais, F., B. Schouten, P. Lugtig, V. Toepoel, J. Arends-Tòth, S. Douhou, N. Kieruj, M. Morren, and C. Vis. 2019. "Can Survey Item Characteristics Relevant to Measurement Error Be Coded Reliably? A Case Study on 11 Dutch General Population Surveys." *Sociological Research Methods and Research* 48 (2): 263–95.

Belson, W. A. 1981. *The Design and Understanding of Survey Questions*. Aldershot, UK: Gower.

Berger, I., and A. Mitchell. 1989. "The Effect of Advertising on Attitude Accessibility, Attitude Confidence, and the Attitude-Behavior Relationship." *Journal of Consumer Research* 16 (3): 269–79.

Bradburn, N., S. Sudman, E. Blair, and C. Stocking. 1978. "Question Threat and Response Bias." *Public Opinion Quarterly* 42 (2): 221–34.

Coutts, E., and B. Jann. 2011. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)." *Sociological Methods and Research* 40 (1): 169–93.

De Leeuw, E. 1992. *Data Quality in Mail, Telephone and Face to Face Surveys*. Amsterdam: Netherlands Organization for Scientific Research.

Edwards, J., M. Thomas, P. Rosenfeld, and S. Booth-Kewley. 1997. *How to Conduct Organizational Surveys: A Step-by-Step Guide*. Thousand Oaks, CA: SAGE Publications.

Faaß, T., L. Kaczmirek, and A. Lenzner. 2008. "Psycholinguistic Determinants of Question Difficulty: A Web Experiment." Paper presented at the seventh International Conference on Social Science Methodology.

Fazio, R. 1986. "How Do Attitudes Guide Behavior?" Chap. 8 in *Handbook of Motivation and Cognition: Foundations of Social Behaviour*, edited by R. Sorrentiono and E. Higgins, 204–43. New York: Guilford Press.

Fazio, R. 1989. "Attitude Accessibility, Attitude-Behavior Consistency, and the Strength of the Object-Evaluation Association." *Journal of Experimental Social Psychology* 18 (4): 339–57.

Fazio, R., and D. Roskos-Ewoldsen. 2005. "Acting as We Feel: When and How Attitudes Guide Behavior." In *Persuasion: Psychological Insights and Perspectives*, edited by T. Brock and M. Green, 41–62. Thousand Oaks, CA: SAGE Publications.

Galletly, C., and S. Pinkerton. 2006. "Conflicting Messages: How Criminal HIV Disclosure Laws Undermine Public Health Efforts to Control the Spread of HIV." *AIDS and Behaviour* 10: 451–61.

Gnambs, T., and K. Kaspar. 2015. "Disclosure of Sensitive Behaviors across Self-Administered Survey Modes: A Meta-analysis." *Behaviour Research Methods* 47: 1237–59.

Haziza, D., and G. Kuromi. 2007. "Handling Item Nonresponse in Surveys." *Journal of Case Studies in Business, Industry and Government Statistics* 1 (2): 102–18.

Heerwegh, D., and G. Loosveldt. 2008. "Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality." *Public Opinion Quarterly* 72 (5): 836–46.

Höglinger, M., B. Jann, and A. Diekmann. 2016. "Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model." *Survey Research Methods* 10 (3): 171–87.

Holbrook, A., Y. Cho, and T. Johnson. 2006. “The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties.” *Public Opinion Quarterly* 70 (4): 565–95.

Just, M., and P. Carpenter. 1992. “A Capacity Theory of Comprehension: Individual Differences in Working Memory.” *Psychological Review* 99 (1): 122–49.

Kim, S., and S. Kim. 2016. “Social Desirability Bias in Measuring Public Service Motivation.” *International Public Management Journal* 19 (3): 293–319.

Knäuper, B., R. Belli, D. Hill, and R. Herzog. 1997. “Question Difficulty and Respondents’ Cognitive Ability: The Effect on Data Quality.” *Journal of Official Statistics* 13 (2): 181–99.

Krosnick, J. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5 (3): 213–36.

Krosnick, J., and S. Presser. 2009. “Question and Questionnaire Design.” In *Handbook of Survey Research*, 2nd edition, edited by J. Wright and P. Marsden. San Diego, CA: Elsevier.

Krumpal, I. 2013. “Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review.” *Quality and Quantity* 47: 2025–47.

Lenski, G., and J. Leggett. 1960. “Caste, Class, and Deference in the Research Interview.” *American Journal of Sociology* 65 (5): 463–67.

Lensvelt-Mulders, G. 2008. “Surveying Sensitive Topics.” In *International Handbook of Survey Methodology*, edited by E. de Leeuw, J. Hox, and D. Dillman, 461–578. New York: Routledge.

Lessler, J., R. Tourangeau, and W. Salter. 1989. *Questionnaire Design in the Cognitive Research Laboratory*. Vital and Health Statistics 6, Cognitive and Survey Measurement 1. DHHS Publication No. (PHS) 89-1076. Hyattsville, MD: US Department of Health and Human Services.

McNeeley, S. 2012. “Sensitive Issues in Surveys: Reducing Refusals While Increasing Reliability and Quality of Responses to Sensitive Survey Items.” In *Handbook of Survey Methodology for the Social Sciences*, edited by L. Gideon, 4377–96. New York: Springer.

OPM (Office of Personnel Management). 2019. *2019 Office of Personnel Management Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: OPM.

Paulhus, D. 1984. “Two-Component Models of Socially Desirable Responding.” *Journal of Personality and Social Psychology* 46 (3): 598–609.

Paulhus, D. 2002. “Socially Desirable Responding: The Evolution of a Construct.” In *The Role of Constructs in Psychological and Educational Measurement*, edited by H. Braun, D. Jackson, and D. Wiley, 49–69. Mahwah, NJ: Erlbaum.

Rässler, S., and R. T. Riphahn. 2006. “Survey Item Nonresponse and Its Treatment.” *Allgemeines Statistisches Arch* 90: 217–32.

Sakshaug, J., T. Yan, and R. Tourangeau. 2010. “Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multimode Survey of Sensitive and Non-sensitive Items.” *Public Opinion Quarterly* 74 (5): 907–33.

Tourangeau, R. 1984. “Cognitive Sciences and Survey Methods.” In *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, edited by T. Jabine, M. Straf, J. Tanur, and R. Tourangeau, 73–100. Washington, DC: National Academy Press.

Tourangeau, R., R. Groves, and C. Redline. 2010. “Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error.” *Public Opinion Quarterly* 74 (3): 413–32.

Tourangeau, R., and K. A. Rasinski. 1988. “Cognitive Processes Underlying Context Effects in Attitude Measurement.” *Psychological Bulletin* 103 (3): 299–314.

Tourangeau, R., L. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Tourangeau, R., and T. Smith. 1996. “Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context.” *Public Opinion Quarterly* 60 (2): 275–304.

Tourangeau, R., and T. Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133 (5): 859–83.

Tversky, A., and D. Kahneman. 1974. “Judgment under Uncertainly: Heuristics and Biases.” *Science* 185 (4157): 1124–31.

World Bank. 2020a. *Final Field Report: Encuesta General de Servidores Públicos y Contratistas del Organismo Ejecutivo y Entitdaded Descentralizadas 2019*. Washington, DC: World Bank.

World Bank. 2020b. *Selecting the Right Staff and Keeping Them Motivated for a High-Performing Public Administration in Romania: Key Findings from a Public Administration Employee Survey*. Washington, DC: World Bank.

Yan, T., and R. Tourangeau. 2008. “Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times.” *Applied Cognitive Psychology* 22 (1): 51–68.

Zaller, J. 1992. *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.

CHAPTER 23

Designing Survey Questionnaires

Should Surveys Ask about Public Servants' Perceptions of Their Organization or Their Individual Experience?

Kim Sass Mikkelsen and Camille Mercedes Parker

SUMMARY

Civil service surveys are often interested in organizational aggregates and comparisons across organizations. Therefore, the choice of question referent is important in questionnaire design. Should survey questions refer to individual employees or to employees' assessments of their organizations? This chapter provides tools for thinking through this choice. Moreover, experimental evidence from representative public service surveys in Romania and Guatemala shows that the choice of referent matters to how employees respond. Finally, the chapter provides evidence that organizational referents can help reduce socially desirable responding, particularly for highly sensitive questions, and that referent effects may be larger for attitudes and behaviors that are uncommon, but that the size of referent effects beyond this is difficult to predict.

ANALYTICS IN PRACTICE

- Many civil service surveys are centrally interested in organizational aggregates. Therefore, the choice of question referent is important in questionnaire design. Should questions refer to individual employees or to employees' assessments of their organizations?
- Inside organizations, perceptions of management practices are often only weakly correlated across respondents, suggesting that they are not organizational *constructs*. Organizational referents can—but often do not—better enable survey questions to reflect organizational constructs.

Kim Sass Mikkelsen is an associate professor of politics and public administration at Roskilde University. Camille Mercedes Parker is an economist at the United States Agency for International Development.

- Experimental evidence from representative public service surveys in Romania and Guatemala shows that the choice of referent matters to how employees respond.
- We provide evidence that organizational referents may help reduce socially desirable responding, particularly for highly sensitive questions.
- We examine, but uncover little systematic evidence for, a set of other factors that could conceivably influence question-referent effects. We conclude that organizational referents may be less useful in situations where attitudes and behaviors are uncommon because respondents may not have the needed information to answer them. Beyond this, however, the size of referent effects is difficult to predict.

INTRODUCTION

Many civil service surveys are centrally interested in organizational aggregates. Which surveyed organization has the highest level of job satisfaction among its employees? Which organizations need additional ethics training to keep up with the ethical awareness of employees in other organizations? Questions such as these are core both to internal government benchmarking and, since aggregates attached to recognizable labels (like organization names) are simple to interpret, to government communication of data from civil service surveys.

The focus on organizational aggregates has an intuitive implication for how questions should be asked in civil service surveys: ask civil servants to evaluate their organizations. Indeed, practitioners and academics alike routinely ask civil servants for such evaluations. For example, the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) asks its respondents to evaluate the extent to which "employees are protected from health and safety hazards on the job" as a measure of workplace safety (OPM 2018). This practice is sensible. If the target of evaluations is the organization, it seems reasonable to align the *level of measurement*—the level in reference to which respondents are asked to provide answers—with the level at which claims are made (Klein, Dansereau, and Hall 1994). *Referents*, the entities to which a survey question refers, can be sensibly chosen to reflect the entities researchers wish to learn about. The question cited above is an example of the use of *organizational referents* in civil service surveys.

It is not always clear, however, that the organization is the most appropriate or most useful level of measurement. While recognizable labels make organizational comparisons simple and appealing, using organizational referents implies one of two claims: that the question measures the respondent's *perceptions* of his or her organizational surroundings, or that the subject of the question is an *organization-level phenomenon* (Klein, Dansereau, and Hall 1994; Klein and Kozlowski 2000). In the first instance, *top-down* claims can be made about how respondents react to their perceptions of organizational characteristics, management practices, leadership, culture, and so on. In the second instance, *bottom-up* claims can be made about organization-level phenomena principally detached from any individual public servant's experiences and beliefs. Both of these claims may be true, but they are infrequently stated explicitly.

Levels of measurement have been subject to contention in leadership research (for example, Schriesheim, Wu, and Scandura 2009), organizational research (for example, Baltes, Zhdanova, and Parker 2009; Chan 1998; Klein and Kozlowski 2000), and survey methods research (for example, Blair, Menon, and Bickart 2004). However, the issue is rarely discussed in the inherently multilevel field focused on civil service surveys. Is the common practice of asking civil service survey questions at the organizational level sensible? Or is the use of *individual referents*, asking respondents to provide information about themselves, a more appropriate strategy? And does the choice matter for survey results?

What is at issue is not whether the level of analysis should match the level at which claims and comparisons are made. There is already good evidence that these levels should match and that the consequence of mismatches is potentially biased results (for example, Gingerich 2013). Instead, we examine the advantages and disadvantages of using individual versus organizational referents in civil surveys. Should the level of *measurement* match the level at which claims and comparisons are made as well? In which situations should the level of measurement be the individual respondent, and in which should it be individuals' assessments of their employing organizations?

Intuitively, the answer parallels the match between claims and levels of analysis. If one is interested in individual public servants, individual referents should be used. If, by contrast, one is interested in organizations, organizational referents should be used. However, this answer is too simple. It underestimates the complexity of the consequences of choosing question referents. Our chapter describes some of this complexity and provides guidelines for understanding what is at stake in choosing question referents and when to choose which referent.

We are—as far as we know—the first to assess the issue of referent choice in civil service survey design. Yet the public organizational setting likely matters. Organizational referents require information from respondents that civil servants may not possess to the same degree private sector employees do, for instance. Public organizations are frequently very large in terms of both personnel and budget and are often hierarchically organized into relatively segmented and informationally insular parts (for example, Eggers 2007). This can make organizational-referent questions difficult for a public official in one part of an organization to answer due to a simple lack of knowledge about other parts of that same organization (cf. Homburg et al. 2012).

For organizations like ministries, this problem may even grow with managerial reforms that further segment and fragment the ministerial hierarchy into deliberately insular agencies (Dunleavy et al. 2006). In a sense, organizational referents in civil service surveys may have to grapple with issues similar to those that whole-of-government approaches to public sector organizations were intended to solve: information and knowledge can have a hard time traversing the organizations that respondents are asked to evaluate (for example, Christensen and Lægreid 2007).

Our advice to civil service survey designers is not to abandon one question referent in favor of another. Instead, we provide a set of important considerations that designers can use when choosing referents. In particular, designers should consider:

- Whether what they are measuring is, conceptually, an organizational phenomenon. Does it make sense to think of all respondents within an organization rating the same entity when responding? If designers are not measuring an organizational phenomenon, organizational referents are less attractive.
- How sensitive their measures are. Respondents tend to respond as they believe is socially desirable when questions are sensitive, and this effect is more pronounced for questions about them as individuals. If questions are very sensitive, organizational referents may be more attractive.
- How easily accessible to respondents the information required for the measure is. Respondents often have better access to their own experiences, beliefs, and attitudes than those of their colleagues. If questions require information that is not readily available to respondents, organizational referents are less attractive.

These conclusions are based partly on a conceptual discussion and partly on empirical evidence. Empirically, we use experiments embedded in two civil service surveys. We embedded experiments in a survey of more than 6,000 civil servants in Romania's central government, randomly assigning respondents to answer questions about human resource management practices—specifically, recruitment, promotion, dismissal, and turnover intent—using individual or organizational referents. We embedded a similar experiment in a survey of more than 3,000 civil servants in Guatemala's central government.

The basic thrust of the experiment is that, if referent choice matters, otherwise similar questions using different referents will result in different average responses. If referents do not matter, the average employee evaluation of the organization will correspond to the average of the employees' evaluation of themselves. Thus, the experiment can provide evidence that referents matter, the core interest of this chapter. The drawback is that the sources of divergences are harder to determine. We do conduct a series of tests attempting to determine sources, but the question we can answer most clearly is whether referents matter. Despite its simplicity, a strong answer to this question is useful to survey designers, many of whom do not seem to know whether referents matter or how they matter to the responses they get.

Beyond the questions it can answer, this experimental approach is valuable for the strength of our conclusions. And it sets our study apart from previous examinations of the use of referents in organizational surveys (for example, Baltes, Zhdanova, and Parker 2009; Klein et al. 2001). Prior examinations of referent issues have asked the same respondents to provide information both about themselves (using individual referents) *and* about their organizations (using organizational referents). This is needed, of course, to show that each of the two referents contributes separate information (Klein et al. 2001). However, it creates the risk that respondents anchor their responses to one set of questions to their answers to the other set of questions in order to appear consistent, or that they respond to both sets of questions relative to one another, either to maintain that they are "above average" on relevant metrics (Guenther and Alicke 2010) or because they form their answers relative to comparisons with significant colleagues (Baltes, Zhdanova, and Parker 2009). Thus, responses to questions with individual referents can affect subsequent responses to question sets with organizational referents and vice versa. Our experimental design avoids this issue, permitting a causal assessment of the relative differences between responses stemming from the two referents.[1]

We proceed in four steps. In section 2, we discuss what difference organizational as opposed to individual referents might make theoretically. We focus particularly on concept levels, socially desirable responding, and information availability. In section 3, we describe our survey experiments. Section 4 presents our results. Section 5 contains our discussion of these results for the design of civil service surveys.

WHAT IS AT STAKE?

In this section, we provide a more detailed account of the already-noted reasons why the choice of question referent might matter. This takes us into the psychology of survey response and questions about levels of theory and measurement from organizational studies. But the point is not the theory. Rather, we aim to provide readers with a rough and simple understanding of the stakes in choosing between individual and organizational referents. Table 23.1 provides an overview of the arguments we discuss. These fall along three main lines: the match between the measure and the target entity of interest, socially desirable responding, and the informational requirements placed on respondents.

Do Analyses Concern Individuals or Organizations?

At base, the choice of referent should reflect the interest of subsequent analyses. If the interest is in measuring, comparing, or benchmarking organizations, organizational referents appear to be the obvious choice because they create a clear match between the entities in subsequent analyses (the target) and the measure. However, this is not as obvious as it would at first appear. Table 23.2 provides an overview of the discussion.

Table 23.2 distinguishes between referents, the entities referred to in survey questions, and target entities, the entities the survey aims to learn something about. The intuition is that referents should be chosen to match the downward diagonal of the table. Inquiries with an individual focus should ask individual-referent questions, while organizational inquiries should use organizational referents.

The first half of this intuition holds. Inquiries with an individual focus should likely ask about individuals. But the second half of the intuition is more complicated. There are three ways of thinking about settings where questions either use organizational referents or aim to learn about organizations: the *top-down* perspective, which asks about organizations to learn about individual employees, the *bottom-up* perspective, which asks about organizations to learn about organizations (when possible) (Klein and Kozlowski 2000),

TABLE 23.1 Advantages and Disadvantages of Organizational and Individual Referents When Used for Calculating and Analyzing Organizational Aggregates

Type of cost or benefit	Organizational referents	Individual referents
Conceptual	+ Match between target and measure − Disagreement	+ No agreement requirement − Possible mismatch between target and measure
Measurement	+ Decreased social-desirability bias − Informational requirements	+ Fewer informational requirements − Social-desirability bias

Source: Original table for this publication.
Note: This table shows a summary of the discussion in the three following subsections. Columns represent question referents (organizational vs. individual). Rows are divided into conceptual concerns (discussed in the first subsection) and measurement concerns (discussed in the second and third subsections). Plus signs indicate competitive advantages relative to the referent in the other column; minus signs indicate competitive disadvantages. Advantages and disadvantages are relative to data used for calculating organizational aggregates. Some points are not relevant in other contexts (for example, “match between target and measure” is not a competitive advantage for organizational-referent questions if an individual’s beliefs are the target, as in the top-down perspective).

TABLE 23.2 Question Referents and Target Entities

		Target entity	
		Organization	*Individual employee*
Question referent	*Organizational*	Bottom-up	Top-down
	Individual	Summary bottom-up	Individual focus

Source: Original table for this publication.

and finally, what we call the *summary bottom-up* perspective, which asks about individuals to learn about organizations through data summaries.

The *top-down* perspective interprets organizational-referent questions as asking about respondents’ perceptions of their working environment. Even questions that appear to be intrinsically at the organizational level may be best thought of at the individual level in terms of definitions, causal efficacy, or both. For instance, Parker et al. (2003, 390) define the *psychological climate* in organizations—a term that, intuitively, has a clear organizational focus, though it does not have this connotation in the relevant literature—as “an individual’s psychologically meaningful representations of proximal organizational structures, processes, and events.” Such representations—including perceptions of management practices and attributions related to those perceptions—are often proposed as causally efficacious for important employee outcomes (for example, Nishii, Lepak, and Schneider 2008). They are related to organizational practices, but they are not themselves organization-level phenomena. Rather, it is employee perceptions or experiences that matter for outcomes. From this perspective, organizational-referent questions are not asking respondents to rate the same entity—indeed, they are, in a sense, not organizational at all. Instead, they are asking about individuals’ representations, beliefs, or experiences. From this perspective, answers to the FEVS question about whether “employees are protected from health and safety hazards on the job” can be interpreted as reflecting individual respondents’ beliefs about health and safety in their workplaces—which can be relevant to understanding their commitment to their workplaces, their job satisfaction, or their turnover intent—but not, strictly, as offering descriptions of their workplaces as they are.

The *bottom-up* perspective is more complicated. It involves interpreting respondents’ evaluations as reflecting genuine organizational constructs—that is, features of the organization—over and above the perspective of the individual respondent. It is not perceptions but features of the organization that are the target of organizational-referent questions, from this perspective. Respondents within an organization are all seen as rating the same entity with the same characteristics.[2] The bottom-up perspective on organizational referents assumes that the characteristic of concern in a question is a characteristic of the organization, not of the respondent. From this perspective, answers to the question about whether "employees are protected from health and safety hazards on the job" are ratings of the organization; they ask the responding employee to evaluate the organization (principally) as a whole. Consequently, since respondents within an organization are rating the same entity, the bottom-up perspective assumes a substantial level of agreement among respondents in the same organization.

Based on the assumptions behind the bottom-up perspective, it seems reasonable to believe that using organizational referents furthers agreement on responses within organizations because individual respondents are essentially instructed to disregard their personal experiences and report using a *referent shift*. From this perspective, there may be reason to prefer organizational referents because they may further the agreement necessary for the desired bottom-up interpretation of organizational aggregates as reliable descriptions of the organization as one entity evaluated by multiple raters.

But what if respondents within organizations do not agree? The answer can be stated, likely too succinctly: then the measures do not appropriately measure an organization-level characteristic but a construct at a lower level (such as an employee perception). This brings us to the *summary bottom-up* perspective, which construes descriptions of organizations using survey data as summaries of individual perspectives. Employee responses to organizational referents can be thought of as such summaries, but they do not have the advantage of capturing the organization above individual perceptions and experiences. This is because the perspective does not consider employees as rating the organization but as providing their own views.

Uneven implementation within organizations is often proposed as a vehicle for intraorganizational differentiation in civil service management practices when these are measured using organizational referents (Bezes and Jeannot 2018; Meyer-Sahling and Mikkelsen 2016, 2020). From this argument, questions with organizational referents do not necessarily result in organization-level assessments by respondents but rather elicit the experience of respondents in their immediate working environments. The disadvantage is uncertainty about the width of the assessments provided by individual respondents if these are not at the organizational level to which survey items refer. If organizational referents do not prompt consensus on ratings of the same entity, it is not clear what level the questions measure. Instead of capturing their organizational target, organizational aggregates are reduced to *summaries* of features of lower levels, be these sections, teams, or individuals.

Indeed, when organizational aggregates of responses are seen as summaries of individual perspectives, organizations are arguably better described using individual-referent questions: the width of the assessment is determined by the question, and the result is still a useful organizational summary. The cost of this view is that organizational characteristics are redefined to nothing more than aggregates of individual answers. Organizational workplace safety, for example, becomes the proportion of employees who think their work is safe.

In sum, if civil service survey designers are primarily interested in organizational aggregates, should they ask questions with organizational referents to ensure correspondence between levels of measurement and levels of theory? It depends. If respondents' within-organization responses are strongly correlated, individual and organizational referents are both useful measures of organizational characteristics. While they entail different perspectives on interpreting answers, and the bottom-up perspective has a more intuitive appeal, both kinds of referent can be used.

However, if responses do not strongly correlate within organizations, this indicates that the use of individual referents is preferable on a conceptual basis. The bottom-up perspective, in this situation, does not lend as much analytical leverage as the summary bottom-up perspective because responses do not reflect an organization-level construct; instead, organizational aggregates are more readily understood as summaries of employee information.

In sum, if employees' beliefs and perceptions are of central interest—as in the right column of table 23.2—the choice of referent is conceptual, not statistical. In that case, organizational referents should be used if respondents' beliefs about the organization are of central interest, and individual questions should be used if respondents' own experiences and behaviors are of interest. However, if the organization is the target, the preferable choice of referent is, in part, statistical because organizational-referent questions impose the requirement of *interrater agreement* among employees of the same organization, while individual-referent questions do not. Even such statistically based choices have conceptual consequences, however, since the bottom-up and summary bottom-up perspectives use different ideas about the composition of individual responses and hence capture somewhat different ideas about what organizational aggregates are (Chan 1998).

Are Questions Sensitive?

It is often less embarrassing and feels less threatening to respond to a question in a socially undesirable way if the question is not about oneself. "Do you ever steal stationery from work?" is a much more sensitive question on its face than "Do colleagues in your organization ever steal stationery from work?" Consequently, many researchers utilize organizational referents not on conceptual grounds but to limit socially desirable responses. Organizational referents are used to make sensitive questions less sensitive to respondents, on the assumption that they will provide more truthful answers and avoid social-desirability bias (SDB) due to question sensitivity (for example, Graaf, Huberts, and Strüwer 2018; Meyer-Sahling and Mikkelsen 2016).

This assumption is plausible and has been indirectly tested in other fields under labels such as "proxy questioning" (Blair, Menon, and Bickart 2004) and "structured projective questioning" (Fisher 1993). For instance, in marketing, Fisher (1993) studies whether questions that ask for the opinion of others rather than the respondent's own opinion can reduce SDB. Fisher's finding accords with the assumptions made in analyses of civil service survey data: indirect questions reduce SDB on questions subject to social influence. Thematically closer to our purpose, Bardasi et al. (2011) find that reported male labor market participation rates dropped substantially when others provided proxy answers, rather than the men themselves. Like these approaches, the use of organizational referents is sometimes interpreted as an indirect question technique because respondents provide information about others, not about themselves.

Questions engender SDB through several channels. Questions can be intrusive, threatening, or socially undesirable (Tourangeau and Yan 2007). Intrusive questions can be seen as offensive, nosy, or taboo. Threatening questions make respondents worry about the disclosure of their responses and the negative consequences that may ensue. Finally, socially undesirable questions are questions for which certain answers violate social norms.

Disclosure threats and socially undesirable answers are particularly relevant to our discussion. In organizational settings, the disclosure of attitudes and behaviors to which colleagues, management, political superiors, the media, or the public will react negatively is a real concern. This is true of questions for which admitting to behaviors can have negative career consequences—such as admitting to kickbacks (Meyer-Sahling and Mikkelsen 2016). And it is true of questions for which agreeing or disagreeing can be seen as negative by colleagues or management and have negative consequences in terms of careers or ostracization at work. *Sensitive* questions—for example, questions about corruption or absenteeism—thus engender one form of SDB, but not the only form. Socially desirable responding can also occur for questions in which anything but a strong endorsement of the question's content can be seen as undesirable—such as questions about helping colleagues or working hard.

If SDB were all about threats of disclosure, however, anonymity safeguards for individual responses should help the problem. Unfortunately, SDB persists—albeit to varying degrees—even when anonymity is guaranteed (Kreuter, Presser, and Tourangeau 2008).[3] This is why many contemporary studies of very sensitive topics, such as corruption, employ indirect questioning techniques, such as the randomized response method (for example, Gingerich 2013) or conjoint experiments (for example, Schuster, Meyer-Sahling, and Mikkelsen 2020) to protect respondents' answers. When such techniques are too cumbersome or are not available, the use of organizational referents may be an attractive way to combat residual socially desirable responding by asking respondents about sensitive topics less directly. The cost of doing so, as we discuss below, is that organizational-referent questions on sensitive topics often place strong demands on respondents for information that may not be accessible to them.

In situations where SDB is severe and information is at least somewhat readily available to organizational outsiders, organizational aggregates may even be obtained from raters external to the organization. Such individuals will likely be less affected by SDB, although they may have other interests at stake in responding. However, using their answers comes at the cost of losing access to information from inside organizational boundaries, which may make their assessments noisy or inaccurate (for example, Razafindrakoto and Roubaud 2010). And, of course, this problem is likely to be particularly pernicious for sensitive questions, in which information is likely to be deliberately concealed from external assessment.

In sum, question sensitivity is a common reason for the use of organizational referents. There are good reasons to think this is an effective strategy, but, as far as we know, it has not been empirically examined in the context of civil service surveys. We do so below.

Is Organizational Information Available to Respondents?

The third topic we cover concerns information. Specifically, in some circumstances, it may be difficult for respondents to have the information that organizational referents ask them to provide. When asked a question with an individual referent, respondents work to retrieve or recall information about the question (Tourangeau, Rips, and Rasinski 2000). For past behaviors, recall involves respondents' remembering what they have previously done. For beliefs or attitudes, following Zaller (1992), we can think of recall as respondents' process of deciding what beliefs or attitudes they hold, which can be either remembered or formed on the spot based on available information.

Recall and introspection are not perfectly reliable, and respondents tend to "fill in" information they are unsure about or do not recall accurately (Tourangeau, Rips, and Rasinski 2000). Yet the difficulties can multiply when questions are posed using organizational referents. Organizational referents impose an additional challenge for respondents. If organizational referents work as intended, respondents rely on different sources of information when answering questions about themselves or about others (cf. Blair, Menon, and Bickart 2004). It is reasonable to believe that information about aggregates, such as organizations, will often be harder for respondents to access, and perhaps harder to recall, than information obtained by introspection (that is, information about themselves). Consequently, respondents' beliefs about their organizational surroundings may be mistaken or biased, which may influence their responses.

When Meyer-Sahling and Mikkelsen (2016), for instance, ask respondents whether "political parties place their supporters in the ministerial structure" as a measure of personnel politicization, they are asking respondents for an evaluation they may not have sufficient information to provide accurately. Did new recruits get their positions due to political influence? Politicization may be hidden, particularly where—as in Central and Eastern Europe, where the authors collect their data—political influence over recruitment often extends to positions formally codified as career posts (for example, Meyer-Sahling 2011).

Due to these difficulties, respondents who are asked questions using organizational referents may get their answers wrong, with consequences for measurement. The literature on *establishment surveys*—instruments in which one respondent replies on behalf of an organization—assumes that respondents use records from the establishment to counteract these difficulties (Edwards and Cantor 2004). However, it is certainly optimistic to expect respondents in civil service surveys to do the same. Even if they could and were willing, many of the topics of central interest to civil service surveys—like politicization—are often not formally recorded. As such, errors rooted in mistaken beliefs are likely to persist.

Moreover, when respondents lack necessary information, they may default to public sector stereotypes or other heuristic shortcuts to construct an answer. If public servants hold views similar to the general public, for instance, they may default to considering their colleagues as stereotypically caring or dedicated (Willems 2020), irrespective of their own concrete knowledge about the caring or dedication of organizational members beyond their immediate coworkers. Or respondents may extrapolate from anecdotes or stories to a systemic evaluation, particularly if they are asked to evaluate questions on topics they view as threatening or emotionally engaging (Freling et al. 2020).

From the perspective of the response process, the built-in assumption behind the use of organizational referents can easily come to seem somewhat heroic in large and complex organizations. Findings from previous studies do not help. Baltes, Zhdanova, and Parker (2009) propose that respondents may rely on "better-off" or "worse-off" colleagues when responding to questions with organizational referents. This may bias estimates of organizational aggregates because the implicit referents that are actually used are no longer representative of the organization.[4] Similarly, Shah (1998) finds that job-related information is often obtained from people in similar positions, whereas organization-related information is obtained from friends within the organization. This means organizational-referent questions are answered using networked information rather than representative information or simple ratings of features of the organization.

However, organizational information may not be equally difficult to obtain in all organizations or by all public servants. When answering questions about others, respondents may start with themselves and subsequently take in the stories and observed behavior of others (Hoch 1987). This information may be sourced from networks, but there are predictable situations in which it is more likely to accurately represent the organization. In those situations, question referents are likely to matter less for responses, and hence concerns with the information requirements of organizational referents may not matter in practice.

First, drawing information from unrepresentative colleagues, stories, and observed behaviors should matter less when questions concern attitudes or behaviors that are either very rare or very common. In these situations, most colleagues, stories, and observed behaviors will provide the same information: that the attitude or behavior is very rare or very common. This means that while respondents may not, in fact, know the answer to a question using an organizational referent, their assessment is likely to be less affected by how they arrive at it. A similar point holds when most members of an organization hold roughly similar views because networked information in this situation is also more likely to be representative of the common view in the organization.

Learning is another factor that may limit how much questions using organizational referents elicit biased assessments. For instance, years of employment in an organization may improve the accuracy of reports about it (cf. Blair, Menon, and Bickart 2004). That is, respondents may learn to answer questions using organizational referents more accurately after years in an organization because they acquire more information over time.[5]

In sum, questions using organizational referents ask a lot of respondents informationally. Employees are asked to assess the characteristics of large and diverse organizations based on information they may not have. This is concerning because responses may come to rely on unrepresentative information, stories, observed behaviors, networks of colleagues, and social comparisons within public organizations rather than the real features of these organizations. This makes such questions less attractive where information is hard to obtain. The more we know about which respondents in which organizations are most likely to have the necessary information, however, the more we can counteract this disadvantage of organizational referents. In our analysis below, we seek to provide such knowledge, but we find that patterns are difficult to uncover.

To summarize, we arrive at the advantages and disadvantages outlined in table 23.1. Organizational referents have the advantage of matching target to measure when an inquiry is interested in describing organizations. This is the promise of the bottom-up perspective on organizational measurement. The disadvantage is that the perspective underpinning them requires substantial agreement in answers between employees within the same organization. This may not obtain. When agreement does not obtain, as our discussion of the top-down and summary bottom-up perspectives reveals, the conceptual advantage of organizational-referent questions for inquiries interested in organizations diminishes.

Moreover, asking about organizations may decrease SDB but may do so at the cost of placing large informational requirements on responding employees. Conversely, questions about respondents themselves require less external information and no within-organization agreement in responses. But this comes at the cost of greater SDB and of presenting organizational summaries rather than describing organizational features beyond individual respondents' aggregated perspectives.

DATA

We rely on two survey experiments to examine the questions we have raised in the previous section. We first describe the surveys in which these experiments were embedded, then the experiments themselves and how they help us gain strong leverage on question referents.

Surveys

Our experiments were embedded in two surveys of central government public servants. We implemented the first survey in Romania between June 2019 and January 2020. Respondents were randomly assigned to face-to-face or online survey formats and partook in our experiment as part of a longer survey on civil service management practices. In all, we interviewed 3,316 respondents face-to-face (for a response rate of 92 percent) and 2,721 respondents online through Qualtrics (for a 24 percent response rate). The representativeness of our samples and the extent to which it differs according to the survey mode is covered in detail elsewhere in *The Government Analytics Handbook* (chapter 19).

We fielded the second survey in 18 Guatemalan government institutions between October and December 2019. Our experiments were embedded in a longer civil service survey. Respondents were sampled through the sample frame used for the Human Resources National Census, comprising staff lists of 14 central and four decentralized government institutions, and were asked to participate in face-to-face interviews. In all, we interviewed 3,465 respondents (for a 96 percent response rate).[6]

Though both surveys included responses concerning a range of civil service management practices of potential interest for questions surrounding the use of referents, we focus our attention on the analysis of the question-referent experiments. This is, as we explain next, where we get the strongest leverage on question-referent issues.

Experiments

Our experiments all share the same essential strategy. Each survey respondent was randomly assigned to one of two survey flows. In one flow, the respondent was asked a set of questions (see below) that use organizational referents. In the other flow, the respondent was asked a set of questions differing from the first questions only in their use of individual rather than organizational referents.

Assignment to each survey flow was random for reasons of causal identification: random assignment ensures that the respondents who answered questions using individual referents and those who answered otherwise equivalent questions using organizational referents are identical, on average, on all observed and unobserved characteristics. As a result, any difference between average responses in the two flows must be due to the difference between them: whether question referents are individual or organizational. This ensures that we can causally identify the difference referents make to respondents' answers. It is the experimental setup that enables us to say with confidence that referents matter, how much, and for which organizations or groups of people.

The gist of our argument is this: if we ask some respondents a question on, say, salary satisfaction with reference to themselves and other respondents a salary-satisfaction question with reference to their organization, the average response from all respondents in an organization to each question should be the same if the question referent does not matter. The respondents who answered questions with individual referents are a random sample of all respondents and, thus, representative of them. The respondents who answered questions with organizational referents are also a random sample and, thus, representative in their views of their organization.[7] Therefore, any average difference between respondents assigned to different question referents must be due to the question referents.

In both surveys we fielded, we manipulated different sets of questions in this manner. In the survey in Romania, we assigned respondents to individual- or organizational-referent versions of questions surrounding recruitment (two questions), promotions (two questions), turnover (five questions), and dismissals (two questions). All questions were in five-point Likert scales. Additional follow-up questions on the use of various sources of information in recruitment and the questions asked at recruitment interviews were similarly randomized. We include these only in some of our analyses as they are scaled differently than the questions listed above. Questions were assigned to respondents in groups such that respondents either got all questions using individual referents or all questions using organizational referents. For instance, the group of respondents who received organizational-referent questions was asked the question "Please indicate the extent to which you agree or disagree with the following statements: The promotion process in my institution is fair." By contrast, the other group of respondents, who received individual-referent questions, was asked the question "Please indicate the extent to which you agree or disagree with the following statements: The promotion process I have to go through in my institution is fair." Appendix K.1 shows the full lists of questions in both versions.

Similarly, in the survey in Guatemala, we assigned respondents to individual- or organizational-referent versions of questions surrounding promotion confidence (one question), promotion fairness (one question), turnover (three questions), dismissals (two questions), and leadership (nine questions). Appendix K.1 shows the full lists of questions in both versions. Even where themes overlap, questions were formulated somewhat differently in Romania and Guatemala. Consequently, a comparison of results between the two countries should be made with caution.

From a design perspective, the experiments illuminate question-referent effects, but they do share a common drawback: we lack an objective benchmark for the phenomena, behaviors, or attitudes they measure. This means that while we are willing to interpret average higher scores on sensitive questions as diminishing SDB, we are often not strictly able to say whether individual or organizational referents caused the stronger method effect grounded solely in the way the question was posed. This is a weakness shared by most nonlaboratory experiments of this type, but we are still able to examine differences between individual- and organizational-referent questions, which are often informative. With this caveat noted, we proceed to our results.

RESULTS

There is much we can examine within our framework using our data. Within the confines of this chapter, we cannot address every possible question. Instead, we opt to answer four questions directly related to the issues of substantive interest, information availability, and social desirability that we have outlined. Each subsection poses a question, which is immediately answered before detailed results are provided.

Do Organizational-Referent Questions Reliably Reflect Organizational Characteristics?

Organizational-referent questions do not generally reflect organizational characteristics, though they often reflect them better than individual-referent questions do. For this reason, individual-referent questions may be preferred on conceptual grounds in many instances, given that organizational referents—while often resulting in increased agreement—are by no means guaranteed to ensure that questions result in clear ratings of organizational characteristics rather than summaries of individual perspectives.

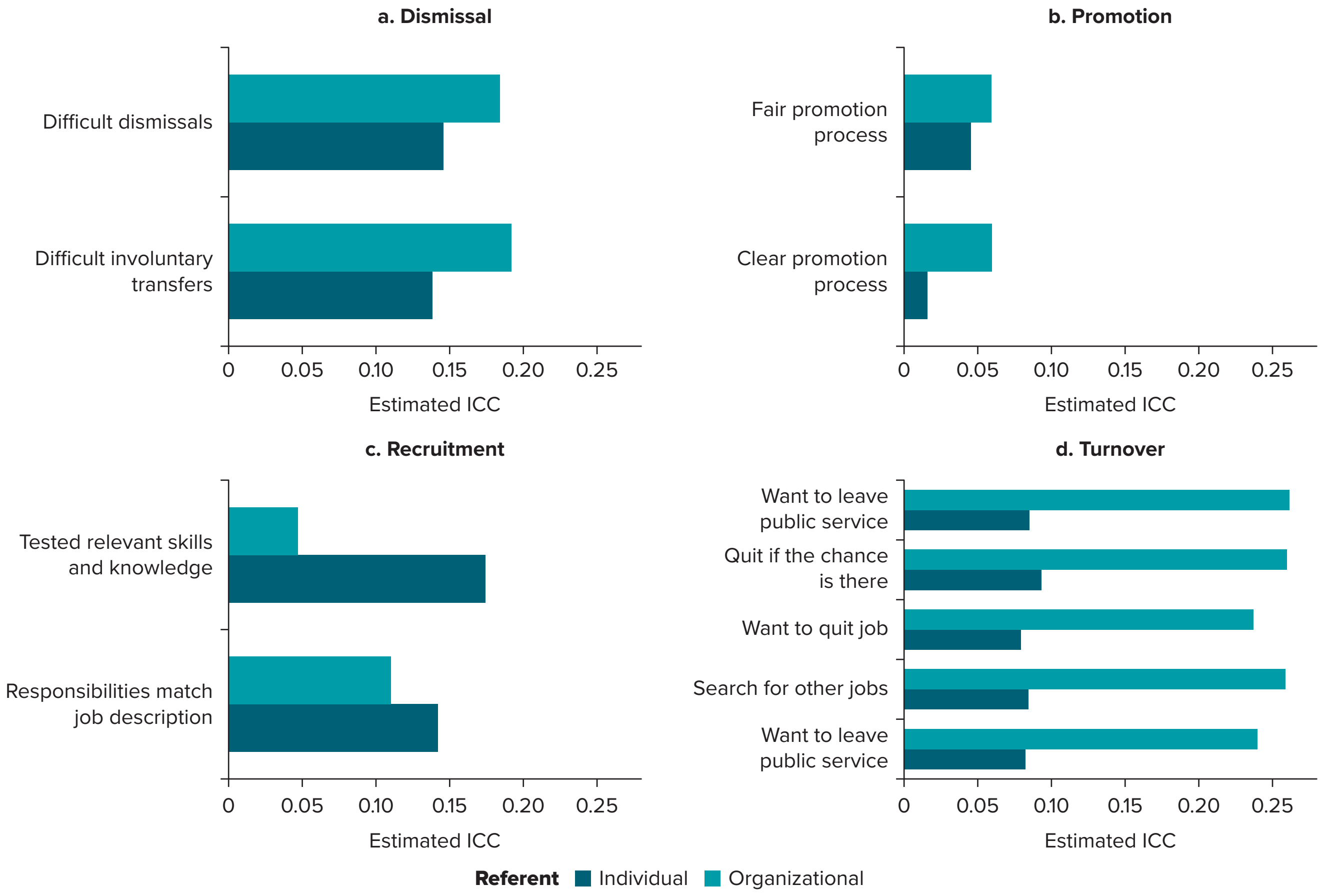
As noted, one central question for the bottom-up perspective on the utility of organizational and individual referents in civil service surveys concerns agreement within organizations. If respondents within an organization tend to agree in their responses to questions about their organization, we can more plausibly claim that their responses evaluate the same organizational phenomenon. If respondents rate the same entity in the world, they should agree in their ratings.

There are many measures of within-group agreement on survey measures. Here we opt for a common and simple measure, intraclass correlation (ICC). ICC is a measure of how much responses to questions rely on respondents' organizational setting. It can be interpreted as the percentage of variation in responses accounted for by organizational level. The higher the ICC, the more responses correlate within organizations—that is, the more respondents within organizations agree on their answers—and the more we can think of measures as reflecting objective organizational characteristics, which are simply observed and reported by respondents.

Our data permit the examination of two questions regarding ICC. First, are responses within organizations correlated to a high enough degree that we can think of the concepts they measure as genuine organization-level constructs? Second, is the correlation affected by the use of organizational or individual referents? If it is, this could indicate that organizational-referent questions can help survey designers elicit answers that characterize organizations from the appealing bottom-up perspective. Other things being equal, responses to questions about organizations *should* correlate more within organizations than responses to individual-referent questions.

