

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.)

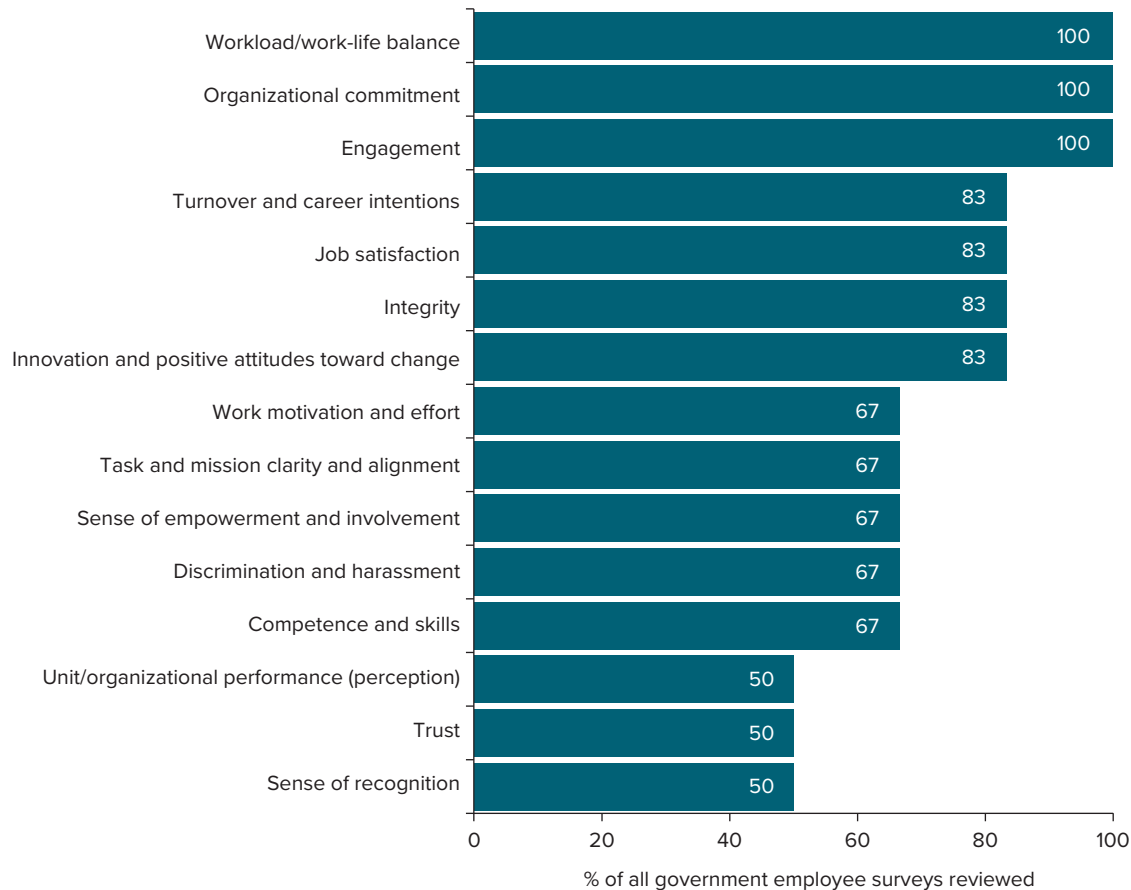
This chapter, therefore, assesses variation in a set of typical indicators derived from data sets of public service surveys to add to the existing literature and provide survey designers and analysts with benchmarks against which to assess their own efforts. The data span administrations in Africa, Asia, 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 Kastle 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.² 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.

