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:

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

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

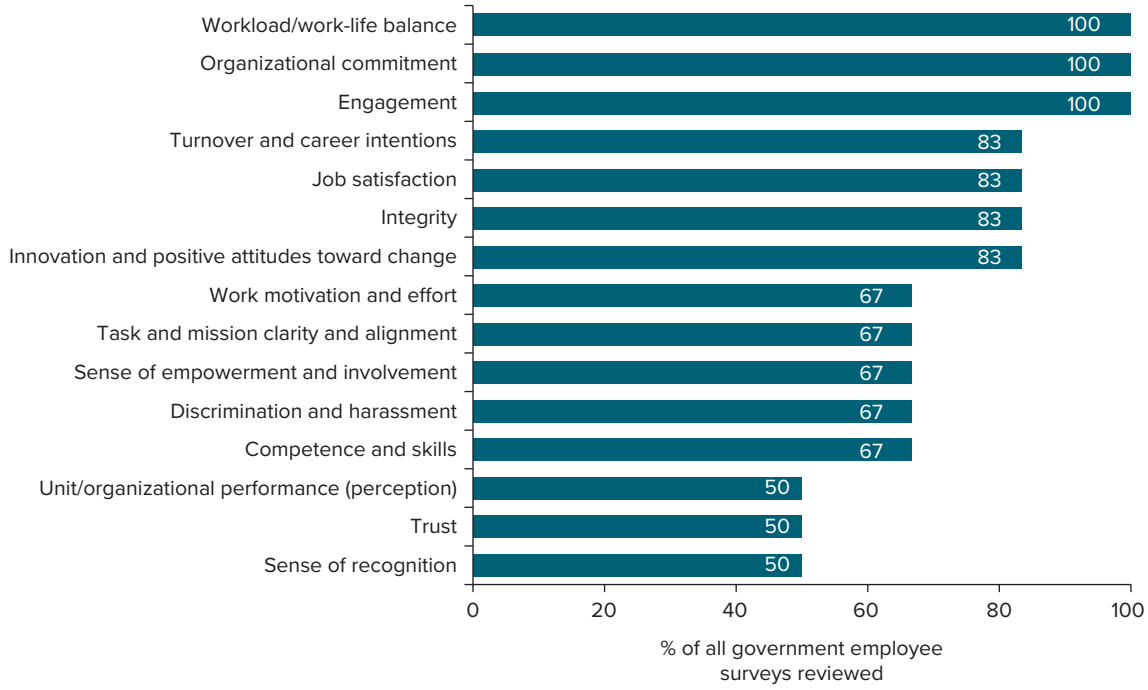
In this chapter, we illustrate what sampling error can be expected based on the variance observed in typical measures used in civil servant surveys and, consequently, what types of inference can be reliably drawn from them.

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.

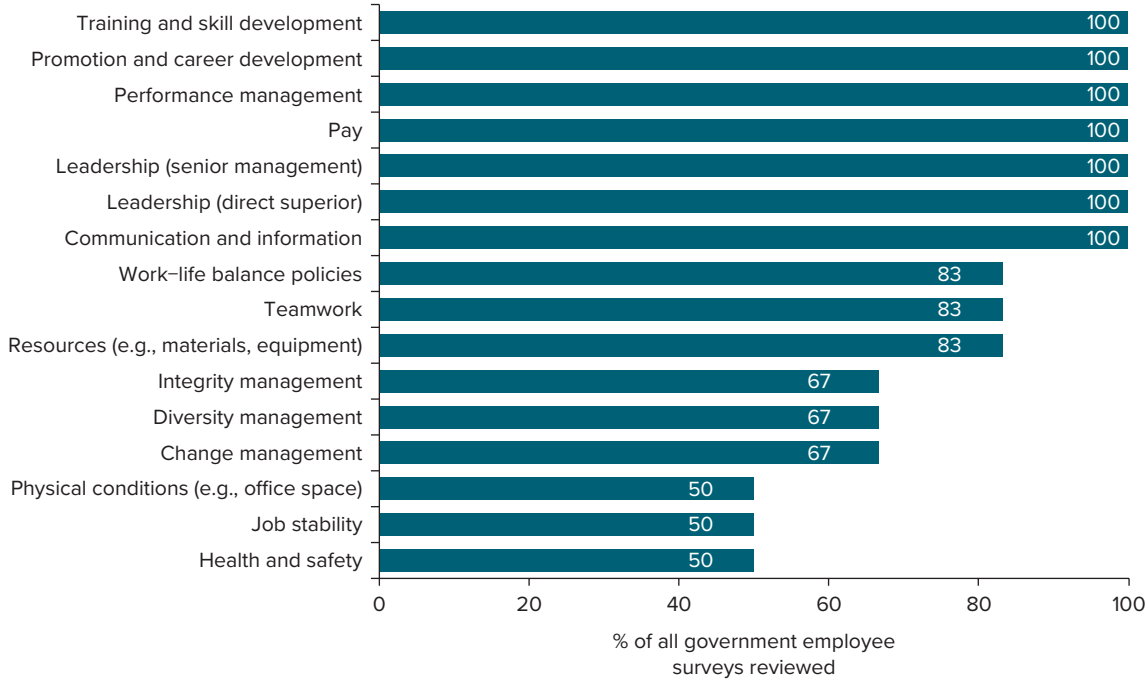
The selected surveys cover four continents and divergent socioeconomic contexts, as well as different sampling approaches and a range of widely used survey questions and indicators. Our analyses focus on a set of questions about job satisfaction, work motivation, performance review (evaluation), and 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)