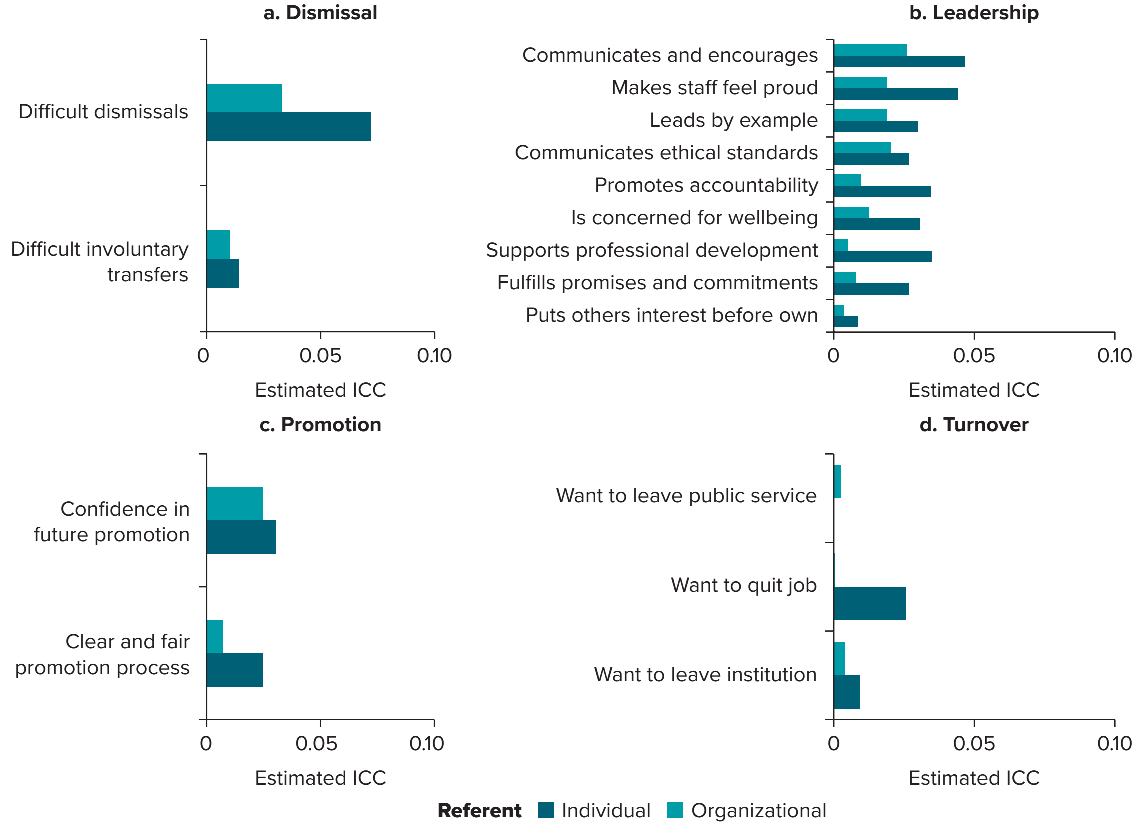
Figures 23.1 and 23.2 examine these questions using the surveys from Romania (figure 23.1) and Guatemala (figure 23.2). Analysis of the Romanian data reveals that there is non-negligible agreement on responses within organizations for several questions but not for others. For some questions, the ICC is low enough that we might ask whether questions using either of the two referents elicit responses that refer to the same underlying phenomenon (rather than reporting two different perspectives).[8]

FIGURE 23.1 Intraclass Correlations for the Romanian Data



Source: Original figure for this publication.
Note: Bars show the calculated organizational ICC for each variable in the survey experiment in Romania, divided by HR area and treatment status. Positive differences between organizational (light blue) and individual (dark blue) referent questions indicate stronger agreement for the former than for the latter. See appendix K.1 for full items and question labels. ICC = intraclass correlation.

FIGURE 23.2 Intraclass Correlations for the Guatemalan Data



Source: Original figure for this publication.
Note: Bars show the calculated organizational ICC for each variable in the survey experiment in Guatemala, divided by HR area and treatment status. Positive differences between organizational (light blue) and individual (dark blue) referent questions indicate stronger agreement for the former than for the latter. The horizontal axis is kept on the same scale as in figure 23.1 for ease of comparison. See appendix K.1 for full items and question labels. ICC = intraclass correlation.

Equally important for our purposes, these data show that organizational-referent questions *do* generally correlate more strongly within organizations than questions with individual referents. The expected agreement effect from organizational referents does emerge for some questions. The ICC for questions using organizational referents is higher for all but two recruitment questions in the Romanian data, though some differences are slight.

Question-referent effects are particularly pronounced for the turnover and recruitment questions. For turnover questions, within-organization agreement climbs by a factor of four. One possible explanation for this is social desirability. If respondents differ in their propensity to provide socially desirable answers more than they differ in their views on turnover intention among their colleagues, we could arrive at the pattern we observe. For now, however, this has to be considered speculative.

Somewhat puzzlingly, referent effects on recruitment items are reversed relative to what we would expect. Respondents to individual-referent questions agree more within organizations than respondents to organizational-referent questions. One possible explanation for this is that the questions using individual referent ask about recruitment processes that may have occurred years ago. This could lead to larger differences within organizations that have changed practices over time. However, our data do not reveal substantial differences in estimated ICCs if we split them along years of service.

Another possible explanation is that respondents' beliefs about public sector recruitment generally lead to the underestimation of differences between organizations, which drives down the ICC for questions using organizational referents, while individual-referent questions capture the diversity in recruitment practices.[9]

This is consistent with the fact that the between-organization variance of organizational-referent questions for recruitment is among the lowest in our data (alongside variables related to career advancement).

Analysis of the Guatemalan data reveals a similar pattern, although with a lower ICC across the board (figure 23.2). This offers two important lessons. First, many of the questions we examine do not appear to be statistically sound measures of bottom-up, organization-level constructs in Guatemala. The lower ICCs are due in part to the larger size of Guatemalan institutions, which leads to more variation within them. But this is precisely the point: respondents in these large organizations may be rating effectively different entities. It seems responses in our Guatemalan data are often better seen as employee perspectives, from the summary bottom-up perspective. Second, organizational referents do sometimes, as expected, help consolidate responses around agreeing ratings of organizational constructs, particularly for leadership and turnover.

What does this mean? From the bottom-up perspective of using survey responses to describe organizational characteristics, these analyses are not generally good news. Instead, they indicate that many organizational aggregates are perhaps better thought of, from the summary bottom-up perspective, as data summaries, particularly in Guatemala. That is, the summary bottom-up perspective appears to have more traction here than the pure bottom-up perspective. As noted, there may be good structural reasons for this. Public organizations are large, segmented, and complex entities in which management practices can vary by team, division, or section—particularly where management and human resources tasks are decentralized to line managers. Expecting consistent organizational characteristics to emerge under these conditions is, perhaps, expecting too much. The use of organizational referents does seem to consolidate a unified description by respondents, but to a limited degree, leaving plenty of disagreement behind.

Conceptually, then, while civil service surveys may benefit from the use of organizational referents, the big prize—the reliable description of organizational phenomena as rated by organizational members above and beyond their individual perspectives—appears elusive in our data. Given this conclusion, the question arises whether the use of organizational or individual referents matters to the data summaries both questions can provide.

Does the Choice of Referent Matter for Responses?

Yes, in most instances, the choice of referent matters for responses, although it matters more for average responses than for relationships between response variables or for the tendency to respond at all. Average responses are sometimes higher and sometimes lower for organizational-referent questions, depending on the question. Similarly, nonresponse is sometimes more common and sometimes less common for organizational-referent questions, depending on the question. There is little systematic evidence that question referents matter to associations between different measures and less evidence still that associations are systematically stronger or weaker.

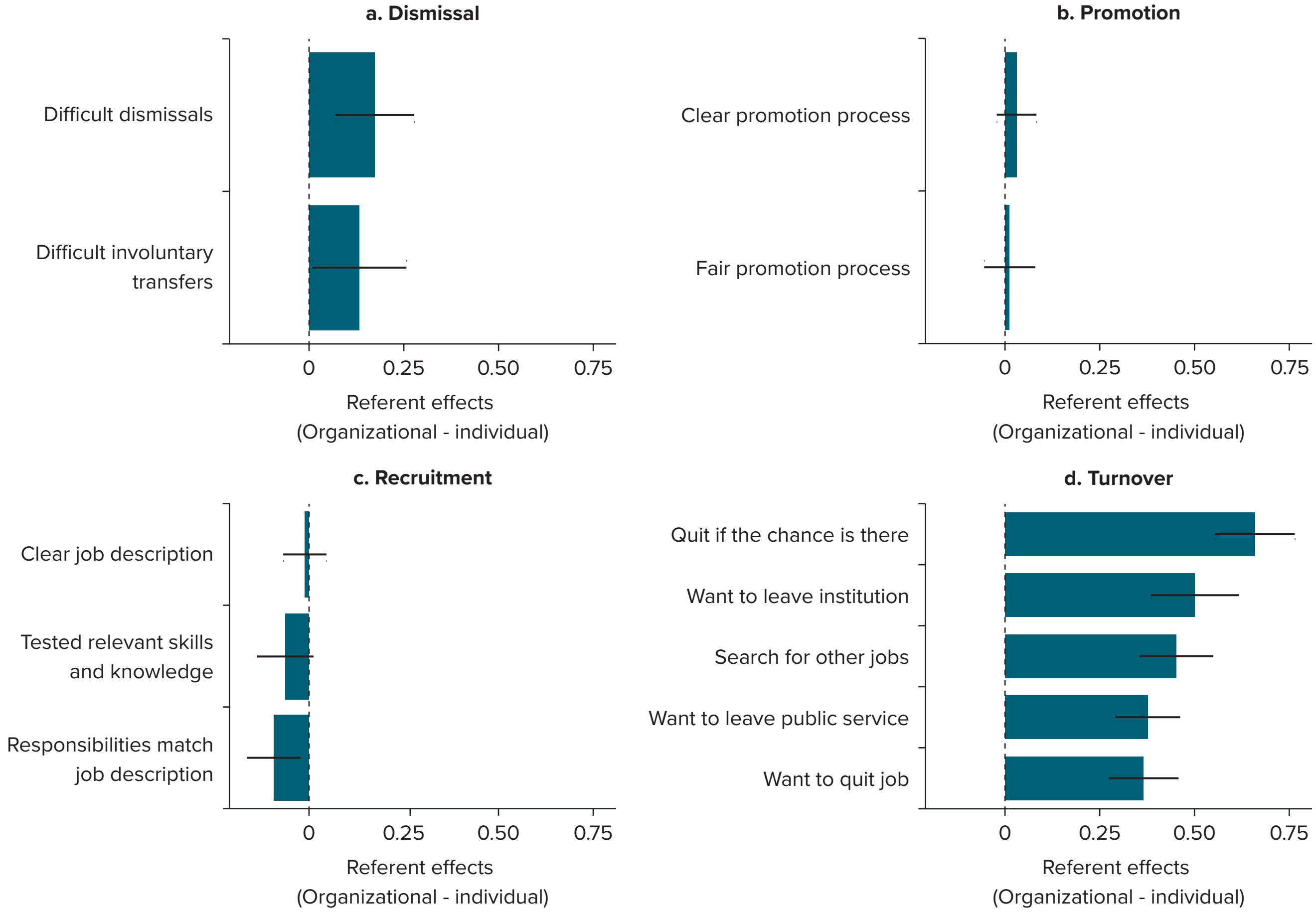
In figure 23.3, we show, using the Romanian data, the differences in average responses to questions on recruitment, turnover, dismissals, and promotion, varying only the use of organizational versus individual referents. As the figure shows, respondents who were asked about themselves rather than their colleagues are, on average,

- More convinced that they are difficult to dismiss or transfer,
- Less convinced that their responsibilities match their job descriptions, and, most markedly,
- Less willing to quit their jobs, organizations, or the public service.

Notably, two recruitment questions and both promotion questions do not show clear evidence that referents matter to responses.

These results provide the minimally expected result that different question referents result in different responses. Moreover, they are our first indication that the use of organizational referents really does make respondents more willing to admit to sensitive attitudes and behaviors, such as turnover intentions, as well as

FIGURE 23.3 Organizational and Individual Referents in the Romanian Data



Source: Original figure for this publication.
Note: Bars show estimated differences between organizational- and individual-referent questions in the survey experiment in Romania with 95 percent confidence intervals based on cluster-robust standard errors. Bars left of zero on the horizontal axis indicate higher scores on the individual-referent version of the question, whereas bars right of zero indicate higher scores on the organizational-referent version. All variables are scaled on the same 1–5 Likert scale. See appendix K.1 for full items and question labels.

slightly less prone to exaggerate their views on dismissals and their job descriptions. We return to this issue in more detail below.

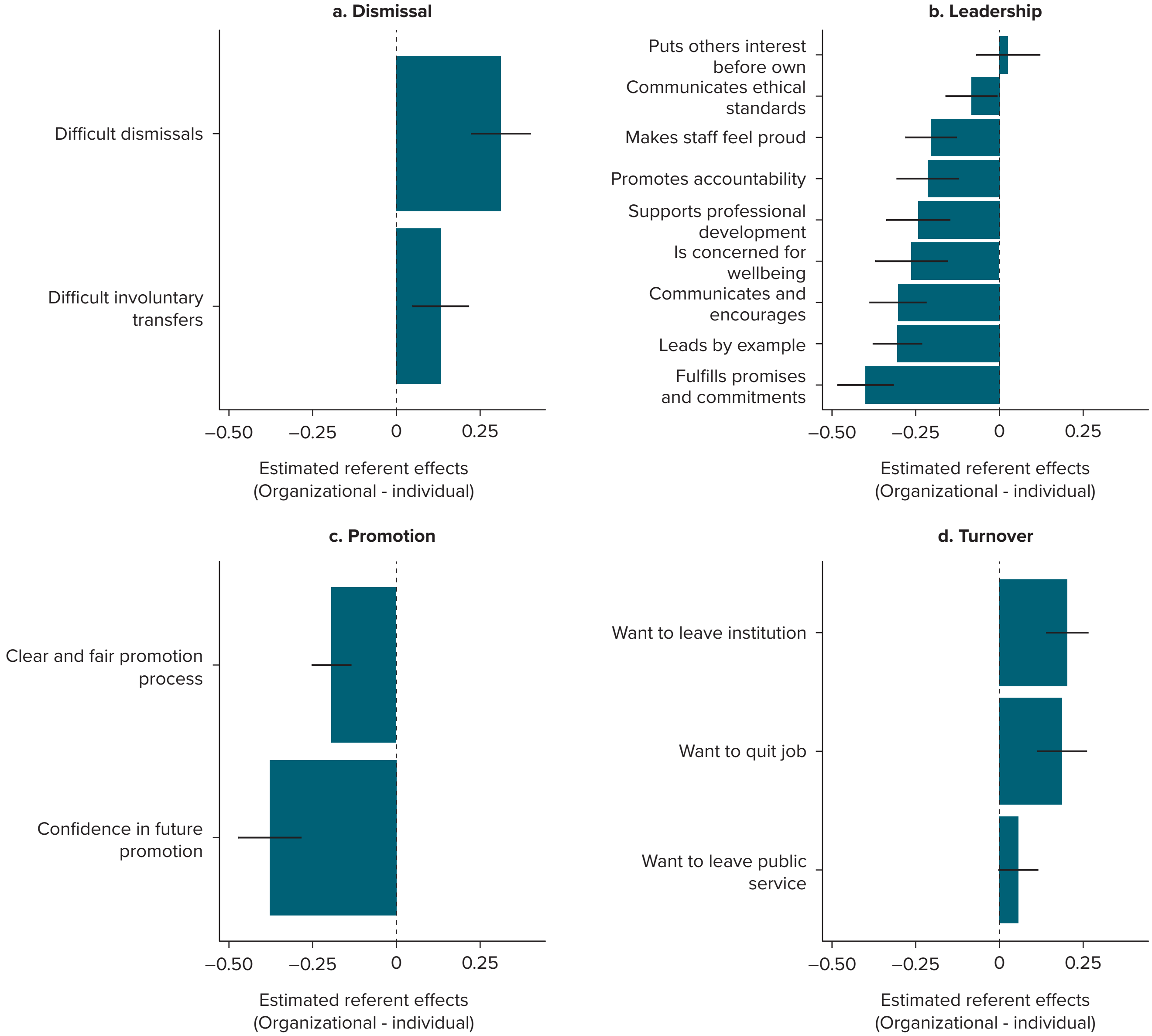
Figure 23.4 shows the results of a similar analysis using the Guatemalan data. In these data, the use of organizational versus individual referents matters more to average responses than in the Romanian data. Individual referents make respondents

- More likely to report that their direct managers are more transformational and ethical in their leadership styles on nearly any measure,
- Less prone to report turnover intentions,
- Less concerned about involuntary dismissals and transfers, and
- More convinced that promotions are within reach and that the process for achieving promotion is fair.

We can conclude at this stage that the choice of referent often matters to average responses—sometimes not a lot, but substantially for some questions. We return to plausible determinants of when referent choice matters below. Qualitatively speaking, however, we can already establish that referents do matter.

The average responses provided to survey questions matter a great deal, not least because they feed the organizational descriptive statistics commonly used in benchmarking organizations (about which,more shortly).

FIGURE 23.4 Organizational and Individual Referents in the Guatemalan Data

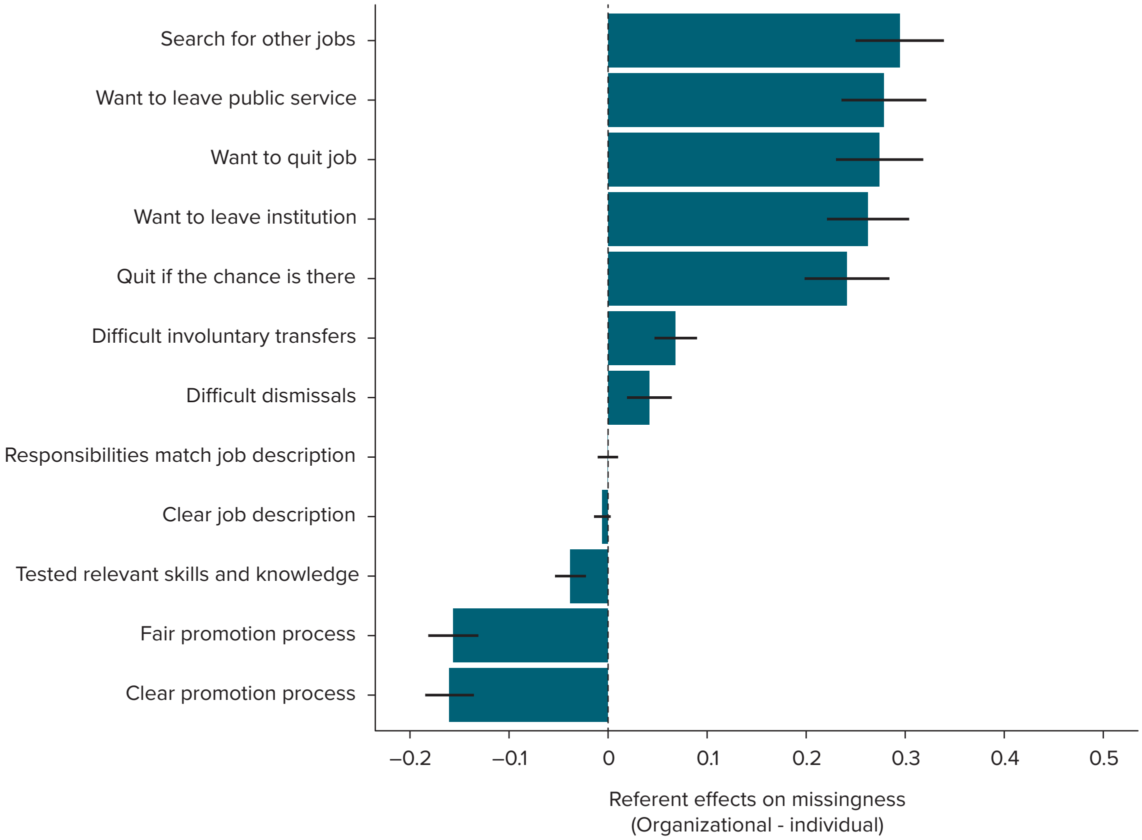


Source: Original figure for this publication.
Note: Bars show estimated differences between organizational- and individual-referent questions in the survey experiment in Guatemala with 95 percent confidence intervals based on cluster-robust standard errors. Bars left of zero on the horizontal axis indicate higher scores on the individual-referent version of the question, whereas bars right of zero indicate higher scores on the organizational-referent version. All variables are scaled on the same 1–5 Likert scale. See appendix K.1 for full items and question labels.

Yet average responses are not the only quantity that question referents may affect. It is possible, for instance, that organizational-referent questions are harder for respondents to understand, prompting item nonresponse—that is, respondents' not responding to individual items (see chapter 22).

Figure 23.5 examines this question using our Romanian data. Using a set of linear probability models with institution fixed effects, we find evidence of substantial nonresponse effects, particularly for questions relating to turnover. For each turnover question, the estimated probability of respondents not responding to individual-referent questions is increased by more than 20 percent relative to otherwise identical organizational-referent questions. This effect is substantial and worth considering. It is also worth noting, however, that less-sensitive questions on dismissal show much smaller effects, and questions on recruitment show no clear evidence of an effect at all. Moreover, individual referents substantially *reduce* nonresponse

FIGURE 23.5 Estimates of Referent Effects on the Likelihood of Item Nonresponse



Source: Original figure for this publication.
Note: Bars show linear probability estimates of the differences between organizational- and individual-referent questions in the survey experiment in Romania with 95 percent confidence intervals based on cluster-robust standard errors. Bars left of zero on the horizontal axis indicate a higher probability of missingness on the individual-referent version of the question, whereas bars right of zero indicate a higher probability of missingness on the organizational-referent version. All variables are scaled on the same 0–1 scale, where 1 indicates "missing." See appendix K.1 for full items and question labels.

relative to organizational referents for questions relating to promotion. One explanation for this finding may be that questions surrounding promotion processes are difficult to answer on behalf of the organization as a whole, leading respondents to nonresponse as a way of indicating they do not know the answer (see chapter 22). We return to the consequences of these findings below.

A final question we can examine is whether there are referent effects not on responses to individual survey variables but on relationships between survey variables. It is possible, for instance, that respondents fall back on their general opinions about the organization when asked for specific information about it, forming their attitudes as they go. This could result in increased statistical relationships between variables because they all tap into the same overarching attitude.

Table 23.3 examines this question using the leadership questions from the Guatemalan data. The table shows differences in statistical association between respondents who answered individual-referent questions and respondents who answered organizational-referent questions. Positive values indicate that organizational-referent questions correlate more strongly than similar individual-referent questions.

Table 23.3 does give some indication that variables covary differently when using organizational- rather than individual-referent questions. The effects we find indicate that the relevant relationships are generally—though not always—stronger when organizational referents are used. The differences vary in size, and not all are substantial. However, qualitative conclusions about the relationships between factors do sometimes hinge on the choice of referent. For example, when using our leadership, recruitment, and promotion variables to

TABLE 23.3 Estimated Differences in Relationships between Leadership Variables for Different Referents, Guatemala (Organizational—Individual)

	Communicates and encourages	Communicates ethical standards	Fulfills promises and commitments	Is concerned for wellbeing	Leads by example	Makes staff feel proud	Promotes accountability	Puts others interest before own	Supports professional development
Communicates and encourages		–0.064* (0.028)	–0.046 (0.033)	–0.027 (0.022)	–0.085*** (0.019)	–0.037 (0.022)	–0.069* (0.031)	0.095* (0.036)	–0.042‡ (0.021)
Communicates ethical standards	–0.092** (0.026)		–0.122** (0.034)	–0.130*** (0.028)	–0.125*** (0.030)	–0.108*** (0.021)	–0.138*** (0.031)	0.059 (0.034)	–0.128*** (0.029)
Fulfills promises and commitments	0.012 (0.040)	–0.027 (0.043)		0.020 (0.041)	–0.055 (0.032)	–0.017 (0.033)	–0.007 (0.044)	0.179*** (0.037)	0.020 (0.026)
Is concerned for wellbeing	–0.033 (0.025)	–0.117** (0.036)	–0.035 (0.026)		–0.086* (0.034)	–0.021 (0.021)	–0.056‡ (0.031)	0.161*** (0.031)	–0.038* (0.017)
Leads by example	–0.037 (0.036)	–0.044 (0.037)	–0.073* (0.034)	–0.032 (0.028)		0.000 (0.025)	–0.046 (0.036)	0.154** (0.042)	–0.028 (0.027)
Makes staff feel proud	–0.038 (0.031)	–0.078* (0.028)	–0.086* (0.033)	–0.022 (0.026)	–0.056* (0.022)		–0.075** (0.022)	0.126*** (0.033)	–0.035 (0.032)
Promotes accountability	–0.013 (0.028)	–0.032 (0.027)	–0.009 (0.038)	–0.003 (0.025)	–0.035 (0.030)	–0.007 (0.023)		0.144*** (0.036)	–0.003 (0.028)
Puts others interest before own	0.137** (0.047)	0.117* (0.048)	0.208** (0.056)	0.191*** (0.044)	0.194** (0.056)	0.162*** (0.038)	0.176** (0.049)		0.156** (0.042)
Supports professional development	–0.019 (0.030)	–0.089* (0.038)	–0.009 (0.021)	–0.004 (0.027)	–0.049 (0.028)	–0.006 (0.025)	–0.032 (0.031)	0.150*** (0.037)	

Source: Original table for this publication.

Note: Results from ordinary least squares models with institution fixed effects and standard errors clustered by institution. Each cell in the table is the estimated interaction between our experimental treatment and the question in the cell's row in a model predicting the question in the cell's column. All variables are scaled on the same 1–5 Likert scale. See appendix K.1 for full items and question labels. *p*-values: ‡ *p* < 0.100, * *p* < 0.050, ** *p* < 0.010, *** *p* < 0.001.

predict turnover variables in the Guatemalan data, 13 percent of estimated associations have different signs depending on the referent used.[10]

In sum, the choice of question referent matters. We find often small but sometimes substantial referent effects on the average responses to most questions we examine. Given our experimental setup, these differences must be due to the way we pose our questions. Hence, average differences are, in most cases, plausibly interpreted as being due to the question referent. We also find substantial referent effects on nonresponse patterns, but without a single direction of the effect. Whether referents make people respond more or less often appears to hinge on the question, its sensitivity, and how difficult it is to respond to. Finally, we find referent effects on relationships between some variables, but not in any clear direction.

Can Organizational Referents Limit Social-Desirability Bias?

Yes, organizational referents limit SDB, but mostly for strongly sensitive items. We find evidence that more-sensitive questions show larger differences between individual- and organizational-referent questions in our experiment. This likely indicates that organizational referents can help limit SDB in civil service surveys. We find indications that this effect may be particularly pronounced for very sensitive questions.

As noted above, combatting SDB is a sensible reason for the use of organizational referents. To examine this question in more depth, we coded our individual questions in the Romania and Guatemala experiments for their sensitivity (see chapter 22 for details on the procedure). For the sake of statistical power in the analyses that follow, we now include the follow-up questions on the use of various sources of information in recruitment and the questions asked at recruitment interviews from the Romania questionnaire we have excluded from our analysis up to this point.

We regress this measure on the absolute difference between average responses to questions using individual and organizational referents (the referent effect), which we standardize to make our different response scales comparable. We run regressions with two sets of observations: one (model 1 in table 23.4) in which each observation is a question—from either survey—with its associated referent effect and sensitivity score, and one (model 2) in which each observation is an organizational aggregate for a question.

If organizational referents guard against SDB, we would expect a positive association between the sensitivity of questions and referent effect sizes because a reduction in SDB for sensitive questions would increase the difference between responses using organizational and individual referents. In our analysis, we find evidence for this assertion. In model 1, the expected positive association is significant only at the 10 percent level due to a low number of observations. In the more well-powered model 2, the expected association is highly significant. As expected, sensitive questions see larger question-referent effects, indicating that organizational referents may diminish socially desirable responding. It is worth noting, however, that this analysis

TABLE 23.4 Standardized Question-Referent Effects, by Sensitivity

	Model 1 (Questions as observations)	Model 2 (Institution aggregates as observations)
Sensitivity	0.107‡ (0.056)	0.079*** (0.016)
(Intercept)	0.201*** (0.043)	0.328*** (0.013)
N	40	2,664
R-squared adjusted	0.065	0.008

Source: Original table for this publication.
Note: Results from ordinary least squares models. Each observation in model 1 is a question; each observation in model 2 is a question aggregate from an institution. The dependent variable is the absolute referent effect—the absolute difference in average responses between individual- and organizational-referent questions—standardized to account for the different scales of the included variables. See appendix K.2 for model results using other measures.
‡ $p < 0.100$; * $p < 0.050$; ** $p < 0.010$; *** $p < 0.001$.

cannot leverage randomization to the same extent that our previous analyses do and, consequently, that it cannot be conclusively established whether the associations we document are due to sensitivity.

However, this analysis masks an additional finding: some very sensitive questions do appear to display larger differences than less sensitive questions. To see this, consider the violin plot in figure 23.6, showing the distribution of standardized referent effects for nonsensitive and sensitive questions (the thicker the "violin" at a certain height, the more questions have referent effects at the corresponding value on the second axis).

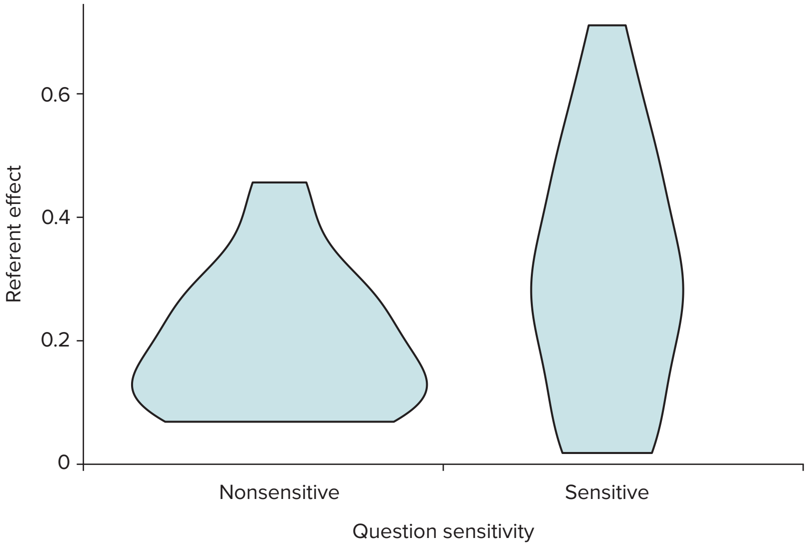
As the figure shows, the top of the referent effects distribution is far above the sensitivity effect observed in the table above, indicating that some sensitive questions have larger-than-predicted referent effects. On a qualitative inspection, these turn out to be very sensitive questions—particularly concerning turnover. This is a valuable conclusion. When examining sensitive issues—particularly highly sensitive issues such as corruption, politicization, or absenteeism—organizational referents appear to be able to combat SDB. For nonsensitive issues, the difference organizational referents make is more limited. The implication is that if the use of individual referents is preferred on other grounds, shifting to organizational referents may be justified on the grounds of SDB if questions are highly sensitive.

Does Information Availability Matter?

Yes, information availability matters, but not in all the ways one might think. We find evidence that referent effects are smaller for very common attitudes and behaviors. However, we find no statistically clear evidence that respondents who have served longer in their organizations are less prone to referent effects.

As discussed, the availability of information may determine how much question referents matter if they are partly rooted in information availability. In these instances, we would expect smaller referent effects for questions about attitudes or behaviors that are either very common or very uncommon in respondents' surroundings. Respondents are less (more) likely to report rare (common) behaviors about themselves by definition, but they are also less (more) likely to report rare (common) behaviors about their organizations because they encounter them rarely (commonly). By contrast, attitudes and behaviors that some hold but others do not can give rise to substantial referent effects, particularly if they are unevenly distributed within organizations.

FIGURE 23.6 Distributions of Referent Effects for Sensitive and Nonsensitive Questions



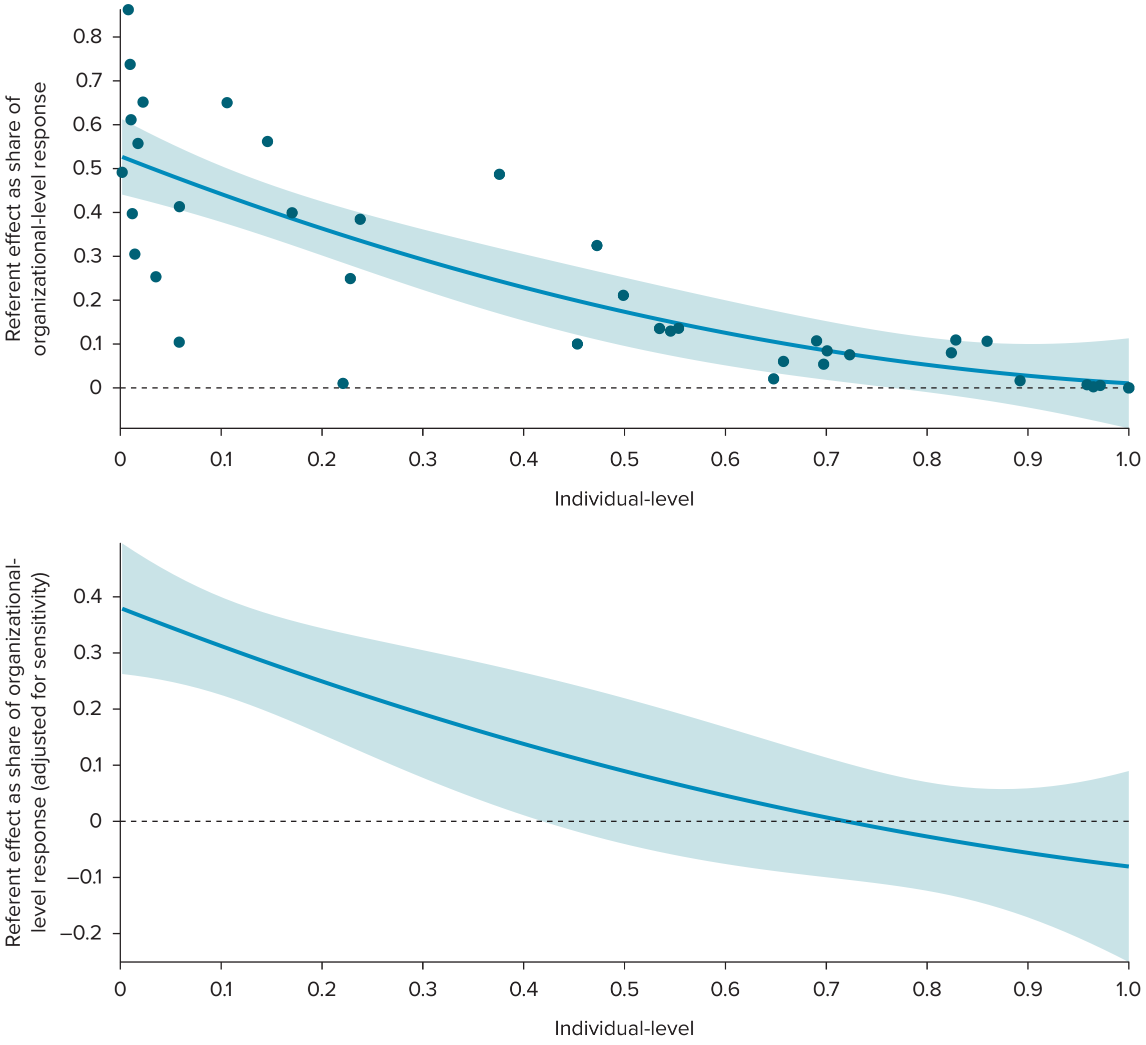
Source: Original figure for this publication.
Note: The figure shows the distributions of referent effects split by question sensitivity. The width of the "violins" indicates the number of referent effects at or around the size indicated on the vertical axis. Thus, sensitive questions have a smaller range of referent effects, with the largest and smallest referent effects larger and smaller than for nonsensitive questions in our sample.

We examine this prediction by looking at patterns in referent effects. If very common or very rare attitudes and behaviors give rise to smaller referent effects, we should expect the referent effects, relative to reported commonality on organizational- (individual-) referent measures, to depend on how commonly the attitude or behavior in question is reported by respondents who are asked individual- (organizational-) referent questions. Specifically, we would expect an inverted-U relationship, in which referent effects are smaller for very rare or very common attitudes or behaviors.

Figure 23.7 speaks to this prediction. The figure plots the organizational proportion of affirmative responses to each question in our Romanian experiment (using individual referents) against the absolute difference between individual- and organizational-referent questions as a proportion of responses to the organizational-referent question.[11] Affirmative responses are interpreted as responses scoring on the upper two quintiles of the possible answers for scale questions (for example, "Strongly agree" and "Agree" on a Likert scale) and affirmative answers to follow-up questions, where respondents could indicate "yes" or "no."

As the top panel in the figure shows, differences are not generally smaller for questions where scores are generally very low or very high, behaviors or practices are very rare or very common, and information is

FIGURE 23.7 Response Score and Question-Referent Effects in the Romanian Data



Source: Original figure for this publication.
Note: The figure shows the average referent effect—here defined as the absolute difference between organizational- and individual-referent versions of each question in the experiment as a proportion of the score of the organizational-referent version—as a polynomial regression function of responses to the individual-referent version of the questions. The top panel shows the raw association, with individual questions plotted as points. The lower panel shows the association adjusted for question sensitivity.

more readily available. Instead, it shows referent effects declining as a function of commonality. One interpretation of this aligns, albeit asymmetrically, with information availability: it is easier for respondents to provide information about their organizations if they experience the relevant attitudes or behaviors around them, and this renders relative referent effects smaller for questions about common attitudes and behaviors than for questions where attitudes and behaviors are less common.

One obvious objection to this finding is that sensitive questions often result in indications that behaviors are rare, either because the behaviors in question *are* rare or because of SDB. As a result, the association depicted in the top panel of figure 23.7 could reflect sensitivity rather than information availability. To examine this issue, the lower panel in figure 23.7 shows the same association adjusted for question sensitivity. Indeed, the identified referent effects are weaker, but the pattern holds: referent effects appear to be smaller for questions targeting attitudes and behaviors that are common. Of course, one cannot definitively conclude from this simple analysis that greater information availability to respondents either will or will not result in smaller question referent effects. But the analysis does suggest that the use of organization-level referents may require caution when targeting rare behaviors or attitudes.

In our data, at least, we are not able to further pin down plausible determinants of information availability that give rise to the predicted changes in referent effects. To exemplify, further analysis of our two data sets (not shown) shows that organization size does not appear to matter for respondents' reactions to organizational versus individual referents, although one might expect smaller organizations to be easier to rate for respondents who use the information available to them, all else being equal. Moreover, as shown in table 23.5, the effect of using organizational referents in our Romanian sample does not generally vary with years of service. The exception is recruitment, where a negative referent effect grows with years of service (contrary to the idea that organizational experience would facilitate learning and diminish information-based referent effects). This effect could reflect changing recruitment practices over time, which would be consistent with the finding not being recovered when limiting the sample to relatively recently recruited public servants (model 6 in table 23.5).

However, some organizational characteristics do matter for referent effect sizes. If split by organization, the average referent effect size in the Romanian data is 0.15 (standardized across all experimental questions), but effect sizes range widely from one organization to another, from 0.02 to 0.31, the latter being a moderately sized effect, whereas the former is negligible.

The conclusion, then, is that if information availability matters in our data, we are not able to get very far in pinpointing its determinants. We can offer two suggestive conclusions, however. First, referent effects appear smaller for questions targeting attitudes that are very common. Second, arguing when information is available is no simple matter and is not a function of simple structural characteristics, such as organization size, or respondent characteristics, such as years of service.

TABLE 23.5 Question-Referent Effects, by Years of Service, Romania

	Model 3 (Dismissals)	Model 4 (Recruitment)	Model 5 (Turnover)	Model 6 (Recruitment, <5 years)
Organizational level	0.148 (0.091)	0.005 (0.041)	0.453*** (0.062)	−0.009 (0.073)
Years of service	−0.000 (0.004)	0.001 (0.002)	−0.003 (0.002)	−0.002 (0.023)
Organizational level × Years of service	−0.001 (0.005)	−0.005* (0.002)	0.002 (0.004)	−0.011 (0.021)
N	3,016	3,298	2,898	656
R-squared adjusted	0.137	0.088	0.216	0.212

Source: Original table for this publication.
Note: Results from ordinary least squares models with cluster-robust standard errors by institution. Each observation is a respondent in the Romania data set. The dependent variable is the indexes for our experimental measures of dismissals (model 3), recruitment (models 4 and 6), and turnover (model 5). All are kept on the same 1–5 scale as their items. Years of service is a single-item measure of how long respondents have served in public administration (measured in years). See appendix K.2 for model results using other measures.
‡ $p < 0.100$; * $p < 0.050$; ** $p < 0.010$; *** $p < 0.001$.

DISCUSSION AND CONCLUSION

Where do our experiments leave us? What do we learn from them? While they do give valuable insights on the effects of individual versus organizational referents in civil service surveys, they also raise new and interesting questions to which we do not yet have the answers.

The primary lesson is that the choice of referent matters. Using organizational referents often leads to more agreement between respondents in the same organization—a finding consistent with Glick (1985) and Klein et al. (2001). Yet in our measures, this agreement is often too low for responses to reliably track bottom-up, organization-level features above and beyond the perspective of respondents. The use of organizational referents can, however, provide summaries of respondents' perspectives and experiences, which are also, per the summary bottom-up perspective, valuable organizational metrics.

Moreover, average responses to survey questions often change when question referents change. For some questions, respondents report stronger agreement when asked about their organizations than when asked about themselves. For other questions, the pattern is reversed. In general, these effects are of modest size, but for some questions, they are substantial—and predicting for which questions referents will matter the most is not straightforward. Similarly, we find substantial question-referent effects on nonresponse but without uncovering one clear direction. For some questions, organizational referents substantially reduce nonresponse; for others, they exacerbate it. We also find some evidence that relationships between variables are affected by referents. But not all associations between variables are clearly impacted by the choice of referent, and we cannot propose a general direction of effects when they are.

We have examined the determinants of referent effect sizes: when does the choice between individual and organizational referents matter the most? From our analyses, we can draw only a few lessons about the question of referent effect size. First, referent effects seem to be larger for (highly) sensitive questions. This is consistent with organizational referents' ability to mitigate SDB for sensitive questions. Second, referent effects seem to be larger for attitudes, behaviors, and practices that are not common among respondents. This is consistent with the view that organization-level questions can pose higher informational demands than respondents can meet. It is also notable that question-referent effects are stronger in some organizations than others, but it is not clear which organizational characteristics drive these differences. And question-referent effects are not negatively associated with experience in the organization, suggesting that learning may have limited consequences for their size.

What does all this mean for civil service survey designers? It means they must be aware of the referents used in the questions they include in their surveys. Using organizational referents, as is common practice today, is not uniformly preferable on conceptual grounds—since responses often track but do not directly reflect organizational characteristics over and above respondents' perspectives. However, using individual referents is not uniformly preferable either. Particularly on measurement grounds, there is evidence that individual referents may suffer from SDB both for sensitive questions and for questions for which respondents wish to positively manage impressions.

Beyond awareness, we can make a few recommendations for more specific situations. First, a survey designer including very sensitive questions in a survey should consider posing these questions using organizational referents to combat SDB. It is important to recognize the limitations of this advice, however. Our analysis shows that more sensitive behavior is reported when using organizational referents. Yet this does not mean organizational referents provide an accurate estimate of how frequently the sensitive behavior or attitude occurs.

Moreover, using organizational referents comes at a heavy conceptual cost if the survey is interested in anything more than organizational aggregates. Predicting individual behaviors and attitudes with individual responses to sensitive organizational-referent questions implies a shift in what is studied (Klein and Kozlowski 2000). There is a difference between saying that a respondent's manager is abusive and that managers in the organization generally are abusive. Predicting sensitive organizational-referent questions with individual attitudes and experiences is often problematic because it tends to operationally conflate beliefs about the organizational collective with individual attitudes and behaviors. If survey designers want to know why individual public servants behave and think as they do, the conceptual cost of organizational referents may be higher than the measurement gain, even for sensitive topics.

Second, survey designers should consider how the information needed to answer a question will be acquired by respondents. If using an organizational referent, can individual respondents reasonably be expected to know the answer? Individual referents are preferable if introspection provides more or more-reliable information than beliefs and available information about the organization. Our findings indicate few systematic patterns in which questions are most affected by this or in which respondents are most prone to provide the needed information accurately, rather than information infused with impressions, rumors, and beliefs. However, this does not mean that information availability can be glossed over by survey designers. Instead, it highlights the need for more measurement studies specifically targeting information availability and its determinants.