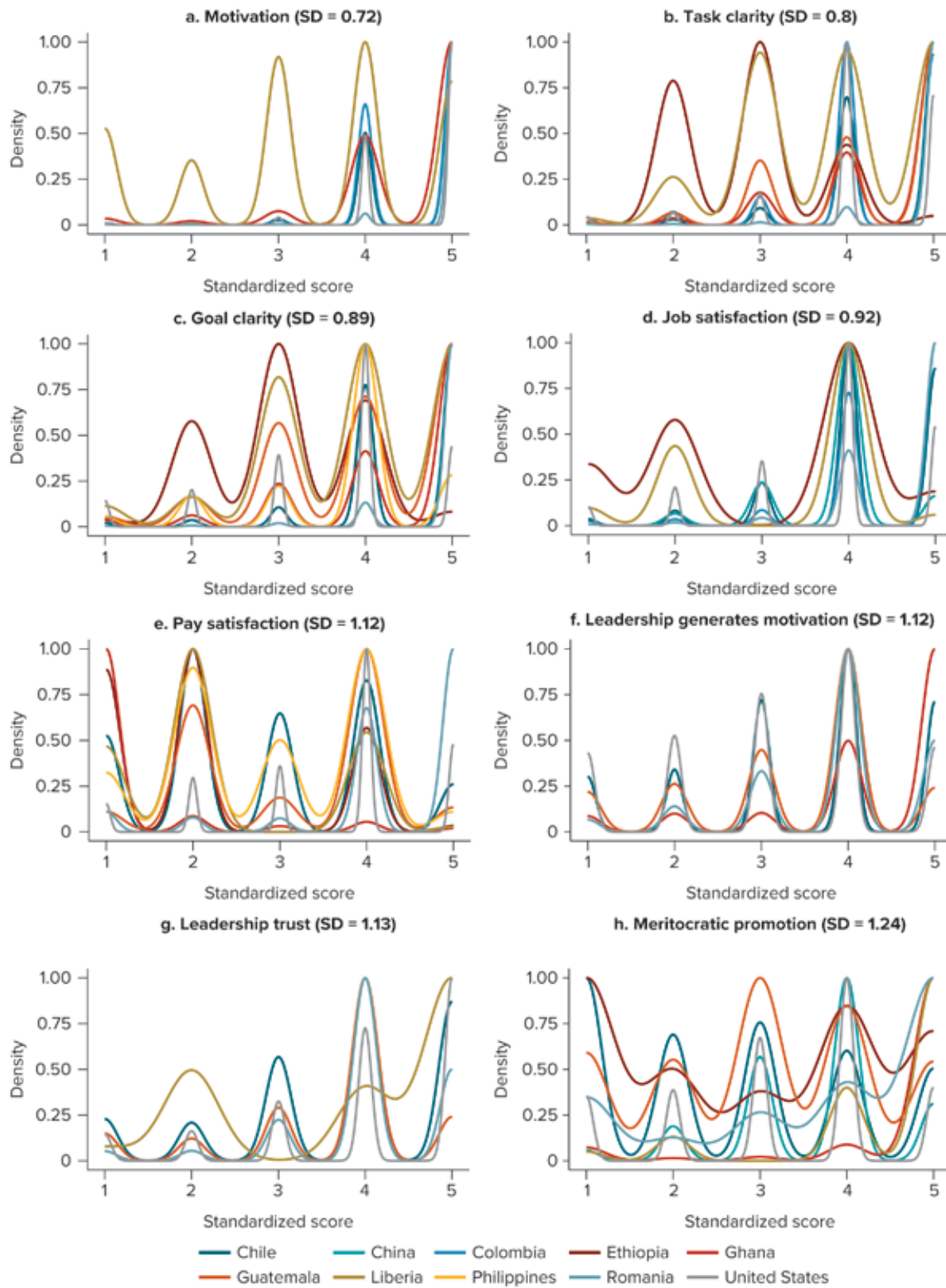
on leadership and promotion. Thus, current practice in measure design compacts a significant proportion of the variation into a minority of the response categories. This limits discriminating variation and the effectiveness of the measures.

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.

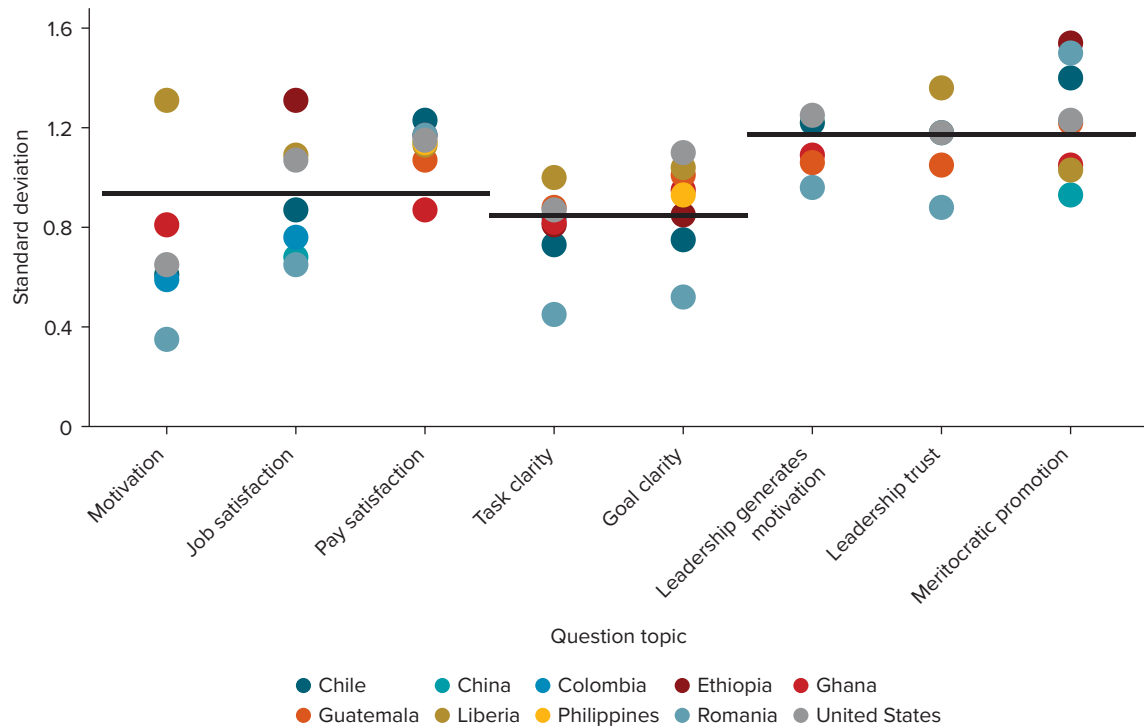
FIGURE 21.2 Distributions of Standardized Scores



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

- Affi, 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/lms/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](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 Kastle. 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.