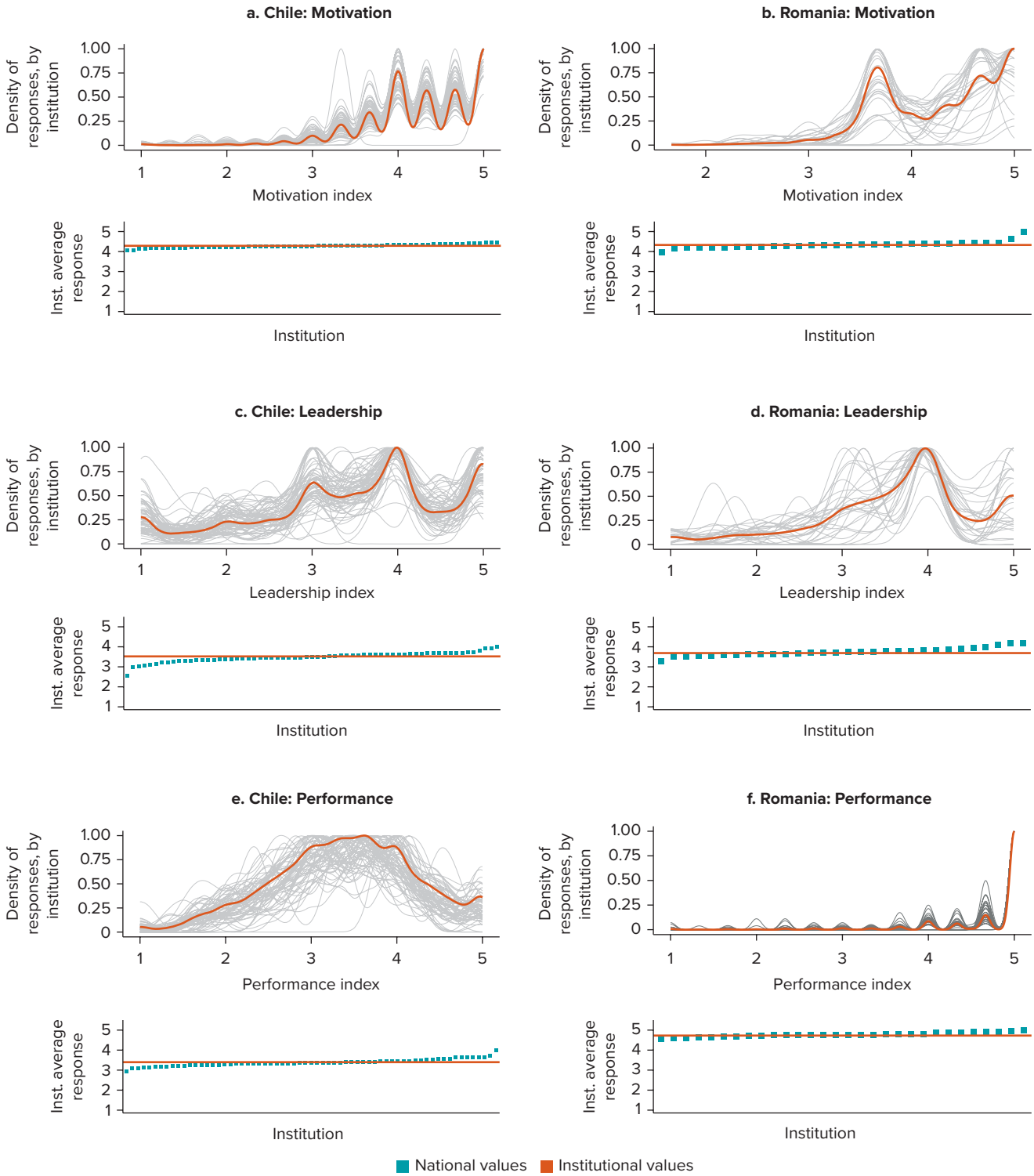
(continues on next page)

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)