Third, our results may help survey designers think about utilizing other levels of measurement than individual or organizational. Of course, this implication is somewhat speculative, and more data are needed. Consider the conceptual issue with an organizational-referent question that elicits low levels of intraorganizational agreement in response. This means that respondents perceive their organization differently even though they all work within it. As noted, the usual interpretation of this occurrence is that organizational practices differ, that implementation of policies and procedures is uneven, and that management and leadership matter to how organizational practices are felt by public servants. There is nothing intrinsically wrong with this interpretation—but it is uncertain. After all, respondents were asked about their organization, not their section, team, manager, or other lower-level entities. It is not clear from our responses which level respondents draw on the most for information. This is an important weakness of organizational-referent questions in such a situation.

The interpretation gives rise to a question we cannot examine in detail using our data. Would it be a better strategy to use team referents or section referents rather than organizational or individual ones? Is it possible that using team referents would combat socially desirable responding without posing too high of informational demands on respondents? Our results cannot speak directly to this question. They do suggest that the answer is likely contingent on the type of question. Teams are often psychologically closer to people than whole organizations (Riketta and Dick 2005), which might mean that for some questions, team referents will do little to combat SDB. Similarly, some information can be difficult or impossible to access even within teams. Yet is likely to be more easily accessible within teams than for the entire organization. As such, team referent measures may be preferred to organizational-referent questions on measurement grounds if questions are not too sensitive. On the other hand, civil service survey designers may be less interested in reporting team aggregates to decision-makers or other audiences. And aggregating team aggregates to the organizational level is not likely to resolve the issues we discuss in this chapter.

Let us end with a few open questions for which both research and practice would benefit from systematic answers. We know much, both in conceptual and measurement terms, about multilevel theory, measurement, and modeling (Humphrey and LeBreton 2019; Klein and Kozlowski 2000). However, the literature on referent choice is limited, seemingly on the assumption that matching to the level of stated claims is all there is to it. This is sensible enough if one requires organizational measures to reflect organizational characteristics over and above respondents' perspectives in order to be useful. Yet such a perspective is overly limiting, not least for the practice of civil service survey design. For many variables, including management practices, perspectives on leadership, human resources functions, and more, data summaries of employee views and perspectives—interpretable from what we have referred to as the summary bottom-up perspective—can be valuable forms of decision support.

If we accept that organizational—or other higher-level—measurement referents can be useful even if respondents do not strongly align in response to them, our analyses point to a series of underexamined questions. First, which questions are particularly exposed to referent effects? We have found very sensitive questions to be affected, but much more knowledge is needed to reliably provide the type of advice survey designers want. Second, we have scarcely any evidence on whether the choice of referent affects different survey respondents differently. We have not found any such effects in a few exploratory analyses, but this does not mean they do not exist. Third, it appears in our data that organizations affect the size of referent effects. We note that organization size does not appear to matter systematically, but we can see in our data that *something* about organizations does. Yet again, much more knowledge is needed on this issue.

The fact that our findings are not straightforward should highlight for both interested academics and survey designers that the choice of levels of measurement is a complicated issue, and, as we have shown, it is a choice that matters more than current practice seems to be aware.

NOTES

1. An alternative design could randomly assign respondents a question order, with one group being asked individual-referent questions before organizational-referent questions and another group being asked organizational-referent questions before individual-referent questions. This would permit estimation of the average anchoring effect. However, as there is likely to be substantial heterogeneity in this effect, adjusting for the effect can become challenging. For this reason, we opt to ask each respondent only one set of questions.
2. For attitudinal variables, the equivalent of this perspective is that survey aggregates capture shared attitudes in the organization (Chan 1998).
3. This is true, in part, because impression management—wanting to control how one is viewed normatively—concerns both others (impression management proper) and oneself (self-deception) (for example, Millham and Kellogg 1980; Paulhus 1986).
4. The findings of Baltes, Zhdanova, and Parker (2009) suggest that organizational aggregates may depend on the unknown mixture of respondents using "upward" and "downward" comparisons to arrive at their answers. (Their findings also suggest that downward comparison is more common in their sample, but they are unable to assess the specific mixture.)
5. Respondents who have served longer in organizations have been shown in previous studies to be less prone to using heuristics in their decision-making because they can substitute their experience (cf. Pedersen, Stritch, and Thuesen 2018). Translated into the survey-response setting, more experienced personnel may not need to rely on stories and other heuristic devices when assessing their organizations.
6. Some respondents were interviewed even though they were not included on the original staff lists, meaning this number is somewhat inflated relative to those staff lists.
7. Note the assumption behind this null hypothesis is that respondents aggregate information in a way that approximates averaging when responding with reference to their organization. If this assumption does not hold, it poses an additional problem for organizational-referent questions because the aggregation used by respondents is then both unknown and does not approximate common-sense (though not the only sensible) aggregation procedures. Theoretically, this simply adds complexity to the information-processing discussion already noted.
8. This is because the organizational construct assessment of the ICC treats it as a measure of reliability. One way to think of this is to consider each respondent a rater of his or her organization. From this perspective, if at most 15 percent of variance is accounted for by organizations, for an ICC of 0.15, and at least 85 percent is accounted for by the raters, this does not indicate a reliable assessment of organizational characteristics. Raters affect responses too much.
9. We are grateful to an external reviewer for pointing us to this possibility and regret we have no better options available for examining it.
10. This figure includes only associations where effects in at least one direction are statistically significant at the 5 percent level. In none of the included cases are effects in both directions both statistically different from zero.
11. We thank a reviewer for pointing us in this direction. We originally considered simply presenting the absolute differences between answers to questions using different referents, but this created downward trends on the extremes of figure 23.7, consistent both with the prediction and a methodological artifact related only to question scaling.

REFERENCES

Baltes, Boris B., Ludmila S. Zhdanova, and Christopher P. Parker. 2009. "Psychological Climate: A Comparison of Organizational and Individual Level Referents." *Human Relations* 62 (5): 669–700. https://doi.org/10.1177/0018726709103454.

Bardasi, Elena, Kathleen Beegle, Andrew Dillon, and Pieter Serneels. 2011. "Do Labor Statistics Depend on How and to Whom the Questions Are Asked? Results from a Survey Experiment in Tanzania." *The World Bank Economic Review* 25 (3): 418–47. https://doi.org/10.1093/wber/lhr022.

Bezes, Philippe, and Gilles Jeannot. 2018. "Autonomy and Managerial Reforms in Europe: Let or Make Public Managers Manage?" *Public Administration* 96 (1): 3–22. https://doi.org/10.1111/padm.12361.

Blair, Johnny, Geeta Menon, and Barbara Bickart. 2004. "Measurement Effects in Self vs. Proxy Response to Survey Questions: An Information-Processing Perspective." In *Measurement Errors in Surveys*, edited by Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 145–66. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118150382.ch9.

Chan, David. 1998. "Functional Relations among Constructs in the Same Content Domain at Different Levels of Analysis: A Typology of Composition Models." *Journal of Applied Psychology* 83 (2): 234–46. https://doi.org/10.1037/0021-9010.83.2.234.

Christensen, Tom, and Per Lægreid. 2007. "The Whole-of-Government Approach to Public Sector Reform." *Public Administration Review* 67 (6): 1059–66. https://doi.org/10.1111/j.1540-6210.2007.00797.x.

Dunleavy, Patrick, Helen Margetts, Simon Bastow, and Jane Tinkler. 2006. "New Public Management Is Dead—Long Live Digital-Era Governance." *Journal of Public Administration Research and Theory* 16 (3): 467–94. https://doi.org/10.1093/jopart/mui057.

Edwards, W. Sherman, and David Cantor. 2004. "Toward a Response Model in Establishment Surveys." In *Measurement Errors in Surveys*, edited by Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 211–33. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118150382.ch12.

Eggers, William D. 2007. *Government 2.0: Using Technology to Improve Education, Cut Red Tape, Reduce Gridlock, and Enhance Democracy*. Lanham, MD: Rowman & Littlefield.

Fisher, Robert J. 1993. "Social Desirability Bias and the Validity of Indirect Questioning." *Journal of Consumer Research* 20 (2): 303–15. https://doi.org/10.1086/209351.

Freling, Traci H., Zhiyong Yang, Ritesh Saini, Omar S. Itani, and Ryan Rashad Abualsamh. 2020. "When Poignant Stories Outweigh Cold Hard Facts: A Meta-Analysis of the Anecdotal Bias." *Organizational Behavior and Human Decision Processes* 160: 51–67. https://doi.org/10.1016/j.obhdp.2020.01.006.

Gingerich, Daniel W. 2013. "Governance Indicators and the Level of Analysis Problem: Empirical Findings from South America." *British Journal of Political Science* 43 (3): 505–40. https://doi.org/10.1017/S0007123412000403.

Glick, William H. 1985. "Conceptualizing and Measuring Organizational and Psychological Climate: Pitfalls in Multilevel Research." *Academy of Management Review* 10 (3): 601–16. https://doi.org/10.2307/258140.

Graaf, Gjalt de, Leo Huberts, and Tebbine Strüwer. 2018. "Integrity Violations and Corruption in Western Public Governance: Empirical Evidence and Reflection from the Netherlands." *Public Integrity* 20 (2): 131–49. https://doi.org/10.1080/10999922.2017.1350796.

Guenther, Corey L., and Mark D. Alicke. 2010. "Deconstructing the Better-Than-Average Effect." *Journal of Personality and Social Psychology* 99 (5): 755–70. https://doi.org/10.1037/a0020959.

Hoch, Stephen J. 1987. "Perceived Consensus and Predictive Accuracy: The Pros and Cons of Projection." *Journal of Personality and Social Psychology* 53 (2): 221–34. https://doi.org/10.1037/0022-3514.53.2.221.

Homburg, Christian, Martin Klarmann, Martin Reimann, and Oliver Schilke. 2012. "What Drives Key Informant Accuracy?" *Journal of Marketing Research* 49 (4): 594–608. https://doi.org/10.1509/jmr.09.0174.

Humphrey, Stephen E., and James M. LeBreton, eds. 2019. *The Handbook of Multilevel Theory, Measurement, and Analysis*. Washington, DC: American Psychological Association.

Klein, Katherine J., Amy Buhl Conn, D. Brent Smith, and Joann Speer Sorra. 2001. "Is Everyone in Agreement? An Exploration of Within-Group Agreement in Employee Perceptions of the Work Environment." *Journal of Applied Psychology* 86 (1): 3–16. https://doi.org/10.1037/0021-9010.86.1.3.

Klein, Katherine J., Fred Dansereau, and Rosalie J. Hall. 1994. "Levels Issues in Theory Development, Data Collection, and Analysis." *Academy of Management Review* 19 (2): 195–229. https://doi.org/10.5465/amr.1994.9410210745.

Klein, Katherine J., and Steve W. J. Kozlowski, eds. 2000. *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*. San Francisco: Jossey-Bass.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity." *Public Opinion Quarterly* 72 (5): 847–65. https://doi.org/10.1093/poq/nfn063.

Meyer-Sahling, Jan-Hinrik. 2011. "The Durability of EU Civil Service Policy in Central and Eastern Europe after Accession." *Governance: An International Journal of Policy, Administration, and Institutions* 24 (2): 231–60. https://doi.org/10.1111/j.1468-0491.2011.01523.x.

Meyer-Sahling, Jan-Hinrik, and Kim Sass Mikkelsen. 2016. "Civil Service Laws, Merit, Politicization, and Corruption: The Perspective of Public Officials from Five East European Countries." *Public Administration* 94 (4): 1105–23. https://doi.org/10.1111/padm.12276.

Meyer-Sahling, Jan-Hinrik, and Kim Sass Mikkelsen. 2020. "Codes of Ethics, Disciplinary Codes, and the Effectiveness of Anti-Corruption Frameworks: Evidence from a Survey of Civil Servants in Poland." *Review of Public Personnel Administration* 42 (1): 142–64. https://doi.org/10.1177/0734371X20949420.

Millham, Jim, and Richard W. Kellogg. 1980. "Need for Social Approval: Impression Management or Self-Deception?" *Journal of Research in Personality* 14 (4): 445–57. https://doi.org/10.1016/0092-6566(80)90003-3.

Nishii, Lisa H., David P. Lepak, and Benjamin Schneider. 2008. "Employee Attributions of the 'Why' of HR Practices: Their Effects on Employee Attitudes and Behaviors, and Customer Satisfaction." *Personnel Psychology* 61 (3): 503–45. https://doi.org/10.1111/j.1744-6570.2008.00121.x.

OPM (Office of Personnel Management). 2018. *Governmentwide Management Report: Results from the 2018 Federal Employee Viewpoint Survey*. Washington, DC: US Office of Personnel Management, US Government. https://www.opm.gov/fevs/reports/governmentwide-reports/governmentwide-reports/governmentwide-management-report/2018/2018-governmentwide-management-report.pdf.

Parker, Christopher P., Boris B. Baltes, Scott A. Young, Joseph W. Huff, Robert A. Altmann, Heather A. Lacost, and Joanne E. Roberts. 2003. "Relationships between Psychological Climate Perceptions and Work Outcomes: A Meta-Analytic Review." *Journal of Organizational Behavior* 24 (4): 389–416. https://doi.org/10.1002/job.198.

Paulhus, Delroy L. 1986. "Self-Deception and Impression Management in Test Responses." In *Personality Assessment via Questionnaires*, edited by Alois Angleitner and Jerry S. Wiggins, 143–65. Berlin: Springer-Verlag. https://doi.org/10.1007/978-3-642-70751-3_8.

Pedersen, Mogens Jin, Justin M. Stritch, and Frederik Thuesen. 2018. "Punishment on the Frontlines of Public Service Delivery: Client Ethnicity and Caseworker Sanctioning Decisions in a Scandinavian Welfare State." *Journal of Public Administration Research and Theory* 28 (3): 339–54. https://doi.org/10.1093/jopart/muy018.

Razafindrakoto, Mireille, and François Roubaud. 2010. "Are International Databases on Corruption Reliable? A Comparison of Expert Opinion Surveys and Household Surveys in Sub-Saharan Africa." *World Development* 38 (8): 1057–69.

Riketta, Michael, and Rolf van Dick. 2005. "Foci of Attachment in Organizations: A Meta-Analytic Comparison of the Strength and Correlates of Workgroup versus Organizational Identification and Commitment." *Journal of Vocational Behavior* 67 (3): 490–510. https://doi.org/10.1016/j.jvb.2004.06.001.

Schriesheim, Chester A., Joshua B. Wu, and Terri A. Scandura. 2009. "A Meso Measure? Examination of the Levels of Analysis of the Multifactor Leadership Questionnaire (MLQ)." *The Leadership Quarterly* 20 (4): 604–16. https://doi.org/10.1016/j.leaqua.2009.04.005.

Schuster, Christian, Jan-Hinrik Meyer-Sahling, and Kim Sass Mikkelsen. 2020. "(Un)principled Principals, (Un)principled Agents: The Differential Effects of Managerial Civil Service Reforms on Corruption in Developing and OECD Countries." *Governance: An International Journal of Policy, Administration, and Institutions* 33 (4): 829–48. https://doi.org/10.1111/gove.12461.

Shah, Priti Pradhan. 1998. "Who Are Employees' Social Referents? Using a Network Perspective to Determine Referent Others." *Academy of Management Journal* 41 (3): 249–68. https://doi.org/10.2307/256906.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859–83. https://doi.org/10.1037/0033-2909.133.5.859.

Willems, Jurgen. 2020. "Public Servant Stereotypes: It Is Not (At) All about Being Lazy, Greedy and Corrupt." *Public Administration* 98 (4): 807–23. https://doi.org/10.1111/padm.12686.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.

CHAPTER 24

Interpreting Survey Findings

Can Survey Results Be Compared across Organizations and Countries?

Robert Lipinski, Jan-Hinrik Meyer-Sahling, Kim Sass Mikkelsen, and Christian Schuster

SUMMARY

With the rise in worldwide efforts to understand public administration by surveying civil servants, issues of survey question comparability become paramount. Surveys can rarely be understood in a void but rather require benchmarks and points of reference. However, it is not clear whether survey questions, even when phrased and structured in the same manner, measure the same concepts in the same way and, therefore, can be compared. For multiple reasons, including work environment, adaptive expectations, and cultural factors, different people might understand the same question in different ways and adjust their answers accordingly. This might make survey results incomparable, not only across countries but also across different groups of civil servants within a national public administration. This chapter uses results from seven public service surveys from across Europe, Latin America, and South Asia to investigate the extent to which the same survey questions measure the same concepts similarly—that is, are measurement invariant—using as an example questions related to *transformational leadership*. To ascertain measurement invariance, models of a hypothesized relationship between questions measuring transformational leadership are compared across countries, as well as along gender, educational, and organizational lines within countries. Solid evidence of metric invariance and tentative evidence of scalar invariance is found in cross-country comparisons. Moreover, factor loadings can be judged equal (*metric invariance*) across gender, education level, and organization in most countries, as can latent factor means (*scalar invariance*). Our results suggest that groups of public servants within countries—delineated, for instance, by gender, education, or organization—can typically be benchmarked without invariance concerns. Across countries, evidence for valid benchmarking—that is, scalar invariance—is strongest for countries in similar regions and at similar income levels. It is weaker—though still suggestive—when comparing all countries in the sample. Our chapter concludes that less culturally contingent concepts may be plausibly benchmarked with care across countries.

Robert Lipinski is a consultant in the World Bank's Development Impact Evaluation (DIME) Department. Jan-Hinrik Meyer-Sahling is a professor at the University of Nottingham. Kim Sass Mikkelsen is an associate professor at Roskilde University. Christian Schuster is a professor at University College London.

ANALYTICS IN PRACTICE

- Many theoretical insights and practical lessons from surveys of civil servants depend on the ability to draw comparisons between countries and demographic groups. This chapter focuses on the comparability of the concept of *transformational leadership* across different contexts and groups—a premise known as *measurement invariance*. Transformational leadership measures the extent to which managers lead by setting a good example, making employees proud, and generating enthusiasm about an organization's mission.
- Equality in the understanding of a single overarching concept, such as transformational leadership, across different countries and groups can be conceptualized in three ways. The same concept can be measured by the same set of questions (*configural invariance*), those questions can have the same strength of a relationship with the underlying concept (*metric invariance*), and the concept can have the latent mean structure (*scalar invariance*).
- When comparing survey measures and concepts across countries, practitioners should consider the extent to which different cultural interpretations of a concept (such as leadership), different social-desirability biases, different pressures in the work environment, and even differences in language may lead to differences in survey means across countries that do not reflect substantive differences in the underlying concept (such as the quality of leadership).
- When empirically assessing the measurement invariance (and thus the cross-country comparability) of a concept that is arguably culturally specific—transformational leadership—we find that cross-country comparisons can be undertaken, although with caution. There is evidence that the concept of transformational leadership is understood in a comparable way across the seven countries included in the analyses. As we find suggestive evidence that cross-country comparisons are possible with even a relatively culturally contingent concept (leadership), cross-country comparisons of more factual questions (for example, "Did you have a performance evaluation last year?") are plausibly often possible in a valid manner.
- Grouping countries by region and income level removes many of the differences across countries. This suggests that comparisons between countries at similar income levels and in the same world regions can be made with greater confidence.
- Within-country comparisons of transformational leadership suffer from fewer concerns about lack of comparability. Empirically, we find that they can be reliably made across public servants of different genders and education levels, and in different institutions.

INTRODUCTION

Surveys of civil servants provide insights into core parts of the public administration production function—such as the quality of management and the attitudes (for example, motivation) of employees. As argued by Rogger and Schuster in chapters 1–3 of *The Government Analytics Handbook*, these determinants of public sector productivity are difficult to measure accurately with other data sources. Survey results are typically presented as percentages of public servants who evaluate favorably dimensions of their work environment, management, or themselves—for instance, the percentage of public servants who recommend their organization as a great place to work, or the percentage of public servants who evaluate the leadership of their superior favorably.

How can governments know whether certain percentages—such as 75 percent of public servants who are satisfied with their jobs—are strengths or weaknesses of their public service? Interpreting survey results—and understanding areas for development in the public service—is often greatly aided by comparison. By benchmarking themselves with other countries on the same survey response, governments can understand where their strengths and weaknesses lie. This is one of the founding motivations of the Global Survey of Public Servants (GSPS) initiative (Fukuyama et al. 2022). Similarly, benchmarking internally between groups of public servants—for example, by gender, education, or institution—can help governments understand where, inside government, strengths and weaknesses lie.

However, such benchmarking presupposes comparability in measurement and the survey response process. In other words, it presupposes that respondents understand concepts—such as leadership, motivation, and satisfaction—in the same manner across different countries, government institutions, or groups (for example, men and women) in public service, and that they face similar biases (for example, social-desirability bias) when responding to survey questions. If the same concepts mean different things to different public servants or trigger different response biases in different public servants, valid comparisons are no longer possible, as differences in means might stem from differences in understanding or bias rather than differences in the underlying concept (for example, differences in actual work motivation).

Many public service survey questions are filtered through cultural factors (for example, "My direct superior leads by setting a good example"), individual-level characteristics, like gender ("I am paid at least as well as colleagues who have job responsibilities similar to me"), or both ("I feel sympathetic to the plight of the underprivileged"). If that is the case, then the survey measure lacks *measurement invariance*, which is "a property of a measurement instrument (in the case of survey research, a questionnaire), implying that the instrument measures the same concept in the same way across various subgroups of respondents" (Davidov et al. 2014, 58).

Past research suggests that measurement invariance might affect some measures in surveys of public servants and, in particular, public service motivation (PSM) (Kim et al. 2013; Mikkelsen, Schuster, and Meyer-Sahling 2020). However, what it *means* to be motivated to serve the public in all its dimensions—such as commitment to public values or compassion—is, arguably, highly dependent on cultural factors. As such, concerns about PSM's lack of cross-country comparability might not travel to other survey questions with less-cultural and more-factual content.

To assess this empirically, this chapter assesses what is arguably a key determinant of public administration effectiveness: the quality of leadership and, in particular, the concept of *transformational leadership*, a style of leadership that inspires and motivates subordinates to go beyond their self-interest and expectation of pecuniary rewards to achieve their goals and an organization's targets (Jensen et al. 2019; Pearce et al. 2002). Transformational leadership has been found to positively affect performance in public sector organizations across multiple contexts (Hameduddin and Engbers 2021; Pandey et al. 2016; Schuster et al. 2020).

Methodologically, we follow Mikkelsen, Schuster, and Meyer-Sahling (2020, 740) and undertake a measurement-invariance analysis given that "systematic cross-cultural and cross-national measurement-invariance analyses are central to gauge the comparability and generalizability." We apply the measurement-invariance analysis to an original seven country survey of public servants, in which transformational leadership is measured with exactly the same measurement scale across countries. We assess measurement invariance across countries and within countries across government institutions, as well as across public servants with different genders and education levels.

Our chapter is organized as follows. The chapter begins with a review of the measurement-invariance literature, with a particular focus on its application in the field of public service surveying and on the concept of transformational leadership within the civil service. It then proceeds to describe the approach taken to analyze the measurement invariance of the concept of transformational leadership, including the data set used and the method of analysis: multigroup confirmatory factor analysis (MGCFA). After that, we present our results—first, for cross-country comparisons and then for within-country comparisons, for civil servants grouped by gender, education level, and organization. We then discuss the theoretical and practical implications of our results and conclude.

LITERATURE REVIEW

The Concept of Measurement Invariance

It is common for surveys to aggregate individual questions into larger, overarching constructs. For example, the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) calculates three subindexes pertaining to some key aspects of public service functioning ("leaders lead," "supervisors," and "intrinsic work experience"), each composed by averaging positive responses to five survey questions. These are, in turn, aggregated into an "employee engagement index" (OPM 2019). To take another example, the United Kingdom's Civil Service People Survey also calculates an "employee engagement index," tabulated over five questions selected based on factor analysis from pilot surveys (Cabinet Office 2019). Despite similarities in their names and their high-income, English-speaking country settings, however, these two measures cannot be directly compared with each other, due to differences in wording and survey methodology. However, another long-standing concern of survey researchers is the possibility that even exactly the same questions can be interpreted differently by various groups of respondents. Engagement measured with the same battery of questions could still be conceived differently by civil servants in the United States and the United Kingdom due to cultural differences, institutional context, or socioeconomic factors.

Therefore, in order to meaningfully compare a statistical construct, like engagement, motivation, or leadership, and related statistical quantities, like means and regression coefficients, across different groups (or time periods), the construct should first be tested for measurement invariance. Demonstrating the measurement invariance (sometimes also termed equivalence) of a given construct entails showing that it is interpreted in a comparable manner by different sets of respondents. In contrast, "measurement *non*-invariance suggests that a construct has a different structure or meaning to different groups or on different measurement occasions in the same group, and so the construct cannot be meaningfully tested or construed across groups or across time" (Putnick and Bornstein 2016, 71; emphasis added).

Three basic levels of measurement invariance are usually distinguished: *configural*, *metric*, and *scalar* (Vandenberg and Lance 2000). They represent progressively stricter tests for comparability between groups. Figure 24.1, below, provides a schematic representation of the generalized idea behind these concepts by illustrating how each of them hypothesizes the relationship between manifest variables and underlying latent constructs. A more detailed visualization is provided by figures L.1, L.2, and L.3 in appendix L. They demonstrate different levels of invariance using examples of models of transformational leadership in public service that are analyzed throughout this chapter.

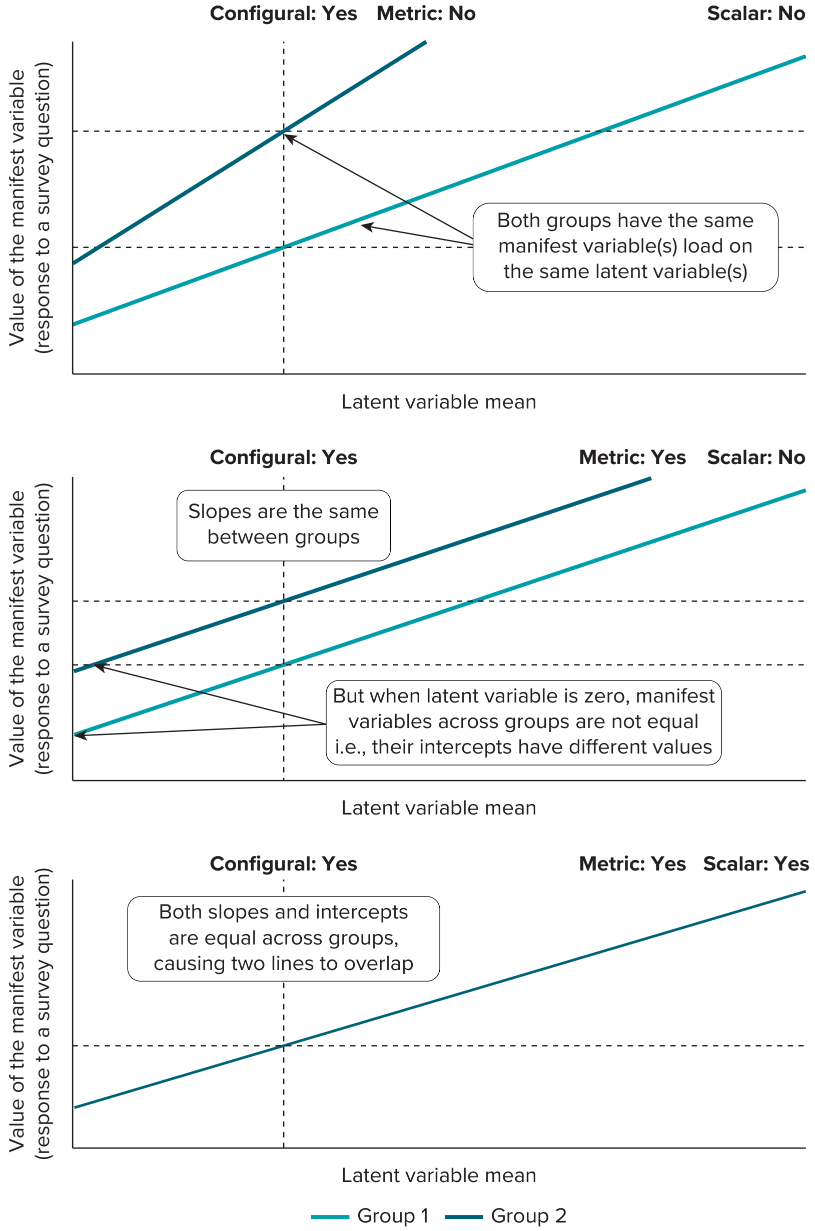
Configural Invariance

In the first step toward establishing the comparability of a statistical concept, it needs to be ascertained that it has the same *factor structure* across all groups being compared. This means that in all groups, the same sets of questions are linked to the same sets of underlying constructs, typically termed *latent factors* or *variables* (Kim et al. 2013). This is schematically represented by the top panel of figure 24.1 and, on the example of transformational leadership, by figure L.1 in appendix L. If, in both groups of interest, it can be shown that a model with all observed survey questions loading onto a single latent variable fits the data well, then configural invariance is deemed to hold (upper panel of figure L.1). However, if this hypothesized model is found not to fit the data, then configural invariance cannot be said to hold. One example of such a situation could be where data from one group display a two-latent-factor structure, as in the bottom panel of figure L.1.

Metric Invariance

Once the same structure of the items and factors is confirmed across groups, researchers might turn their attention to the equality of *factor loadings* between the groups. Factor loadings can be understood

FIGURE 24.1 Schematic Visual Representation of the Three Levels of Measurement Invariance: Configural, Metric, and Scalar



Source: Adapted from Cieciuch et al. 2019, 179.
Note: "The X axis represents the latent variable mean; the Y axis represents the response to a survey question item measuring the latent variable. The diagonal represents the function relation between the latent variable and the response to the survey question item in two countries (in unstandardized terms)" (Cieciuch et al. 2019, 179). Here, *countries* is replaced by the more generic term *groups*.

as measures of the strength of the relationship between the observed survey items and the latent factors. Confirming metric invariance is a pre-requisite that "ensures that *structural regression estimates* are comparable across groups" (Mikkelsen, Schuster, and Meyer-Sahling 2020, 4; emphasis added). This is because, in metric-invariant models, differences between survey items are linked to differences in the underlying latent-factor models in the same fashion across all the groups included in the analyses (Steenkamp and Baumgartner 1998). The equal slopes of the lines in the middle panel of figure 24.1 demonstrate this point. In other words, one unit change in the *x*-axis value of the latent variable is associated with a change in manifest variable values that is the same for both groups. Consequently, only with metric invariance can regression with observed survey items be compared in a meaningful manner (Hong, Malik, and Lee 2003). This focus of metric invariance on the equality of factor loadings across groups is also shown in figure L.2 in appendix L.

Scalar Invariance

To ensure scalar invariance, not only factor loadings but also the means of the *item intercepts* must be shown to be equal between groups (Vandenberg and Lance 2000). Even when factor loadings suggest that the latent constructs have the same impact upon the value of observed items in all the groups considered, as in the definition of metric invariance, it is still possible for groups to have different values of the intercept—that is, the value of observed items when the latent variable is zero—due to some unobservable characteristics.

Only when the intercepts are the same in all groups, as in the bottom panel of figure 24.1, can the model be said to be scalar invariant. This is also the situation presented in figure L.3 in appendix L. Thus, establishing scalar invariance should precede any attempt at comparing the latent means and intercepts of observed items between the groups. Only once it is established can it can be assumed that "cross-national differences in the means of the observed items are due to differences in the means of the underlying construct(s)" (Steenkamp and Baumgartner 1998, 80).

Measurement Invariance

Although first developed in the mid-20th century in the field of psychology (see, for example, Meredith 1964; Struening and Cohen 1963), measurement invariance has since become a concern in multiple other disciplines. In the field of education, researchers have applied it to better understand the comparability of concepts such as the time management of US undergraduate students (Martinez 2021) or the different subscales of school climate measured in the Georgia School Climate Survey (La Salle, McCoach, and Meyers 2021). The Organisation for Economic Co-operation and Development (OECD) has used it to gauge the comparability of latent factors measured by the Programme for International Student Assessment (PISA) and several other cross-country surveys (Van De Vijveri et al. 2019). It has also been analyzed in other contexts as diverse as consumer research (De Jong, Steenkamp, and Fox 2007) and sociology—for example, to better understand concepts such as attitudes toward granting citizenship rights in the International Social Survey Program (Davidov et al. 2018) and German adolescents' life attitudes (Seddig and Leitgöb 2018). Given its importance and the widespread academic interest in it, public administration researchers, in the face of increasing surveying efforts, have also come to analyze measurement invariance in public service surveys.

Measurement Invariance in Public Service Surveys

In recent years, multiple researchers have emphasized the importance of studying public administration from a comparative perspective, both to improve researchers' theoretical understanding and to draw practical lessons (Fitzpatrick et al. 2011; Jreisat 2005). At the same time, surveys have become one of the key methods used to better understand public administration. As emphasized throughout this part of the *Handbook*, surveys allow researchers and policy practitioners to gain insights into dimensions of public administration's functioning that would otherwise be unmeasurable. Concepts such as job satisfaction or attitudes toward management could scarcely be gauged otherwise. Surveys can also be used to anonymously ask about aspects of civil servants' work that might otherwise not be talked about—such as perceptions of corruption or workplace harassment. However, the comparability of survey results—both across countries and across demographic groups within the civil service—cannot be taken for granted. Challenges to comparability stem from several sources, like differences in the mode of survey delivery (see chapter 19) or perceived question sensitivity (see chapter 22). Another obstacle is the different phrasing of questions—an issue that initiatives such as the GSPS have recently begun to address (Fukuyama et al. 2022). However, even with all these problems solved, it is not certain that the same survey concepts would be understood in the same way by different groups of civil servants.

This explains the recent turn of several public administration scholars toward analyzing measurement invariance across public service surveys. Kim et al. (2013) test for measurement invariance in PSM—one of the frequently recurring parts of many public service survey questionnaires, which aims to measure respondents' motivation and willingness to serve society. The authors use a PSM index containing questions asked using a 1–5 Likert scale. The tests for configural invariance suggest that PSM has the same structure in 8 out of the 12 countries studied. However, neither metric nor scalar invariance can be detected, meaning that the construct has a different meaning and different levels across countries (apart from the sample restricted to the culturally similar Australia, the United Kingdom, and the United States, for which metric invariance can be detected). Mikkelsen, Schuster, and Meyer-Sahling (2020) expand these results by using a more diverse sample of countries and larger sample sizes within countries. Using survey results from over 23,000 civil servants across 10 countries, they demonstrate that a 16-item PSM scale displays first- and second-order partial metric invariance, apart from the case of two Asian countries studied (Bangladesh and Nepal). Still, their study finds that PSM levels cannot be compared across countries due to a lack of scalar invariance.

The Concept of Transformational Leadership

In the face of the expansion of measurement-invariance studies within the field of public administration, a relative lack of attention to concepts other than PSM can be discerned. Although PSM is of clear importance and is commonly measured, many other dimensions of work affect civil servants' performance and are regularly included in public service surveys. One key concept, measured in some form in virtually every public service survey, is leadership. Its measurement is most frequently based on past research, scales, and wording from the management science and psychology literature (Tummers and Knies 2016). In particular, the idea of *transformational leadership* has gained traction with public administration researchers (see, for example, Kroll and Vogel 2014; Pandey et al. 2016). It was first developed in the 1970s by Downton (1973) and more fully by Burns (1978), who applied it, together with the contrasting idea of *transactional leadership*, to study political leaders. Whereas transactional leadership is conceived as a leadership style focused on tangible benefits obtained via exchange between a leader and followers (for example, jobs for votes), transformational leadership is chiefly focused on motivating and engaging potential followers to move in a desired direction by conveying a sense of mission, employing compelling argumentation, and using one's own example. Under transformational leadership, followers are inspired to maximize their performance and achieve set goals for the sake of "higher level needs such as self-actualization" (Pearce et al. 2002, 281), attention, and personal development (Nguyen et al. 2017). Bass (1985) extended both conceptions of leadership to the management of organizations. Since then, transformational leadership has been found to be one of the key factors explaining improvements in many dimensions of performance in multiple settings in the private sector, including increased agreement on strategic goals in a large Israeli telecommunications firm (Berson and Avolio 2004), satisfaction with supervisors in Turkish boutique hotels (Erkutlu 2008), and knowledge management in Spanish firms (García-Morales, Lloréns-Monte, and Verdú-Jover 2008). A meta-analysis of 113 primary studies on the topic by Wang et al. (2011) finds transformational leadership to be associated with better performance across the individual, team, and organizational levels.

Moreover, a recent meta-analysis of the PSM and leadership literature, conducted by Hameduddin and Engbers (2021), has found that 50 percent ($n = 20$) of publications concerned with leadership rely on the concept of transformational leadership, making it the most common conceptualization of leadership by public administration scholars. Following this approach, Park and Rainey (2008) establish a positive relationship between transformational leadership and outcomes such as job satisfaction, quality of work, and perceived performance across US federal agencies. Pandey et al. (2016) find its direct and indirect impacts on normative public values. Donkor, Sekyere, and Oduro (2022) further find that higher transformational leadership is linked to higher organizational commitment across 16 Ghanaian public sector organizations. In a survey of over 21,000 civil servants in Chile, Schuster et al. (2020) similarly find transformational leadership to be correlated with higher job satisfaction, motivation, and engagement. Hameduddin and Engbers' (2021) review of 40 studies finds a link between transformational leadership and PSM—a relationship that holds across a diverse set of countries analyzed.

Transformational leadership was therefore chosen as the survey instrument of focus in the present chapter because of its solid theoretical development, extensive academic research pedigree, and practical importance for public sector performance. Two further reasons can be adduced to explain this choice. First, it is a concept that can usually be mapped onto a single underlying construct. In other words, survey questions about transformational leadership are all aimed at measuring different but related aspects of the same latent factor. This is often not the case with many other sections of public service surveys, like salaries or performance management, which measure many divergent subdimensions—including administrative (for example, salary amount and participation in performance evaluations), motivational (for example, satisfaction with salary and usefulness of performance evaluations), and ethical (for example, salary and performance evaluations' fairness) subdimensions.

Second, there exists a relative imbalance between the large number of studies relying on measures of transformational leadership in the public sector and the lack of research investigating the measurement invariance of this concept. To the best of the authors' knowledge, the only analysis of measurement invariance focused on transformational (and transactional) leadership is a paper by Jensen et al. (2019). However, it presents only a limited test of measurement invariance for transformational leadership, as it focused on full configural and metric invariance, without tests of partial metric invariance or scalar invariance. Jensen et al. (2019) also do not engage in cross-country or cross-cultural analysis of invariance because their sample is composed of respondents from Denmark. The authors focus on invariance across time, sector (including public vs. private), and randomized training groups but not demographic variables, like gender or education, or organizations within the public sector—a focus of the present chapter. Thus, although transformational leadership has gained a well-established position within the public administration literature, only limited attention has been paid to testing the measurement invariance of this concept, which provides the rationale for the analyses contained in the pages below.

METHODOLOGY

Data Set

The data used for the analysis in this chapter come from the GSPS initiative. The GSPS is a combined effort of researchers at the World Bank's Bureaucracy Lab, University College London (UCL), the University of Nottingham, and Stanford University that aims to better understand the attitudes and behaviors of civil servants around the globe. Part of the GSPS is focused on making public administration survey questionnaires more comparable. It strives to achieve this by developing and promoting the inclusion of a "core" survey module, which would ask the same set of questions about the principal dimensions of civil service work, such as job satisfaction, work motivation, and leadership, to all the civil servants surveyed.

Seven public service surveys are included in the analyses below. They come from the following countries: Albania, Bangladesh, Brazil, Chile, Estonia, Kosovo, and Nepal.[1] Together, the surveys gathered responses from over 21,000 civil servants. Surveys were delivered both online and in person between 2017 and 2018 and included an extensive set of questions pertaining to multiple aspects of civil service functioning.[2] Importantly for present purposes, the phrasing of questions was exactly the same across countries. In order to ensure that respondents' understanding of the questions would remain unaffected by translation into local languages, the questions were pretested using cognitive interviews with civil servants and iteratively revised (Mikkelsen, Schuster, and Meyer-Sahling 2020). Moreover, each survey strove to include a comparable sample of respondents—that is, central government civil servants who perform general administrative duties.[3] Due to incomplete personnel records on civil servants, the samples are not fully representative. Furthermore, in the in-person surveys, informal quota sampling and in-person surveys based on information from individual public administration organizations were used (see Mikkelsen, Schuster, and Meyer-Sahling 2020). When possible, the demographics of the survey samples were compared to servicewide values (see table 24A.1), and those comparisons reveal broadly aligned values.

The final advantage of the present choice of surveys is that they represent a diverse set of regional and economic groupings: from South America through Europe to Asia, and from lower-middle-income countries, like Bangladesh and Nepal, through upper-middle-income Albania, Brazil, and Kosovo to high-income Chile and Estonia (see table 24.1). This allows analyses in this chapter to not only focus on differences between groups within each civil service but also to compare invariance across the cross-cultural contexts of different regions and countries.

In the present sample of civil servant surveys, the concept of transformational leadership was measured using the level of agreement with the following three questions, all starting with the prompt "To what extent do you agree with the following statements?":

1. My direct superior articulates and generates enthusiasm for my organization's vision and mission (abbreviated as *enthusiasm*).
2. My direct superior leads by setting a good example (abbreviated as *good example*).
3. My direct superior says things that make employees proud to be part of this organization (abbreviated as *pride*).

The responses were measured using a 1–5 Likert scale, where 1 signified "strongly disagree" and 5 "strongly agree." The basic statistics on each of the variables are presented in table 24.2. A majority of the respondents agree with the question prompts, confirming that their direct superiors generate enthusiasm about the organization's vision and mission, lead by setting a good example, and make them proud to be a part of the organization. Correlations between the three variables are also very high (>0.75), which could be interpreted as an early indication that they indeed measure one underlying concept of transformational leadership.

TABLE 24.1 Summary of the Seven Public Servant Surveys Used in the Chapter

	Albania	Bangladesh	Brazil	Chile	Estonia	Kosovo	Nepal
Respondents	3,690	1,049	3,992	5,742	3,555	2,465	1,249
Response rate	47%	Convenience sample	11%	37%	25%	14%	Convenience sample
Mode of delivery	Online	In-person	Online	Online	Online	Online	In-person
Year	2017	2017–18	2018	2016–17	2017	2017	2017–18
Language	Albanian	English, Bangla	Portuguese	Spanish	Estonian	Albanian, Serbian	English, Nepali
Report	Meyer-Sahling et al. (2018d)	Meyer-Sahling et al. (2019)	Pereira et al. (2021)	Schuster et al. (2017)	Meyer-Sahling et al. (2018a)	Meyer-Sahling et al. (2018b)	Meyer-Sahling et al. (2018c)
Region[a]	ECA	South Asia	LAC	LAC	ECA	ECA	South Asia
Income group[a]	Upper-middle income	Lower-middle income	Upper-middle income	High income	High income	Upper-middle income	Lower-middle income
GDP per capita (current US$)[a]	$5,246	$1,967	$6,797	$13,232	$23,027	$4,347	$1,155

Source: Original table for this publication.
Note: ECA = Europe and Central Asia; LAC = Latin America and the Caribbean.
a. Based on World Bank data and groupings.

TABLE 24.2 Basic Statistics on the Three Questions Aiming to Measure Transformational Leadership

Statistic	Variable		
	Enthusiasm	Good example	Pride
Mean	3.59	3.74	3.40
Median	4.00	4.00	4.00
SD	1.29	1.27	1.28
Skew	−0.63	−0.81	−0.42
Kurtosis	2.29	2.61	2.11
Corr. with *enthusiasm*	1.00	0.80	0.84
Corr. with *good example*	0.80	1.00	0.79
Corr. with *pride*	0.84	0.79	1.00

Source: Original table for this publication.
Note: All variables are measured on a 1–5 Likert scale, where higher values indicate greater agreement. The values shown in the table are aggregated across countries. SD = standard deviation.

