CHAPTER 19

Determining Survey Modes and Response Rates

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

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

SUMMARY

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

ANALYTICS IN PRACTICE

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

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

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

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

INTRODUCTION

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

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

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

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

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

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

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

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

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

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

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

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

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

LITERATURE REVIEW

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

Response Rates

Much of the existing research on survey modes has focused on the difference in response rates between modes. In general, online surveys have been found to have significantly lower response rates compared to all other survey modes, including face-to-face (Biemer et al. 2018; Lozar Manfreda et al. 2008; Shih and Fan 2008). While not specific to public administration, a recent meta-analysis conducted by Daikeler, Bošnjak, and Lozar Manfreda (2020) summarizes the results of 114 experimental studies conducted among many different populations (students, the general public, businesses, and employees), on diverse topics (public opinion, technology, lifestyle, job, etc.), by various sponsors (academic, governmental, and commercial),

both with and without participation incentives, and with varying recruitment strategies, prenotification methods, and solicitation methods. They found that in aggregate, online surveys have response rates that are 12 percentage points lower than all other survey modes.[3]

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

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

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

These differences between online respondents and nonrespondents to government surveys suggest that variation in response rates may significantly impact the degree to which online surveys provide unbiased estimates of public employees' perceptions and behaviors. In addition, the proclivity of managers and researchers to compare survey responses across organizations or other subgroups means that variation in response rates may lead to the comparison of differential subgroups of staff (Groves 2006). The self-selection issues in public administration surveys are less of a concern in the face-to-face mode because most surveys of this type record response rates close to 100 percent. For example, the Romanian face-to-face survey analyzed here collected responses from 3,316 out of 3,592 sampled individuals,

yielding a response rate of 92 percent. Similar surveys in different settings give comparably high response rates: for example, Guatemala (96 percent) and Ethiopia (94 percent). Assuming successful random sampling, the almost-perfect response rate minimizes any issues arising from differences between survey respondents and nonrespondents in the face-to-face mode.

Survey Breakoff

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

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

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

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

Survey Responses

Finally, a substantial portion of the existing survey research literature has focused on the degree to which survey modes may impact the magnitude of survey responses directly. In general, online survey respondents tend to exhibit lower levels of motivation to answer survey questions and often pay less attention when answering questions compared to face-to-face respondents (Kaminska and Foulsham 2014; Krosnick 1991). Several studies have found that online surveys are associated with higher rates of satisficing behaviors, including selecting "I don't know" or "N/A" response options, providing less differentiation across groups of responses, and providing more neutral responses (for example, "Neither agree nor disagree" or "Neutral") than face-to-face surveys (Duffy et al. 2005; Haan et al. 2017; Heerwegh and Loosveldt 2008). Online surveys

are also more likely to produce noncontingent responses (NCR), wherein there is a substantial difference between survey items that are expected to be highly correlated with each other (Heerwegh and Loosveldt 2008; Krosnick and Presser 2010). These kinds of responses imply that respondents may have simply selected answers at random or read through survey items carelessly in order to quickly complete the survey (Anduiza and Galais 2017). Taken together, these satisficing behaviors can reduce the validity and reliability of online responses (Baumgartner and Steenkamp 2001; Podsakoff et al. 2003).

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

A SURVEY MODE EFFECTS EXPERIMENT

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

How Does the Survey Mode Impact Response Rates?