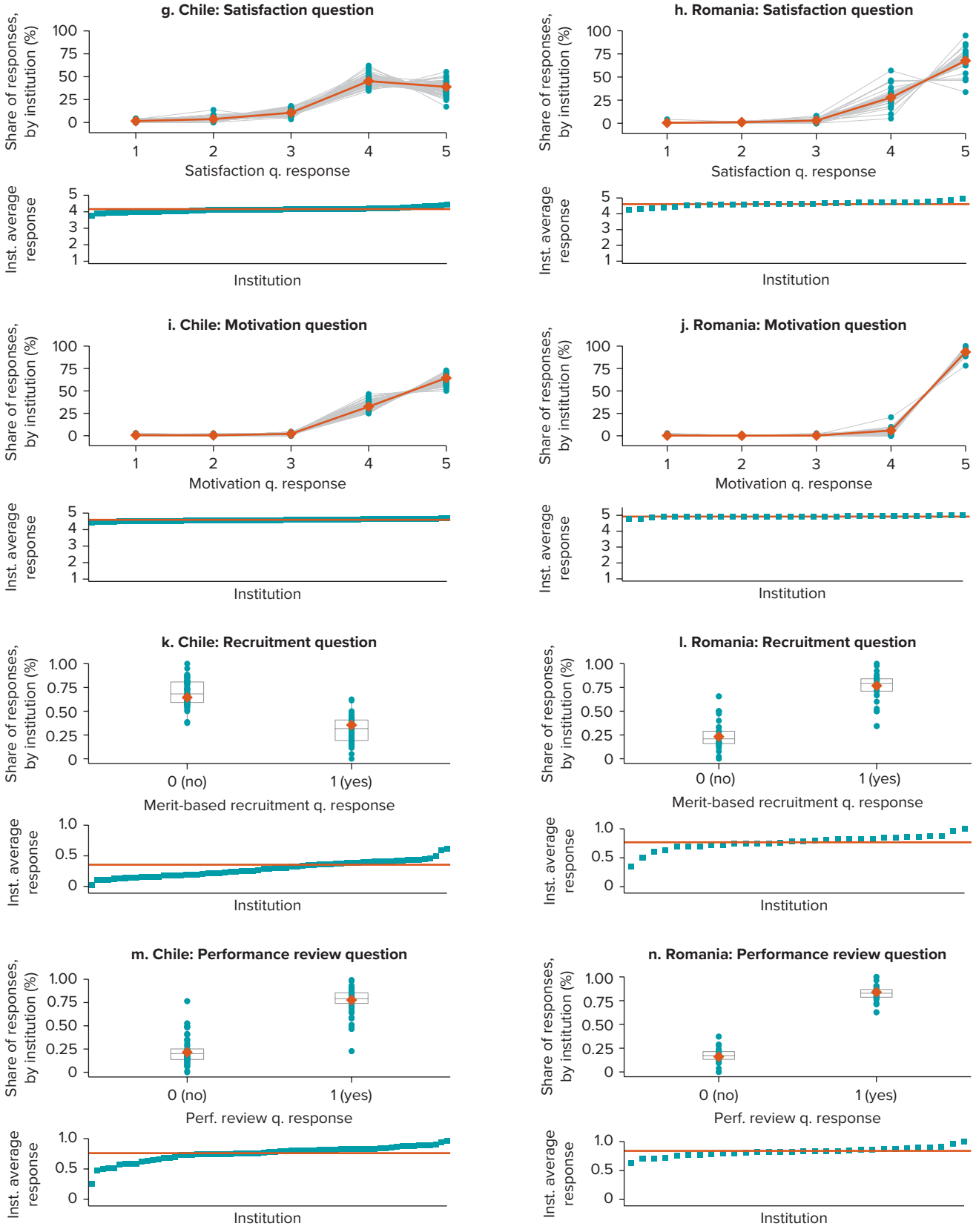
Source: Original table for this publication.

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



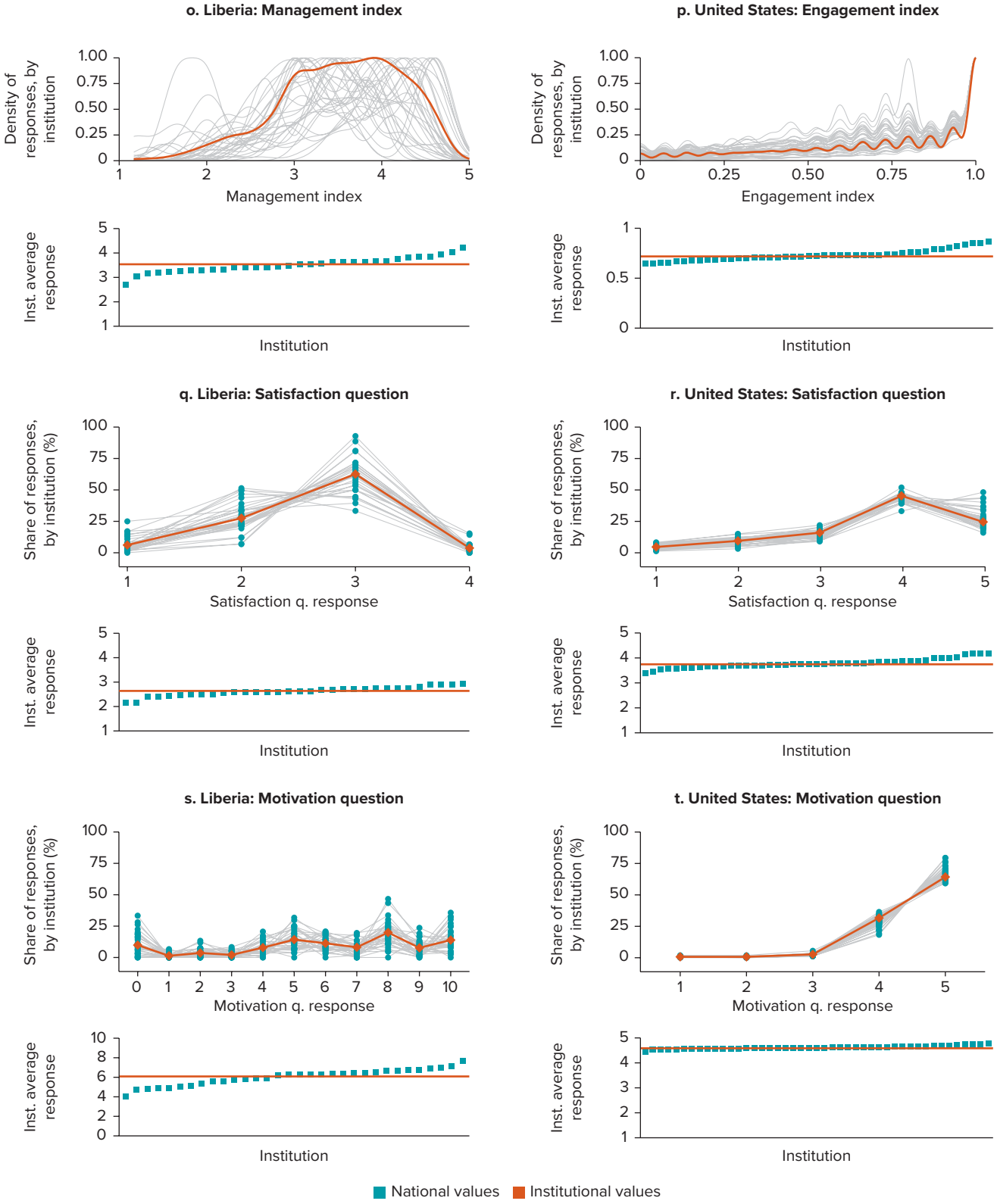
(continues on next page)