Measuring Invariance

As discussed more broadly in the literature review section, invariance can be measured on three key levels: configural, metric, and scalar. These levels of invariance are tested here using MGCFA. This has been the main method of testing measurement invariance in the past three decades (see, for example, Hofman, Mathieu, and Jacobs 1990; Mikkelsen, Schuster, and Meyer-Sahling 2020; Putnick and Bornstein 2016). It is carried out by setting progressively stricter constraints upon the parameters of the model being evaluated. First, the same model structure is imposed on all groups tested. If the model fit proves satisfactory (see the subsection below for criteria on this), metric invariance is tested by restricting factor loadings to be equal across groups. If the results from the comparison of model fit show that the constrained model is not performing significantly worse than the unconstrained one, then metric invariance can be inferred. Upon finding evidence of metric invariance, the means of the latent construct can be set to be equal across the groups, and, if this extra restriction also does not result in significantly worse model fit, then scalar invariance can be ascertained.

The first set of MGCFA tests pertains to measurement invariance across the seven countries included in the study. First, models for a set of countries grouped by region and income level are fit, before moving to full cross-country models. The results therefore demonstrate the extent to which national context determines how civil servants understand the concept of transformational leadership. The second set of analyses turns toward demographic groups within countries and evaluates whether respondents of different genders (female vs. male), education levels (below university vs. university), and organizations within the public administration interpret the questions about transformational leadership in the same manner. These two levels of analysis—inter- and intracountry—have been the key focus of measurement-invariance research (Vandenberg and Lance 2000).

Model Fit Indexes

To compare the progressively more restricted measurement-invariance models, one has to calculate how well they fit the data. Three measures are relied upon for this purpose. The first one is chi-square (χ^2). This is a likelihood ratio test that calculates how well the specified model and the associated expected distributions fit the observed data distributions. The χ^2 value, combined with the model's degrees of freedom, can be used to calculate the *p*-value—the likelihood that the observed deviation from the perfect model is due to chance. However, researchers are in agreement that, because the mathematical formula for its derivation is dependent on the sample size (*N*), this statistic is highly sensitive in large samples and might show statistically significant differences in model fit even when only small deviations from perfect fit are present (Byrne, Shavelson, and Muthèn 1989; Cheung and Rensvold 2002; French and Finch 2006; Putnick and Bornstein 2016).

For this reason, two further fit indexes are consulted when comparing model fit. One is the comparative fit index (CFI). Its value is scaled between 0 and 1 and is specifically designed to deal with the limitations of χ_2, including its oversensitiveness in large samples (Bentler 1990). The model might be assumed to fit well already when the CFI is above 0.90 (Cheung and Rensvold 2002), but a more restrictive threshold of 0.95 is typically used (Hooper, Coughlan, and Mullen 2008; Hu and Bentler 1999). However, in the measurement-invariance literature, if restricting model parameters leads to a decrease in the CFI of more than 0.01, the invariance is typically rejected (Cheung and Rensvold 2002).

The third and final fit index consulted throughout the analyses is the standardized root mean squared error (SRMR) (see Bentler 1995). The SRMR is calculated on a range from 0 to 1 and can be viewed "as the average standardized residual covariance" of the model variables (Shi, Maydeu-Olivares, and Rosseel 2020, 2). It can range from 0 to infinity, and, typically, absolute SRMR values below 0.05 are indicative of good model fit, although values up to 0.08 are deemed satisfactory (Hu and Bentler 1999). When the fit of models is compared for invariance, increases in the SRMR of more than 0.03 and 0.01 are taken as signaling significant model deterioration in metric- and scalar-invariance models, respectively (Chen 2007). Given the large sample sizes used here, the concern with the overrejection of invariant models by the SRMR raised by Chen (2007) is largely ameliorated.[4]

Therefore, when discussing model fit below, whether in absolute terms or when comparing its fit to another model, changes (indicated with Δ) in all fit indexes are reported (the *p*-value of $\Delta\chi^2$, ΔCFI, and ΔSRMR). The models are estimated in RStudio using the lavaan::cfa() function. Given a nonsymmetrical distribution and the ordinal nature of the data (see table 24.2), a diagonally weighted least squares (DWLS) estimator is used for model estimation (Li 2016; Rosseel 2012). Comparisons of model fit (χ_2, CFI, and SRMR values), are made using the semTools::compareFit() function.

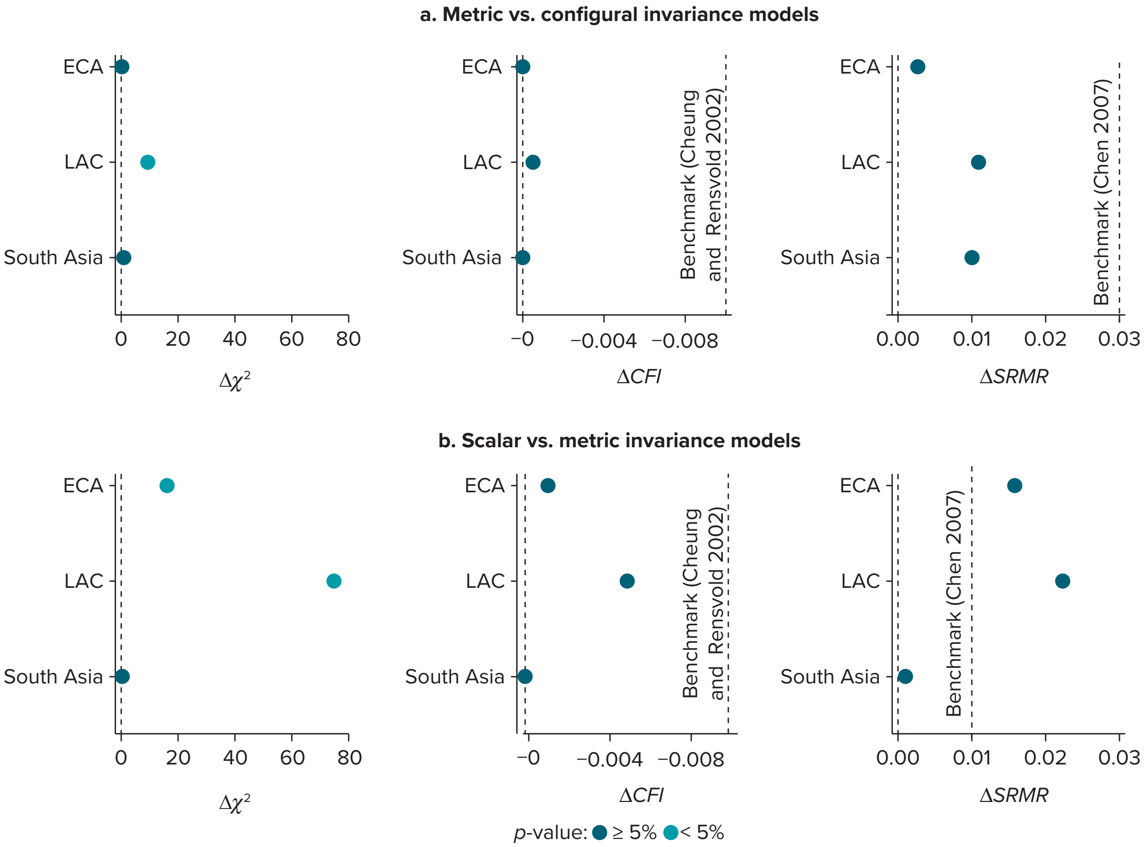
RESULTS

Measurement invariance is tested first in the cross-country context before moving to within-country invariance across demographic groups (gender and education level) and public administration organizations. We start by fitting the cross-country comparison in groupings of countries based on their income and region before moving to compare individual countries to each other. In each case, configural-, metric-, and scalar-invariance models are tested sequentially, provided that the acceptable fit of a higher-level model is first confirmed.

Cross-Country Comparison

The analysis begins with models comparing groups of like countries against each other. It is expected that civil servants in similar countries—that is, those at comparable levels of development or in a single geographical region—are more likely to conceive of transformational leadership in the same manner. Such grouping of countries ensures that the inevitable cultural and socioeconomic differences between countries are minimized. By contrast, comparing a high-income European country, like Estonia, and a large, upper-middle-income country in the heart of Latin America, like Brazil, is a much more demanding test of the invariance concept. Therefore, we move to the latter only after establishing that invariance holds within broader groupings of like countries.

FIGURE 24.2 Measurement Invariance across Countries Classified by Region: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: Regional classifications are based on World Bank data. The Europe and Central Asia (ECA) Region includes Albania, Kosovo, and Estonia; Latin America and the Caribbean (LAC) includes Brazil and Chile; and South Asia includes Bangladesh and Nepal. CFI = comparative fit index; SRMR = standardized root mean squared error.

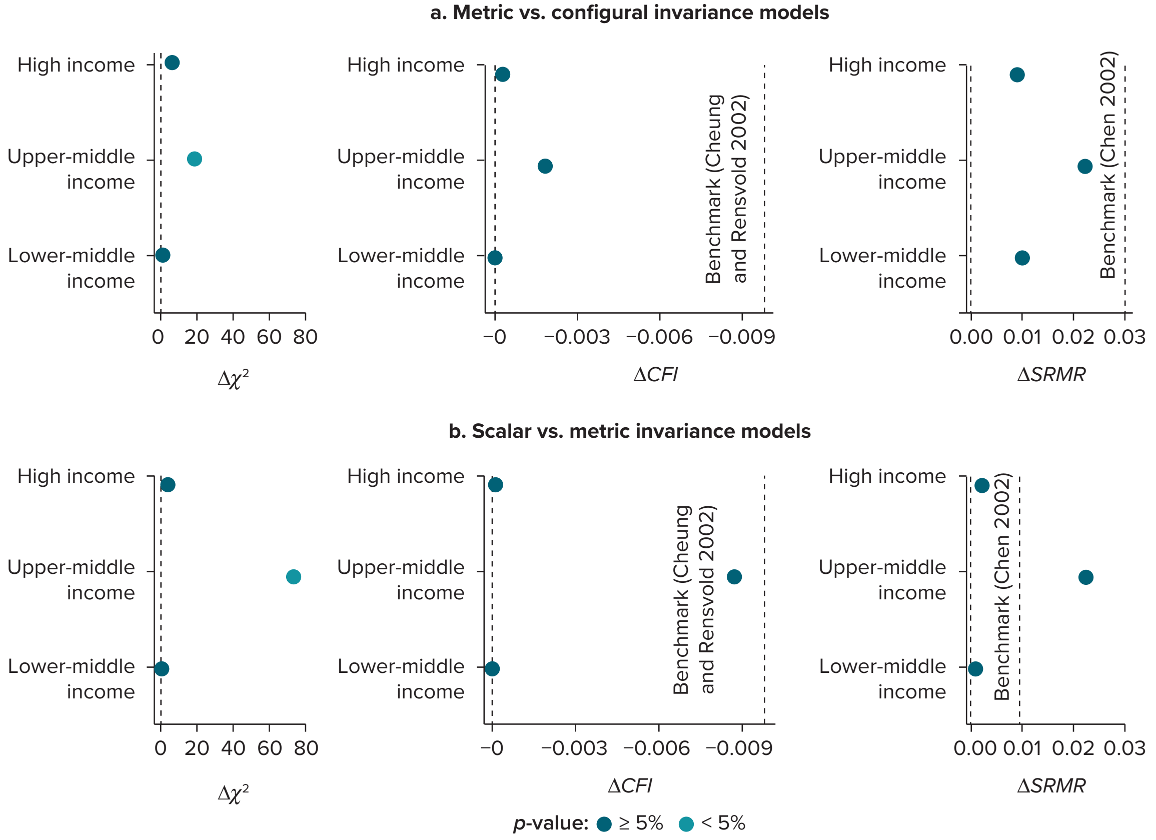
The results of comparing measurement invariance across countries within the same geographical region are shown in figure 24.2. Since all configural-invariance models with only three manifest variables and one latent factor have, by definition, a perfect level of fit, what the figure shows is the change (Δ) in key fit statistics—χ^2, CFI, and SRMR—between configural- and metric-invariance models and between metric- and scalar-invariance ones. For consistency, the changes in the CFI are reversed, meaning that model fit is deteriorating when going right along the *x* axis.

From the figure, it can be seen that metric invariance models fitted across countries from the Europe and Central Asia region (Albania, Estonia, and Kosovo) and for the South Asia region (Bangladesh and Nepal) do not exhibit significantly worse fit on all three indexes. For Latin America and the Caribbean (LAC) (Brazil and Chile), only the $\Delta\chi^2$ is statistically significant, but, given very low changes in the other two indexes, metric invariance can still be inferred.

Taking the metric-invariant models to the next level and imposing scalar invariance, model fit remains fully acceptable for South Asia. For ECA and LAC, both the $\Delta\chi^2$ and the $\Delta SRMR$ point to a significantly worse fit, and, therefore, as with the full cross-country model, scalar invariance can be only tentatively inferred based on the fact that the $\Delta CFI < 0.01$.

Figure 24.3 shows the results of the same analyses replicated across income groupings rather than regions. On the basis of the ΔCFI and the $\Delta SRMR$, all three income groupings exhibit metric invariance. Only a high *p*-value for the upper-middle-income group (Albania, Brazil, and Kosovo) points toward a

FIGURE 24.3 Measurement Invariance across Countries Classified by Income Group: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: Income groupings are based on World Bank data. High-income countries include Chile and Estonia; upper-middle-income countries include Albania, Brazil, and Kosovo; and lower-middle-income countries include Bangladesh and Nepal. CFI = comparative fit index; SRMR = standardized root mean squared error.

different conclusion, but, as explained in the methodology section, this is not taken as sufficient evidence to overrule a good fit based on the CFI and the SRMR.

If such an interpretation is adopted, scalar-invariance models can be estimated for all income groupings. According to all fit indexes, scalar invariance can be inferred across high-income (Chile and Estonia) and lower-middle-income (Bangladesh and Nepal) countries. For the upper-middle-income countries, the $\Delta\chi^2$ is statistically significant, and the $\Delta SRMR$ is well above the threshold of 0.01. The absolute value of the SRMR of the scalar-invariance model is also only borderline acceptable, at 0.046. The ΔCFI, standing just below 0.009, similarly approaches the threshold of significant deterioration. Therefore, a conclusion of scalar invariance can be drawn only on the basis of the CFI, and, even then, it is not strong. (It should also be noted that the results and conclusions for countries in the lower-middle-income category are exactly the same as for the South Asia category above because those two groups happen to contain the same pair of countries: Bangladesh and Nepal.)

Given the relatively robust evidence of metric and scalar invariance within groupings of comparable countries, we now move to compare all seven countries against each other. When the metric-invariance model is fitted by restricting the factor loadings to be equal across all seven countries, the absolute model fit is still good according to all three fit indexes. The value of χ^2 is 47.1 (df = 12), and the associated p-value is close to 0. The CFI drops to 0.998, and the SRMR increases to 0.019. Therefore, the change in the latter two fit indexes is well within the limits recommended by the literature. Although the $\Delta\chi^2$ with a p-value below 5 percent points toward significantly worse fit, the large sample size and perfect fit of the unrestricted model make this a less reliable measure. Therefore, metric invariance can be inferred for cross-country comparisons of transformational leadership.

Given this conclusion, a scalar-invariance model can be fitted. It represents a borderline case of significant deterioration. The *p*-value of the $\Delta\chi^2$ is close to 0, and the $\Delta SRMR$ is 0.018, which is above the threshold recommended for scalar-invariance models by Chen (2007). However, the difference is small, and the absolute model fit (the SRMR = 0.037) is still good. Furthermore, the ΔCFI of 0.008 can be viewed as acceptable. Therefore, a tentative conclusion of scalar invariance can be reached.

To summarize the above analyses—in a full cross-country analysis, no significant deterioration in the model of metric invariance suggests that researchers and policy practitioners should be able to compare factor loadings and structural regression coefficients across countries. Item intercepts and means of the indicators can also be compared, although cautiously, given that not all fit indexes suggest that the model with equal intercepts fits the data well.

However, comparisons of this type might be more warranted within groups of like countries. There is evidence that cross-cultural differences in understanding of the idea of transformational leadership are (largely) removed by grouping countries according to their geographical regions. Within such groupings, there is clear evidence of metric invariance. With the same caveat as in the full cross-country models, scalar invariance can also be demonstrated for those models. Invariance is even stronger when countries are grouped by their income level. Both high- and lower-middle-income groups exhibit full metric and scalar invariance. The only group where the conclusion of scalar invariance has very little backing is upper-middle-income countries. This is perhaps unsurprising, given that this group can be viewed as the most heterogeneous, and, therefore, differences in the understanding of concepts such as leadership remain substantial.[5]

Within-Country Comparison: Gender

Turning to intracountry comparisons, gender is the key demographic measure in all public service surveys and also, typically, one of the first lines along which survey results are broken down. The distribution of respondents by gender in the survey sample used here is reported in table 24.3. As can be observed, the gender distribution of civil servants who respond to the surveys varies highly by country. In three out of seven countries, women form the majority of the respondents. The female-to-male ratio varies from approximately 3:1 in Estonia to less than 1:3 in Bangladesh. These cross-country differences are largely consistent with the variation in gender balance across survey populations in countries where personnel records are available (see table 24A.1).

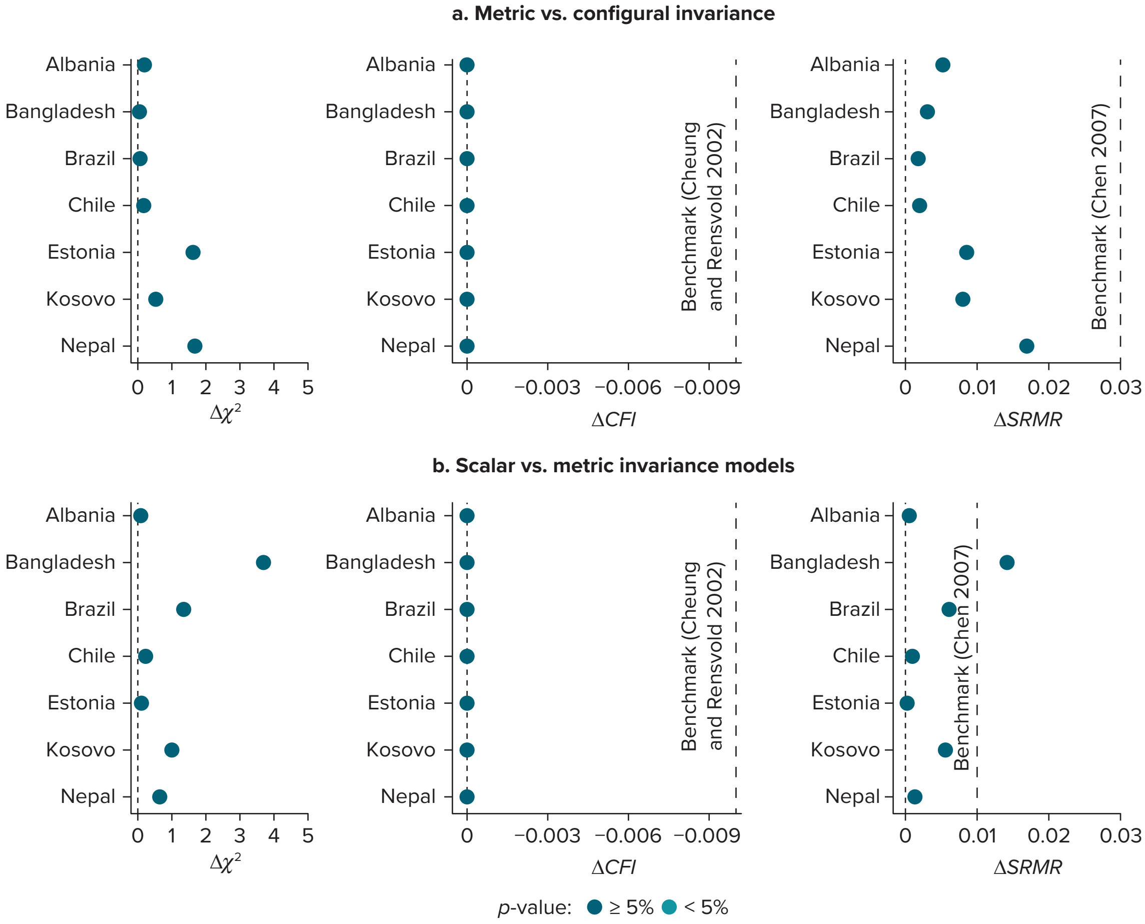
The results of measurement-invariance analyses across gender groups within countries are presented in figure 24.4. Metric invariance—the equality of factor loadings between genders—is obtained with little

TABLE 24.3 Distribution of Respondents, by Gender

Country	Male	Female	Missing
Albania	1,374 (37.2%)	2,261 (61.3%)	55 (1.5%)
Bangladesh	801 (76.4%)	224 (21.4%)	24 (2.3%)
Brazil	2,268 (56.8%)	1,701 (42.6%)	23 (0.6%)
Chile	2,502 (43.6%)	3,155 (54.9%)	85 (1.5%)
Estonia	845 (23.8%)	2,462 (69.3%)	248 (7.0%)
Kosovo	1,363 (55.7%)	1,028 (42.0%)	57 (2.3%)
Nepal	817 (65.4%)	421 (33.7%)	11 (0.9%)

Source: Original table for this publication.

FIGURE 24.4 Measurement Invariance across Gender within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: CFI = comparative fit index; SRMR = standardized root mean squared error.

space for doubt for all seven countries. In none of them are changes in χ^2 statistically significant, nor are the changes in the CFI and SRMR above their respective thresholds.[6]

As a next step, scalar-invariance models are fitted and compared. Here, the arguments and conclusion remain unchanged. All three fit indexes point to good fit and no significant deterioration of the model after adding equality constraints on item intercepts, which allows us to conclude scalar invariance across genders in all countries considered.

Within-Country Comparison: Education Level

Like with the analyses focused on gender, this subsection concerning education begins with a demographic overview (table 24.4), which presents the distribution of civil servants by their level of education across countries. Here, the heterogeneity across surveys is even more pronounced than in the case of gender. Whereas in Albania, 92.1 percent of civil servants who responded to the survey had university-level education, and only 5.1 percent had below-university-level education, these proportions are equal in Nepal, and in Chile become almost exactly reversed.

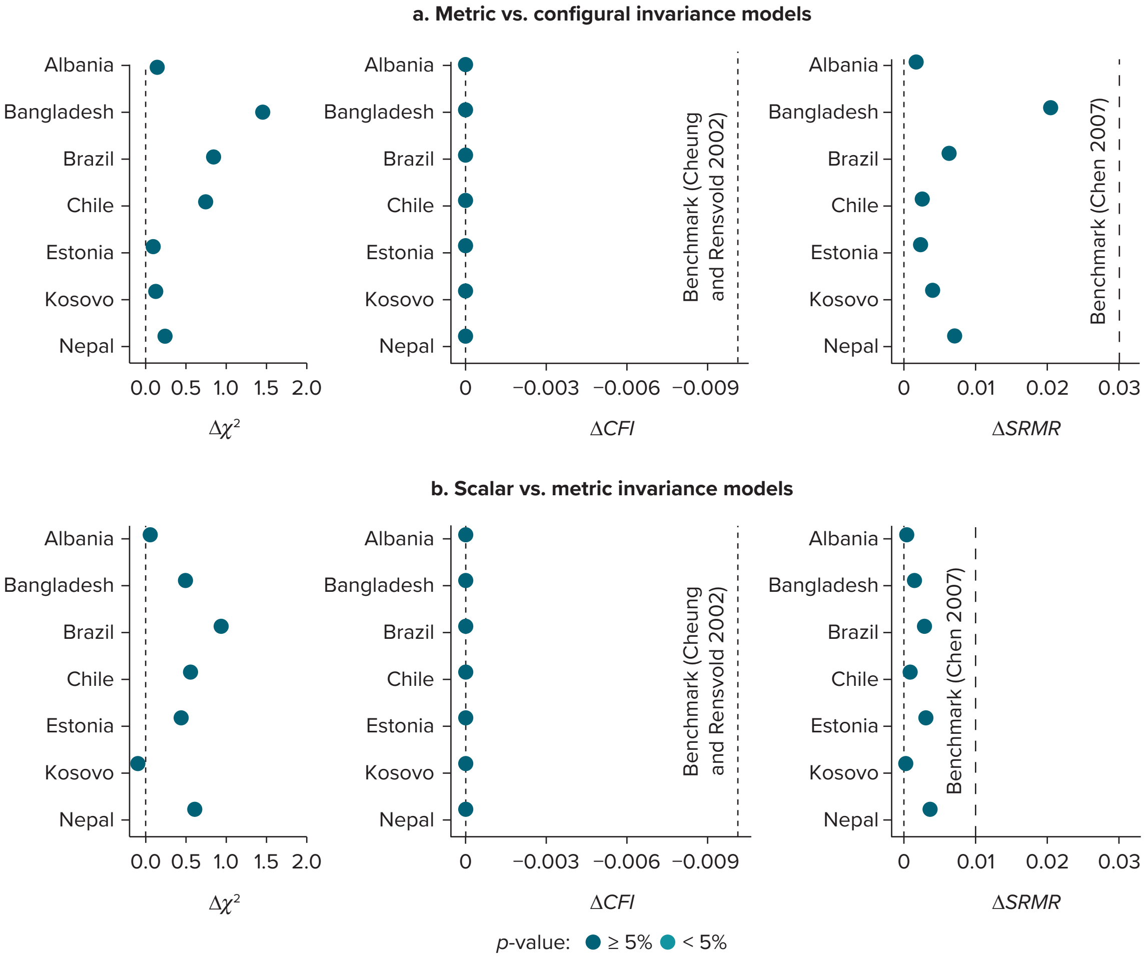
Figure 24.5 demonstrates the results from fitting different levels of invariance models across different education levels in seven countries. As with gender, both metric and scalar invariance can be concluded for all seven countries. None of the changes in the fit indexes come near their respective thresholds.

TABLE 24.4 Distribution of Respondents, by Education Level

Country	University	Below university	Missing
Albania	3,399 (92.1%)	188 (5.1%)	103 (2.8%)
Bangladesh	560 (53.4%)	468 (44.6%)	21 (2.0%)
Brazil	1,964 (49.7%)	1,895 (47.5%)	113 (2.8%)
Chile	586 (10.2%)	5,081 (88.5%)	75 (1.3%)
Estonia	1,898 (53.4%)	1,439 (40.5%)	218 (6.1%)
Kosovo	1,150 (47.0%)	1,261 (51.5%)	37 (1.5%)
Nepal	603 (48.3%)	603 (48.3%)	43 (3.4%)

Source: Original table for this publication.

FIGURE 24.5 Measurement Invariance across Education Levels within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: CFI = comparative fit index; SRMR = standardized root mean squared error.

Within-Country Comparison: Public Administration Organization

The final set of invariance models is fitted across public administration organizations. The surveys analyzed here were conducted among several central government organizations in each country (see table 24.5). The number of organizations with more than 50 respondents ranges from 7 in Bangladesh to 27 in Estonia. Across countries, the mean number of respondents per organization varies between 83.9 in Kosovo and 522 in Chile. In the latter country, standing at 1,520, the largest number of respondents per organization is also observed.

Figure 24.6 replicates the measurement-invariance comparisons discussed above for gender and education level. However, here the results are less clear-cut. For metric-invariance models, there is clear evidence to suggest the equality of factor loadings for five out of seven countries. For Kosovo and Nepal, the $\Delta SRMR$ is, however, just above 0.03, which suggests significant deterioration compared to the configural-invariance model. Yet the $\Delta\chi^2$ remains small in absolute terms and is also not statistically significant, even though this metric tends to be the most sensitive of the fit indexes. Therefore, metric invariance is concluded for these two countries, albeit with a caveat.

We find similar results when scalar-invariance models are fitted, although here it applies to two additional countries: Bangladesh and Estonia. For these countries, a change in the SRMR points toward significant deterioration in model fit, whereas all other measures suggest acceptable deterioration. Overall, the results suggest that both factor loadings and the means of the transformational-leadership latent factor can be compared across organizations within public administration, but this conclusion is tentative for Kosovo and Nepal, as well as for Bangladesh and Estonia in the case of scalar invariance.

DISCUSSION

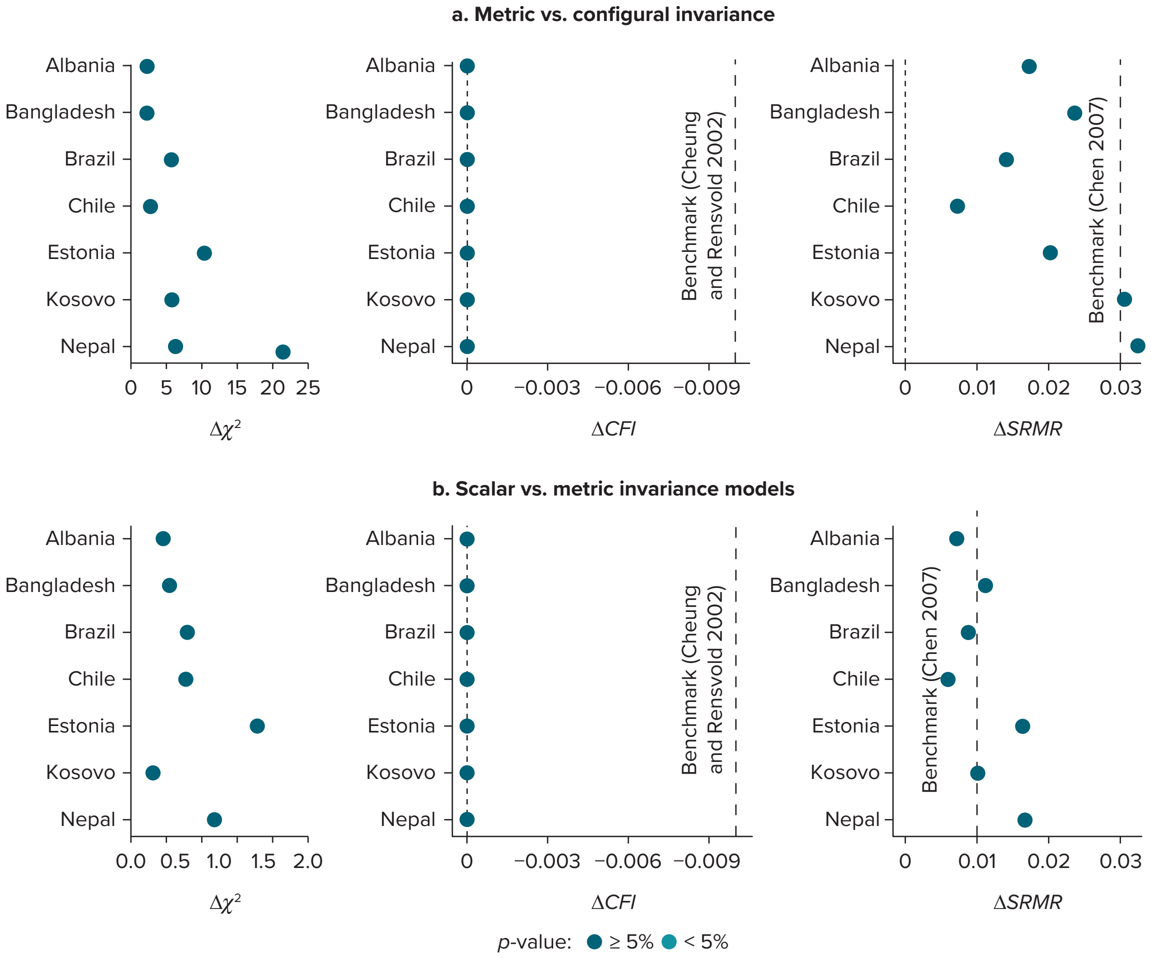
The results of the measurement-invariance analyses of the concept of transformational leadership presented above warrant a tentative two-level conclusion. First, there is strong evidence of metric invariance across countries and tentative evidence of scalar invariance. The latter conclusion can be strengthened if countries are grouped according to region or income level. In that case, full scalar invariance is observed across high-income and South Asian or lower-middle-income countries. Second, transformational leadership appears invariant, both at the level of factor loadings and latent factor means, across gender, broad education level, and organization within public administration in most of the countries studied. The evidence for the

TABLE 24.5 Distribution of Respondents within Public Administration Organizations

	No. of respondents					
Country	Mean	Median	SD	Min.	Max.	No. of organizations
Albania	215.1	183.0	141.9	83	585	15
Bangladesh	120.0	82.0	70.7	52	218	7
Brazil	292.5	165.0	305.9	57	1,062	12
Chile	522.0	382.0	449.4	87	1,520	11
Estonia	108.1	80.0	70.6	64	331	27
Kosovo	83.9	73.5	30.5	54	150	14
Nepal	97.8	83.5	61.1	55	241	8

Source: Original table for this publication.
Note: Only groups with 50+ observations are included in the analyses. SD = standard deviation.

FIGURE 24.6 Measurement Invariance across Public Administration Organizations within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



a. Metric vs. configural invariance

b. Scalar vs. metric invariance models

Source: Original figure for this publication.
Note: Only groups with 50+ observations are included in the analyses. CFI = comparative fit index; SRMR = standardized root mean squared error.

first two groups is clear-cut, and, for invariance across organizations, the only caveat that should be raised is the borderline significance of the Δ*SRMR* values in some models.

It is possible that the relatively stronger evidence of invariance within rather than across countries comes from translation differences of the survey items, rather than their differential interpretation. The language of a survey is known to affect the thought and response-forming process of survey respondents, even when care is taken—for instance, through extensive cognitive interviews—to ensure comparable understanding across languages (Peytcheva 2020). Chen (2008) suggests that such difference could also come from a propensity, observable in some cultures, to skew survey responses toward more-neutral options. These possibilities highlight the need not only to standardize the question wording and response scale, as was done here, but also for researchers to retest measurement invariance in public service surveys across further concepts and country settings, including countries with a shared language. Although safeguarding actions were taken to minimize these differences, the results suggest that some residual variation might stem from them.

This unavoidable limitation can, however, serve as a response to another potential criticism of the analyses presented above—namely, the fact that they were focused on only seven countries and a small number of questions concerned with transformational leadership. Including a larger sample of countries would be admittedly desirable, yet it is problematic precisely because the question wording and the wider context of different public service surveys are too dissimilar to warrant inclusion. Most surveys of civil servants include a leadership section, but they are not directly focused on transformational leadership, and in the cases when they are, their phrasing or response scales make comparisons difficult (chapter 18). Despite including only seven countries, this chapter is, to the best of the authors' knowledge, the first to look at the measurement invariance of the concept of transformational leadership in the public sector in a cross-country context.

CONCLUSION

Surveying civil servants is often the only feasible way to learn more about their attitudes, behaviors, and work environment. Yet survey results are challenging to interpret without context—comparisons to other countries or previous surveys, or between different demographic groups within the public service. However, researchers and policy practitioners only occasionally pause to statistically assess whether the attitudes and behaviors they want to measure—be they engagement, motivation, or leadership—are understood in the same way by different groups of civil servants.

Drawing on the concept of measurement invariance, this chapter was able to show that when it comes to the differential understanding of the concept of transformational leadership, differences in gender, education level, and organization have a very small impact. In most countries, the three leadership questions relate in the same way to the underlying concept of transformational leadership, and the mean levels of transformational leadership are also comparable across groups. The same can be said when looking across countries, even if the conclusion of the comparability of the mean levels is weakened, to an extent, when comparing countries at different income levels and, in particular, in different regions. This suggests, tentatively, that global benchmarking exercises like the GSPS have a legitimate empirical foundation. Contrasting our results with those of measurement-invariance analyses of PSM—a concept which is arguably even more culturally specific than leadership—suggests, in particular, that questions that are less culturally loaded and more factual—for instance, about management practices rather than culturally specific attitudes—might have a stronger empirical basis for comparison.

Of course, we assess only one measurement scale in our analysis and draw on data from seven countries. Thus, much fertile empirical ground for future cross-country work on measurement invariance remains to further solidify claims about what can be compared across countries and what cannot. At least four further contributions would be especially welcome in the future. First, future investigations should extend analyses of measurement invariance to other recurring topics in public service surveys. Perceptions about work environment, engagement, teamwork, compensation, turnover, performance, meritocratic practices, and harassment are components of many public service surveys. Yet the extent to which they measure the same underlying concepts across different groups of civil servants and across countries is uncertain. A second possible extension of the present analyses would include more countries in the analyses, preferably with heterogeneous geographical and economic features. A third avenue for future work would address the fact that at present, only a limited number of groupings of civil servants have been compared for measurement invariance. This chapter focused on gender, education level, and organization. Including further groupings—by age and tenure level, managerial position, and contract type—would be warranted in future studies. A fourth type of analysis would ascertain intertemporal measurement invariance. Just as the same question can measure divergent concepts across different countries, cultures, or demographic groups, it can be measurement variant across different time periods. Due to changes in social, economic, political, and, in the longer term, cultural conditions, the same survey question might come to be interpreted differently in different time periods, even when asked to the same population.

Along with further investigations of measurement invariance, researchers and practitioners wishing to compare the results of surveys of public servants would be well served by relying, at least in part, on a standardized questionnaire. One such effort is the GSPS initiative, which catalogs 20+ sets of public service survey results, along with their respective questionnaires, section names, and metadata. Including some of the standardized questions would allow for survey results to be more readily compared with other countries' results and, ultimately, for the establishment of international benchmarks against which civil servants' attitudes and behaviors could be reliably compared. Even when such comparisons are tentative, given concerns with measurement invariance, this certainly trumps comparisons of core concepts (for example, employee engagement) across countries using different measures.

NOTES

We are grateful to Daniel Rogger and Galileu Kim for helpful comments.

1. Unless justified for other reasons, in all instances countries are listed in alphabetical order. See the GSPS website (https://www.globalsurveyofpublicservants.org/) and Mikkelsen, Schuster, and Meyer-Sahling (2020) for further details on the surveys included.
2. Technical limitations, like limited access to electricity, computer, or the Internet, as well as incomplete databases of email records for civil service officials, made online surveying unfeasible in the two Asian countries (Bangladesh and Nepal) included in this chapter.
3. State or local government officials and nonadministrative public sector employees, like teachers, nurses, doctors, policemen, and the military, were thus excluded.
4. It was decided that another commonly employed measure of model fit, the root mean square error of approximation (RMSEA), would not be used in the analyses presented here. The RMSEA was introduced by Steiger and Lind (1980) and extended by, among others, Browne and Cudeck (1993) and Steiger (1998). However, it can be unreliable when comparing just-identified with overidentified models, as is done here. Using Monte Carlo simulations, Kenny, Kaniskan, and McCoach (2015) find that in models with few degrees of freedom, the RMSEA tends to be overinflated and, therefore, falsely points to bad model fit. Moreover, in close-fit models, more restricted models might counterintuitively show a decrease in the RMSEA—that is, better fit—because of the increased number of degrees of freedom (Shi, Lee, and Maydeu-Olivares 2019). Notwithstanding the above, RMSEA values point toward the same broad conclusions as the other three fit indexes consulted in the text.
5. In contrast, the lower-middle-income group is relatively homogeneous, since it is comprised of two South Asian countries.
6. In fact, it can be observed the ΔCFI is 0 across all intracountry comparisons. This is because the model fit is close to perfect, and, as a result, χ^2 is low enough as to be smaller than the number of degrees of freedom. Given the formula used to calculate the CFI, the resulting value of this fit index will always be 1 in those cases (see Bentler 1990).

REFERENCES

Bass, B. M. 1985. *Leadership and Performance beyond Expectations*. New York: Free Press.

Bentler, P. 1990. "Comparative Fit Indices in Structural Models." *Psychological Bulletin* 107 (2): 238–46. https://doi.org/10.1037/0033-2909.107.2.238.

Bentler, P. 1995. *EQS 5* [Computer program]. Encino, CA: Multivariate Software.

Berson, Y., and B. J. Avolio. 2004. "Transformational Leadership and the Dissemination of Organizational Goals: A Case Study of a Telecommunication Firm." *The Leadership Quarterly* 15 (5): 625–46. https://doi.org/1016/j.leaqua.2004.07.003.

Browne, M. W., and R. Cudeck. 1993. "Alternative Ways of Assessing Model Fit." In *Testing Structural Equation Models*, edited by K. A. Bollen and J. S. Long, 136–62. Newbury Park, CA: Sage.

Burns, J. M. 1978. *Leadership*. New York: Harper & Row.

Byrne, B. M., R. H. Shavelson, and B. Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105 (3): 456–66. https://doi.org/10.1037/0033-2909.105.3.456.

Cabinet Office. 2019. *Civil Service People Survey 2019: Technical Guide*. London: Cabinet Office, United Kingdom Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachmentdata/file/867302/Civil-Service-People-Survey-2019-Technical-Guide.pdf.

Chen, F. F. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 14 (3): 464–504. https://doi.org/10.1080/10705510701301834.

Chen, F. F. 2008. "What Happens If We Compare Chopsticks with Forks? The Impact of Making Inappropriate Comparisons in Cross-Cultural Research." *Journal of Personality and Social Psychology* 95 (5): 1005–18. https://doi.org/10.1037/a0013193.

Cheung, G. W., and R. B. Rensvold. 2002. "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 9 (2): 233–55. https://doi.org/10.1207/S15328007SEM0902_5.

Cieciuch, J., E. Davidov, P. Schmidt, and R. Algesheimer. 2019. "How to Obtain Comparable Measures for Cross-National Comparisons." *Kolner Zeitschrift fu¨r Soziologie und Sozialpsychologie* 71 (S1): 157–86. https://doi.org/10.1007/s11577-019-00598-7.

Davidov, E., H. Dülmer, J. Cieciuch, A. Kuntzm, D. Seddig, and P. Schmidt. 2018. "Explaining Measurement Nonequivalence Using Multilevel Structural Equation Modeling: The Case of Attitudes toward Citizenship Rights." *Sociological Methods & Research* 47 (4): 729–60. https://doi.org/10.1177/0049124116672678.

Davidov, E., B. Meuleman, J. Cieciuch, P. Schmidt, and J. Billiet. 2014. "Measurement Equivalence in Cross-National Research." *Annual Review of Sociology* 40 (1): 55–75. https://doi.org/10.1146/annurev-soc-071913-043137.

De Jong, M. G., J.-B. Steenkamp, and J.-P. Fox. 2007. "Relaxing Measurement Invariance in Cross-National Consumer Research Using a Hierarchical IRT Model." *Journal of Consumer Research* 34 (2): 260–78. https://doi.org/10.1086/518532.

Donkor, F., I. Sekyere, and F. A. Oduro. 2022. "Transformational and Transactional Leadership Styles and Employee Performance in Public Sector Organizations in Africa: A Comprehensive Analysis in Ghana." *Journal of African Business* 23 (4): 945–63. https://doi.org/10.1080/15228916.2021.1969191.

Downton, J. V. 1973. *Rebel Leadership: Commitment and Charisma in the Revolutionary Process*. New York: Free Press.

Erkutlu, H. 2008. "The Impact of Transformational Leadership on Organizational and Leadership Effectiveness: The Turkish Case." *Journal of Management Development* 27 (7): 708–26. https://doi.org/10.1108/02621710810883616.

Fitzpatrick, J., M. Goggin, T. Heikkila, D. Klingner, J. Machado, and C. Martell. 2011. "A New Look at Comparative Public Administration: Trends in Research and an Agenda for the Future." *Public Administration Review* 71 (6): 821–30. https://doi.org/10.1111/j.1540-6210.2011.02432.x.

French, B. F., and H. W. Finch. 2006. "Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance." *Structural Equation Modeling* 13 (3): 378–402. https://doi.org/10.1207/s15328007sem1303_3.

Fukuyama, F., D. Rogger, Z. Hasnain, K. Bersch, D. Mistree, C. Schuster, K. Mikkelsen, K. Kay, and J. Meyer-Sahling. 2022. *Global Survey of Public Servants Indicators*. https://www.globalsurveyofpublicservants.org.

García-Morales, V. J., F. J. Lloréns-Montes, and A. J. Verdú Jover. 2008. "The Effects of Transformational Leadership on Organizational Performance through Knowledge and Innovation." *British Journal of Management* 19 (4): 299–319. https://doi.org/10.1111/j.1467-8551.2007.00547.x.