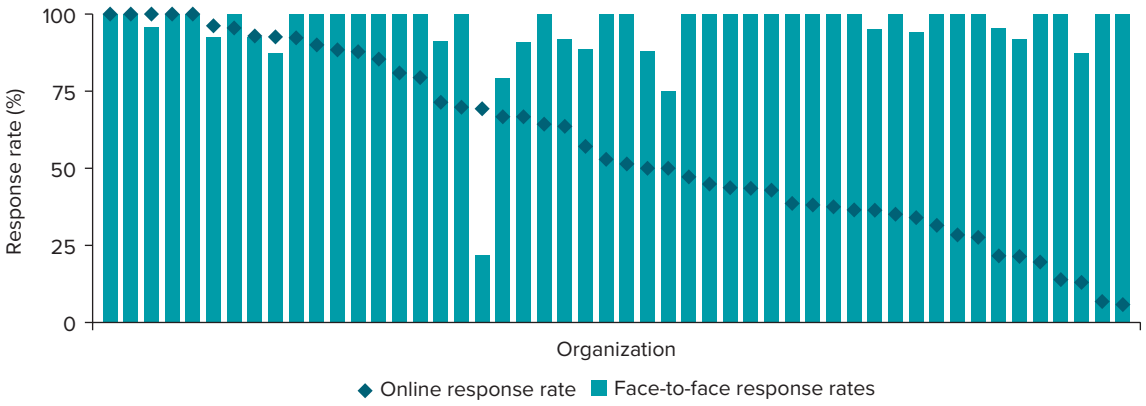
Our face-to-face survey has high response rates across most government institutions, with an average of 92.5 percent, while our online response rate—consistent with other online government employee surveys—varies widely across government institutions and ranges from a maximum value of 100 percent (5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

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

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

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

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



Source: Original figure for this publication.

TABLE 19.1 Balance in Demographic Characteristics between Surveys

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

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

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

Consistent with the existing literature, we find many statistically significant (at the 1 percent level) deviations from the population's values in the sample of online survey respondents. Most noticeably, 31 percent of face-to-face survey respondents are male, compared to only 26 percent female.[8] They are also relatively

less educated, with 47.4 percent having an undergraduate degree and 48.8 percent having a Master's degree or PhD, whereas, for online respondents, these numbers stand at 43.3 percent and 51.8 percent, respectively.[9] Moreover, 87.3 percent of face-to-face respondents are civil servants (as opposed to contractors), compared to 91 percent of online respondents. We also find statistically significant differences in average tenure, but being below one year, these differences appear to be of limited magnitude.

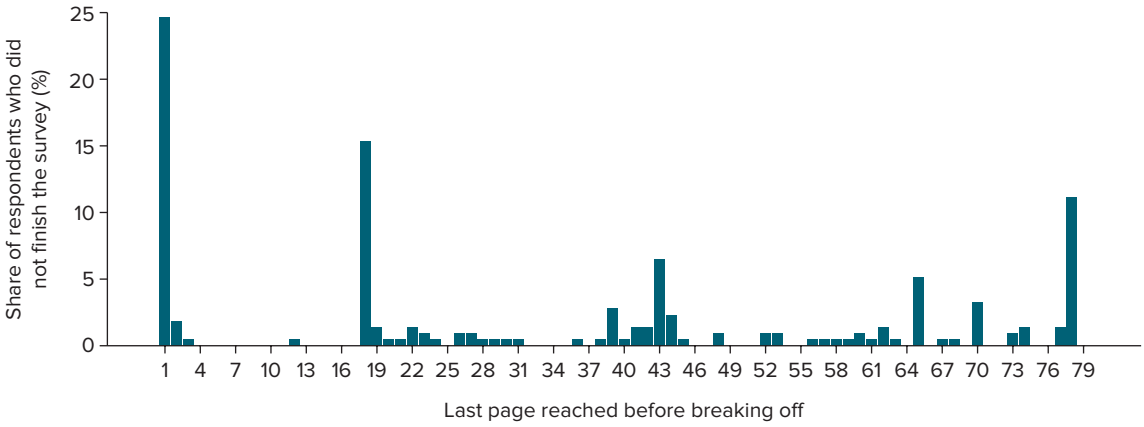
Overall, our data reflect the frequent finding that face-to-face and online samples differ along a range of margins. As many of these variables, like gender, education, and contract status, can affect survey responses, table 19.1 provides an initial rationale to look deeper into the differences between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

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

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

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

FIGURE 19.2 Online Survey Breakoff, by Page



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

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

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

SURVEY MODE EFFECTS ON THE VALIDITY AND RELIABILITY OF DATA

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

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

Does the Survey Mode Make a Difference to Question Values?