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.

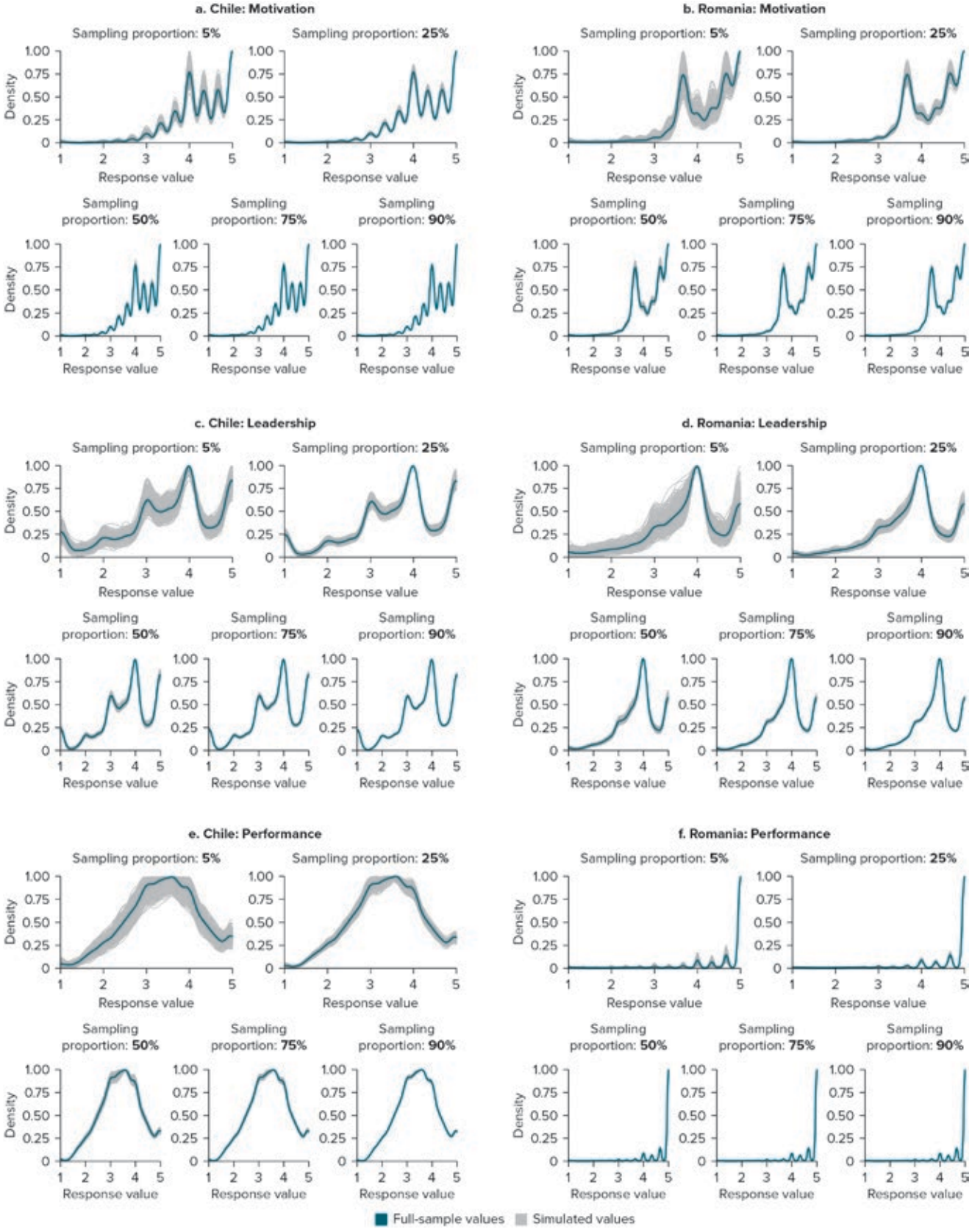
EMPIRICAL FINDINGS ON SAMPLE SIZE REQUIREMENTS FOR PUBLIC ADMINISTRATION SURVEYS

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.