Hameduddin, T., and T. Engbers. 2021. "Leadership and Public Service Motivation: A Systematic Synthesis." *International Public Management Journal* 25 (1): 86–119. https://doi.org/10.1080/10967494.2021.1884150.

Hofman, D. A., J. E. Mathieu, and R. Jacobs. 1990. "A Multiple Group Confirmatory Factor Analysis Evaluation of Teachers' Work Related Perceptions and Reactions." *Educational and Psychological Measurement* 50 (4): 943–55. https://doi.org/10.1177/0013164490504024.

Hong, S., M. L. Malik, and M.-K. Lee. 2003. "Testing Configural, Metric, Scalar, and Latent Mean Invariance across Genders in Sociotropy and Autonomy Using a Non-Western Sample." *Educational and Psychological Measurement* 63 (4): 636–54. https://doi.org/10.1177/0013164403251332.

Hooper, D., J. Coughlan, and M. R. Mullen. 2008. "Structural Equation Modelling: Guidelines for Determining Model Fit." *The Electronic Journal of Business Research Methods* 6 (1): 53–60.

Hu, L.-T., and P. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1): 1–55. https://doi.org/10.1080/10705519909540118.

Jensen, U. T., L. B. Andersen, L. L. Bro, A. Bøllingtoft, T. L. Eriksen, A.-L. Holten, C. B. Jacobsen, et al. 2019. "Article Conceptualizing and Measuring Transformational and Transactional Leadership." *Administration & Society* 51 (1): 3–33. https://doi.org/10.1177/0095399716667157.

Jreisat, J. G. 2005. "Comparative Public Administration Is Back in, Prudently." *Public Administration Review* 65 (2): 231–42. https://doi.org/10.1111/j.1540-6210.2005.00447.x.

Kenny, D. A., B. Kaniskan, and B. D. McCoach. 2015. "The Performance of RMSEA in Models with Small Degrees of Freedom." *Sociological Methods & Research* 44 (3): 486–507. https://doi.org/10.1177/0049124114543236.

Kim, S., W. Vandenabeele, B. E. Wright, L. B. Andersen, F. P. Cerase, R. K. Christensen, C. Desmarais, et al. 2013. "Investigating the Structure and Meaning of Public Service Motivation across Populations: Developing an International Instrument and Addressing Issues of Measurement Invariance." *Journal of Public Administration Research and Theory* 23 (1): 79–102. https://doi.org/10.1093/jopart/mus027.

Kroll, A., and D. Vogel. 2014. "The PSM–Leadership Fit: A Model of Performance Information Use." *Public Administration* 92 (4): 974–91. https://doi.org/10.1111/padm.12014.

La Salle, T. P., D. B. McCoach, and J. Meyers. 2021. “Examining Measurement Invariance and Perceptions of School Climate across Gender and Race and Ethnicity.” *Journal of Psychoeducational Assessment* 39 (7): 800–15. https://doi.org/10.1177/07342829211023717.

Li, C.-H. 2016. “Confirmatory Factor Analysis with Ordinal Data: Comparing Robust Maximum Likelihood and Diagonally Weighted Least Squares.” *Behavioral Research Methods* 8 (3): 936–49. https://doi.org/10.3758/s13428-015-0619-7.

Martinez, A. J. 2021. “Factor Structure and Measurement Invariance of the Academic Time Management and Procrastination Measure.” *Journal of Psychoeducational Assessment* 39 (7): 891–901. https://doi.org/10.1177/07342829211034252.

Meredith, W. 1964. “Notes on Factorial Invariance.” *Psychometrika* 29: 177–85. https://doi.org/10.1007/BF02289699.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, C. Pesti, and T. Randma-Liiv. 2018a. *Civil Service Management in Estonia: Evidence from a Survey of Civil Servants and Employees*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. Last updated October 2018. https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Pesti%20Randma-Liiv%20Estonia%20Report%20FINAL.pdf.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, H. Qeriqi, and F. Toth. 2018b. *Towards a More Professional Civil Service in Kosovo: Evidence from a Survey of Civil Servants in Central and Local Government*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Qeriqi%20Toth%20Kosovo%20Report%20FINAL.pdf.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, T. Rahman, K. M. Islam, A. S. Huque, and F. Toth. 2019. *Civil Service Management in Bangladesh: Evidence from a Survey of More Than 1,000 Civil Servants*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. https://christianschuster.net/2019.03.10.%20Bangladesh%20FOR%20PUB LICATION.pdf.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, S. K. Shrestha, B. Luitel, and F. Toth. 2018c. *Civil Service Management in Nepal: Evidence from a Survey of More than 1,200 Civil Servants*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. https://christianschuster.net/2018.12.20.%20Nepal%20FOR%20PUBLICATION.pdf.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, and A. Shundi. 2018d. *The Quality of Civil Service Management in Albania: Evidence from a Survey of Central Government Civil Servants and Public Employees*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Shundi%20Albania%20Report%20FINAL.pdf.

Mikkelsen, K. S., C. Schuster, and J.-H. Meyer-Sahling. 2020. “A Cross-Cultural Basis for Public Service? Public Service Motivation Measurement Invariance in an Original Survey of 23,000 Public Servants in Ten Countries and Four World Regions.” *International Public Management Journal* 24 (6): 739–61. https://doi.org/10.1080/10967494.2020.1809580.

Nguyen, T. T., L. Mia, L. Winata, and V. K. Chong. 2017. “Effect of Transformational-Leadership Style and Management Control System on Managerial Performance.” *Journal of Business Research* 70: 202–31. https://doi.org/10.1016/j.jbusres.2016.08.018.

OPM (Office of Personnel Management). 2019. *2019 Office of Personnel Management Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: US Office of Personnel Management, US Government.

Pandey, S. K., R. S. Davis, S. Pandey, and S. Peng. 2016. “Transformational Leadership and the Use of Normative Public Values: Can Employees Be Inspired to Serve Larger Public Purposes?” *Public Administration* 94 (1): 204–22. https://doi.org/10.1111/padm.12214.

Park, S. M., and H. G. Rainey. 2008. “Leadership and Public Service Motivation in U.S. Federal Agencies.” *International Public Management Journal* 11 (1): 109–42. https://doi.org/10.1080/10967490801887954.

Pearce, C. L., H. P. Sims Jr., J. F. Cox, G. Ball, E. Schnell, K. A. Smith, and L. Trevino. 2002. “Transactors, Transformers and Beyond. A Multi-Method Development of a Theoretical Typology of Leadership.” *Journal of Management Development* 22 (4): 273–307. https://doi.org/10.1108/02621710310467587.

Pereira, A. K., R. A. Machado, P. L. Costa Cavalcante, A. De Avila Gomide, A. Gomes Magalhaes, I. De Araujo Goellner, R. R. Coelho Pires, K. Bersch, F. Fukuyama, and A. R. Da Silva. 2021. “Government Quality and State Capacity: Survey Results from Brazil.” CDDRL Working Paper, Center on Democracy, Development, and the Rule of Law, Stanford University, Stanford, CA. https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/governancereportbrazil.pdf.

Peytcheva, E. 2020. “The Effect of Language of Survey Administration on the Response Formation Process.” In *The Essential Role of Language in Survey Research*, edited by M. Sha and T. Gabel, 3–22. Research Triangle Park, NC: RTI Press.

Putnick, D. L., and M. H. Bornstein. 2016. “Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research.” *Developmental Review* 41: 71–90. https://doi.org/10.1016/j.dr.2016.06.004.

Rosseel, Y. 2012. “Lavaan: An R Package for Structural Equation Modeling.” *Journal of Statistical Software* 48 (2): 1–36. https://doi.org/10.18637/jss.v048.i02.

Schuster, C., J. Fuenzalida, J.-H. Meyer-Sahling, K. Mikkelsen, and N. Titelman. 2020. *Encuesta Nacional de Funcionarios en Chile: Evidencia para un servicio público más motivado, satisfecho, comprometido y ético* [National Survey of Civil Servants in Chile: Evidence for a More Motivated, Satisfied, Engaged, and Ethical Public Service]. Santiago: Dirección Nacional del

Servicio Civil. https://www.serviciocivil.cl/wp-content/uploads/2020/01/Encuesta-Nacional-de-Funcionarios-Informe-General-FINAL-15ene2020-1.pdf.

Schuster, C., J. Meyer-Sahling, K. S. Mikkelsen, and C. González Parrao. 2017. *Prácticas de gestión de personas para un servicio público más motivado, comprometido y ético en Chile: Evidencia de una encuesta con 20.000 servidores públicos en Chile y otros países*. Santiago: Dirección Nacional del Servicio Civil. https://documentos.serviciocivil.cl/actas/dnsc/documentService/downloadWs?uuid=60fcd3de-fa9e-4906-9396-c7637b4cd167%20.

Seddig, D., and H. Leitgöb. 2018. "Approximate Measurement Invariance and Longitudinal Confirmatory Factor Analysis: Concept and Application with Panel Data." *Survey Research Methods* 12 (1): 29–41.

Shi, D., T. Lee, and A. Maydeu-Olivares. 2019. "Understanding the Model Size Effect on SEM Fit Indices." *Educational and Psychological Measurement* 79 (2): 310–34. https://doi.org/10.1177/0013164418783530.

Shi, D., A. Maydeu-Olivares, and Y. Rosseel. 2020. "Assessing Fit in Ordinal Factor Analysis Models: SRMR vs. RMSEA." *Structural Equation Modeling: A Multidisciplinary Journal* 27 (1): 1–15.

Steenkamp, J.-B., and H. Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25 (1): 78–90. https://doi.org/10.1086/209528.

Steiger, J. H. 1998. "A Note on Multiple Sample Extensions of the RMSEA Fit Index." *Structural Equation Modelling* 5 (4): 411–19. https://doi.org/10.1080/10705519809540115.

Steiger, J. H., and J. C. Lind. 1980. "Statistically Based Tests for the Number of Common Factors." Paper presented at the annual Spring Meeting of the Psychometric Society, Iowa City, IA.

Struening, E. L., and J. Cohen. 1963. "Factorial Invariance and Other Psychometric Characteristics of Five Opinions about Mental Illness Factors." *Educational and Psychological Measurement* 23: 289–98. https://doi.org/10.1177/001316446302300206.

Tummers, L., and E. Knies. 2016. "Measuring Public Leadership: Developing Scales for Four Key Public Leadership Roles." *Public Administration* 94 (2): 433–51. https://doi.org/10.1111/padm.12224.

Vandenberg, R. J., and C. E. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3 (1): 4–70. https://doi.org/10.1177/109442810031002.

Van De Vijveri, F. J. R., F. Avvisatiii, E. Davidoviii, M. Eidiv, J.-P. Foxv, N. Le Donnéii, K. Lekvi, B. Meulemanvii, M. Paccagnellaii, and R. Van De Schoot. 2019. "Invariance Analyses in Large-Scale Studies." OECD Education Working Paper 201, OECD, Paris. https://doi.org/10.1787/254738dd-en.

Wang, G., I.-S. Oh, S. H. Courtright, and A. E. Colbert. 2011. "Transformational Leadership and Performance across Criteria and Levels: A Meta-Analytic Review of 25 Years of Research." *Group & Organization Management* 36 (2): 223–70. https://doi.org/10.1177/1059601111401017.

CHAPTER 25

Making the Most of Public Servant Survey Results

Lessons from Six Governments

Christian Schuster, Annabelle Wittels, Nathan Borgelt, Horacio Coral, Matt Kerlogue, Conall Mac Michael, Alejandro Ramos, Nicole Steele, and David Widlake

SUMMARY

Governments around the world increasingly implement governmentwide surveys of public servants. How can they make the most of them to improve civil service management? This chapter first develops a self-assessment tool for governments that lays out the range of potential uses and benefits of public servant survey findings, arguing that public servant survey results can improve civil service management by providing tailored survey results to four key types of users (the government as a whole, individual public sector organizations, units within organizations, and the public, including public sector unions); holding government organizations accountable for taking action in response to survey results; and complementing descriptive survey results with actionable recommendations and technical assistance for how to address the survey findings to each user type. To substantiate the tool, the chapter then assesses the extent to which six governments—Australia, Canada, Colombia, Ireland, the United Kingdom, and the United States—make use of public servant survey findings. It finds that five out of six governments provide tailored survey results at both the national and agency levels, yet no government fully exploits all the potential uses and benefits of public servant surveys. For instance, not all governments provide units inside government organizations with their survey results or complement survey results with accountability or recommendations for improvement. Many governments could thus, at a low cost, significantly enhance the benefits they derive from public servant surveys for improved civil service management.

Christian Schuster is at University College London. Annabelle Wittels is an independent researcher. Nathan Borgelt is in the Australian Public Service Commission. Horacio Coral is in the National Administrative Department of Statistics Colombia. Matt Kerlogue is in the UK Cabinet Office. Conall Mac Michael is in the Department of Public Expenditure, Government of Ireland. Alejandro Ramos is in the National Administrative Department of Statistics Colombia. Nicole Steele is in the Australian Public Service Commission. David Widlake is in the UK Cabinet Office.

ANALYTICS IN PRACTICE

- Public servant data can provide important evidence for management improvements in government, but how impactful it is depends on what governments do with it. This chapter contains self-assessment tools for governments conducting surveys of public servants, with a number of relatively low-cost actions governments can take to support evidence-based reforms based on insights from public servant surveys.
- Reporting results has two core aims. The first aim is to make salient key takeaways about the strengths and weaknesses of particular organizations or units. Reporting should thus include coded management reports or appropriately coded front pages of dashboards, which provide an overview of strengths and areas for development. Second, reporting aims to enable users to explore the survey results in a bespoke manner (while ensuring the anonymity of responses). This can be done, for example, through dashboards that allow users to split questions by demographic groups—for instance, by gender or age.
- Reporting results is more impactful when it reaches the different groups that can take action based on them in a tailored manner. These groups include central government agencies (for example, the civil service agency), individual public sector organizations, individual units (or their managers) within public sector organizations, and the public, including public sector unions. Tailored results reports can enable better management responses. For instance, by providing individual public sector organizations and units with tailored survey results, public managers can more easily identify appropriate actions to tackle the specific problems of their organizations or units.
- Reporting results is also more impactful when it includes recommendations to users—such as the managers of units or organizations—on how best to address survey findings, as well as action plans for users to develop their own actions. At low cost, recommendations can be automated at the unit and organizational levels—for instance, by linking training offerings to specific survey results or providing management “checklists” to managers with certain survey results. Moreover, action plan templates can be provided to units and organizations, with suggested methodologies to develop actions based on survey results. Where more resources are available, automated recommendations and action plan templates can be complemented by tailored technical assistance—or human resource management (HRM) consultancy—provided either by a central human resource (HR) unit or an external provider to help managers turn survey findings into improvements.
- To foster the use of results, governments can introduce accountability mechanisms—for instance, through central oversight of actions taken in response to survey findings by government organizations and units, by making (anonymized) survey data available to the public and other users (such as unions) to construct “best place to work” indexes and enhance transparency around staff management in public sector institutions generally, or by introducing survey measures that capture employee perceptions of the extent to which government organizations take action in response to survey findings.

INTRODUCTION

How can governments make the most of public servant survey results for management improvements? Understanding this challenge is important. Governments around the world increasingly implement governmentwide employee surveys (see chapter 18). Implementing surveys is often costly to governments, not least in terms of the opportunity cost of staff time to respond to the survey (chapter 20). This puts a premium on making the most of public servant survey results—in other words, maximizing the benefits governments derive from public servant survey results for civil service management improvements. Yet the results from surveys of public servants do not themselves engender change. They require effective dissemination, as well as the capacity and motivation to improve civil service management based on them. This translation process is challenging. In the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS), for instance, only a minority of public servants believe that survey results will be used to make their agency a better place to work (OPM 2021).

How, then, can governments tackle this translation challenge more effectively? This chapter complements the in-depth exploration of the FEVS in chapter 26 of *The Government Analytics Handbook* with a self-assessment framework for governments to use and a case comparison of six governments to identify the range of potential approaches governments can take to maximize management improvement benefits from public servant survey results.

The conceptual starting point for the self-assessment framework consists of a series of theories of change linking public servant survey results to civil service management. The framework posits that public servant survey results can improve civil service management by enhancing the *informational basis* for civil service management improvements, the *capacity* of managers to improve civil service management, and the *motivation* of managers to improve civil service management. Tailored survey results—in the form of dashboards and reports—can improve the informational basis for management improvements for the government as a whole, for individual organizations, and for units within organizations. Publishing survey findings can provide both internal central oversight stakeholders and external stakeholders—such as the public and unions—with information to hold public managers accountable for management improvements, thus motivating managers to act on findings. Finally, complementing descriptive survey results with actionable recommendations and technical assistance in addressing the survey findings can enhance the capacity and ability of managers to pursue management improvements.

The chapter then assesses empirically the extent to which six governments—Australia, Canada, Colombia, Ireland, the United Kingdom, and the United States—make use of this range of potential uses of public servant survey findings.[1] It finds that most governments provide tailored survey results at both the national and agency levels, yet no government fully exploits all the potential uses and benefits of public servant surveys. For instance, not all governments provide units inside organizations with their survey results or complement survey results with accountability or recommendations for improvement. Many governments could thus, at a very low cost, significantly enhance the benefits they derive from public servant surveys for civil service management improvements.

INFORMATION, MOTIVATION, AND CAPACITY: HOW PUBLIC SERVANT SURVEY RESULTS CAN IMPROVE CIVIL SERVICE MANAGEMENT

The core purpose of implementing public servant surveys is to improve employee management to, ultimately, attain a stronger workforce. For instance, the United Kingdom Civil Service People Survey seeks to inspire action "to increase and maintain . . . levels of employee engagement, and staff wellbeing" (UK Government 2018). How can public servant survey results attain this aim? From a theory-of-change perspective, three mechanisms stand out.

Survey results can enhance the *informational basis* for management improvements, the *motivation* of managers to pursue management improvements, and the *capacity* of managers to pursue improvements.

These mechanisms provide a broad framework for centralized entities to assess their own efforts at inducing public sector action from surveys of public servants. In relation, chapter 26 in the *Handbook* highlights how a complementary architecture within each agency supports these actions. Thus, the two chapters can be seen together as a framework against which public sector analysts interested in generating action can benchmark the institutional environment in which their survey results are disseminated.

Business Intelligence: Improving the Informational Basis for Management Improvements through Survey Results

Better business intelligence—a stronger informational basis for management decisions—is the first and most obvious use of public servant survey results. As the Australian Public Service Commission puts it, the "results also help target strategies to build Australian Public Service (APS) workplace capability now and in the future" (Australian Public Service Commission 2021b). Or, as the government of Canada lays out:

> The objective of the Public Service Employee Survey is to provide information to support the continuous improvement of people management practices in the federal public service. The survey results will allow federal departments and agencies to identify their areas of strength and concern related to people management practices, benchmark and track progress over time, and inform the development and refinement of action plans. Better people management practices lead to better results for the public service, and in turn, better results for Canadians. (Government of Canada 2021)

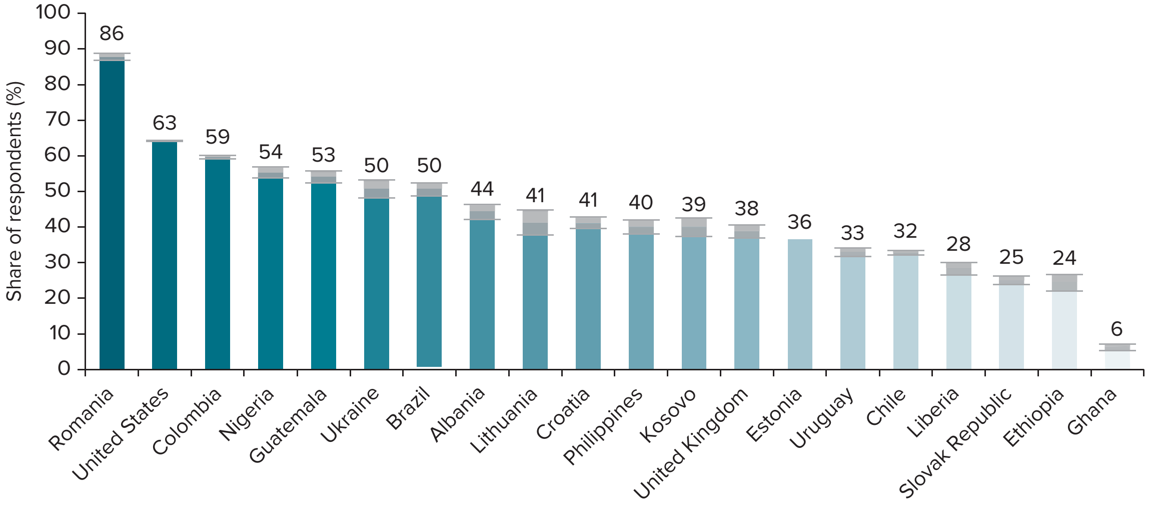
Public servant surveys can provide business intelligence on several aspects of the public administration production function (see chapter 2). They can help in understanding key public servant attitudes and how civil servants experience their work—for example, their job satisfaction or intent to stay in or leave their organization. And they can help in understanding management practices and the organizational environments shaping these public servant attitudes and experiences, such as the quality of leadership or performance management. Having data on both can also help in understanding the drivers of employee attitudes, such as engagement (namely, which management practices are statistically most important to improve engagement). In some countries where personnel databases of the civil service are highly decentralized (and centralized demographic data about the civil service are not available), surveys have also been used to create an overview of the demographic structure of the civil service (for example, India's Civil Services Survey of 2010), by asking about gender, age, or education, for instance.

A number of users can benefit from this business intelligence. First, this business intelligence can enable governmentwide reforms. Governmentwide public servant survey results can spur improvements to specific management functions if particular government shortcomings are identified. For instance, upon finding in its National Survey of Public Servants that a third of public servants indicated that they entered public service through personal or political connections, the government of Chile drafted new legislation to strengthen the merit basis of public service (Briones and Weber 2020). Governmentwide survey results can also highlight the need to improve management of and for particular groups—for instance, to track diversity and inclusion progress, as in New Zealand's government (Te Kawa Mataaho Public Service Commission 2021).

Understanding strengths and weaknesses governmentwide is often aided by international benchmarking, when survey measures across governments are comparable. For instance, if a government wants to understand whether it needs to act upon the low pay and benefits satisfaction of its staff, one potential point of reference is the pay and benefits satisfaction of public servants in other countries. The Global Survey of Public Servants (GSPS) enables such benchmarking, as illustrated below (figure 25.1). In Ghana, for instance, 6 percent of public servants are satisfied with their pay, compared to between 24 percent and 86 percent of public servants in other countries, suggesting that pay satisfaction might constitute a particular challenge in Ghana (rather than merely reflecting the general discontent of public servants with their salaries around the world).

For business intelligence from public servant surveys to be intelligible and actionable, it needs to be presented in a manner that increases awareness and understanding of key areas measured by the survey and, in particular, the key priority areas for action in light of the survey results. It also needs to allow governmentwide users to explore topics of interest, such as how survey responses differ by key groups of public servants—for instance, between men and women (cf. Pandey and Garnett 2007). Understanding key areas for action requires reporting results either in a management report or in appropriately coded dashboards, which front-page key areas of strength and development. Complementing management reports with

FIGURE 25.1 Share of Public Servants Satisfied with Their Pay and/or Total Benefits, Various Countries



Source: Fukuyama et al. 2022.
Note: Years of measurement vary by country. Colors denote the extent of job satisfaction, with darker shades signifying greater job satisfaction. The gray vertical bars denote 95% confidence intervals.

dashboards allows users to easily explore aggregate data splits—for instance, by demographic group. User exploration is also aided by allowing ad hoc requests from central government agencies (such as ministries of finance) for particular tailored survey data analyses that go beyond what is displayed in a dashboard—for instance, particular regression analytics to understand the drivers of gender gaps in different organizations. Finally, central business intelligence is further strengthened when public servant survey results are integrated with other human resources (HR) data sources—for instance, in an HR dashboard that places survey results side-by-side with indicators such as retention, sick leave, number of applicants for public sector jobs, and gender pay gaps.

Second, public servant survey results can enable reforms at the organizational level by disaggregating results to organizational averages, benchmarking organizations in the public sector against each other, and allowing organizations to understand differences in the experiences and responses of different groups inside an organization. Providing organization-level business intelligence matters because differences in employees' experiences between public sector organizations inside a government are often larger than differences between governments (Meyer-Sahling, Schuster, and Mikkelsen 2018). Governmentwide reforms alone thus often miss priorities for improvement in particular public sector organizations. Drawing on its organization-level results, to cite just one example, the Primary Care Division of the Scottish government identified key areas for improvement (including empowerment of staff and team spirit) in its 2012 Civil Service People Survey—in which it scored 54 percent in engagement—and it acted upon the survey findings to increase engagement to 78 percent in 2014 (Cabinet Office 2015). Management reports for each organization, appropriately coded dashboards, which front-page key areas of strength and development for each organization, and dashboards to allow organizations to explore aggregated responses of different demographic groups inside the organization can provide the business intelligence for such organizational improvements.

Third, public servant survey results can enable improvements at the level of units or divisions inside organizations by disaggregating results to the unit level and making them accessible to unit managers through management reports and dashboards.[2] Unit-level reporting is important because differences in key indicators between units inside organizations—such as in the quality of leadership and employee engagement—are often as large as differences between organizations (see chapter 20). The UK Cabinet Office's Social Investment and Finance Team (SIFT), for instance, excelled relative to other teams inside the Cabinet

Office in employee engagement through “tight-loose” leadership—tightness around the mission but delegation in allowing members of the team autonomy to achieve the mission (Cabinet Office 2016).

Capacity: Enhancing the Ability of Managers to Undertake Management Improvements

Descriptive survey results can identify key strengths and weaknesses in staff management in the government, a particular government organization, a unit inside an organization, or a particular demographic group of public servants. By themselves, however, survey results are not prescriptive: they do not identify how best to address survey findings. In other words, they identify strengths and weaknesses but not managerial actions for improvement. It is thus important to complement survey results with either a process to identify improvements or the identification of specific substantive improvements.

Approaches that focus on an improved process can take the form of methodologies to develop action plans, with templates and, potentially, technical assistance (for example, from a civil service agency or a management consultancy) to help government organizations or units undertake improvements. This approach is typical of employee engagement consultancies, which have developed standardized toolkits based on staff survey results (see, for example, Gallup 2022).

The substantive approach couples the presentation of survey results with specific recommendations for improvement based on the survey results to facilitate turning results into action. In country-level reports, these can be qualitative and detailed, based on inferring key management improvements from the data (see, for example, Schuster et al. 2020). At lower levels of disaggregation—for organizations and, in particular, units where hundreds of results reports are needed—recommendations can be automatically coded to be added to the results presentation. For instance, Google’s approach to people analytics flags specific training offerings to managers based on survey results for their units (Penny 2019).

Accountability: Motivating Managers to Undertake Management Improvements

Public servant survey results can make transparent the quality of management in specific units or organizations or in the government as a whole. Where transparency is coupled with accountability for management improvements, it can provide additional motivation to managers to pursue improvements (beyond their intrinsic motivation).

Accountability can come, first of all, from the bottom up: public servant surveys provide employees with a voice to raise concerns about their experiences with and perceptions of management, their team, and their organizational environment. For employees—or public sector unions as their representatives—to hold government organizations accountable for management improvements, results need to be published, at least at an aggregate level. Providing employees with a voice is an explicit objective of most public servant surveys. For instance, the Australian government stresses that their survey “is an opportunity for employees to tell the Australian Public Service Commissioner and Agency heads what they think about working in the APS” (Australian Public Service Commission 2021b). Accountability to employees can be fostered by measuring employee perceptions of the extent to which their organization is taking action to respond to survey findings. For instance, the UK Civil Service People Survey asks respondents about their agreement with the statement “Where I work, I think effective action has been taken on the results of the last survey” (Cabinet Office 2019).

Accountability can also come from the outside—the media, public sector watchdogs, and researchers—when data, including organization-level data, are made public.[3] For instance, the Partnership for Public Service—a US nonprofit—generates the Best Places to Work in the Federal Government index based on published US public servant survey results, benchmarking public sector organizations in the United States and rendering salient organizations that perform poorly (Partnership for Public Service 2021). This type of transparency and publicity about poor performance may, in a poorly performing organization, motivate action to improve its ranking.

Similarly, the media can act as an external accountability mechanism to motivate improvements when data are made public. For instance, in Australia, low staff morale and dissatisfaction with leadership in the Department of Home Affairs made headlines in main news outlets (Doran 2019). Similarly, in Ireland, the media reported that only a small fraction of civil servants thought that poor staff performance was adequately addressed in their departments (Wall 2021).

Researchers can add a further layer of accountability, particularly when anonymized microdata from survey respondents are made available. This precludes the selective reporting of results by allowing researchers to analyze the anonymized raw data. It can thus further improve the aforementioned informational basis for management improvements by fostering a body of research work about a government's public service. To illustrate, a recent review identified 48 research articles using published microdata from the FEVS (Resh et al. 2019). Among these studies, a number have assessed diversity management in the US government based on these microdata. They have found, for instance, that employees in organizations with greater racial diversity tend, all else being equal, to report lower job satisfaction. Yet they have also found that when diversity is managed well, employees in organizations with more racial diversity report greater job satisfaction (Choi 2009; Choi and Rainey 2010). This makes transparent both a potential challenge in the US government (lower job satisfaction in more diverse institutions) and the effectiveness of diversity management as a solution.

Accountability and oversight can, of course, also be internal. For instance, heads of organizations can hold managers of units inside their organizations accountable for improvements based on their results, and central oversight agencies (such as ministries of finance or civil service agencies) can hold public sector organizations accountable for improvements. As detailed below, in the Irish government, a dashboard tracks the actions of each government organization in response to the public servant survey, while Canada uses a management accountability framework (MAF) to assess the progress made by organizations in management practices, including those identified in the employee survey.

In short, public servant survey results can foster management improvements through better business intelligence, greater managerial motivation, and an increased capacity to improve. Governments can maximize each of these uses by generating customized reports for the government as a whole, each organization, and each unit, ensuring that users can both explore aggregate data easily and access key findings for their organization/unit/government.

Governments can also complement descriptive survey results with recommendations, action plans, and methodologies to turn survey results into improvements and accountability mechanisms inside the government and externally—including publishing results and data—to motivate action. The next section will compare the extent to which six governments with long-standing public servant surveys have made use of these approaches to maximize the benefits of public servant survey results.

TO WHAT EXTENT ARE GOVERNMENTS MAKING FULL USE OF PUBLIC SERVANT SURVEY RESULTS? BENCHMARKING SIX GOVERNMENTS

To what extent are governments making full use of public servant survey results? This section compares the approaches taken by six governments with long-standing (at least three iterations) governmentwide public servant surveys: Australia, Canada, Colombia, Ireland, the United Kingdom, and the United States. It does so by benchmarking the actions taken by each government against each of the potential uses of public servant survey results identified in the previous section of this chapter. Table 25.1 summarizes this comparison and the self-assessment framework, which can be used by other governments to identify actions that could further enhance their use of public servant survey results. Of course, there may be variations within each category across the six governments we have reviewed. For simplicity, we code each country in the framework for each category according to a binary: *exists* vs. *does not exist*.

Looking first at business intelligence, the comparison shows that governments generally produce country-level results reports. With one exception, they also produce agency-level reports (that is,

TABLE 25.1 Comparing Country Approaches to Making the Most of Survey Results

	Australia	Canada	Colombia[a]	Ireland	United Kingdom	United States
Information provided to central government						
National results report	🟩	🟩	🟩	🟩	🟩	🟩
Dashboard for customized queries	🟩	🟩	🟥	🟥	🟩	🟩
Ad hoc analyses on topics of interest to central government	🟩	🟩	🟥	🟩	🟩	🟩
Survey results integrated in HR business intelligence platform or regular report with other HR data (for example, turnover or mobility)	🟩	🟩	🟥	🟥	🟥 Only ad hoc in select agencies	🟩
Information provided to government organizations						
Results report for each agency	🟩	🟩	🟩	🟩	🟥	🟩
Dashboard with results of agency and internal comparisons	🟩	🟩	🟥	🟥	🟩	🟥
Rapid-response analyses on topics of interest in response to requests from particular agencies	🟩	🟩	🟥	🟥	🟥	🟥
Information provided to units inside government organizations						
Results report for each unit within the agency	🟩	🟩	🟥	🟥	🟥	🟩
Dashboard with results of units and customized queries	🟩	🟩	🟥	🟥	🟩	🟥
Capacity to take action based on survey results						
National results report with recommendations for management improvement	🟥 In accompanying reports	🟥 In accompanying reports	🟥	🟥	🟥	🟥
Organizational reports with recommendations for improvement	🟥	🟥	🟥	🟥	🟥	🟥
Action plan templates and methodologies to help organizations take	🟩	🟥	🟥	🟩	🟩	🟩
action based on survey findings	🟩	🟥	🟥	🟩	🟩	🟩
Results presentations and technical assistance to help agencies take action based on survey results	🟩	🟩	🟥	🟩	🟥	🟥
Accountability: Information made available to the public						
National results report or table	🟩	🟩	🟩	🟩	🟩	🟩
Dashboard for customized queries	🟩	🟩	🟥	🟥	🟥	🟥 Previously the Unlock Talent dashboard

(continues on next page)

TABLE 25.1 Comparing Country Approaches to Making the Most of Survey Results *(continued)*

	Australia	Canada	Colombia[a]	Ireland	United Kingdom	United States
Institutional results reports or dashboards	🟩	🟩	🟩	🟥	🟩 In a spreadsheet	🟩
Anonymized individual-level microdata	🟩	🟥 On request	🟩	🟥 On request	🟥	🟩
Bottom-up and top-down accountability for using survey results						
Central government mechanism to hold organizations accountable for acting on results	🟥	🟩	🟥	🟩	🟥	🟩
Survey measuring whether public servants perceive their organization is taking action to address results	🟥	🟥	🟥	🟥	🟩	🟩

Source: Original table for this publication.

Note: In the table, green cells indicate Yes and red cells indicate No. To make the analysis tractable, the authors have delineated a binary conception of whether countries undertake the focal practices or not. Though there may be variation within each category and country, this provides a generalized assessment of the information available from public data and clarifications received from countries.

a. Colombia counts on a comprehensive management dashboard that covers human resource management, enables comparisons over time, and contains recommendations for each organization and action plans (DAFP 2022). However, this dashboard currently does not integrate results from Colombia's public servant survey. HR = human resources.

reports for individual government organizations), enabling each organization to understand its strengths and weaknesses based on survey results. There is a greater divergence when it comes to unit-level reports. Australia, Canada, the United Kingdom, and the United States disaggregate data to the unit level, enabling heads of units or divisions inside a government organization to understand their strengths and weaknesses. As this disaggregation to unit-level reports or dashboards multiplies the number of potential users of the data, it is an important low-cost avenue for greater management impact of the survey in countries that currently lack this disaggregation. Governments also differ in the extent to which they create dashboards that allow users to easily explore the results along the margins most interesting to them—for instance, by splitting indicators by demographic groups (such as gender) for the government as a whole or particular organizations. As the creation of such dashboards need not be costly—for instance, if free online platforms such as Tableau Public are used—this represents a second low-cost way for many governments to enhance the business intelligence users derive from survey results. All governments, with one exception, also undertake bespoke analyses of the data for users—for instance, in response to requests from the ministry of finance or other particular organizations with specific interests. Finally, Australia, Canada, and the United States integrate public servant survey results systematically with other HR data—such as data on turnover—in their reporting to generate a more comprehensive overview of HR strengths and weaknesses.

In terms of enhancing the capacity to turn survey results into actions at the national level, only Australia and Canada accompany their descriptive survey results with specific management improvement recommendations in accompanying briefings and reports (though not in the survey results directly). At the agency level, two countries rely on action plan templates to help organizations with a process to turn survey results into action. Finally, in four of the countries, the center of government provides results presentations or technical assistance to individual public sector organizations to help them turn survey results into action.

In terms of external accountability, all countries publish country-level results. All governments except for one also make institution-level reports public. However, only Australia and Canada provide the public with access to a dashboard to explore the data, while three countries publish the anonymized microdata (and a further two countries make the data available upon request to researchers under certain conditions). Similarly, three governments have institutionalized center-of-government mechanisms for holding government organizations accountable for improvements based on survey results, and only a minority of governments measure the extent to which civil servants believe that their organizations are taking effective action based on survey results. In many countries, stronger external transparency and internal accountability mechanisms to motivate managers to take action based on survey results could thus be considered.

Table 25.1 highlights both commonalities and variations between countries in the extent to which survey results are used—and opportunities to further this use. To make these opportunities more actionable, the next subsections showcase specific examples of how governments approach each of these uses.

First, a brief note on the capacity to undertake these actions is due. While this chapter does not focus on *why* different governments do not adopt some of the potential uses of survey results, a plausible conjecture is the differential organizational setup of public servant surveys across countries. This differential organizational setup generates differences in, for instance, organizational capacity to deliver management reports, dashboards, and bespoke analyses. In the United Kingdom and Australia, data collection is contracted out, as is, for instance, the production of results dashboards. In Colombia, the national statistical agency handles the process, while Canada and Ireland use a hybrid approach whereby surveys are conducted through a partnership between civil service departments and the national statistics agency. In the United States, the survey is conducted by the OPM, which is the US federal civil service department. Where surveys are conducted in-house, the ability to deliver dashboards and coded reports is conditioned by the data analytics staff's capacity in the government agency in charge of the survey.

Information Provided to the Central Government

As noted above, all governments generate national results. They do so in different ways, however. In the United Kingdom, a slide deck is produced for the most senior officials (the cabinet secretary, the civil service's chief operating officer, and departmental permanent secretaries) and HR directors in departments. They are also given access to the interactive dashboards so they can explore the results in more detail. In some previous years, a slide deck visually highlighting key findings and showing the progression from the past year was also made public (figure 25.2), and the head of the civil service provided a write-up of highlights (for example, Heywood 2017). This is not currently the case, however. The Colombian government, similarly, presents national results in a slide deck together with a press release with key findings (DANE 2022).

By contrast, Ireland and the United States present national results *reports*. Both highlight up front the most positive and the most challenging results. The Irish report does this by theme (figure 25.3); the US report lists items with the highest and lowest agreement (as key areas of strength and development) (OPM 2021).

As a further means of highlighting key strengths and weaknesses, the Irish report also contains international comparators (figure 25.4)—a practice otherwise underutilized by governments, in light of the comparator data available through the GSPS (Fukuyama et al. 2022).

Finally, Australia presents results not only in a slide deck (Australian Public Service Commission 2021c) and a summary write-up of results (Australian Public Service Commission 2021a) but also in an annual State of the Service Report that integrates employee survey results with other workforce data—for instance, on gender pay gaps, diversity, and mobility—to provide a comprehensive HR diagnostic, often focused on key themes (Australian Public Service Commission 2021d). Figure 25.5 showcases an example figure from the State of the Service Report, which integrates findings from the country's public servant survey with external labor market data to better understand skills shortages in the public sector. Similarly, Canada and the United States integrate HR and survey data in their reporting. For instance, in the United States, employee survey results are, as part of the President's Management Agenda (PMA), provided to the White House together with HR metrics, such as staffing and quit rates. Survey and HR data were also integrated into a

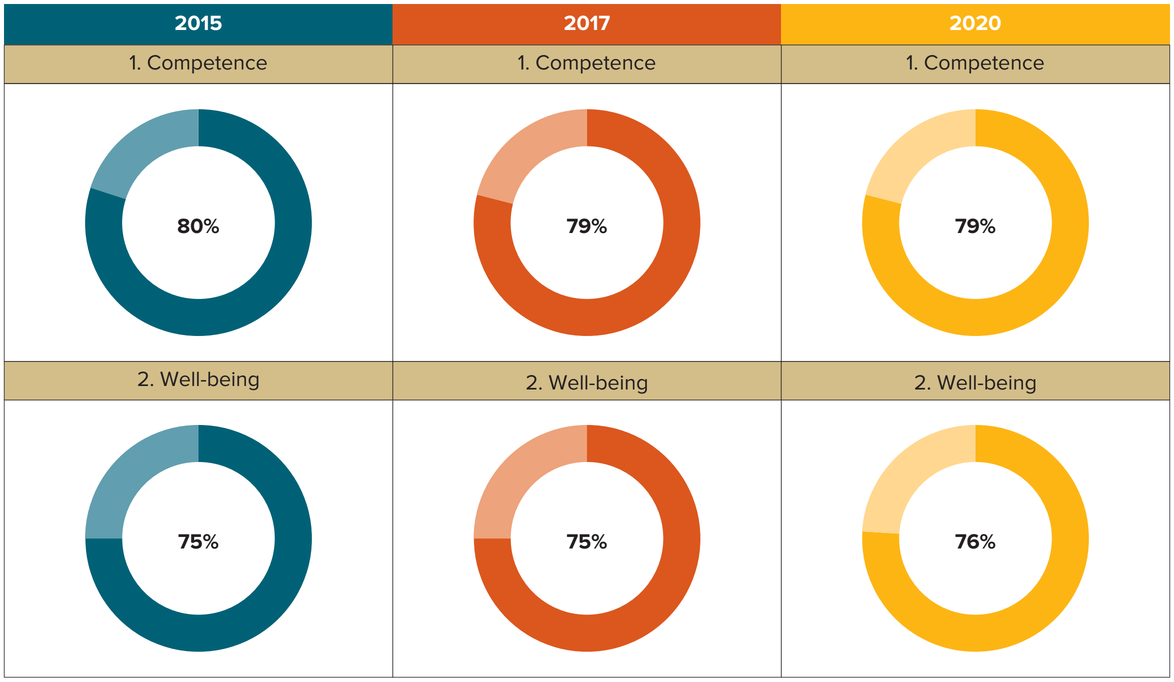
FIGURE 25.2 Results Report from the UK Civil Service People Survey

Employee Engagement Index	63%	▲	Difference from CSPS 2018 +1
My work	77%	▲	Difference from CSPS 2018 +1
Organisational objectives and purpose	83%	=	Difference from CSPS 2018 0
My manager	71%	▲	Difference from CSPS 2018 +1
My team	82%	▲	Difference from CSPS 2018 +1
Learning and development	55%	▲	Difference from CSPS 2018 +1
Inclusion and fair treatment	79%	▲	Difference from CSPS 2018 +1
Resources and workload	74%	▲	Difference from CSPS 2018 +1
Pay and benefits	34%	▲	Difference from CSPS 2018 +3
Leadership and managing change	49%	▲	Difference from CSPS 2018 +2

Source: Cabinet Office 2019.
Note: CSPS = Civil Service People Survey.

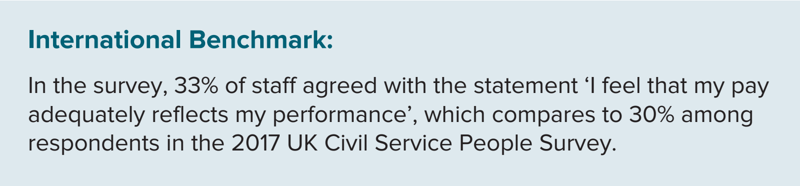
FIGURE 25.3 Results Report from Ireland, Top 5 Positive Results

Positive Results – Top 5



Source: Department of Public Expenditure and Reform 2020.

FIGURE 25.4 International Benchmarking in Results Report from Ireland



International Benchmark:

In the survey, 33% of staff agreed with the statement ‘I feel that my pay adequately reflects my performance’, which compares to 30% among respondents in the 2017 UK Civil Service People Survey.

Source: Department of Public Expenditure and Reform 2017.

dashboard—Unlock Talent—that allowed users to compare agencies and units in survey results (for example, engagement) and HR data. Funding for the dashboard has run out and, at the time of the writing of this chapter, the US government is developing a replacement.

In short, all countries report national results. Four of the six countries do well to visualize highlights up front, giving stakeholders a sense of key strengths and areas for improvement. Ireland also uses international comparisons to further contextualize strengths and challenges, while Australia and Canada are the only countries to systematically integrate employee survey and other workforce data for a more comprehensive, regular HR diagnostic.

Governments also differ in the extent to which they enable government users to further explore data beyond the results report—by making a results dashboard available or conducting on-demand, bespoke analysis of the data. Australia, Canada, the United Kingdom, and the United States use dashboards to enable users to explore the (aggregate) data in a more customized way—for instance, by comparing responses of different demographic groups in different state institutions. These dashboards can be relatively low cost, as in the case of the Employee Viewpoint Survey Analysis and Results Tool (EVS ART) in the United States (see chapter 9, case study 9.3 in chapter 9, and chapter 26) or Canada’s Power BI dashboard (figure 25.6). Canada’s Power BI dashboard allows users to compare indicators, organizations, and trends over time. All data are aggregated as percentages for each response option (for example, the percentage of respondents who answered “strongly agreed” or “agreed”) (Government of Canada 2020b).

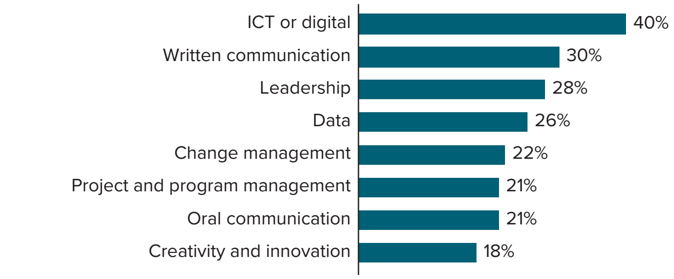
Canada also produces dashboards focused on specific groups of public servants—such as Indigenous people, women, persons with disabilities, or LGBTQ+ employees. Figure 25.7, for instance, shows the dashboard for persons with a disability. Canada thus provides users with accessible overviews of results for groups of public servants with particular needs or particularly concerning results.

A subset of governments also conducts more bespoke, on-demand analysis of data. For instance, the Australian Public Service Commission analyzes and reports on employee survey data in bespoke reports for specific purposes. These are typically reports for internal civil service use and consideration but may also comprise reports for public release. Areas from across the civil service that require employee survey results to inform their work and activities can request these from the commission. The commission then prepares responses to these requests for information. In Canada and the United States, analytical reports can be requested by participating agencies. The OPM also publishes a series of special reports—for instance, on women in public service, employee engagement drivers, and millennials in public service (OPM 2022). Ireland also occasionally commissions academics and consultants to provide more in-depth analytical reports to provide further insight into areas that were identified as needing intervention (Department of Public Expenditure, National Development Plan Delivery and Reform 2022).

FIGURE 25.5 State of the Service Report from Australia

APS insights

Top skill shortages identified by employees

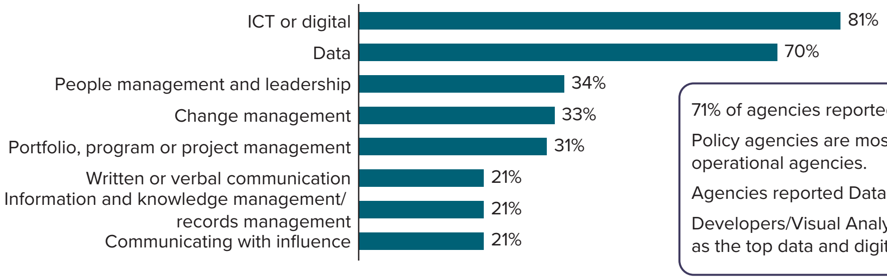


43% of employees reported skills or capability gaps within their immediate workgroup. Employees in Organisation Leadership (62%), Intelligence (55%) and Digital (50%) roles were the most likely to report capability gaps.

SES employees are most likely to report capability gaps (64%), followed by EL (53%) and APS1-6 (37%) employees. Trainees/Graduates are the least likely.

Source: 2021 APS Employee Census

Top skill shortages identified by agencies



71% of agencies reported skill or capability gaps within their agency.

Policy agencies are most likely to report gaps, followed by larger operational agencies.

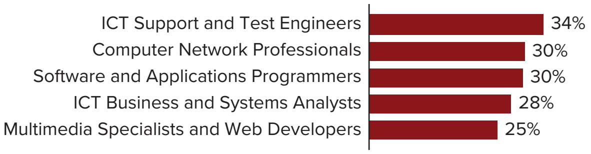
Agencies reported Data Analysts, Business Intelligence

Developers/Visual Analytics Specialists and Cyber Security specialists as the top data and digital skill shortages.

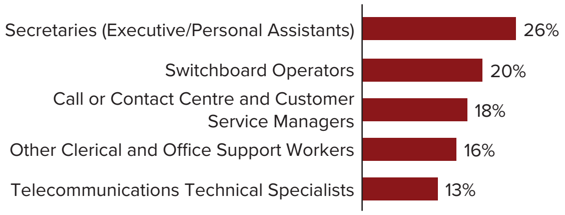
Source: 2021 APS Agency Survey

Labour market insights

Labour market growth rate for emerging roles in demand (+)
(based on 5-year labour market intelligence, roles relevant to the APS)



Labour market role decline rate (-)
(based on 5-year labour market intelligence, roles relevant to the APS)



Top 5 job roles with the highest projected demand growth in Australia all **reflect niche digital professional skillsets.** Economy-wide shortages for these skills are expected to exacerbate to 2025.

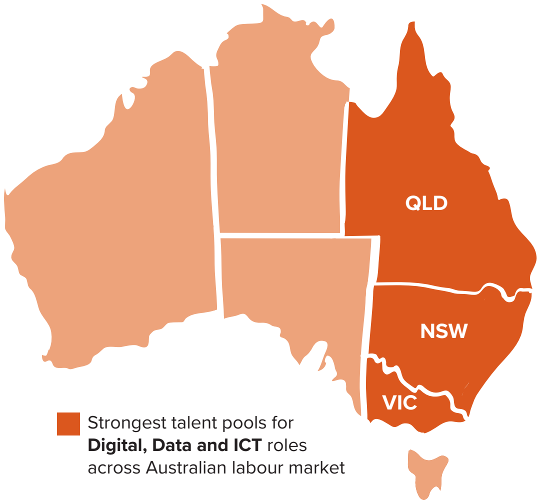
Significant increase in demand on the labour market, and an under-supply of specialists with technical skillsets highlight the need for the APS to build its talent pipelines into these roles.

A strategic workforce planning approach and focused reskilling efforts are required, to transition staff in transactional roles projected to decline.

Source: National Skills Commission employment projections to November 2025

Labour market talent pool for top skill shortages

Digital, Data and ICT talent pools across Australia


QLD
NSW
VIC
Strongest talent pools for Digital, Data and ICT roles across Australian labour market

59% of employees in Digital and ICT roles in **Canberra**. The strongest labour market talent pools for ICT, Data and Digital roles are in **QLD, NSW** and **VIC** making up over 70% of the national talent pool.

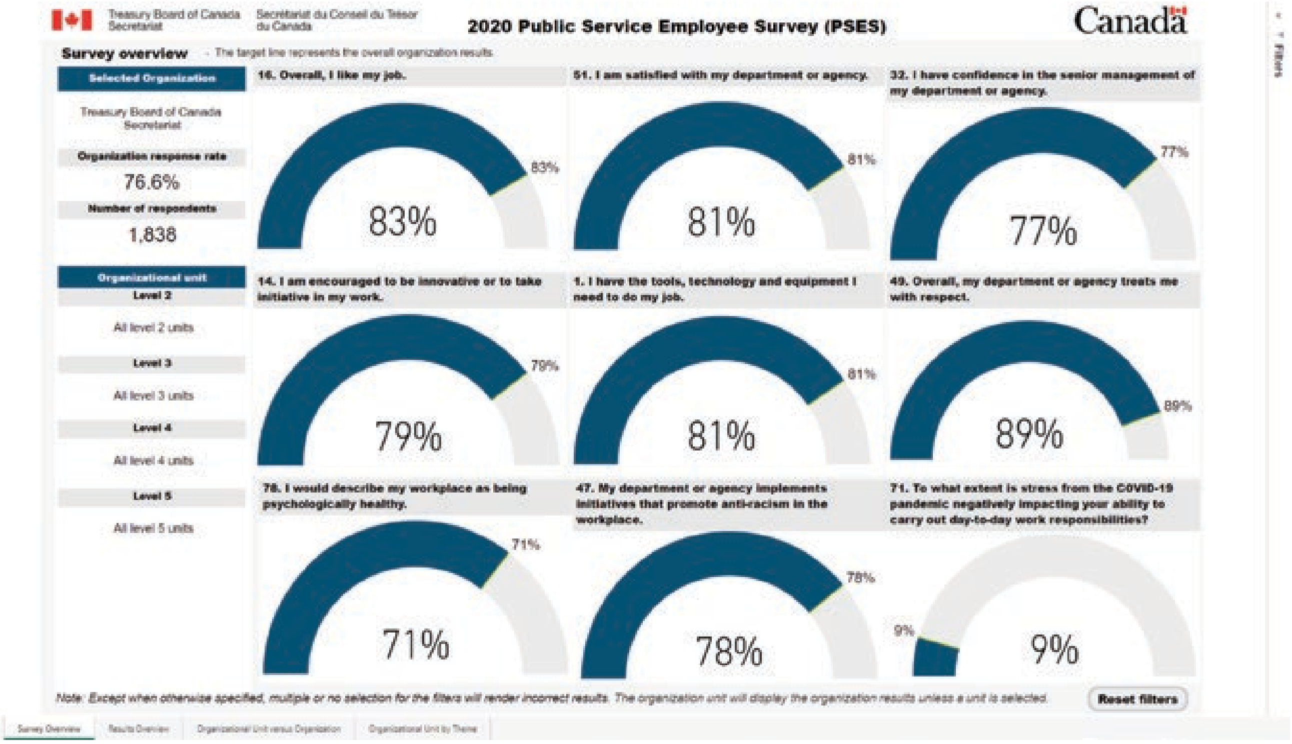
The majority of agencies (70%) experience shortages after trying to recruit for these roles predominantly in Canberra.

APS agencies should consider their current location strategy for specialist roles.

Source: Australian Job Outlook and 2021 APS Employee Census

Source: Australian Public Service Commission 2021d.
Note: APS = Australian Public Service; EL = executive level; ICT = information and communication technology; SES = senior executive service.

FIGURE 25.6 Canada Public Service Employee Survey Dashboard



Treasury Board of Canada Secretariat | Secrétariat du Conseil du Trésor du Canada

2020 Public Service Employee Survey (PSES)

Canadä

Survey overview - The target line represents the overall organization results.

Selected Organization: Treasury Board of Canada Secretariat

Organization response rate: 76.6%

Number of respondents: 1,838

Organizational unit

Level 2: All level 2 units

Level 3: All level 3 units

Level 4: All level 4 units

Level 5: All level 5 units

16. Overall, I like my job. 83%

51. I am satisfied with my department or agency. 81%

32. I have confidence in the senior management of my department or agency. 77%

14. I am encouraged to be innovative or to take initiative in my work. 79%

1. I have the tools, technology and equipment I need to do my job. 81%

49. Overall, my department or agency treats me with respect. 89%

78. I would describe my workplace as being psychologically healthy. 71%

47. My department or agency implements initiatives that promote anti-racism in the workplace. 78%

71. To what extent is stress from the COVID-19 pandemic negatively impacting your ability to carry out day-to-day work responsibilities? 9%

Note: Except when otherwise specified, multiple or no selection for the filters will render incorrect results. The organization unit will display the organization results unless a unit is selected.

Source: Government of Canada 2020b (example screenshot).

Information Provided to Government Organizations

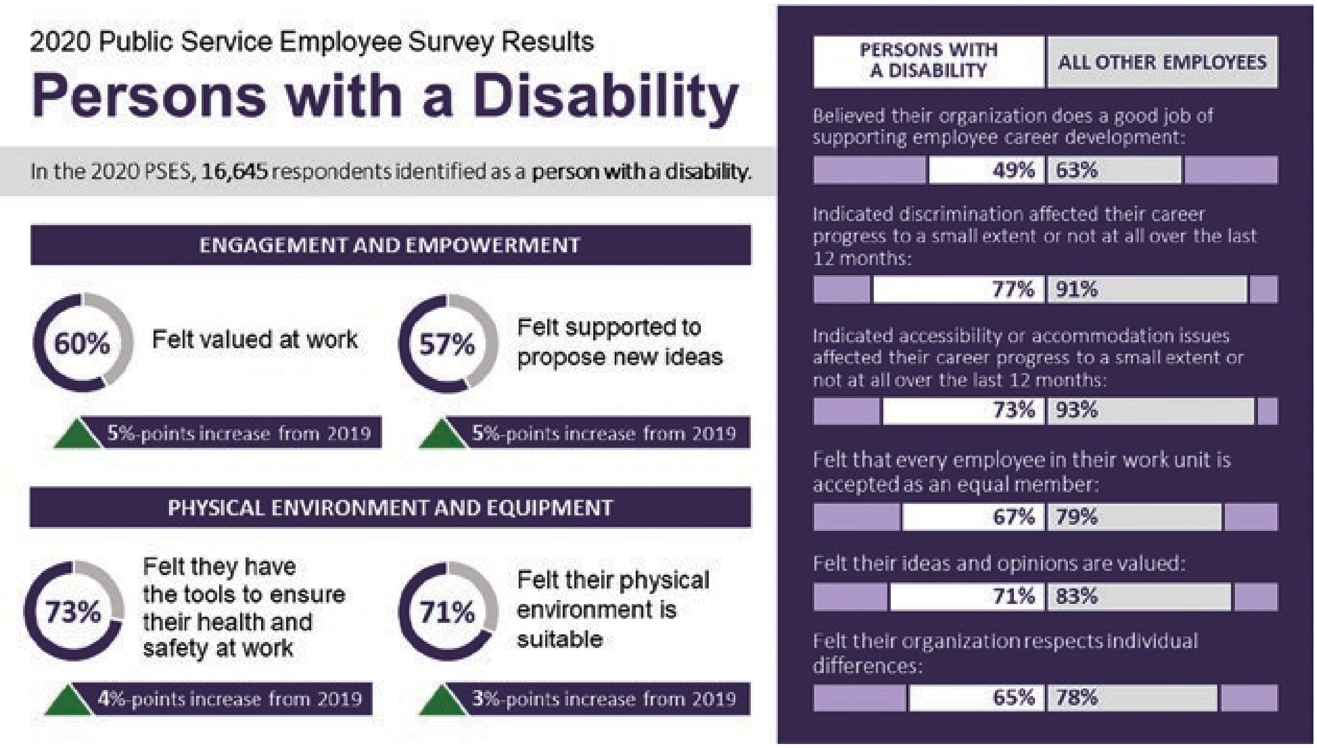
All governments provide results data at the organizational level to participating government organizations. The format and accessibility of these agency-level results, however, differ. In Colombia and the United States, data are presented in tables or data files (see chapter 9, case study 9.3 in chapter 9, and chapter 26 for greater detail on the EVS ART approach). Ireland produces bespoke reports for each agency, accompanied by an "at a glance" dashboard (see more on this below). These reports are descriptive. Agencies are encouraged to draw their own conclusions for programs of change based on the results.

Australia, Canada, and the United Kingdom offer online dashboards to agencies through which they can filter results and explore the parts of the data relevant to them. Dashboards have privacy protection safeguards programmed into them, such as not allowing for cross-tabulations below a certain number of employees or only providing a subset of open-ended responses for teams that are very small. Figure 25.8 visualizes the UK Civil Service People Survey's (internal) dashboard, which the United Kingdom contracts from Qualtrics. Australia, similarly, uses a contractor to generate an easily accessible online dashboard that allows splits at the agency and subdivision level by, for instance, gender and technical expertise for each agency and subdivision. Canada built its own dashboard with Power BI (figure 25.6).

Information Provided to Units inside Government Organizations

Canada, Australia, the United Kingdom, and the United States also generate unit-level results—for instance, by generating team-level reports accessible to each team, as in the United Kingdom's (figure 25.8) and

FIGURE 25.7 Canada Public Service Employee Survey Dashboard for Persons with a Disability



2020 Public Service Employee Survey Results

Persons with a Disability

In the 2020 PSES, **16,645** respondents identified as a **person with a disability.**

ENGAGEMENT AND EMPOWERMENT

- **60%** Felt valued at work — ▲ 5%-points increase from 2019
- **57%** Felt supported to propose new ideas — ▲ 5%-points increase from 2019

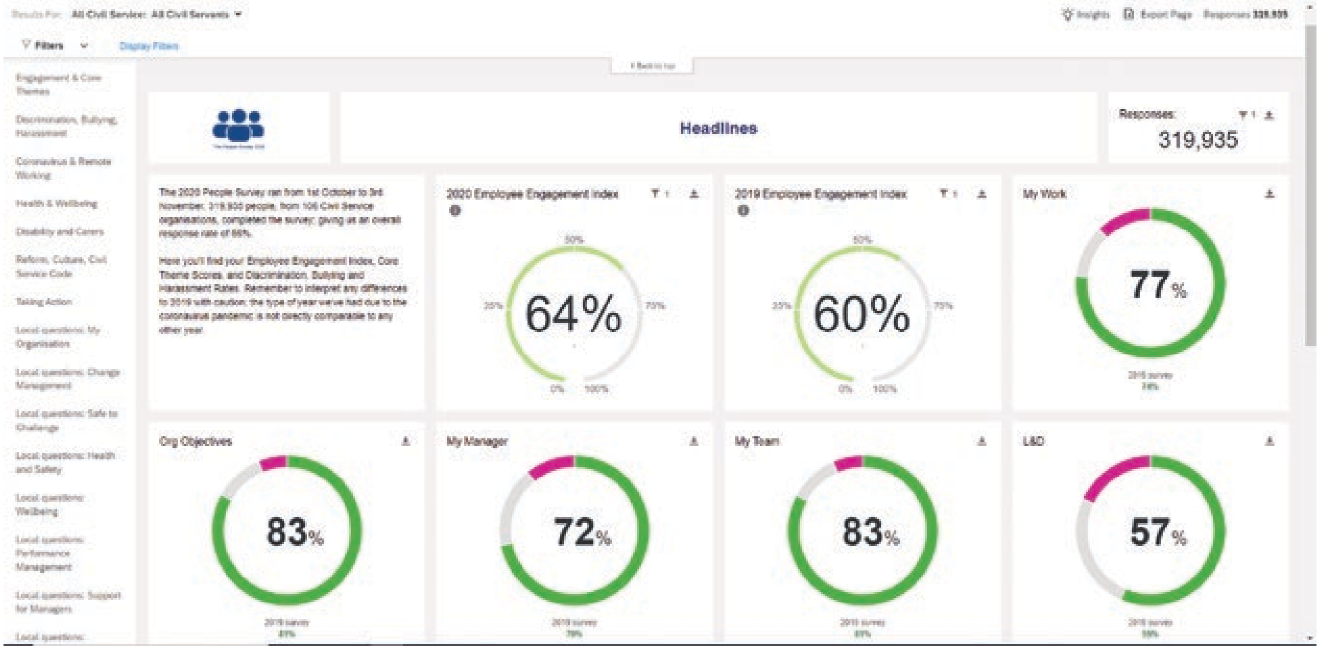
PHYSICAL ENVIRONMENT AND EQUIPMENT

- **73%** Felt they have the tools to ensure their health and safety at work — ▲ 4%-points increase from 2019
- **71%** Felt their physical environment is suitable — ▲ 3%-points increase from 2019

	PERSONS WITH A DISABILITY	ALL OTHER EMPLOYEES
Believed their organization does a good job of supporting employee career development:	49%	63%
Indicated discrimination affected their career progress to a small extent or not at all over the last 12 months:	77%	91%
Indicated accessibility or accommodation issues affected their career progress to a small extent or not at all over the last 12 months:	73%	93%
Felt that every employee in their work unit is accepted as an equal member:	67%	79%
Felt their ideas and opinions are valued:	71%	83%
Felt their organization respects individual differences:	65%	78%

Source: Government of Canada 2020b (example screenshot).
Note: PSES = Public Service Employee Survey.

FIGURE 25.8 United Kingdom Civil Service People Survey Results Dashboard for Organizations and Teams


Results For: All Civil Service: All Civil Servants
Insights
Export Page
Responses 319,935
Filters
Display Filters
Back to top
Engagement & Core Themes
Discrimination, Bullying, Harassment
Coronavirus & Remote Working
Health & Wellbeing
Disability and Carers
Reform, Culture, Civil Service Code
Taking Action
Local questions: My Organisation
Local questions: Change Management
Local questions: Safe to Challenge
Local questions: Health and Safety
Local questions: Wellbeing
Local questions: Performance Management
Local questions: Support for Managers
Local questions:
Headlines
Responses:
319,935
2020 Employee Engagement Index
50%
25%
64%
75%
0%
100%
2019 Employee Engagement Index
50%
25%
60%
75%
0%
100%
My Work
77%
2019 survey
78%
Org Objectives
83%
2019 survey
81%
My Manager
72%
2019 survey
70%
My Team
83%
2019 survey
83%
L&D
57%
2019 survey
58%

The 2020 People Survey ran from 1st October to 3rd November. 319,935 people, from 106 Civil Service organisations, completed the survey; giving us an overall response rate of 66%.

Here you'll find your Employee Engagement Index, Core Theme Scores, and Discrimination, Bullying and Harassment Rates. Remember to interpret any differences to 2019 with caution: the type of year we've had due to the coronavirus pandemic is not directly comparable to any other year.

Source: Screenshot of the headlines page of the internal dashboard used by the Civil Service People Survey Team.

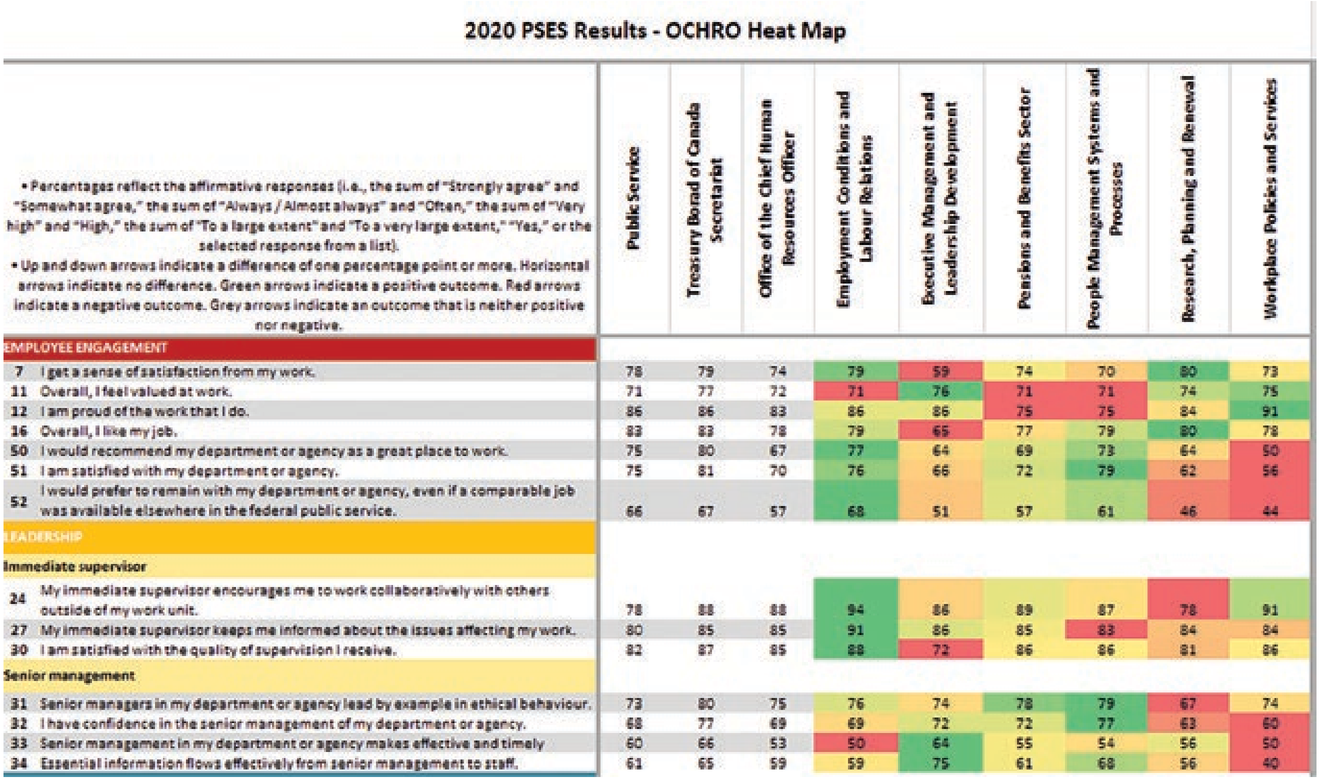
Canada (figure 25.6) dashboards. Canada also provides heat maps for each unit to crystallize strengths and areas for development (figure 25.9).

Generating unit-level results requires generating unit- or division-level identifiers in each organization, which are either linked to the unique survey ID of a survey respondent or selected by survey respondents. These can be collected from central human resources management information systems (HRMIS), where they exist, collected from institution-level HRMIS and appended to the email addresses used to disseminate the survey, or gathered from each government organization manually, with respondents then selecting the unit in which they work when completing the survey.

For instance, in Australia, several agencies choose to map their Australian Public Service Employee Census respondents to their organizational hierarchies. Where an organizational hierarchy has been included, analysis and reporting of results are possible for individual work units within an agency. This includes analysis and reporting for demographic and other groups within an agency or organizational unit. In 2020, just over 60 percent of agencies included an organizational hierarchy in the Australian Public Service Employee Census. How far down an agency chooses to disaggregate its hierarchy typically depends on its size and structure. Most, however, will disaggregate their hierarchies to the lowest practicable level while safeguarding anonymity (for instance, by not reporting results for work units with fewer than 10 respondents).

When agencies provide such disaggregation, reports for agencies and their organizational units are developed and released to those agencies. Representatives within individual agencies have access to the online dashboard, in which they can source their prepared summary reports but also analyze, filter, and compare results for their agency and its constituent organizational units. This portal allows for more interactive descriptive analysis and exploration of results and enables agencies to source more survey results than are made available in the static reports.

FIGURE 25.9 Canada's Heat Maps for Survey Results of Units



2020 PSES Results - OCHRO Heat Map

• Percentages reflect the affirmative responses (i.e., the sum of "Strongly agree" and "Somewhat agree," the sum of "Always / Almost always" and "Often," the sum of "Very high" and "High," the sum of "To a large extent" and "To a very large extent," "Yes," or the selected response from a list).
• Up and down arrows indicate a difference of one percentage point or more. Horizontal arrows indicate no difference. Green arrows indicate a positive outcome. Red arrows indicate a negative outcome. Grey arrows indicate an outcome that is neither positive nor negative.

		Public Service	Treasury Borad of Canada Secretariat	Office of the Chief Human Resources Officer	Employment Conditions and Labour Relations	Executive Management and Leadership Development	Pensions and Benefits Sector	People Management Systems and Processes	Research, Planning and Renewal	Workplace Policies and Services
EMPLOYEE ENGAGEMENT										
7	I get a sense of satisfaction from my work.	78	79	74	79	59	74	70	80	73
11	Overall, I feel valued at work.	71	77	72	71	76	71	71	74	75
12	I am proud of the work that I do.	86	86	83	86	86	75	75	84	91
16	Overall, I like my job.	83	83	78	79	65	77	79	80	78
50	I would recommend my department or agency as a great place to work.	75	80	67	77	64	69	73	64	50
51	I am satisfied with my department or agency.	75	81	70	76	66	72	79	62	56
52	I would prefer to remain with my department or agency, even if a comparable job was available elsewhere in the federal public service.	66	67	57	68	51	57	61	46	44
LEADERSHIP										
Immediate supervisor										
24	My immediate supervisor encourages me to work collaboratively with others outside of my work unit.	78	88	88	94	86	89	87	78	91
27	My immediate supervisor keeps me informed about the issues affecting my work.	80	85	85	91	86	85	83	84	84
30	I am satisfied with the quality of supervision I receive.	82	87	85	88	72	86	86	81	86
Senior management										
31	Senior managers in my department or agency lead by example in ethical behaviour.	73	80	75	76	74	78	79	67	74
32	I have confidence in the senior management of my department or agency.	68	77	69	69	72	72	77	63	60
33	Senior management in my department or agency makes effective and timely	60	66	53	50	64	55	54	56	50
34	Essential information flows effectively from senior management to staff.	61	65	59	59	75	61	68	56	40

Source: Screenshot of unit report heat map by the Treasury Board of Canada Secretariat.

Similarly, in the United States, disaggregation occurs up to the ninth level of hierarchy in some organizations, multiplying the number of units and teams benefiting from survey results. Lower levels of government are provided with "subagency breakout reports," which display results for an individual office (the lowest level of the agency) for all core and demographic survey items, and "subagency comparison reports," which compare all work units within a breakout for all core and demographic survey items.

Capacity to Take Action on the Basis of Survey Results

As noted, turning survey results into action is facilitated by accompanying descriptive survey results with prescriptive recommendations at the national, organizational, or unit level, where appropriate (for instance, by linking training offers to managers to certain survey results in leadership quality); by presenting results in person to organizations to help them understand them and consider actions in response; and by offering action plan methodologies to agencies or units to take action based on results.

In national survey results reports, governments typically do not include prescriptive recommendations, though recommendations or actions are sometimes included in accompanying publications—for instance, in a blog by the chief executive of the UK's civil service (Manzoni 2020), a press release by Colombia's Public Service Department (DAFP 2016), or, perhaps most directly, in Australia's State of the Service Report, which, as mentioned before, integrates public servant survey data with other HR data sources to analyze key HR themes and suggest ways forward (Australian Public Service Commission 2021a). In Canada, presentations and briefings by the Treasury Board of Canada Secretariat include recommendations.

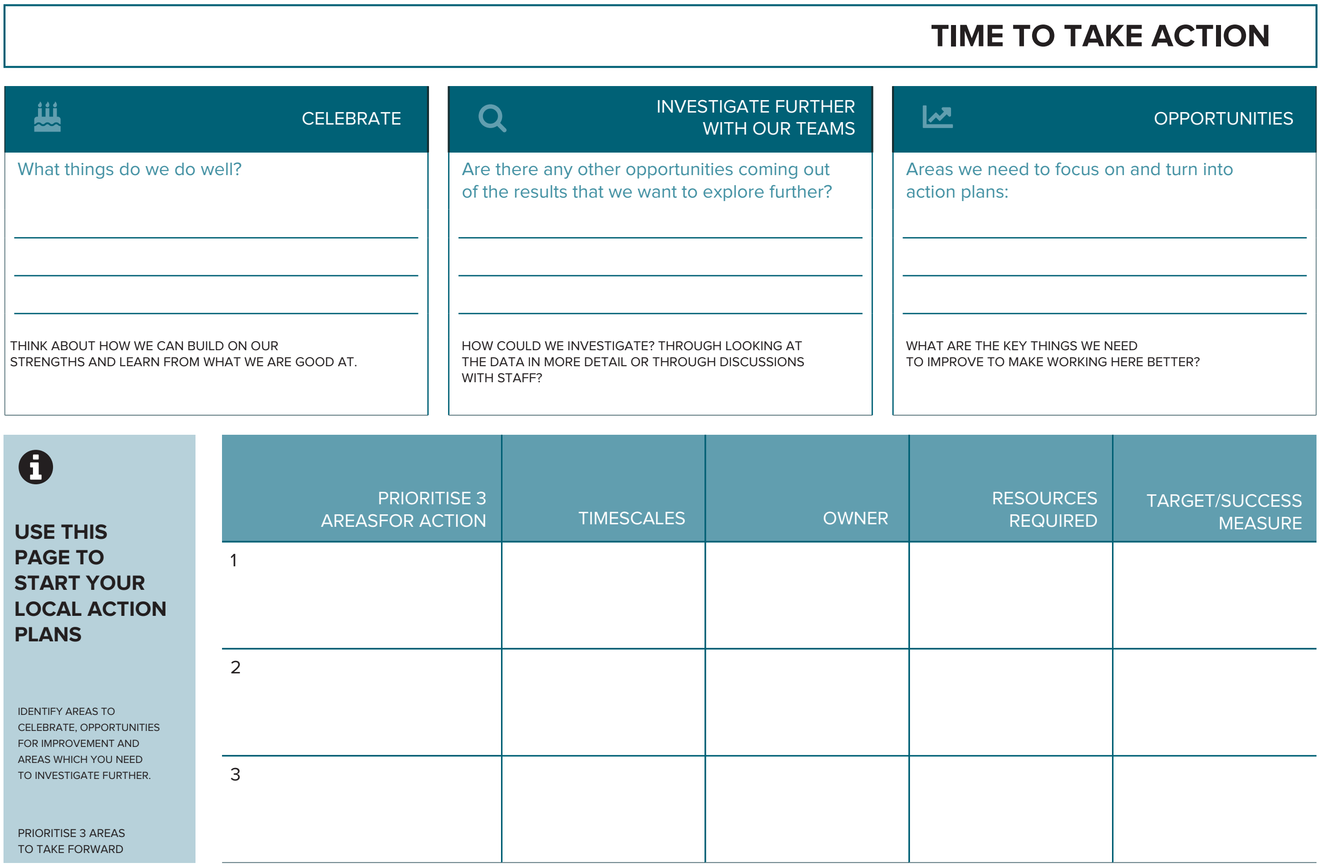
Australia also explicitly offers organizations action plan templates and methodologies to help them take action based on survey findings. Each agency report includes an action template that encourages managers to map actions against survey outcomes (figure 25.10).[4] This is encouraged by tying the release of survey results to the Australian State of the Service Report, which sets out a strategic mission for the civil service. Senior executives from the national commission are asked to present key points of the report to employees in their state and territory. These presentations typically give a high-level overview of the perspectives and direction of the commission and also include Australian Public Service Employee Census results. Each year, focus groups are held with representatives of agencies, during which the use of the results is discussed. Canada, in turn, has an interdepartmental committee in which best practices are shared and organizations are provided guidance on how to create their plans; however, specific templates are not provided. In the United Kingdom, the Cabinet Office shares with departments a guide to running a workshop to discuss the results as a team and take action, while, in the United States, senior accountable officers have been appointed in past years within agencies, and experts in the OPM have worked closely with them to support the interpretation of employee survey results and develop and assess action plans.

Bespoke consultancy by a central agency to help individual organizations improve management based on survey results remains less systematized across governments. Results presentations at the organizational level occur but are not universal or part of a systematic intervention program by a central government agency to boost management practices and employee engagement based on survey results across line agencies. As mentioned before, follow-up consultancy is a cornerstone of the work of engagement consultancy firms—and thus a missed opportunity—but, of course, also resource intensive. At the same time, governments are not currently making use of lower-cost, automated recommendations based on survey results for organizations or units—for instance, by showing specific training offerings to managers with scores in need of improvement in certain areas. More could thus be done to help organizations and managers turn survey results into management improvements.

External Accountability: Information Made Available to the Public

All countries make country-level reports or statistics publicly available. Australia and Canada provide dashboards to enable the public to explore data. Colombia and the United Kingdom provide statistical

FIGURE 25.10 Australian Public Service Commission Action Plan Template



TIME TO TAKE ACTION

CELEBRATE

What things do we do well?

THINK ABOUT HOW WE CAN BUILD ON OUR STRENGTHS AND LEARN FROM WHAT WE ARE GOOD AT.

INVESTIGATE FURTHER WITH OUR TEAMS

Are there any other opportunities coming out of the results that we want to explore further?

HOW COULD WE INVESTIGATE? THROUGH LOOKING AT THE DATA IN MORE DETAIL OR THROUGH DISCUSSIONS WITH STAFF?

OPPORTUNITIES

Areas we need to focus on and turn into action plans:

WHAT ARE THE KEY THINGS WE NEED TO IMPROVE TO MAKE WORKING HERE BETTER?

USE THIS PAGE TO START YOUR LOCAL ACTION PLANS

IDENTIFY AREAS TO CELEBRATE, OPPORTUNITIES FOR IMPROVEMENT AND AREAS WHICH YOU NEED TO INVESTIGATE FURTHER.

PRIORITISE 3 AREAS TO TAKE FORWARD

PRIORITISE 3 AREASFOR ACTION	TIMESCALES	OWNER	RESOURCES REQUIRED	TARGET/SUCCESS MEASURE
1				
2				
3				

Source: Australian Public Service Commission 2021c.

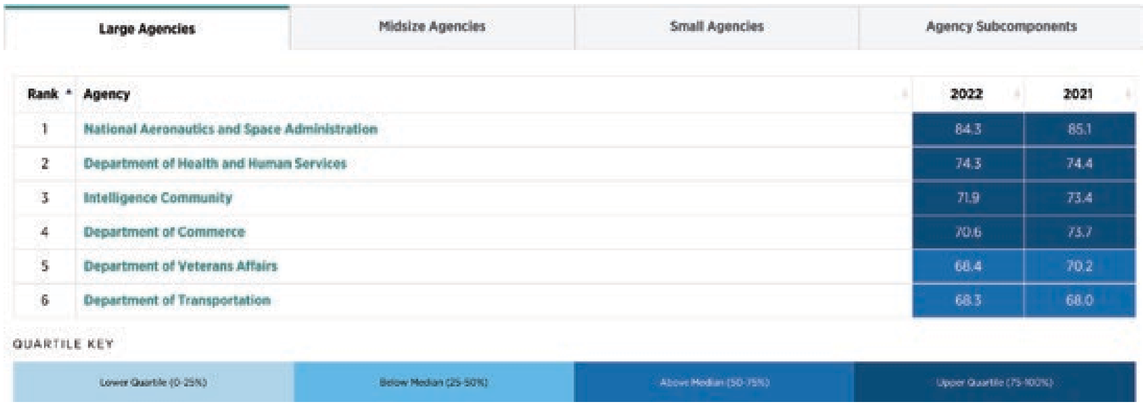
summaries, which might not be easily accessible for audiences unfamiliar with statistics. The British dashboard is not available to the public. Data for Colombia and the United Kingdom can be accessed in an aggregated format by agency on a government website and downloaded as Excel files. Australia, Ireland, and the United States also publish written reports with overall findings. In Australia, Canada, the United Kingdom, and the United States, the availability of publicly available written reports of individual agencies depends on the participating agencies' willingness to publish them. Ireland does not publish organization-level reports (Australian Public Service Commission 2021b; Cabinet Office 2021; Government of Canada 2020a; OPM 2020).

In terms of transparency to the public, Australia, Colombia, and the United States publish individual-level microdata to enable researchers and other interested users to explore the data. Canada and Ireland provide these data to researchers upon request (and with certain requirements).

Australia and the United Kingdom provide statistics aggregated at the response and agency levels that can be downloaded, and Ireland provides summary statistics in report form that can be publicly accessed.

Only in the United States is public information from the employee survey drawn on by external actors. In the United States this is the Partnership for Public Service, which compiles the Best Places to Work in the Federal Government rankings of public sector organizations as a means to generate further external accountability and motivation for improvement in survey scores for public sector organizations (figure 25.11).

FIGURE 25.11 Best Places to Work in the US Federal Government, 2022 Rankings



Rank	Agency	2022	2021
1	National Aeronautics and Space Administration	84.3	85.1
2	Department of Health and Human Services	74.3	74.4
3	Intelligence Community	71.9	73.4
4	Department of Commerce	70.6	73.7
5	Department of Veterans Affairs	68.4	70.2
6	Department of Transportation	68.3	68.0

QUARTILE KEY

Lower Quartile (0-25%)	Below Median (25-50%)	Above Median (50-75%)	Upper Quartile (75-100%)

Source: Partnership for Public Service 2023 (screenshot, https://bestplacestowork.org/rankings/?view=overall&size=large&category=leadership&).

In short, there remain significant opportunities for greater transparency and external accountability for public servant survey results, particularly at the organizational level, in many governments—for instance, by replicating "best place to work" rankings and presenting survey results at the national and organizational levels to stakeholders in a more accessible way.

Internal Accountability for Using Survey Results

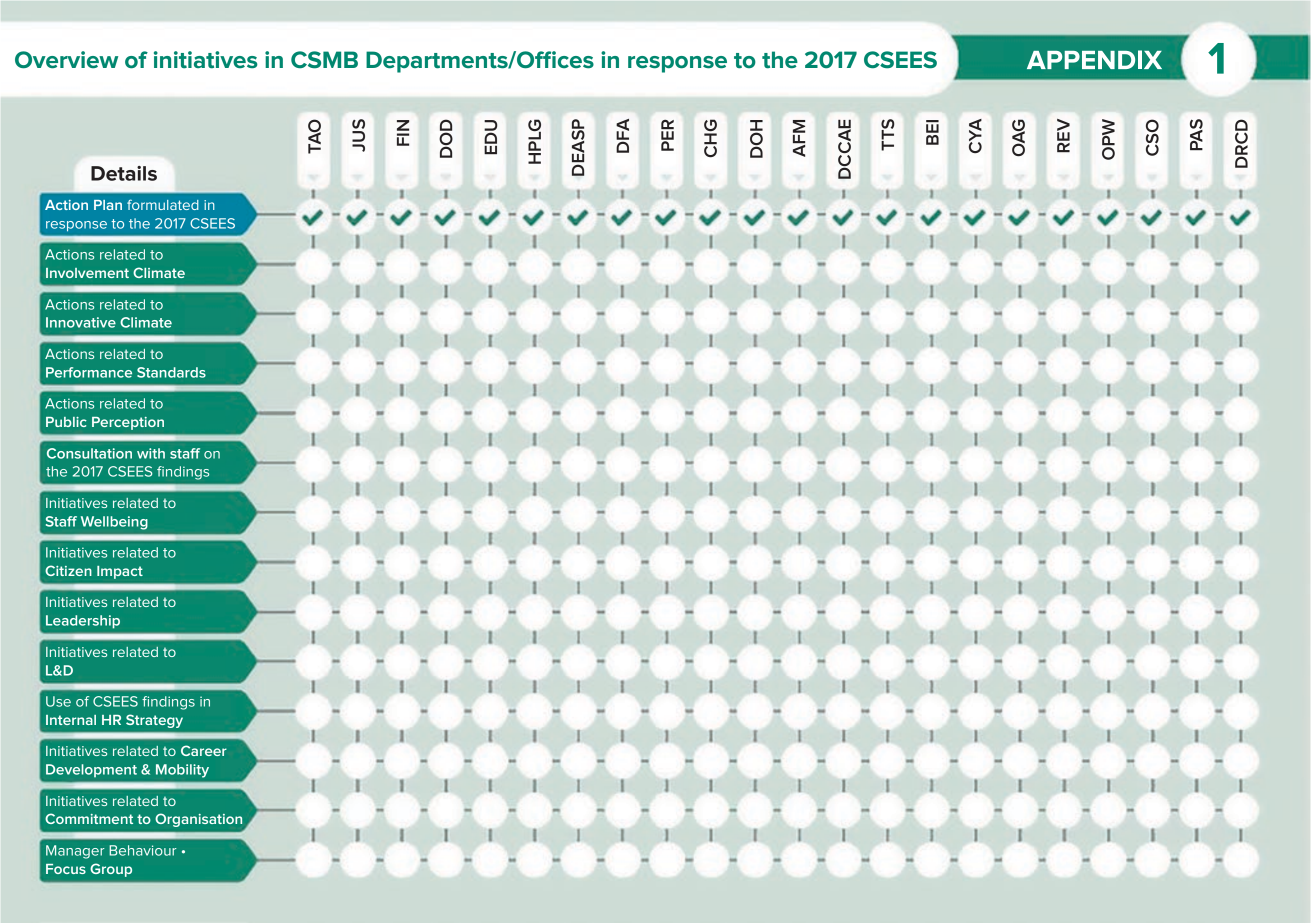
Internal accountability can be top-down (through central oversight) or bottom-up (by employees). Among the countries studied, Ireland has the most-established formal top-down accountability mechanism: it obliges all government departments to map actions taken in response to survey outcomes. After each survey, departments are asked to produce an action plan detailing how they will respond to challenging results within their organizations. The report is organized by thematic area, requiring organizations to state the issue, state the statistic underlying the problem identified, list agreed-upon actions, and list the processes put in place to address them.