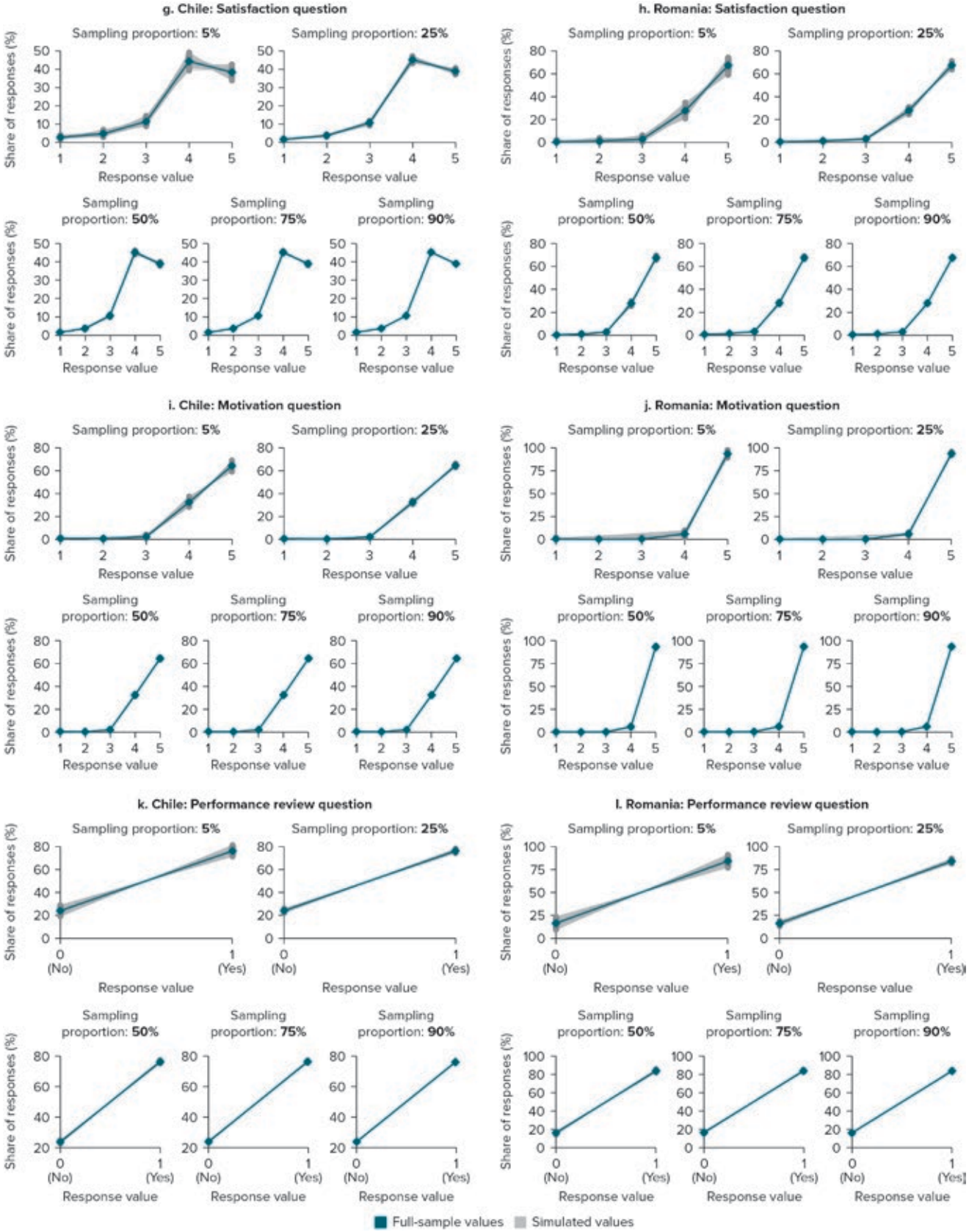
While our first metric suggests that countries oversample, our second and third metrics yield different conclusions. Consider, first, the results on the benchmarking of organizations. We find, as expected given the limited variation in many civil service survey indicators, that individual ranks are highly 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



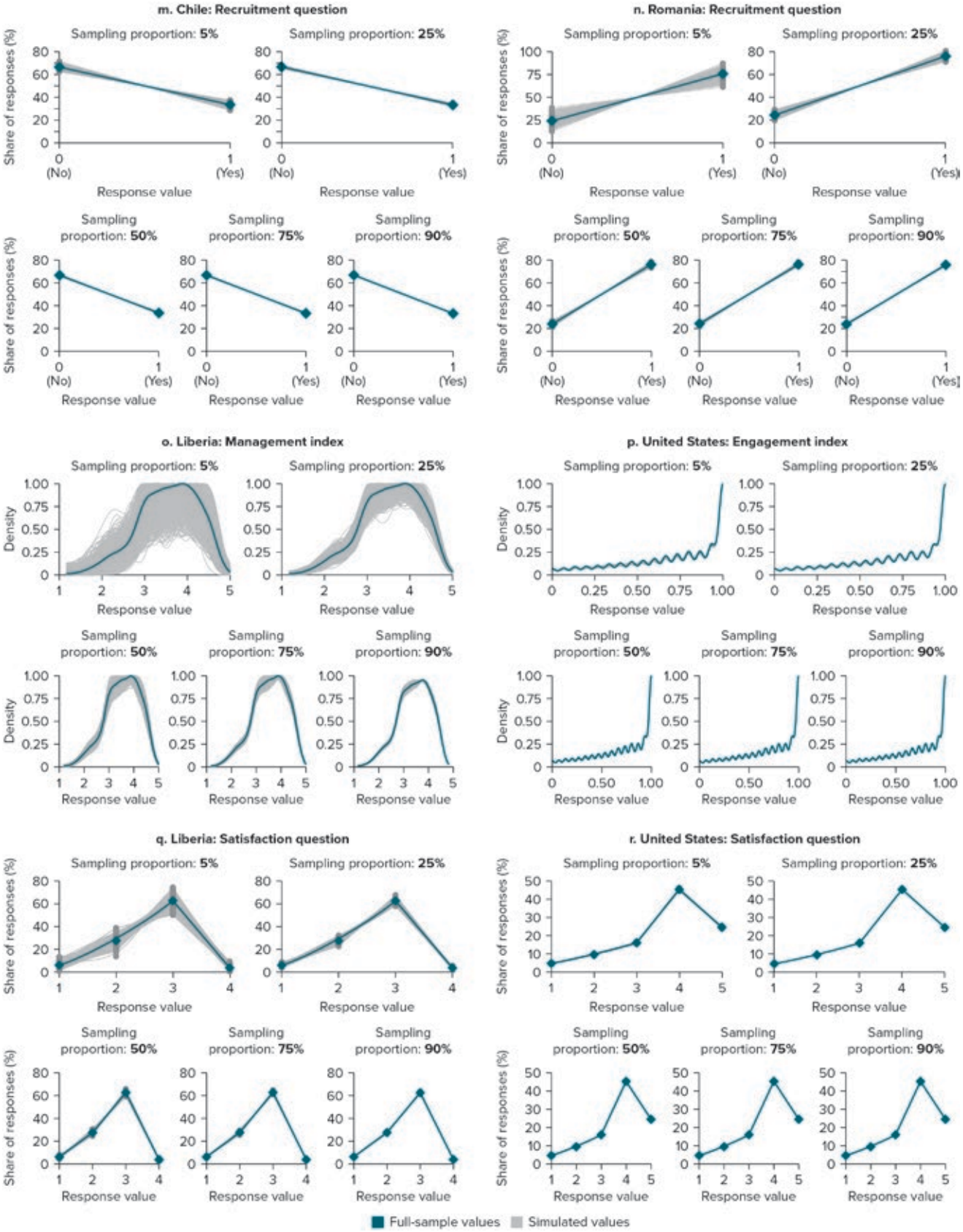
(continues on next page)

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



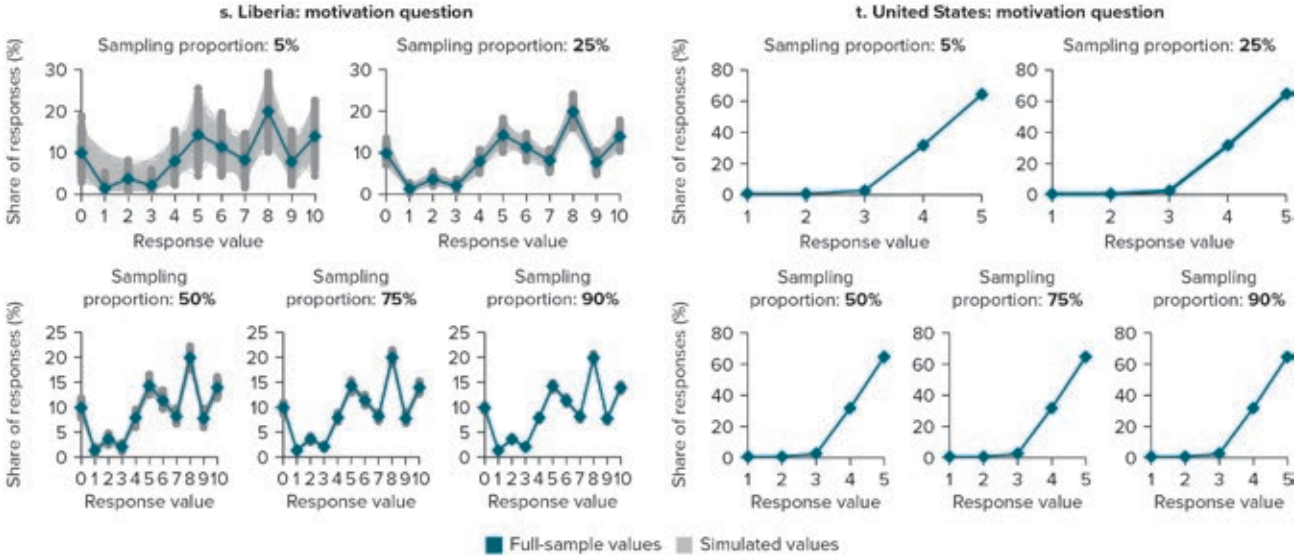
(continues on next page)

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.