A quarterly update is prepared by the Civil Service Renewal Programme Management Office and then relayed to the Civil Service Management Board. An "at a glance" dashboard allows each head of office or secretary general to chart the progress of his or her organization. An interdepartmental working group provides officials with a forum to share experiences and best practices regarding survey management, driving strong response rates, and responding to their organizational results. In nonsurvey years, the group meets on a quarterly basis to share feedback on responding to departmental results. In survey years, the group meets on a more frequent basis to ensure milestones and targets are met in the run-up to the launch of the survey. Figure 25.12 visualizes the "at a glance" dashboard, which tracks actions taken by departments in response to survey results.

Canada, in turn, leverages the MAF to assess the progress made by an organization in its management practices (in seven areas identified by the survey). The MAF involves three key stakeholders (deputy heads of organizations, the HR community, and the Treasury Board of Canada Secretariat) and enables the Treasury Board to "monitor trends and identify gaps in policy compliance across departments," among other things, such as including accountability for improvement in poorly performing indicators.

In the United States, survey results are included in the PMA, and agencies are held accountable for action toward organizational change, including employee engagement and related issues, such as diversity, equity, inclusion, and accessibility (see, for example, Donovan et al. 2014). Other governments lack a similarly institutionalized reporting and accountability mechanism for actions taken.[5]

FIGURE 25.12 "At a Glance" Dashboard, Government of Ireland



Overview of initiatives in CSMB Departments/Offices in response to the 2017 CSEES

APPENDIX 1

Details	TAO	JUS	FIN	DOD	EDU	HPLG	DEASP	DFA	PER	CHG	DOH	AFM	DCCAE	TTS	BEI	CYA	OAG	REV	OPW	CSO	PAS	DRCD
Action Plan formulated in response to the 2017 CSEES	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Actions related to **Involvement Climate**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Actions related to **Innovative Climate**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Actions related to **Performance Standards**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Actions related to **Public Perception**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Consultation with staff on the 2017 CSEES findings	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Initiatives related to **Staff Wellbeing**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Initiatives related to **Citizen Impact**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Initiatives related to **Leadership**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Initiatives related to **L&D**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Use of CSEES findings in **Internal HR Strategy**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Initiatives related to **Career Development & Mobility**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Initiatives related to **Commitment to Organisation**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Manager Behaviour • **Focus Group**	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Source: Screenshot of Civil Service Management Board "at a glance" dashboard, Government of Ireland.
Note: CSMB = Civil Service Management Board; L & D = learning and development.

In terms of bottom-up accountability, in Ireland, Canada, and the United States, questionnaires are typically shared with key employee representative groups and unions before the launch of a survey. For instance, in Canada, extensive consultative engagements with key stakeholders—such as participating departments and agencies and unions—inform questionnaire development, and stakeholders are kept apprised of progress before the survey launch. The United Kingdom and the United States also measure in their surveys whether public servants perceive that their organizations are taking action to address survey results, thus making transparent whether organizations are—in the perception of their staff—acting on the results (and facilitating accountability where staff members do not perceive that their organization is taking effective action). For instance, the US survey inquires whether respondents "believe the results of this survey will be used to make [their] agency a better place to work" (OPM 2021). In short, there remains leeway to strengthen both bottom-up and top-down accountability mechanisms across countries.

DISCUSSION AND CONCLUSION

This chapter has developed a self-assessment framework to enable governments to identify which additional uses of public servant survey results they could contemplate to maximize the impact on civil service management. It then benchmarked six governments against the self-assessment framework to showcase the use of the framework, provide further qualitative detail on each of the potential uses of public servant surveys, and provide a state of play for how governments are currently using (or not using) results from public servant surveys.

Our case selection focused on countries with regular governmentwide employee surveys—which, as of now, tend to be Organisation for Economic Co-operation and Development (OECD) member governments. Our findings about the prevalence of different practices should be interpreted accordingly. Non-OECD governments implement governmentwide employee surveys less frequently, though many of the practices we identify in the chapter would certainly be attainable and low in cost for them as well (for example, publishing anonymized microdata from survey results in an Excel file).

The case comparison has shown that all countries we surveyed provided country-level results—including for public consumption—and, for the most part, results to participating agencies.

Reports that provide information on a subagency (that is, a unit or division) level are less common, as are dashboards that allow government organizations and units to explore and filter the data in the way most relevant to them.

Most reports also remain descriptive. Strategic advice and consulting services are typically not included as part of the mission of survey administration teams, nor are automated recommendations tying survey results to specific management actions. However, as some countries (Australia and Ireland) have acknowledged, the demand for bespoke results and advice has increased, and some countries have at least provided action plan templates for organizations to take action.

Countries also differ in the extent of their external and internal accountability mechanisms. Publication of organization-level results is voluntary and selective in most countries, and some do not publish them at all. Three countries (Australia, Colombia, and the United States) publish anonymized individual-level microdata (with Ireland and Canada making the data available upon request). Internal oversight and accountability for taking actions based on results are only formally institutionalized in a dashboard system in Canada's MAF, Ireland, and the United States' PMA, while the United Kingdom and the United States track the extent to which employees believe effective survey action has been taken.

In conjunction, our results suggest that many governments could, at very low cost, significantly enhance the benefits they derive from public servant surveys for civil service management improvements, including by

- Ensuring that results are disaggregated and disseminated to suborganizational hierarchical levels (for example, divisions and units);
- Creating simple dashboards to allow users at different levels of government—and the public, for national and organization-level results—to explore and filter the data according to their needs;
- Coding management reports (or dashboard front pages) such that the key strengths and areas for development of a particular organization or unit are easily identifiable;
- Including action plan methodologies and automated recommendations to users—such as the managers of units or organizations—about how to best address survey findings (automated recommendations can, for instance, contain training offerings tied to specific survey results or management "checklists" for managers with certain survey results);
- Strengthening accountability for results (for instance, through central oversight of actions taken in response to survey findings by government organizations and units, by enabling third parties—or the government itself—to construct "best place to work" league tables of government organizations, and by capturing employee perceptions of the extent to which government organizations take action in response to survey findings);
- Publishing anonymized microdata to encourage research and insight creation by third parties; and
- Standardizing questions to increase comparability with other countries or industry surveys to create better benchmarks of national scores (for example, through the GSPS).

Where further resources are available, governments may also

- Complement agency-level reports with bespoke presentations and consultancy services to agencies to help them improve in response to survey findings,
- Provide insight reports centered around key strategic topics to move the dial on key HR topics with survey results, and
- Integrate staff surveys with other workforce data to generate more holistic HR dashboards and reports on the public service as a whole, as well as particular strategic themes.

NOTES

1. By *surveys of public servants*, we refer to surveys of employees of government organizations. The coverage of these surveys extends, variously across countries, to the civil service, the public service as a whole—including organizations outside the civil service—or a combination of the two.
2. As with the publication of (anonymized) survey microdata, care needs to be taken to protect the anonymity of survey respondents when disaggregating data to units—for instance, by not reporting unit- or group-level averages with fewer than 10 respondents (cf. OPM 2021).
3. Providing transparency to citizens about the operations of government—including by publishing public servant survey results—is, of course, also an important part of democratic accountability more broadly.
4. The template can be accessed at https://www.apsc.gov.au/initiatives-and-programs/workforce-information/aps-employee-census-2020.
5. In Australia, each organization also has a "champion" who fosters survey participation and the use of results from the survey.

REFERENCES

Australian Public Service Commission. 2021a. "The 2021 APS Employee Census Overall Results." Australian Public Service Commission, Australian Government, November 30, 2021. https://www.apsc.gov.au/initiatives-and-programs/workforce-information/aps-employee-census-2021/2021-aps-employee-census-overall-results.

Australian Public Service Commission. 2021b. "APS Employee Census 2020." Australian Public Service Commission, Australian Government. https://www.apsc.gov.au/initiatives-and-programs/workforce-information/aps-employee-census-2020.

Australian Public Service Commission. 2021c. *Highlights Report: APS Overall*. Canberra: Australian Public Service Commission, Australian Government. https://www.apsc.gov.au/sites/default/files/2021-12/APS00878%20-%20APS%20Overall.pdf.

Australian Public Service Commission. 2021d. *State of the Service Report 2020–21: Reform in the Shadow of COVID*. Canberra: Australian Public Service Commission, Australian Government. https://www.apsc.gov.au/sites/default/files/2021-11/APSC-State-of-the-Service-Report-202021.pdf.

Briones, Ignacio, and Alejandro Weber. 2020. "Un mejor empleo público para un mejor Estado." *El Mercurio*, February 19, 2020. https://t.co/u4x6WTzi79.

Cabinet Office. 2015. "Case Study: Employee Engagement and Wellbeing: Scottish Government, Primary Care Division." Cabinet Office and Civil Service, United Kingdom Government, July 2, 2015. https://www.gov.uk/government/case-studies/employee-engagement-and-wellbeing-scottish-government-primary-care-division.

Cabinet Office. 2016. "Case Study: Employee Engagement and Wellbeing: Cabinet Office's Social Investment and Finance Team." Cabinet Office and Civil Service, United Kingdom Government, February 18, 2016. https://www.gov.uk/government/case-studies/employee-engagement-and-wellbeing-cabinet-offices-social-investment-and-finance-team.

Cabinet Office. 2019. *Civil Service People Survey: Civil Service Benchmark Scores 2009 to 2019*. London: Cabinet Office, United Kingdom Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/876879/Civil_Service_People_Survey_2009_to_2019_Median_Benchmark_Scores_-_final.pdf.

Cabinet Office. 2021. "Civil Service People Survey: 2020 Results." Cabinet Office, United Kingdom Government, May 7, 2021. https://www.gov.uk/government/publications/civil-service-people-survey-2020-results.

Choi, Sungjoo. 2009. "Diversity in the U.S. Federal Government: Diversity Management and Employee Turnover in Federal Agencies." *Journal of Public Administration Research and Theory* 19 (3): 603–30.

Choi, Sungjoo, and Hal G. Rainey. 2010. "Managing Diversity in U.S. Federal Agencies: Effects of Diversity and Diversity Management on Employee Perceptions of Organizational Performance." *Public Administration Review* 70 (1): 109–21.

DAFP (Departamento Administrativo de la Función Pública). 2016. "Sala de prensa: Noticias." [Press release, December 12, 2016]. Departamento Administrativo de la Función Pública, Government of Colombia. https://www.funcionpublica.gov.co/noticias/-/asset_publisher/mQXU1au9B4LL/content/el-97-5-de-los-servidores-consideran-que-su-trabajo-contribuye-al-logro-de-los-objetivos-de-su-entidad-encuesta-edi.

DAFP (Departamento Administrativo de la Función Pública). 2022. "Modelo Integrado de Planeación y Gestión." Departamento Administrativo de la Función Pública, Government of Colombia. https://www.funcionpublica.gov.co/web/mipg.

DANE (Departamento Administrativo Nacional de Estadística). 2022. "Comunicado de prensa: Encuesta sobre ambiente y desempeño institucional Nacional y departamental (EDI-EDID) 2021" [Press Release, June 7, 2022]. Departamento Administrativo Nacional de Estadística, Government of Colombia. https://www.dane.gov.co/files/EDI_nal/2021/Comunicado_prensa_EDIEDID_2021.pdf.

Department of Public Expenditure and Reform (Department of Public Expenditure, National Development Plan Delivery and Reform). 2017. *Civil Service Employee Engagement Survey*. Dublin: Department of Public Expenditure and Reform, Government of Ireland. https://www.gov.ie/en/collection/5e7009-civil-service-employee-engagement-survey/#2017.

Department of Public Expenditure and Reform (Department of Public Expenditure, National Development Plan Delivery and Reform). 2020. *Civil Service Employee Engagement Survey*. Dublin: Department of Public Expenditure and Reform, Government of Ireland. https://www.gov.ie/en/collection/5e7009-civil-service-employee-engagement-survey/#2020.

Department of Public Expenditure, National Development Plan Delivery and Reform. 2022. "What Is the Civil Service Employee Engagement Survey?" *Civil Service Employee Engagement Surveys*. Department of Public Expenditure, National Development Plan Delivery and Reform, Government of Ireland. https://www.gov.ie/en/collection/5e7009-civil-service-employee-engagement-survey/#what-is-the-civil-service-employee-engagement-survey.

Donovan, Shaun, Beth Cobert, Katherine Archuleta, and Meg McLaughlin. 2014. "Strengthening Employee Engagement and Organizational Performance." Memorandum for Heads of Executive Departments and Agencies, December 23, 2014. https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2015/m-15-04.pdf.

Doran, Matthew. 2019. "Bleak Outlook for Home Affairs Morale, as Staff Report Dissatisfaction with Work and Leadership." *ABC News*, August 29, 2019. https://www.abc.net.au/news/2019-08-29/bleak-outlook-for-home-affairs-staff-morale/11461442.

Fukuyama, Francis, Daniel Rogger, Zahid Husnain, Katherine Bersch, Dinsha Mistree, Christian Schuster, Kim Sass Mikkelsen, Kerenssa Kay, and Jan-Hinrik Meyer-Sahling. 2022. *Global Survey of Public Servants Indicators*. https://www.globalsurveyofpublicservants.org/.

Gallup. 2022. "What Is Employee Engagement and How Do You Improve It?" Workplace, Gallup. https://www.gallup.com/workplace/285674/improve-employee-engagement-workplace.aspx.

Government of Canada. 2020a. "2020 Public Service Employee Survey Results." Government of Canada. https://www.tbs-sct.gc.ca/pses-saff/2020/results-resultats/en/bq-pq/index.

Government of Canada. 2020b. "Public Service Employee Data Analytics." Government of Canada. https://hrdatahub-centrededonneesrh.tbs-sct.gc.ca/PSES/Home/Index?GoCTemplateCulture=en-CA.

Government of Canada. 2021. "About the 2020 Public Service Employee Survey." Government of Canada, January 22, 2021. https://www.canada.ca/en/treasury-board-secretariat/services/innovation/public-service-employee-survey/2020/about-2020-public-service-employee-survey.html.

Heywood, Jeremy. 2017. "Civil Service People Survey 2017—The Results." *Civil Service* (blog). November 16, 2017. United Kingdom Government. https://civilservice.blog.gov.uk/2017/11/16/civil-service-people-survey-2017-the-results/.

Manzoni, John. 2020. "Civil Service People Survey 2019—The Results." *Civil Service* (blog). March 26, 2020. United Kingdom Government. https://civilservice.blog.gov.uk/2020/03/26/civil-service-people-survey-2019-the-results/.

Meyer-Sahling, Jan-Hinrik, Christian Schuster, and Kim Sass Mikkelsen. 2018. *Civil Service Management in Developing Countries: What Works?* London: UK Department for International Development. https://christianschuster.net/Meyer%20Sahling%20Schuster%20Mikkelsen%20-%20What%20Works%20in%20Civil%20Service%20Management.pdf.

OPM (Office of Personnel Management). 2020. "Data Reports." OPM Federal Employee Viewpoint Survey, US Office of Personnel Management, United States Government. https://www.opm.gov/fevs/reports/data-reports.

OPM (Office of Personnel Management). 2021. *Federal Employee Viewpoint Survey Results: Governmentwide Management Report.* Washington, DC: US Office of Personnel Management, United States Government. https://www.opm.gov/fevs/reports/governmentwide-reports/governmentwide-management-report/governmentwide-report/2021/2021-governmentwide-management-report.pdf.

OPM (Office of Personnel Management). 2022. “Special Reports.” OPM Federal Employee Viewpoint Survey, US Office of Personnel Management, United States Government. https://www.opm.gov/fevs/reports/special-reports/.

Pandey, Sanjay, and James Garnett. 2007. “Exploring Public Sector Communication Performance: Testing a Model and Drawing Implications.” *Public Administration Review* 66: 37–51. https://doi.org/10.1111/j.1540-6210.2006.00554.x.

Partnership for Public Service. 2023. *2022 Best Places to Work in the Federal Government Rankings*. Washington, DC: Partnership for Public Service. https://bestplacestowork.org/rankings/?view=overall&size=large&category=leadership&.

Penny, Charlotte. 2019. “Case Study: How Google Uses People Analytics.” *Sage*, December 1, 2019. https://www.sage.com/en-au/blog/case-study-how-google-uses-people-analytics/.

Resh, William, Tima Moldogaziev, Sergio Fernandez, and Colin Angus Leslie. 2019. “Reversing the Lens: Assessing the Use of Federal Employee Viewpoint Survey in Public Administration Research.” *Review of Public Personnel Administration* 41 (1): 132–62. https://doi.org/10.1177/0734371X19865012.

Schuster, Christian, Javier Fuenzalida, Jan Meyer-Sahling, Kim Sass Mikkelsen, and Noam Titelman. 2020. “Encuesta Nacional de Funcionarios en Chile.” Chile Civil Service. https://www.serviciocivil.cl/wp-content/uploads/2020/01/Encuesta-Nacional-de-Funcionarios-Informe-General-FINAL-15ene2020-1.pdf.

Te Kawa Mataaho Public Service Commission. 2021. “Workforce Data—Diversity and Inclusion.” Te Kawa Mataaho Public Service Commission, New Zealand Government (accessed December 7, 2021). https://www.publicservice.govt.nz/research-and-data/workforce-data-diversity-and-inclusion/.

UK Government. 2018. “Civil Service People Survey Hub.” Civil Service, United Kingdom Government, October 15, 2018. https://www.gov.uk/government/collections/civil-service-people-survey-hub.

Wall, Martin. 2021. “Most Civil Servants Happy with Conditions but Not Promotional Access.” *Irish Times*, May 14, 2021. https://www.irishtimes.com/news/ireland/irish-news/most-civil-servants-happy-with-conditions-but-not-promotional-access-1.4565595.

CHAPTER 26

Using Survey Findings for Public Action

The Experience of the US Federal Government

Camille Hoover, Robin Klevins, Rosemary Miller, Maria Raviele, Daniel Rogger, Robert Seidner, and Kimberly Wells

SUMMARY

Generating coherent public employee survey data is only the first step in using staff surveys to stimulate public service reform. The experiences of agencies of the United States federal government in using the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) provide lessons in the translation of survey results to improvements in specific public agencies and public administration in general. Architecture at the agency level that supports this translation process is critical and typically includes a technical expert capable of interpreting survey data, a strong relationship between this expert and a senior manager, and the development of a culture or reputation for survey-informed agency change and development initiatives. This chapter outlines the way that the FEVS, its enabling institutional environment, and corresponding cultural practices have been developed to act as the basis for public sector action.

ANALYTICS IN PRACTICE

- Generating coherent survey data that describe the state of the public administration is a vital foundation for inspiring effective reform of the public service. But it is only the first step. Complementary efforts to stimulate the use of that survey data are vital for achieving corresponding change.

Camille Hoover is an executive officer and Robin Klevins is a senior management analyst at the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health. Rosemary Miller is a psychologist and Maria Raviele is a program analyst at the US Office of Personnel Management. Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Robert Seidner was formerly a performance manager at the US Office of Management and Budget. Kimberly Wells is a managing research psychologist at the US Office of Strategic Workforce Planning.

- Survey questions should aim at action from the beginning by asking about topics that staff and senior leaders find most challenging to the achievement of their mission. Designing questions with the chain of policy influence and action in mind prevents the survey process from being weakened at inception by a poor focus on what is important to public sector stakeholders.

- Public action from surveys of public employees requires at least one technical expert capable of analyzing and interpreting survey results and translating them into a clear action plan for the improvement of a specific unit or organization. This may simply entail using the survey as a launchpad to learn more about the issue from people in the organization. At the scale of many public sector organizations, this survey analyst should be embedded within the individual organization. This will provide them with sufficient time and focus to promote, digest, and translate survey findings as a core component of their work program. Given that many of the improvements in public sector organizations have capable personnel driving them, providing an official with the time required to anchor relevant discussions with colleagues is a necessary component of reform.

- Rich survey data and technical expertise to digest their implications are insufficient for public action. Any survey analyst or team must have a strong relationship with a senior manager who sees the value of the survey data for agency reform. Such a manager acts as a bridge between the technical translation of the survey into a form usable by an organization and the strategic processes required to build momentum for change. Rich survey data can generate political will by identifying or making salient significant inequities, opportunities for improved performance, or problematic parts of the agency. However, the case study outlined in this chapter, concerning the United States federal government, implies that a senior manager must champion change for substantial public action to occur. While the skills of the technical expert are important, the accountability and responsibility for developing a sustainable action plan rest on supervisors and leaders.

- For reform to be sustained, the technical staff and the leaders who are the “change champions” must inculcate and manage a culture of using survey data for public service reform. The easiest way to do this is to rapidly respond to issues identified by surveys, with leaders transparently sharing survey results with the workforce and emphasizing the results they deem most important. Leaders should then show staff how they are further exploring the results and creating initiatives that speak directly to the findings. Visible leadership responses to survey results will generate broader buy-in from agency staff, which will strengthen the credibility of the survey process and catalyze the impact of managerial responses. Changes in public administration typically require a coproduction approach, with both managers and staff moving toward improvements. For example, if staff feel that their capacity to perform is not being sufficiently developed, management must make opportunities for capacity development available and feasible to take up, while staff must take those opportunities and put in the effort required for learning.

- A centralized, governmentwide office in charge of survey design and implementation is useful for several reasons. First, there are important methodological decisions that affect all survey users equally but are costly to negotiate. A centralized team can ensure that surveys are effectively implemented and respond to changing service requirements, relieving frequently overburdened agency analytics teams. Second, ensuring a common platform for comparison catalyzes the usefulness of an agency survey by allowing for cross-agency benchmarking. Interagency comparisons rely upon a set of common measures, with data collected using consistent methodologies and under the same conditions and timeframe. Third, such an office can make choices that serve the public service as a whole, independent of any individual agency manager. For example, publishing data on all units, rather than selectively sharing results, ensures a more accurate representation of reality.

- When this central office does not have the capacity to address the demands of all managers in the public service, the case of the US federal government indicates that complementary efforts from individual officers strengthen the possibility that surveys incite public action. For example, the National Institutes of Health (NIH) Employee Viewpoint Survey Analysis and Results Tool (EVS ART) has facilitated granular analytics of the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS), allowing managers across the public service to better understand the implications of the survey for their work.

INTRODUCTION

Generating coherent survey data that describe the state of public administration is a vital foundation for inspiring effective reform of the public service. One of the best-known influential surveys of public officials is the Federal Employee Viewpoint Survey (FEVS), which is administered by the United States federal government's Office of Personnel Management (OPM).[1] The survey was first fielded in 2002 and has been repeated annually since 2011, generating a panel of agency- and unit-level variables including measures of employee engagement, satisfaction, welfare, cooperation, development, leadership, and performance management. Technically an organizational climate survey, it functions to "assess how employees jointly experience the policies, practices, and procedures characteristic of their agency and its leadership" (OPM 2022). FEVS data are made available to managers of units, and an aggregated and anonymized version of the data is made public.[2]

The continuous, comparable, and public nature of the FEVS data has been a boon to the United States government analysts and researchers alike. As Janelle Callahan (2015, 399) states, "Ten years ago, few in government were talking about what federal employees thought or how the survey information could be used to improve employee satisfaction and commitment, and the performance of federal agencies." Various qualities of the FEVS have made it influential in debates within the federal government, Congress, and society more broadly. The Government Accountability Office (GAO) frequently uses FEVS data in its assessments of federal agencies.[3] Similarly, the Partnership for Public Service, a nonprofit organization focused on strengthening the US civil service, uses the FEVS to publish its Best Places to Work in the Federal Government index.[4] This index frequently stimulates substantial debate on the public service labor market and its relationship to analogous private sector jobs (see, for example, Brust 2021; Mullins 2021).[5]

The core purpose of the FEVS is to provide public sector managers with direct but anonymized feedback from employees on the "state" of their work units. This may be at the agency level or in teams as small as 10 people (it is FEVS policy not to release data on any subagency work units with fewer than 10 respondents in order to protect the identity of individuals). It provides managers with a snapshot of current strengths, opportunities for improvement, and challenges for the organizations they manage, as well as how these have changed since the last survey. As Thevee Gray of the US Department of Agriculture (USDA) stated in an interview for this chapter, "OPM FEVS has been a great tool to ensure everyone has the same collective information on what is happening in our agency—a great starting point." The ability for staff to provide management with anonymous feedback about their current experience concerning their work, work environment, management, and leadership ensures a minimum floor of feedback across the federal government. While agencies have their own surveying efforts, the FEVS provides a consistent platform for comparison across time and agencies. In a setting like the public service, where benchmarks of the work environment in other offices provide a crucial complement to more objective but coarse measures, this is a powerful feature of the survey.

The FEVS team provides agency managers with summary results of FEVS data for their units, with some breakdown by demographics, and FEVS individual-level data are released publicly (though fully anonymized).[6] It does not typically provide custom analysis to individual managers. This is a product of the mismatch in the scale of the federal government and the size of the FEVS team—there is simply not enough capacity to provide full-service analytics on demand. This leaves most managers with rich but unstructured survey data to explore. Initiatives such as the National Institutes of Health (NIH) Employee Viewpoint Survey Analysis and Results Tool (EVS ART) have sprung up to support managers in analyzing the data of particular interest to them. However, generating work-unit-focused or topic-specific knowledge from the FEVS requires an effort to engage with the data themselves, which may seem costly or to be a low priority.

Such data releases provide a platform for public service reform at all levels of government, helping managers to better understand the reality of their management approaches and helping agency heads—who often have relatively short tenures heading large and disparate agencies—to identify priorities for the organization as a whole. Agency responses to the FEVS interact with external stakeholders in three ways. First, agencies can quickly identify their relative performance in personnel management, communicate with and learn from more successful agencies, and feel implicit pressure from their public standing. Second, the OPM, Congress, and the White House can do the same. The GAO explicitly uses the FEVS to make recommendations to Congress about how agencies should be reformed. In both cases, agency managers may feel there are career consequences related to improving their standing in the FEVS.[7] Third, bodies outside the government can monitor the workings of the public service and make recommendations about how it should reform or provide inputs into the change and development process for individual agencies.

Simply producing rich survey data has rarely been sufficient to generate public sector action. Complementary efforts to stimulate the use of these data are vital for achieving corresponding change. This chapter argues that external factors and pressures play a secondary role compared to the internal architecture of an agency's response to the FEVS. The experience of the FEVS in its two-decade-long history is that, though external and internal pressures are highly complementary, three pillars of response are critical for the FEVS to induce public action. First, public action requires a technical expert who is capable of analyzing and interpreting the FEVS data and who has sufficient time and focus to understand the implications of the FEVS for an agency and its work units. The approach of these individuals to promoting, digesting, and translating the FEVS for their agencies has varied, but in all cases, these individuals have been committed to the FEVS as a key tool of management and agency betterment. They can be seen as the spark of public action at the agency level.

Second, the survey analyst must have a strong relationship with a senior leader of the agency who sees the value of and endorses the use of FEVS feedback to inform agency-specific development at all levels of the organization. This is often a frontline senior leader or an executive within an organization below the agency level. Broad change can certainly be initiated at higher levels, but real change must happen on the front lines to create sustained culture change. This "change champion" acts as a bridge between the technical translation of the FEVS into a form usable by an agency and the strategic processes required to build momentum for change. The relationship between the survey analyst and the change champion can be seen as the positive friction that turns the spark into a flame for effective organizational development, change, and, ultimately, public sector reform.

However, without the broader buy-in of agency supervisors and staff working within a culture of responsiveness, such efforts are likely to be in vain. This buy-in begins at the initiation of the survey. If few staff respond to the FEVS, the data will not be seen as representative of broader staff concerns. Similarly, if staff do not believe that management will use their feedback to create change, they will not take the survey seriously. Thus, the credibility of FEVS data as a management tool requires a belief that they will indeed be used as a management tool. Once the data are published, agency change and development initiatives stemming from the FEVS currently require a coproduction approach, with both managers and staff moving toward improvements. For example, if staff feel that their capacity to perform is not being sufficiently developed, management must make opportunities for capacity development available and feasible to take up, and staff must take those opportunities and put in the effort required for learning. A culture of survey-informed action at the agency level is the tinder and kindling of public action.

Where these pillars of action have been in place, the FEVS has become a central pillar of personnel management in the US federal government. Callahan (2015, 399) provides the following example:

> The Department of Commerce had [FEVS] subcomponents with the highest employee satisfaction in government and the lowest in 2013, prompting leaders to ask what was going

> on and to take action. The U.S. Patent and Trademark Office (USPTO) was the number one agency of 300 subcomponents regarding employee satisfaction and commitment, while the Economic Development Administration (EDA), also in Commerce, ranked last. EDA officials said they began consulting with the USPTO and other organizations to gather best practices and work on improving employee satisfaction. In 2014, EDA was the most improved subcomponent, raising its satisfaction score by 11.8 points.

This chapter aims to describe the enabling environments within the US federal government that have been most prevalent in the translation of FEVS results into changes to the way public administration functions. It begins with a discussion of the key uses of public employee surveys through the lens of the use of the FEVS and then presents an overview of experiences using FEVS data and results for public action that stresses the three features of agency environments outlined above that have led to policy changes and improvements in government administration. The arguments presented here are based on the experiences of the authors—many of whom have played a key role in the development of the FEVS or its translation and use at the agency level over the past decade—and interviews with key stakeholders from across the US federal government.

THE USES OF SURVEYS OF PUBLIC OFFICIALS

Most surveys of public sector employees intend to improve the quality of the environment in which they work and the processes that they undertake. In turn, work environment or process improvements are intended to improve the actions of the public sector toward the better delivery of public services. While some surveys target aspects of public administration that have direct impacts on service delivery, their intention is frequently the improvement of the administrative environment itself.

As such, survey content typically focuses on aspects of the administration that are widely regarded as meaningful for the quality of the work environment or administrative processes. The features of the work environment a survey assesses will directly determine the potential uses of its results. To have the best chance of informing or inducing reform of the public service, surveys should be designed with a theory of policy influence in mind.[8]

One use of survey results is as a centralized monitoring tool. A centralized personnel management agency may want to track the motivation of employees across the public service to ensure they are being effectively managed by senior leadership. The FEVS was initially implemented after an act of Congress required each agency to survey its employees annually.[9] The act required the collection of perceptions of leadership practices contributing to agency performance, employee satisfaction with policies and practices, work environment, rewards and recognition, professional development and growth opportunities, and organizational mission supports. The required content is included in the FEVS, so agency participation in the survey satisfies their statutory requirements. Incentivizing agency participation in a governmentwide survey also provides leadership with data for shaping policies intended to support federal employees.

The content of such centralized, standardized monitoring surveys will necessarily focus on aspects of administration that are said to be of importance to the quality of the work environment generally. Agencies and units can then be assessed against each other for comparison across the public service. Centralized stakeholders, including oversight entities, such as Congress, can use relative performance on survey measures to identify the worst-performing agencies in a particular area or to identify areas of strength and needs for improvement for individual—or even all—agencies.

For example, after a series of reports and internal surveys identified systemic problems in several national parks in 2016, congressional hearings were held on misconduct and mismanagement in the public service.[10] The agency responded with a range of reforms, including complementing the FEVS with a series of new pulse surveys.[11] The Department of the Interior now uses agency-specific items on the FEVS to monitor agency reforms related to anti-harassment training and employees' knowledge of their rights and resources related to harassment.

Second, surveys of public officials can be used as a tool for agency personnel management. Without having to rely on centralized intervention or coordination, agency or unit managers can undertake their own assessments of their agencies' work environments. If a survey intends to improve agency management, it will naturally focus on elements of the work environment most relevant to its mission. Some of these elements will overlap with the wider service, but others will deviate. Here lies a key tension of centralized surveys of public employees—between the need for comparability and central control over the focus of the questionnaire, on the one hand, and the contemporary requirements of specific agencies, on the other.[12]

Comparability allows managers to use common benchmarks to better understand where they are performing well or poorly. But if comparability is focused on measures that are not relevant to their current concerns, the value of centralized surveys falls. Within the framework of a standardized survey, the FEVS has looked to counter this by providing agency managers with tailored insights, as resource constraints allow. In 2012, as the utility of providing agencies and units with survey results directly became clear, a series of initiatives were undertaken by the OPM to provide work-unit-level data. The OPM intended to empower agency heads and managers to capitalize on it as a tool for the agency. As the OPM (2012) stated,

> Working with the information from the survey, … an agency can make a thorough assessment of its own progress in its strategic goals and develop a plan of action for further improvement. The OPM FEVS findings allow agencies to assess progress by comparing earlier results with [contemporary] results, to compare agency results with the Governmentwide results, to identify current strengths and challenges, and to focus on short-term and longer-term action targets that will help agencies reach their strategic human capital management goals.

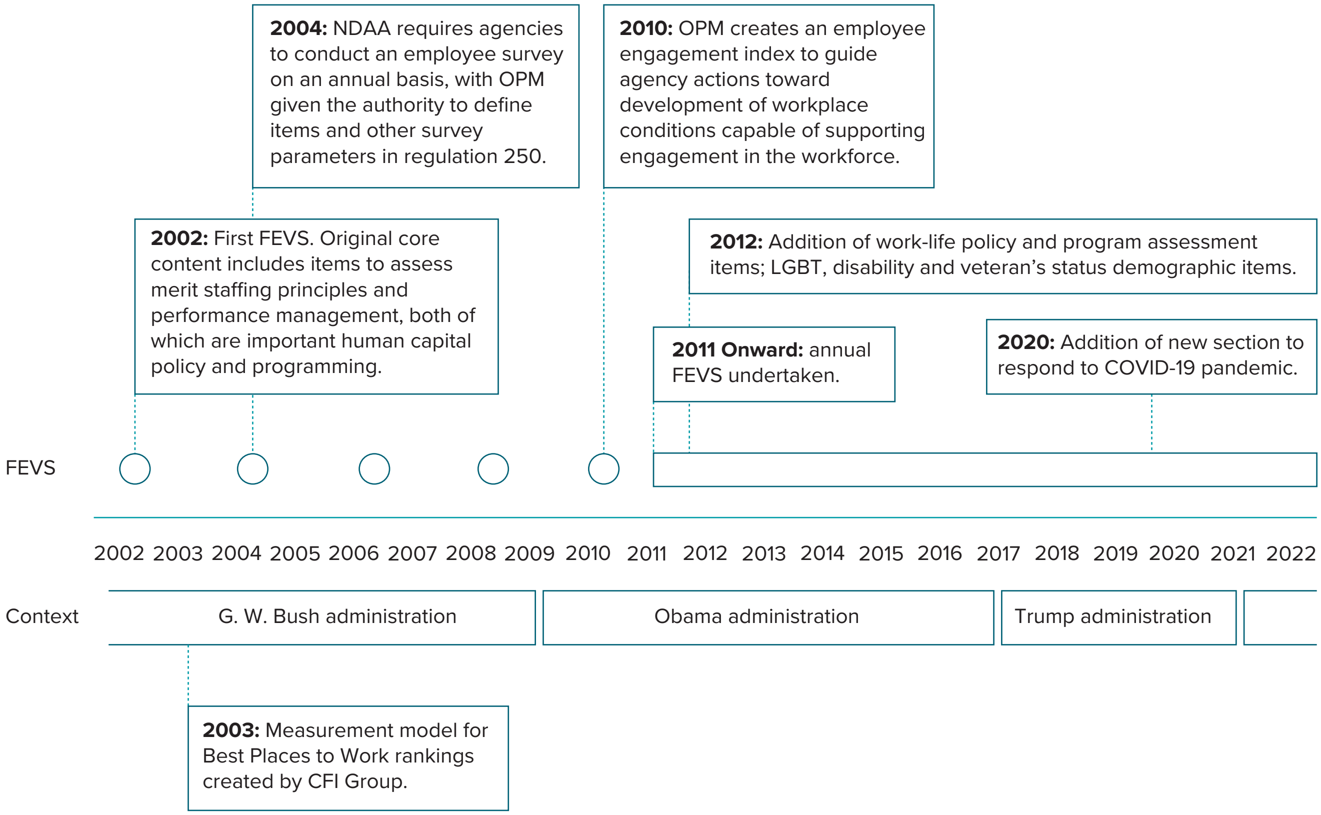
Third, data from public employee surveys can be used as a tool for ensuring the accountability of the government to citizens. In this case, citizens may be less interested in how satisfied or motivated public employees are but more interested in whether they are undertaking their jobs effectively and ensuring the judicious and efficient use of public funds. This implies a third realm of focus for public action to which survey questions may be targeted. Since 2012, the OPM has released anonymized data at the individual level. This has allowed analysts, researchers, the media, and the public to explore the world of the federal government in an unprecedented way.

Figure 26.1 summarizes the use of the FEVS across these three realms over time. The FEVS has at once been an oversight tool for Congress, a key resource for the GAO's large-scale evaluations of government, a means by which the Office of Management and Budget (OMB) can support broad agency functioning, and a rich resource for agencies to use as a core management tool. Each of these drivers of public action has matured and evolved toward an increasingly valuable architecture for the FEVS to impact government functions. These uses of the FEVS have co-evolved, and agency-level responses to the FEVS have been a critical complement to governmentwide policy and program assessments.[13]

The unifying theme of interest across varying federal government stakeholders is organizational effectiveness and performance. In particular, officials must make informed decisions or recommendations, interact with other members of the public service, and effectively deliver their mission to other members of government or the public. Succeeding at these tasks requires sufficient performance. The FEVS is designed as an organizational climate survey—a type of employee survey typically utilized to support organizational change and development initiatives.[14] Climate surveys collect employees' perceptions of management policies and practices, perceptions shown over decades of research to relate to performance. Moreover, employee input on policy enactment provides valuable data for assessing the function of those policies and ensuing practices, serving to guide effective change.

The FEVS contains several variables shown by research to relate to performance. Following an extensive body of research demonstrating the importance of employee engagement to performance, in 2010, the FEVS team introduced an employee engagement index (EEI) (see figure 26.1) that brought together those survey

FIGURE 26.1 Timeline of the Evolution of the Federal Employee Viewpoint Survey



2004: NDAA requires agencies to conduct an employee survey on an annual basis, with OPM given the authority to define items and other survey parameters in regulation 250.
2010: OPM creates an employee engagement index to guide agency actions toward development of workplace conditions capable of supporting engagement in the workforce.
2002: First FEVS. Original core content includes items to assess merit staffing principles and performance management, both of which are important human capital policy and programming.
2012: Addition of work-life policy and program assessment items; LGBT, disability and veteran's status demographic items.
2011 Onward: annual FEVS undertaken.
2020: Addition of new section to respond to COVID-19 pandemic.
FEVS
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
Context
G. W. Bush administration
Obama administration
Trump administration
2003: Measurement model for Best Places to Work rankings created by CFI Group.

Source: Original figure for this publication.
Note: CFI = Claes Fornell International; FEVS = Federal Employee Viewpoint Survey; LGBT = lesbian, gay, bisexual, and transgender; NDAA = National Defense Authorization Act; OPM = Office of Personnel Management.

questions related to different aspects of employee engagement. The EEI was subsequently featured in the President's Management Agenda (PMA) and, accordingly, became a focus for agency change and development initiatives and a central part of the FEVS team's reporting and dissemination efforts.

Importantly, the FEVS EEI measures conditions that can lead to the state of engagement. The 15 EEI questions do not directly measure employees' feelings of engagement but rather assess conditions conducive to engagement (for example, effective leadership, meaningful work, and learning and growth opportunities), in keeping with the frame appropriate to an organizational climate survey. Understanding the engagement potential of federal workplaces along the factors of the measure enables one to identify leverage points for developing and sustaining work conditions capable of supporting employee engagement and, consequently, performance. These work conditions can be targeted for reform and provide policy-relevant variables for data collection. With a common measure, offices undertaking centralized monitoring can search for service-wide engagement trends and for work units that are falling behind others in terms of engagement. Agency managers can work to resolve issues with work conditions flagged by surveys, and stakeholders outside of government can monitor the health of their public service through the engagement of public employees.

The legislative foundation of the FEVS questions has limited change in the survey's content over time, although as figure 26.1 points out, changes have been made. Recently, the regulation governing content has been revised, with the number of required questions being reduced from 45 to 16. With this change, an FEVS modernization initiative has resulted in the addition of new content meant to respond to federal government priorities (for example, Executive Order 14035: Diversity, Equity, Inclusion, and Accessibility in the Federal Workplace) and advances in contemporary management theory and research (for example, innovation and organizational resilience).

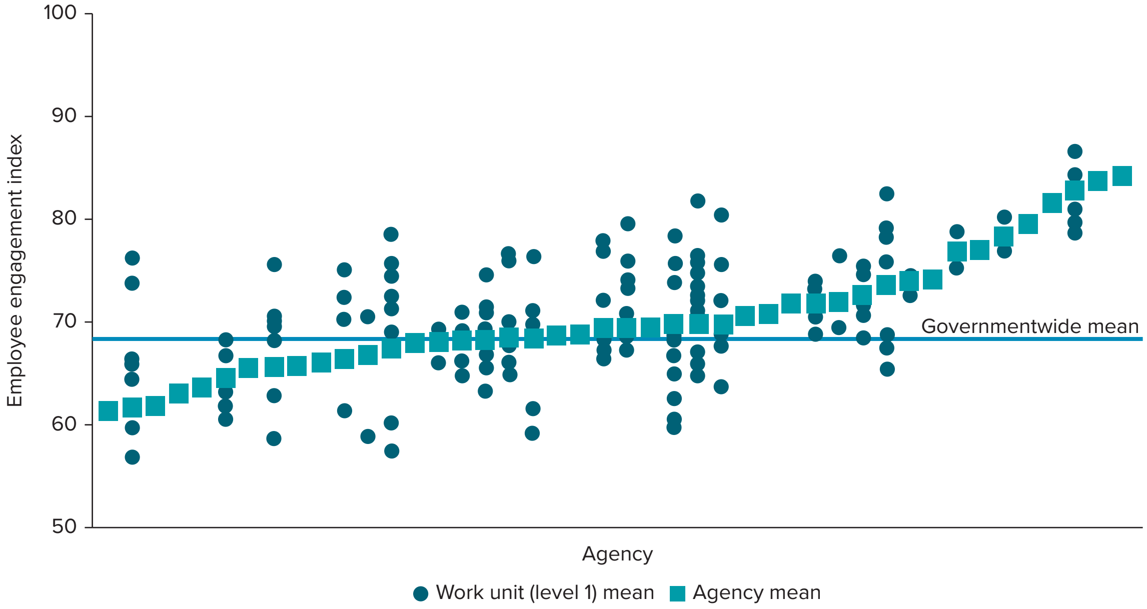
A major goal for the entire FEVS program is to respond to evolving conditions and priorities. When the public service as a whole faced a significant new challenge, the FEVS responded rapidly. An entirely new and substantial section was added to the survey—for the first time since the development of the FEVS nearly two decades ago—due to the COVID-19 pandemic. Given the nature of the FEVS, OPM leadership felt such an addition would be particularly appropriate to understanding the implications of changes made to governmentwide and agency management practices and policies addressing pandemic challenges. The addition of items to assess responses to the pandemic has given the survey another layer of utility, with results critical to determine responses to future emergencies and inform ongoing discussions about the future of work.

THE IMPORTANCE OF THE INSTITUTIONAL ENVIRONMENT

FEVS data have provided a window into the public administration of the US federal government. In contrast to many stereotypes of a monolithic bureaucracy, the experience of working in government is hugely diverse. Figure 26.2 presents the EEI across agencies (the solid squares) for 2018. As can be seen, there are substantial differences in how employees across the government perceive the engagement potential of their agencies. This point is amplified by looking at work units within agencies. Stacked vertically around each agency mean are the scores of the departments/work units (level 1) within that agency. In many agencies, we see that EEI scores can range as widely as they do across the public service as a whole.

Variation across and within agencies is a core reason why the institutional environment of an agency or department is so critical in generating public sector reform. The problems facing individual organizations will differ, requiring an organization- or work-unit-specific response that can only be generated if that organization has the right architecture in place to identify problems and build momentum for solutions. The fact that such variation is found in teams with similar budgets, jobs, senior leadership, and history indicates that differences in employee engagement are likely to have unique causes that require specific attention within the organization.

FIGURE 26.2 Variation across and within Agencies in the Employee Engagement Index



Source: Original figure for this publication based on the FEVS 2019 public data.
Note: Solid squares represent agencies, and dots are scores for department or work units within that agency. FEVS = Federal Employee Viewpoint Survey.