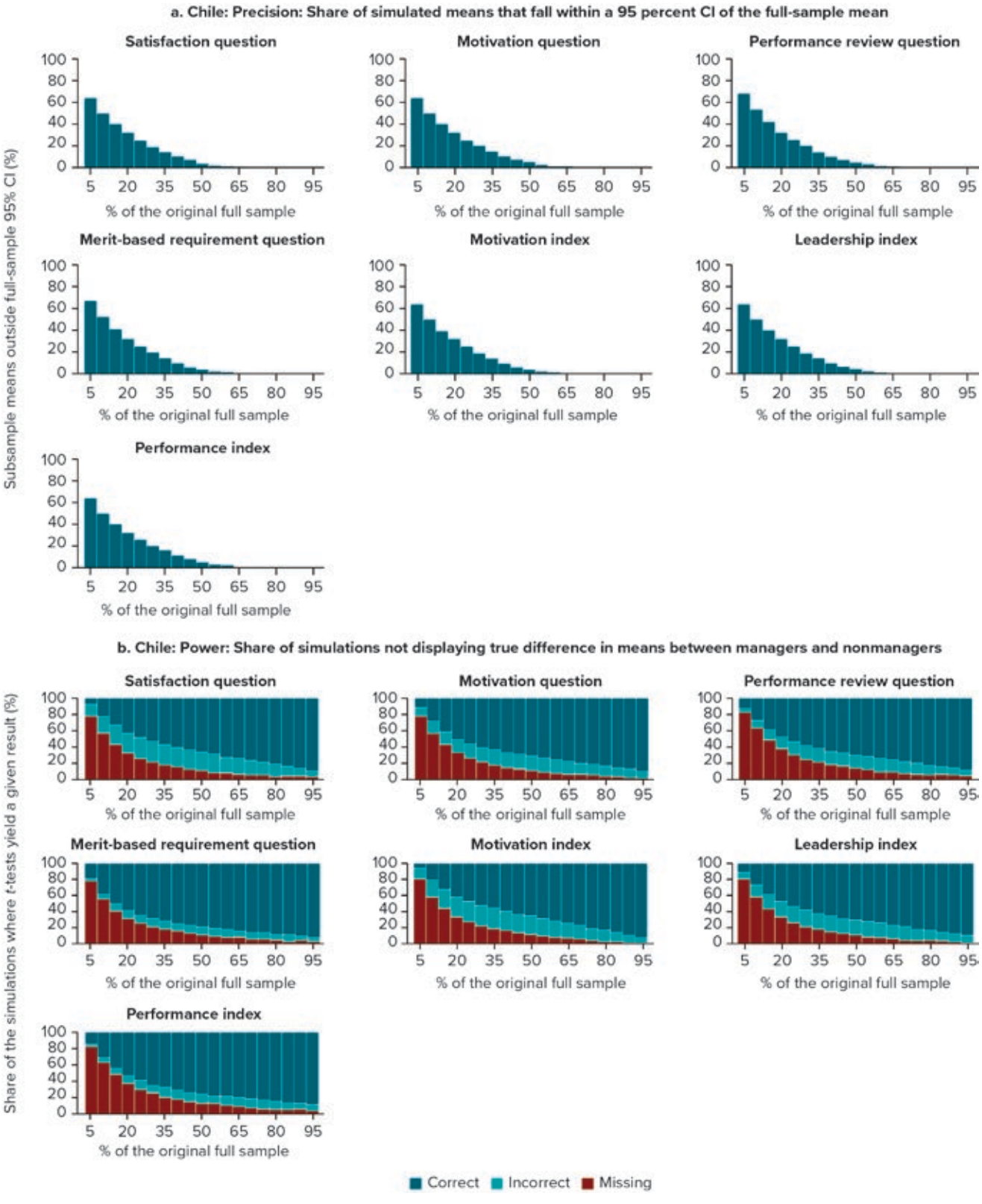
three to four ranks (see appendix H for the variation in institution-level values across simulations). We can also express this in terms of the Kendall's tau rank correlation coefficient, which indicates how well rankings obtained from the original data set correlate with those of the smaller samples. 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.

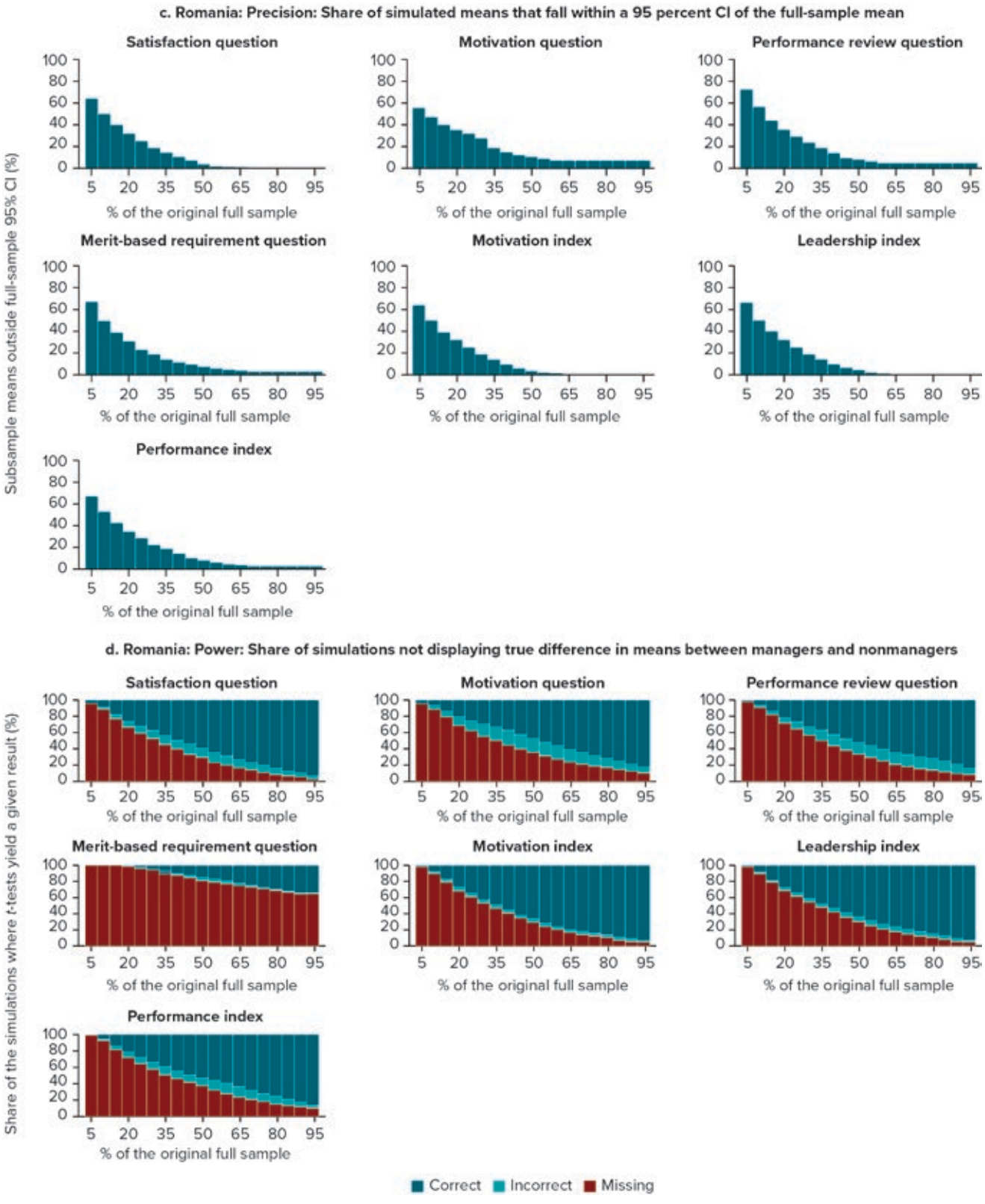
These small differences imply that the proportion of each of these subgroups needs to be rather large to be able to capture differences between the groups. Error rates for most indicators remain at around 20 percent with reduced sample sizes—considerably higher than the widely accepted 5–10 percent—until 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