Variation is also at the core of why survey data are so powerful. Rather than making policies based on the general experience of government (perhaps best represented in figure 26.2 by the governmentwide mean), policies can be targeted at those agencies and departments that are most in need. And lessons can be learned from those that are most successful. Thus, the FEVS aims to improve the quality of the management and work environment across agencies and departments by collecting individual employees' perceptions and experiences of their workplaces.

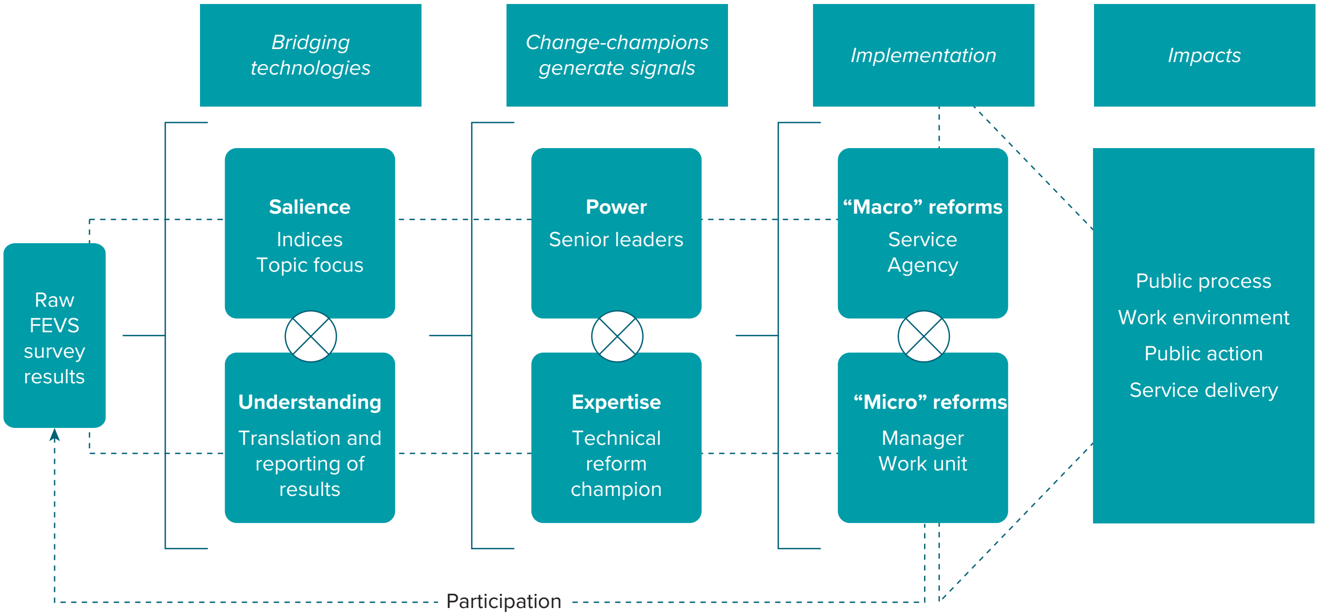
After 20 years of evolution, the FEVS is now built on an increasingly rich institutional scaffold that encompasses statutory requirements for surveying, reporting tools, and centralized initiatives that focus on the weakest performers, as defined by FEVS-based indexes. In many agencies, there are complementary scaffolds that support agency responses to FEVS results, either in reaction to centralized monitoring or as part of an agency-based reform initiative.

The evolutionary process that has occurred in the US federal government has guided agencies toward a structure with a series of key features. Figure 26.3 articulates these features as follows. The first column shows how raw FEVS results require a bridge into the agency where they are translated and their ramifications understood. Given the number of work units in most agencies and the number of questions in the FEVS, the potential complexity of reporting is substantial. Some topics must be made salient, requiring the survey analyst within the agency to appreciate where there is scope for reform and how the survey results might interact with those issues.

That iterative process of mapping results to areas of agency work is not done by the survey analyst alone but happens in collaboration with a senior "change champion." In the second column of figure 26.3, we see how the interaction between *power* and *expertise* within the agency generates the momentum for change, or at least signals it to the wider agency. In the third stage, proposed reforms must be implemented either at a *macro* level, across the public service or agency, or at a *micro* level, by a manager for, perhaps, a single work unit. For many reforms of the public service, a quorum of agency staff must accept the change and invest effort to shift to the new way of working. Together, these columns make up the architecture for impact on public processes, the quality of the work environment, and, eventually, the quality of services delivered.

This static exposition ignores the dynamic nature of these elements. As agency officials observe reactions to the FEVS survey results, the survey itself gains credibility, leading to greater participation in the survey.

FIGURE 26.3 The Architecture of Policy Action for the Federal Employee Viewpoint Survey



Source: Original figure for this publication.
Note: FEVS = Federal Employee Viewpoint Survey.

Greater participation, in turn, makes results more representative of the underlying issues, which, in turn, leads to more relevant reform approaches, increasing the credibility of the wider process. In this way, a virtuous circle can be formed. As Tracey Hilliard of the US Department of Health and Human Services (HHS) stated in an interview for this chapter, "Once everyone responds, more managers get [results specific to their work unit], and this ensures problems are less likely to be hidden in averages—a manager can tell what their particular issues are."

These elements are all necessary to an agency architecture for inducing public action from FEVS raw data. Once these structures are in place, managers receive feedback on their performance and know that senior management is knowledgeable about the areas in which they need to improve. This creates accountability, communication, and a shared understanding for change.

TRANSLATION OF SURVEY RESULTS THROUGH TECHNICAL EXPERTISE

The FEVS contains a substantial amount of information. For each respondent, there are roughly 85 questions/items (depending on the specific survey, with length varying by year and the track respondents follow). These questions can be assessed by a wide range of groupings, compared to previous years' trends, or benchmarked against the dynamics of similar variables and groups. Each of these cuts of data can be made for each work unit or aggregated to the departmental, agency, and service levels. For this reason, the potential complexity of analyzing the FEVS is significant.

Similarly, although FEVS results are presented to senior managers in a series of high-level reports, they are also released in a relatively unstructured form to ensure maximum flexibility for managers to analyze those issues most relevant to their teams. As noted above, trying to provide managers with flexibility is one way the FEVS tries to be useful to a diverse public administration. But this confronts managers—many of whom do not have any background in survey data analysis—with the demanding task of making sense of rich but complicated data. That task must fit into their wider work of managing a work unit and undertaking their own portfolio of activities. Frequently, this combination of complexity and constraint prevents managers from fully engaging with the FEVS data. As Stephen Pellegrino of the US Department of Energy stated in an interview for this chapter, "We get a lot of data from OPM, and managers are not going to tease out what is relevant to them."

Having a colleague whose work program includes time to undertake analysis of the FEVS data and who can identify their relevance for a work unit overcomes the first bottleneck to using the survey to generate reform. Simply having someone who can "translate" the data into practical issues for specific managers ensures the data have meaning for all officials, irrespective of their previous training and inclinations. Mr. Pellegrino noted that he provides his colleagues with simplified answers to the questions they have about the FEVS and only delves into greater detail on methodology for those who request it. As he frames it, "When you get down to a granular level, the statistics don't matter as much as the story."

This is particularly true for more senior members of the administration. As Gonzalo Ferro of the US Securities and Exchange Commission (SEC) argued in an interview, "There is a need to help leaders understand the OPM FEVS data for their organization." For Mr. Ferro, this includes developing data visualizations (such as trend graphs and heat maps) that help managers make sense of the FEVS quickly and efficiently. "At the SEC," Mr. Ferro reported, "we built a dashboard that makes all of our OPM FEVS data (from 2012 to present) accessible to all of our employees."

Having a "technical expert" translate and report the results also consolidates the effort to engage with the FEVS at the agency. Potentially, this makes the analysis of the survey more efficient compared with having each manager do complementary work themselves. Such a survey analyst, aware of the priorities and issues of the organization they work for, may also be able to make salient results that speak to priorities. They can link the results to discussions throughout the organization.

In an interview for this chapter, Thevee Gray of the USDA expressed, "For strategic and effective change to happen, it's important to know how to bridge the gap between the current state and our desired vision. That's when the survey data plays an important role. It's vital to know how to interpret the information, understand the culture and speak to both the grassroots and upper management." In her experience analyzing FEVS data, while a survey expert is critical, so are the data. She continued,

> My team and I leverage the FEVS data to help shine the light on issues within an organization and help managers recognize the importance of understanding the collective feedback from their employees. We presented an activity with them where we wrote on a board what they believed they were doing well and then showed them the FEVS data. It was an "aha" moment for all of them. The challenges they identified were completely opposite from employees' perspective. If we don't have this data to help guide them, management would focus on completely different issues. They would not be able to effectively close the gaps, wonder why the challenges remain and the needle has not moved in a positive direction.
>
> Additionally, survey results are vital because they provide statistically valid information about what employees think. However, I always share with leadership to probe for what lies behind the survey results. Because as you analyze the data, it doesn't explain why employees respond to questions as they do, and the reasons will not always be clear. This is why, when assessing the state of organization, the survey data should be used in conjunction with other information.

Ms. Gray worked for the USDA Farm Service Agency, and her work there provides an example of an agencywide initiative arising from the FEVS results. She used the FEVS analysis to identify staff recognition as an area of the work environment that was particularly challenging throughout the organization. Through a series of focus groups and managerial briefings, she and the wider agency came up with a system of celebratory coins themed with harvest-related features. Though the "USDA is not a coin culture—that comes from the military," it worked effectively in giving managers a low-cost way to recognize excellence within their work units.

Similarly, Tracey Hilliard of HHS argued that it is important to have someone who can work with and interpret the FEVS data at the organizational level: "[The HHS Centers for Disease Control and Prevention] has 10,000 employees, so we created coordinators—two people in each organization that are a point of contact and can interpret the data. They came to meetings twice a month and helped get the data out to the managers to help translation."

PARTNERSHIPS FOR POLICY ACTION

In most hierarchical organizations—which, arguably, most agencies of the US federal government are—expertise in FEVS data analytics is not enough to generate change. To generate change, leaders must appreciate the validity and importance of feedback and use this information to make informed strategic decisions, including providing the necessary resources to affect change. In all of the interviews undertaken for this chapter, and in the broader experience of the FEVS team, change has always required buy-in from senior management and the supervisor of the organization. Without buy-in, power will be a bottleneck rather than an enabler. When discussing her experience of trying to generate responses to FEVS results, Thevee Gray argued that "the leadership buy-in was crucial to help shift the needle in a positive direction … Once you have their buy-in, that cascades through the organization."

As Ms. Gray pointed out earlier, without the FEVS as a diagnostic tool, management might not tackle the right work environment issues. Thus, in figure 26.3, change arises from the interaction between the survey analyst and the senior manager rather than from the manager alone. Tracey Hilliard suggests, "The survey didn't change the organization; the leadership did, but they used the survey as their vehicle." As reflected in many of our interviews, the FEVS data do indeed seem to provide managers with new insights into their strengths and weaknesses and the current environment of their work units. Though this was reported to be truer for less-experienced managers and those with some of the worst results, the FEVS was felt to be instructive in general. It is interesting to note that in figure 26.2, those agencies with higher overall engagement scores are also those with the lowest variation in engagement across work units. This is consistent with the idea that agencies in which management has taken engagement seriously ensure that engagement is consistently addressed across the agency.

Once management begins to respond, the FEVS ensures a feedback loop is built into any reform program so that management can continue to measure their success in making relevant changes in the years following reform efforts. Increasingly, over the past 20 years, the FEVS has become the federal government management's tool for getting feedback on the current state of the administration, allowing those who are implementing reforms to course-correct their efforts. And as more senior managers understand the value of the FEVS as a management tool, the peer pressure on the wider cadre of management increases. The frontline supervisor must also understand the value of the FEVS analysis and be the accountable party to take the necessary next steps for reform. Correspondingly, the demand for survey analysts who can support the increased demand for FEVS analytics also increases.

The FEVS enables the team of the survey analyst and the change champion to develop appropriate reforms beyond just ensuring reforms are informed by the survey. The relationship between the survey analyst and the senior manager ensures reforms are based on evidence, but conversely, the FEVS data and evidence make that relationship possible. Stephen Pellegrino points out that without the FEVS data, a survey analyst may not be able to have conversations about areas of weakness with staff and managers across the organization. He suggests that "it's an easy way to start difficult discussions that are otherwise challenging to have—it's the data."

Once a formal position, team, or office is set up to process the FEVS data, it supports the further strengthening of the relationship with senior management. Like in any administration, having an office and personnel dedicated to a topic—in this case, the analysis of the FEVS—increases the salience and acceptability of a message. Karlease Kelly, formerly of the USDA, has presented across the US public service about the idea that consistently reporting FEVS results to managers makes them increasingly likely to accept the results and associated recommendations. The architecture becomes strengthened over time as managers come to view the FEVS as a standard part of their management approach.

GENERATING A CULTURE OF CREDIBILITY AND RESPONSIVENESS

If the survey analyst provides the spark from the FEVS, and the relationship between the analyst and senior management is the positive friction that turns the spark into a flame for reform, there must be tinder and kindling for public action. Perhaps the key bottleneck to the use of results is the cultural resistance that agencies can have toward capitalizing on diagnostic data, such as the FEVS. The agencies that have been the most successful in using the FEVS for policy action have been those that have created a culture of using the FEVS across the entire staff to complement the basic scaffolding outlined above.

Without the cooperation and effort of the wider body of staff at an agency, any public action is unlikely to succeed. Thus, staff must feel that their efforts—to fill in the survey or support a reform—will be rewarded in some way. This can include simply fulfilling a norm that they expect to fulfill and that they expect others will also conform to. A strong FEVS-based reform culture is one in which all staff members believe that it is the norm to fill in the FEVS, to expect that its results will be responded to by senior management, and to agree that whatever change comes should be adhered to by all staff.

In part, such a culture requires the process of reform to be inclusive. Where cultures of FEVS-based reform have been built, FEVS results and identified focus areas were shared with staff, who were invited to talk about them openly and often in a variety of venues. These meetings were community focused, diverse, and respectful. The FEVS was not presented as a report card but as a platform for discussion about where to go next. Once challenges were identified and generally agreed upon, staff were involved in changes in a positive way, such as through professional development opportunities. Reform leadership opportunities were created for those with a passion for the subject matter at hand to help lead projects, initiatives, and trainings that had been identified as necessary. A culture of responsiveness to the FEVS has arisen, in part, from credibility built over time. Once the relevant survey analyst and both senior management and frontline supervisors articulated to staff the results and what actions would be taken in response, it often took time for staff to believe this would be a systematic approach. Staff belief in action sometimes took years to develop.

The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) provides an example of how an agency has transformed its work environment by focusing on building a culture of responsiveness to FEVS results. In 2015 and 2016, the NIDDK was scoring toward the middle of the distribution of agencies in its sector. Though it was not one of the overall laggards, its senior leadership wanted it to become a stronger and more effective agency. They focused on key themes coming out of the FEVS and created a campaign based around the motto "You speak, we listen, things happen."

The NIDDK formulated an approach based on three principles: share FEVS results and analysis broadly and continuously throughout the year, meet with subgroups to better understand their perspectives and feedback, and undertake focus groups and listening sessions to continue the conversation on FEVS results. The FEVS was no longer viewed as a report card looking backward but, instead, as a launchpad for robust conversation moving forward. A clear outreach strategy was combined with regular reporting on how challenges were being targeted. Actions taken by the agency were communicated and tied back to the FEVS results and, more specifically, to "the voice of the people." Thus, NIDDK staff were given clear signals that senior managers had taken the FEVS seriously and were trying to improve the work environment in response.

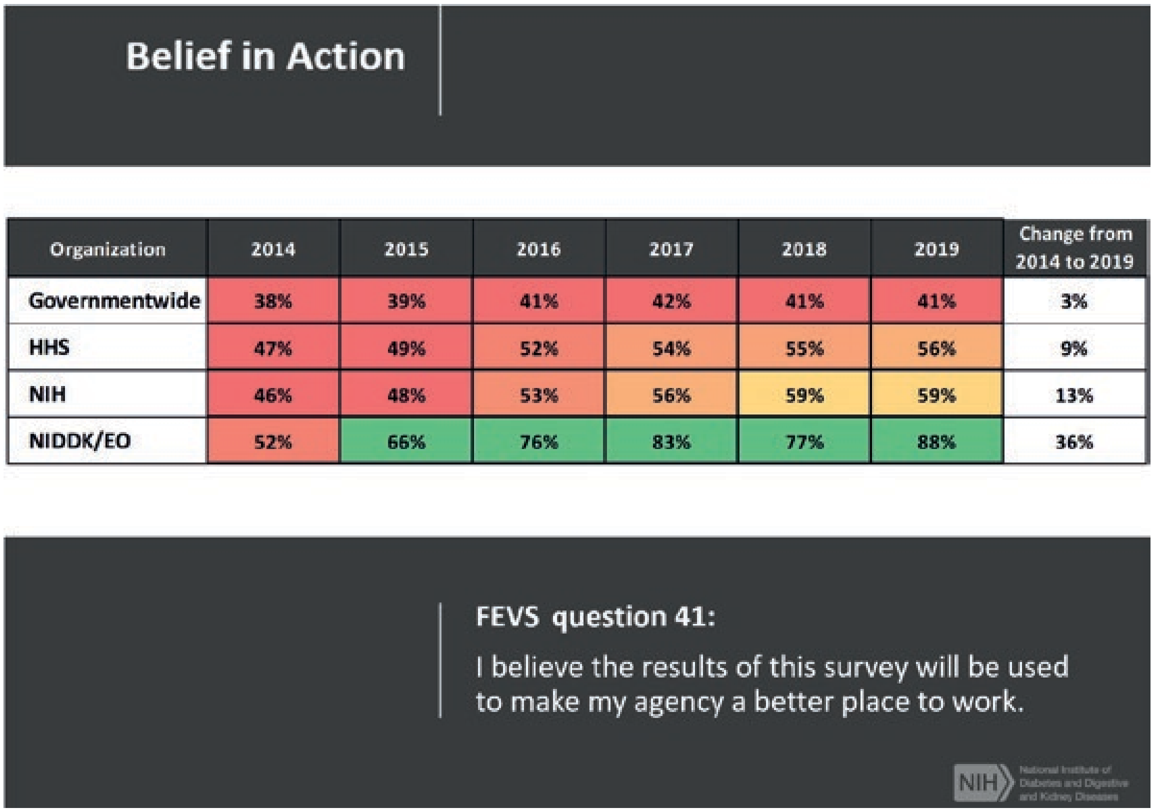
Staff confidence in the efficacy of the FEVS as a management tool grew, and figure 26.4 illustrates the difference this approach made. The figure shows the NIDDK's trajectory of positive responses to the FEVS item "I believe the results of this survey will be used to make my agency a better place to work." In contrast to little change in the government as a whole, the NIDDK's trajectory was substantially positive, changing over 20 percentage points between 2014 and 2019. This also equated to increased survey participation, which climbed from 37 percent in 2014 to 69 percent in 2019, providing even more data for decision-making. Comparator scores are provided for the NIH and for HHS as a whole.

The NIDDK's increased positive scores related to poor performance are also remarkable and are related to putting both standards and accountabilities into place, in addition to several targeted interventions. Figure 26.5 shows the NIDDK's results for one of the lowest-scoring questions across the federal government: "In my work unit, steps are taken to deal with poor performers." In 2015 and 2016, the NIDDK's results were stagnant, like those of the government and the NIH as a whole. However, with the initiation of strategic initiatives and increased transparency through the launch of the "You speak, we listen, things happen" campaign in 2017, the proportion of positive responses jumped and continued to climb in the following years, indicating an increase in employees' positive perception of how their agency dealt with poor performers. The NIDDK, using an architecture representative of that outlined in figure 26.3, successfully transformed its staff's perception of accountability at the organization.

The strategic use of the FEVS has created a ripple effect throughout the institute. As of the 2020 FEVS cycle, the NIDDK's cumulative scores in the areas of employee engagement, global satisfaction, and "leaders lead" were the highest across all 28 NIH institutes and centers, with positive percentage scores of 91 percent, 88 percent, and 90 percent, respectively.

In some ways, the actions undertaken at the NIDDK increasingly echo across the federal government. Using the FEVS as a critical management tool is becoming the norm, and cultures like the NIDDK's are being built more widely. This is partly because the culture of the entire public service is being changed by the FEVS. Once disaggregated FEVS results were shared publicly and members of the government and the

FIGURE 26.4 National Institute of Diabetes and Digestive and Kidney Diseases Staff Responses to Federal Employee Viewpoint Survey "Belief in Action" Question Compared to Organization-Level and Governmentwide Responses



Belief in Action

Organization	2014	2015	2016	2017	2018	2019	Change from 2014 to 2019
Governmentwide	38%	39%	41%	42%	41%	41%	3%
HHS	47%	49%	52%	54%	55%	56%	9%
NIH	46%	48%	53%	56%	59%	59%	13%
NIDDK/EO	52%	66%	76%	83%	77%	88%	36%

FEVS question 41:

I believe the results of this survey will be used to make my agency a better place to work.

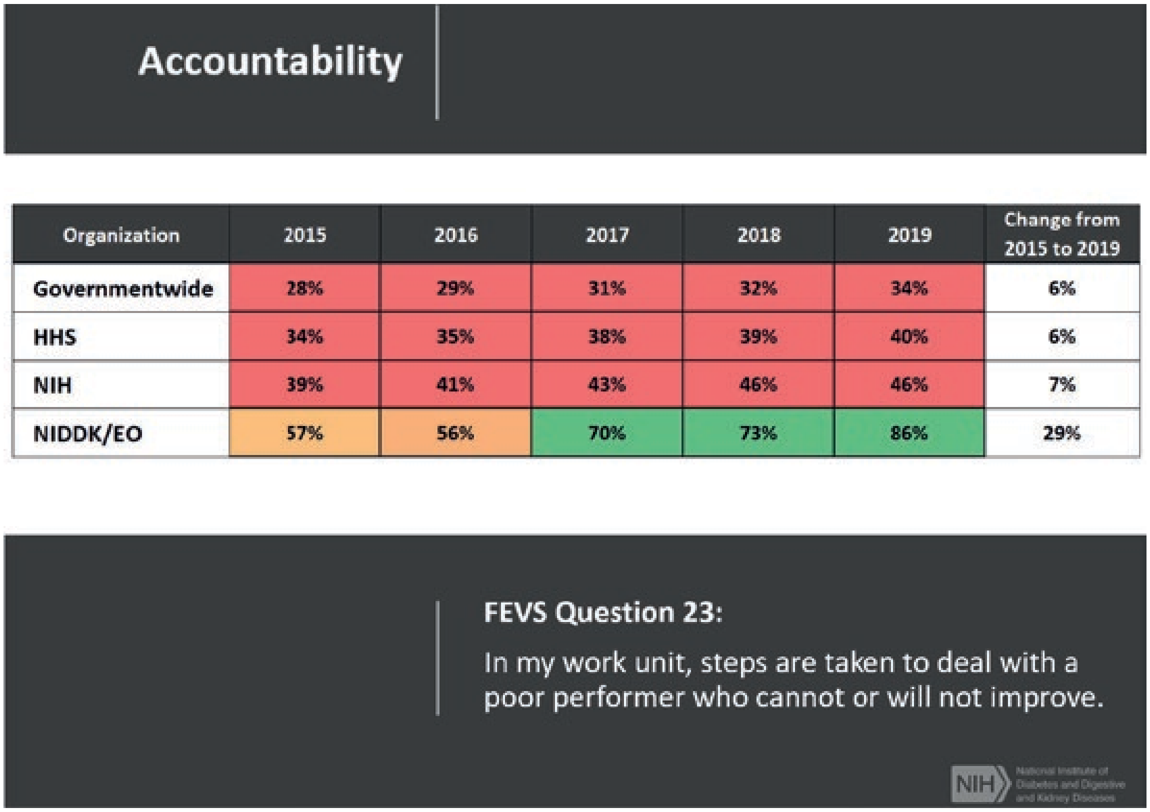
Source: Screenshot of the National Institutes of Health's FEVS dashboard.
Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; HHS = Health and Human Services; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; NIH = National Institutes of Health.

public were able to analyze the raw data by work unit, problem areas in the public service were increasingly difficult to hide within broader averages. This increased peer pressure on managers and changed the nature of recruitment because the quality of the workplace became more transparent. The FEVS data thus generated pressure from within and from outside organizations, making senior leaders more likely to pay attention to its findings.

In contrast to the "top-down" culture change process outlined so far, there are examples of "bottom-up" efforts to build responsiveness to FEVS results. Even when senior management fails to respond to the FEVS, public officials still want to use it as a tool to highlight problems in their organizations. These federal employees may have a passion for improving employee engagement for the benefit of staff and to better support their organizations' missions, or they may want to improve their own work environments and see the FEVS as a tool to do so. In either case, the FEVS provides them with the ability to draw attention to needs, obtain buy-in for proposals, and measure the impact of the work being done.

A constraint to building a "bottom-up" culture for the use of the FEVS is the ability to communicate throughout an organization. There are often significant hurdles to frontline staff's agreeing on the key issues presented in the FEVS data and generating a strategy in response. One such issue is the diversity of challenges faced within a single agency, as illustrated in figure 26.2. Thus, at least to date, much of the culture change around the use of the FEVS has arisen from the actions of survey analysts and senior management.

FIGURE 26.5 National Institute of Diabetes and Digestive and Kidney Diseases Staff Responses to Federal Employee Viewpoint Survey "Accountability" Question Compared to Organization-Level and Governmentwide Responses



Accountability

Organization	2015	2016	2017	2018	2019	Change from 2015 to 2019
Governmentwide	28%	29%	31%	32%	34%	6%
HHS	34%	35%	38%	39%	40%	6%
NIH	39%	41%	43%	46%	46%	7%
NIDDK/EO	57%	56%	70%	73%	86%	29%

FEVS Question 23:

In my work unit, steps are taken to deal with a poor performer who cannot or will not improve.

Source: Screenshot of the National Institutes of Health's FEVS dashboard.
Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; HHS = Health and Human Services; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; NIH = National Institutes of Health.

THE FEVS IN A WIDER ORGANIZATIONAL AND SOCIETAL ENVIRONMENT

The evolution of the FEVS as a tool for public action at the agency level has interacted significantly with the wider organizational and societal environment of public service. Centralized initiatives have driven agencies to better understand and value the survey as a management tool. Grassroots initiatives have generated novel insights and created a community of users. And external actors have supported the development of reform initiatives by presenting the FEVS data through new lenses. This section paints a picture of some of the most influential aspects of the broader environment in which agencies have worked.

Influencing from the Center

Senior leaders are most likely to seek out the sorts of relationships with survey analysts that drive effective reform when they are told to do so from the president's office. In 2014, the FEVS became a part of the PMA and was added to leadership performance plans. This grabbed the attention of any senior manager who had not yet taken their FEVS results seriously and gave the FEVS a stronger accountability role. The PMA also encouraged senior managers to learn from agencies that had been the most successful in developing a quality work environment and personnel management. An initiative collected successful workforce practices from across the federal government and created a platform to share them broadly. As a result, federal leaders, supervisors, and human resources practitioners can now easily review, evaluate, and adopt—or adapt—these proven successful practices with minimal effort. Appendix M.1 provides screenshots from the website.[15]

In a similar vein, the OPM has tried to collate best practices from across the public service so that when agencies determine that they want to improve in a particular area, they have resources to turn to. Appendix M showcases a screenshot of the OPM's "Successful Workforce Practices" webpage that links to the resources outlined in appendix M.2, as well as other resources. The intention is thus to provide learning resources as well as accountability.

Such initiatives clearly complement agency-specific efforts to respond to the FEVS results. By increasing the salience of and incentives for responding to issues highlighted in the FEVS, the center makes senior officials more likely to set up an architecture like that outlined in figure 26.3. The learning resources provide a menu of options for responding to identified issues.

Much of this thinking was brought together by the "20-20-20 initiative." The effort, a pillar of President Donald Trump's first PMA, took aim at the lowest-performing work units in an agency. Trends had shown that while the entire government was improving, these units were falling even further behind. Notably, no one at the leadership level was responsible for focusing on these units. With the culture change now firmly in place, the goal was to improve the lowest 20 percent of an agency's work units by 20 percent by 2020.

Influencing across Agencies

Centralized initiatives to share best practices suffer from many of the same issues as centralized surveys. Topics, best-practice methods, and recommendations are all chosen at the center. But cross-agency collaboration and learning in the face of FEVS results have only grown over the past two decades. This learning is related to the facilitation and analysis of the FEVS as well as potential practices that could be put in place to address challenges identified by the survey.

One example of how agencies have tried to support one another in analyzing and responding to FEVS data is the Employee Viewpoint Survey Analysis and Results Tool (EVS ART).[16] In 2015, a small team at the NIH and the NIDDK worked to create a framework that would allow users to translate the enormous amount of survey data they received from the FEVS in a user-friendly and efficient manner. In many ways, it made the task of the first column of figure 26.3 easier by expanding the set of individuals who could undertake that role and potentially widening the number of managers able to analyze FEVS data themselves.

The resulting Excel-based tool, EVS ART, has provided officials across the government with a no-cost, practical, and easy-to-use resource that allows for the easy identification of focus areas with substantial time and cost savings. The team has gifted EVS ART governmentwide, with no usage or licensing fees, and has provided training and support for using it. This has helped to eliminate the duplication of effort because agencies and supervisors no longer need to conduct supplemental analysis. With a simple "copy and paste" motion, EVS ART takes employees' FEVS feedback and—through a series of hidden pivot tables—translates it into the index measures supplied by the OPM (see screenshots of EVS ART in appendix M.3).

EVS ART contributed to solving two problems. First, it gave survey analysts and the wider community interested in analyzing FEVS data a handy tool for analysis. This reduced the time and cost required to produce disaggregated reports. Second, it showcased the use of the FEVS by other agencies, raising the survey's profile as a management tool. EVS ART has been generally well received across the public service. As Tracey Hilliard states, "When EVS ART came out, that was wonderful." It is an example of how grassroots action can complement agency efforts. However, EVS ART is only part of the wider architecture we have outlined, not a substitute for it.

Influencing from beyond the Public Service

Finally, as hinted at above, external stakeholders have played a role in the development of the FEVS as a tool for public action. Simply by expecting the government to become more analytical, external stakeholders can put pressure on the government to use the FEVS as a tool. However, such an abstract approach is unlikely to gain traction. Instead, most external stakeholders have used the data to draw out interesting perspectives regarding public service that have influenced the debate inside the government about priorities for reform.

The most famous example of this approach is the Partnership for Public Service's Best Places to Work in the Federal Government index. The index ranks government organizations based on FEVS data in terms of the quality of the experience of working for them. As an example of how influential the index was, the December 2015 edition of *MyUSDA: A Progress Report for Employees on USDA's Cultural Transformation* was headlined "USDA Moves Up in Best Places to Work Ranking."

The Best Places to Work index uses a simple idea to motivate agencies to improve their rankings. Agency staff may feel motivated by a desire for their agency to look better in the rankings or to improve the talent pool that seeks employment at the organization. The index brings what is a relatively dry personnel issue in the public sector into the public sphere. Thevee Gray feels that:

> the Partnership for Public Service has assisted in gaining the necessary attention for the FEVS. Senior leadership desires for their agencies to be seen as one of the best places to work in the government. As such, it holds them accountable and increases their responsibility to take the FEVS results seriously, knowing they will be published in an influential index and debated publicly.

CONCLUSION

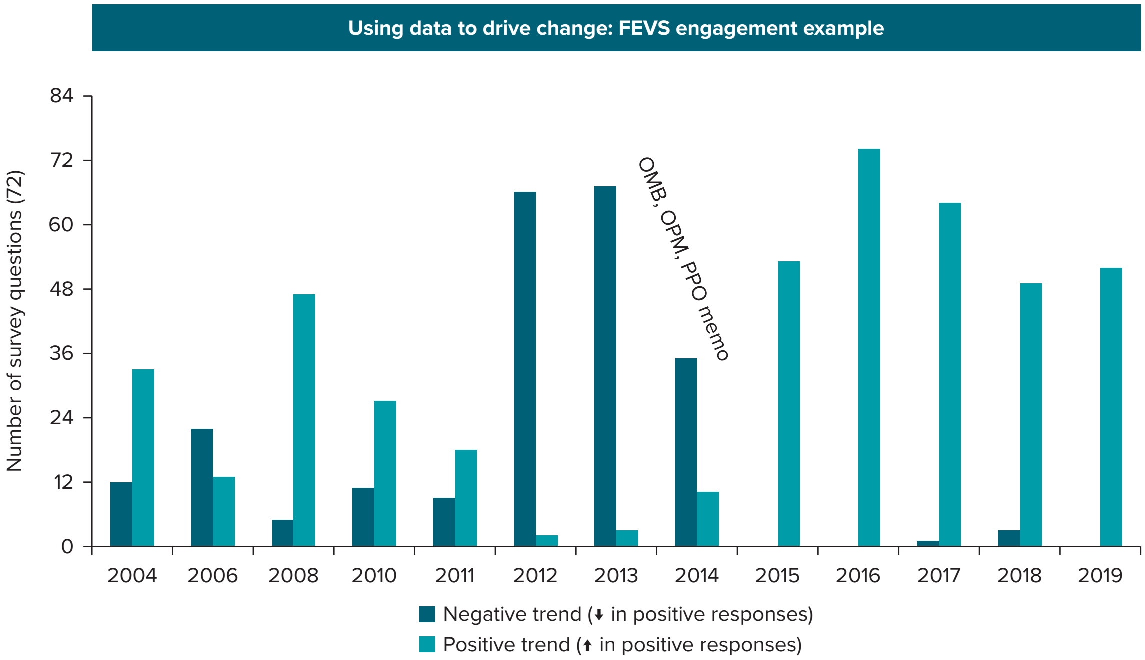
This chapter has argued for the potential power of surveys of public employees for driving public sector reform and improvement in the functioning of government. But it has cautioned that surveys must be embedded within a wider architecture of policy action for impacts to be realized. This architecture requires that a public sector organization have an official capable of translating survey results into actionable advice that a senior manager, working closely with that official, can use to build momentum for reform. And it requires responsiveness to survey results to build a culture that induces wider agency staff to contribute to reform.

Evidence for the arguments laid out in this chapter can be found in a 2014 joint memo from the Executive Office of the President and the OPM (Donovan et al. 2014). In many ways, its guidance closely tracks the arguments made here. After noticing that early adopters, like the Department of Transportation and the Federal Labor Relations Authority, demonstrated rapid improvement in their FEVS scores because of dedicated effort from their senior leaders, the OMB determined that the FEVS and employee engagement deserved elevation as a cross-agency priority goal in the PMA. A joint memorandum signed by OMB, OPM, and White House Presidential Personnel Office leadership with the subject line "Strengthening Employee Engagement and Organization Performance" laid out explicit mandates to agencies that echo the arguments of this paper.

Notably, each agency's career and noncareer leadership needed to take responsibility for changing how they had previously responded or, more realistically, did not respond to employee feedback. For two years after 2014, agencies had senior accountable officials and full-time staff dedicated to analyzing results and creating immediate action plans. The OMB included the results in the "FedStat" meetings held at senior levels as well as in reports to senior White House officials and directly to President Barack Obama.

The results from this effort were quickly apparent in the FEVS data themselves. Whereas the broad trend between 2012 and 2014 had been a declining number of positive responses to the FEVS questions,

FIGURE 26.6 Trends in Negative and Positive Responses to Federal Employee Viewpoint Survey Questions, 2004–19



Source: US Office of Personnel Management.
Note: FEVS = Federal Employee Viewpoint Survey; OMB = Office of Management and Budget; OPM = Office of Personnel Management; PPO = Public Procurement Office.

the publication of the joint memo led to a positive trend in the majority of the FEVS questions in the following years (as shown in figure 26.6). This happened despite a change of administration, a government shutdown that lasted more than a month, below-market pay adjustments, and the start of the COVID-19 pandemic. Thus, as the FEVS became embedded in a wider architecture for public action, it became a stimulus for reform.

In many cases, where FEVS results were not translated into change in practice, it has been due to the lack of architecture at the agency level to support that translation process. Simply producing survey data, of however high a quality, is rarely enough to drive public action.

Despite the qualities of the FEVS and its related successes, there are legitimate criticisms of the relevance and scope of the FEVS survey questions. The FEVS focuses on drivers of staff engagement and thus has a limited scope in terms of topics. Similarly, given the complexity and breadth of work undertaken in the US federal government, it seems natural that a standardized survey would not be the most effective driver of change across all federal agencies all of the time. But the impacts it has inspired showcase the potential of employee surveys in inducing public action for better government.

NOTES

The authors are grateful to Corey Adams, Gonzalo Ferro, Thevee Gray, Tracy Hilliard, and Stephen Pellegrino for discussions. Many of the authors have played a key role in the development of the Office of Personnel Management's Federal Employee Viewpoint Survey or in its translation and use at the agency level over the past decade. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of any of the US federal agencies with which several authors are affiliated, including the US Office of Personnel Management.

1. Discussion of aspects of the FEVS can be found in chapter 9,, case study 9.3 in chapter 9, and chapter 25. FEVS data are also used in chapters 19, 20, 21, and 22.
2. For those interested in the specific details of the survey, the OPM releases technical reports each year to accompany the survey report and data. These are available on the OPM FEVS website at https://www.opm.gov/OPMFEVS/reports/technical-reports/. A special "research synthesis" in the *Public Administration Review* (Callahan 2015) articulated a series of academic perspectives on the strengths and weaknesses of the FEVS for public service reform and research.
3. See, for example, GAO (2015), which focuses on drivers of engagement and implications for various agencies; GAO (2018), which focuses on the OPM's delivery of information on performance management; and GAO (2021), which focuses on the US Department of Homeland Security.
4. The current Best Places to Work index is available on the Partnership for Public Service's website at https://bestplacestowork.org/about/methodology/.
5. The publication of the index by an external entity also allows for independent assessments of what drives improvements in the public service work environment, such as Partnership for Public Service and Deloitte (2013).
6. The full set of reports published by OPM can be found in the "Data Reports" section under "Reports" on the OPM FEVS website at https://www.opm.gov/fevs/reports/data-reports. Releases are limited to groups of at least 10 officers to safeguard against the identification of respondents.
7. The FEVS also provides the OPM itself with insights into weaknesses in the public service system as a whole that can be targeted without direct agency action. However, this chapter will focus on agency-level reform efforts in response to FEVS findings.
8. There are many potential topics on which a survey could focus, including the physical environment, relationships between colleagues, the quality of management, the engagement of the survey respondent with his or her job, and the most significant challenges the respondent finds to undertaking their work effectively. Chapter 18 provides an overview of the topics the world's major public servant surveys focus on.
9. National Defense Authorization Act for Fiscal Year 2004, Public Law 108–136, Nov. 24, 2003, 117 STAT. 1641.
10. *Examining Misconduct and Mismanagement at the National Park Service: Hearing before the Committee on Oversight and Government Reform, House of Representatives*, 114th Cong. (2016). https://www.congress.gov/event/114th-congress/house-event/LC51983/text?s=1&r=100.
11. In a statement in response to the findings of the hearings, Deputy Director of the National Parks Service Michael Reynolds made reference to actions the agency took to try to improve working conditions for park staff (Reynolds 2016). A webpage outlining the response to the harassment issues can be found on the National Park Service website at https://www.nps.gov/aboutus/transparency-accountability.htm.
12. An alternative perspective is that centralized surveying generates greater awareness and appetite for surveys of public servants, thus increasing the likelihood of complementary efforts by managers. Frequently, the FEVS has inspired follow-up surveys by agencies seeking to better understand an area in which they are performing relatively poorly. And the structured survey approach of the FEVS can be complemented by deeper-dive focus groups and listening sessions, which can more deeply explore red flags relevant to a particular agency.
13. The FEVS has gradually changed its content over time to meet the evolving demands of officials using it for policy assessment and organization development initiatives. As outlined in figure 26.1, in 2012, a series of items were added to the FEVS to improve how well it could inform OPM policy evaluations, reports to Congress, and oversight functions, as well as workforce development initiatives within agencies. In 2020, items were added to support policy assessments, including military spouse items and new leave policies for COVID-19 pandemic response. Simultaneously, the performance confidence index was added to support change and development initiatives and action in agencies.
14. Conceptually, organizational climate is a surface manifestation of culture: employees' perceptions of management practices and policies speak to the values and norms embodied in a culture.
15. The OPM highlights the main features of the successful workforce practices initiative on the "Successful Workforce Practices" page of its website at https://www.opm.gov/policy-data-oversight/human-capital-management/successful-workforce-practices/. The website housing the full collection of successful practices is accessible by US government officials at https://community.max.gov/display/HumanCapital/PMA+Successful+Workforce+Practices+Home.
16. EVS ART can be accessed by all US government officials at https://community.max.gov/display/HHS/EVS+ART. A fuller exposition of EVS ART is provided in chapter 9 and case study 9.3 in chapter 9.

REFERENCES

Brust, Amelia. 2021. "After 15 Years of Best Places to Work, Data Findings Consistently Point to Engagement Needs." *Federal News Network*, August 5, 2021. https://federalnewsnetwork.com/hiring-retention/2021/08/after-15-years-of-best-places-to-work-data-findings-consistently-point-to-engagement-needs/.

Callahan, Janelle. 2015. "From Results to Action: Using the Federal Employee Viewpoint Survey to Improve Agencies." *Public Administration Review* 75 (3): 399–400.

Donovan, Shaun, Beth Cobert, Katherine Archuleta, and Meg McLaughlin. 2014. "Strengthening Employee Engagement and Organizational Performance." Memorandum for Heads of Executive Departments and Agencies, December 23, 2014. https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2015/m-15-04.pdf.

GAO (US Government Accountability Office). 2015. *Federal Workforce: Additional Analysis and Sharing of Promising Practices Could Improve Employee Engagement and Performance*. Report to Congressional Requesters, GAO-15-585. Washington, DC: US Government Accountability Office. https://www.gao.gov/products/gao-15-585.

GAO (US Government Accountability Office). 2018. *Federal Workforce: Opportunities Exist for OPM to Further Innovation in Performance Management*. Report to the Chairman, Committee on Homeland Security and Governmental Affairs, US Senate, GAO-19-35. Washington, DC: US Government Accountability Office. https://www.gao.gov/products/gao-19-35.

GAO (US Government Accountability Office). 2021. *DHS Employee Morale: Some Improvements Made, but Additional Actions Needed to Strengthen Employee Engagement*. Report to the Chairman, Committee on Homeland Security, House of Representatives, GAO-21-204. Washington, DC: US Government Accountability Office. https://www.gao.gov/products/gao-21-204.

Mullins, Luke. 2021. "The Best Places to Work within the Federal Government, Ranked." *The Washingtonian*, June 29, 2021. https://www.washingtonian.com/2021/06/29/the-best-places-to-work-within-the-federal-government-ranked/.

OPM (Office of Personnel Management). 2012. *2011 Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: US Office of Personnel Management. Accessed April 10, 2022. https://www.opm.gov/fevs/reports/technical-reports/.

OPM (Office of Personnel Management). 2022. "About." OPM Federal Employee Viewpoint Survey, US Office of Personnel Management, United States Government (accessed March 10, 2022). https://www.opm.gov/fevs/about/.

Partnership for Public Service and Deloitte. 2013. *Ten Years of the Best Places to Work in the Federal Government Rankings: How Six Federal Agencies Improved Employee Satisfaction and Commitment.* Boston: Deloitte. https://www.opm.gov/policy-data-oversight/training-and-development/reference-materials/online-courses/maximizing-employee-engagement/content/common/cw/data/Ten_Years_of_BPTW_Rankings.pdf.

Reynolds, Michael. 2016. "NPS Misconduct: Examining Misconduct and Mismanagement at the National Parks Service." Statement of Michael Reynolds, Deputy Director, National Park Service, Department of the Interior, before the House Committee on Oversight and Government Reform, on the National Park Service response to incidents of employee misconduct, September 22, 2016. Office of Congressional and Legislative Affairs, US Department of the Interior. https://www.doi.gov/ocl/nps-misconduct.