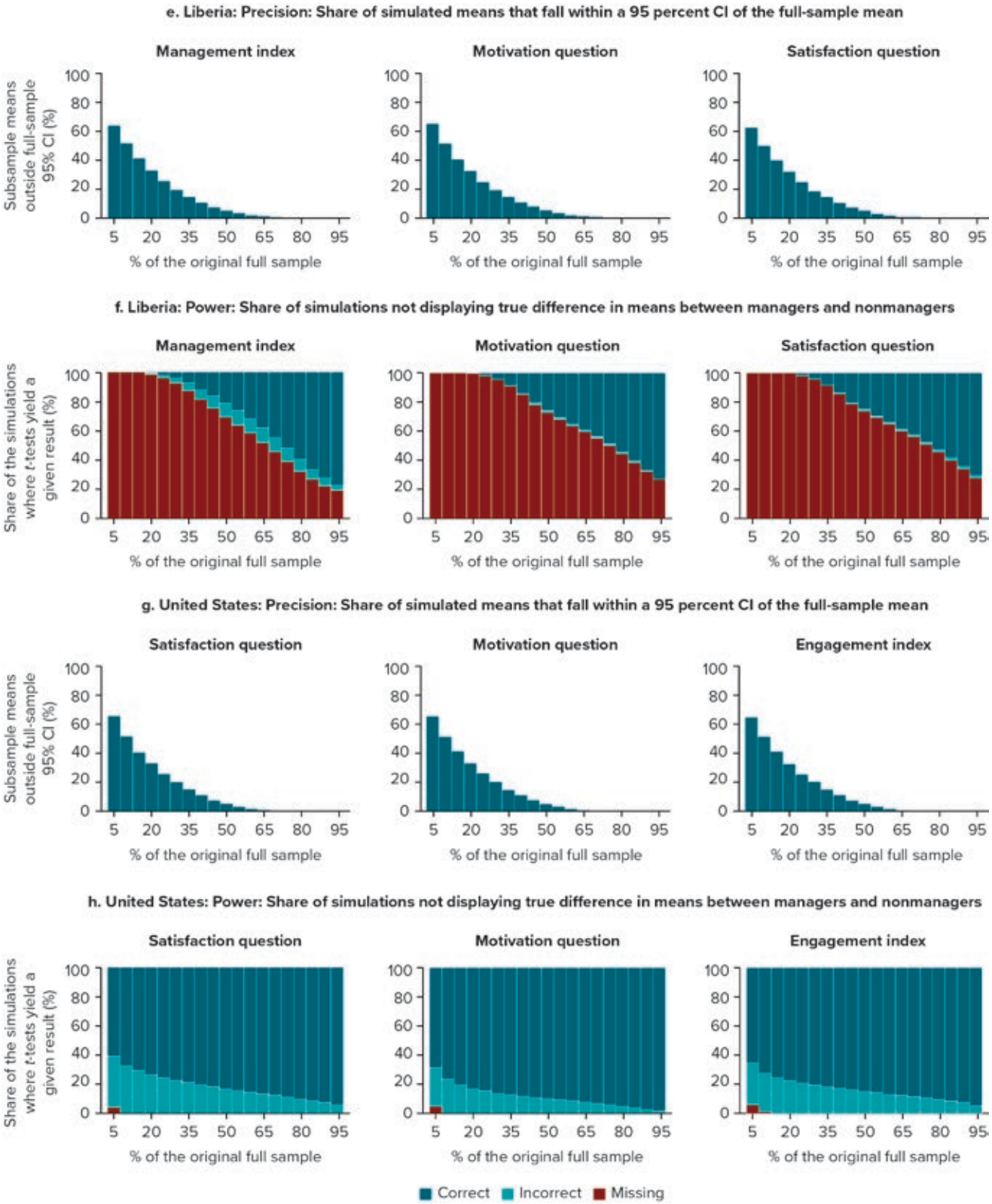
(continues on next page)

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



(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: $\tau = \frac{2\ (n_{concordant} - n_{discordant})}{n\ (n-1)}$.

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.