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

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

TABLE 19.2 Mean Modal Difference, by Level of Analysis

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

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

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

Country-Level Quantities

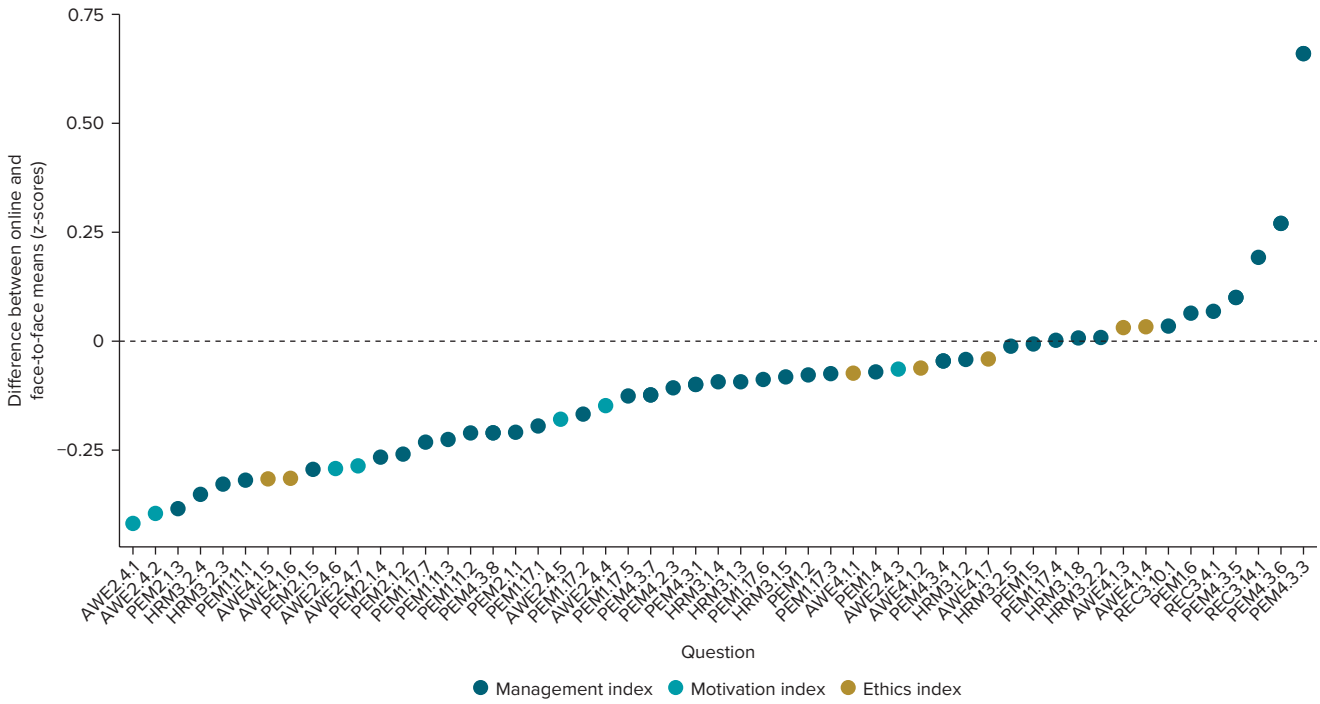
At the national level (panel 1 of table 19.2), we calculate the mean difference across all civil servants as the average score of the index in the online sample minus the average score on the same index in the face-to-face sample.

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

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

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

FIGURE 19.3 Average Item Difference, by Survey Topic



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

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

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

Organization-Level Quantities

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

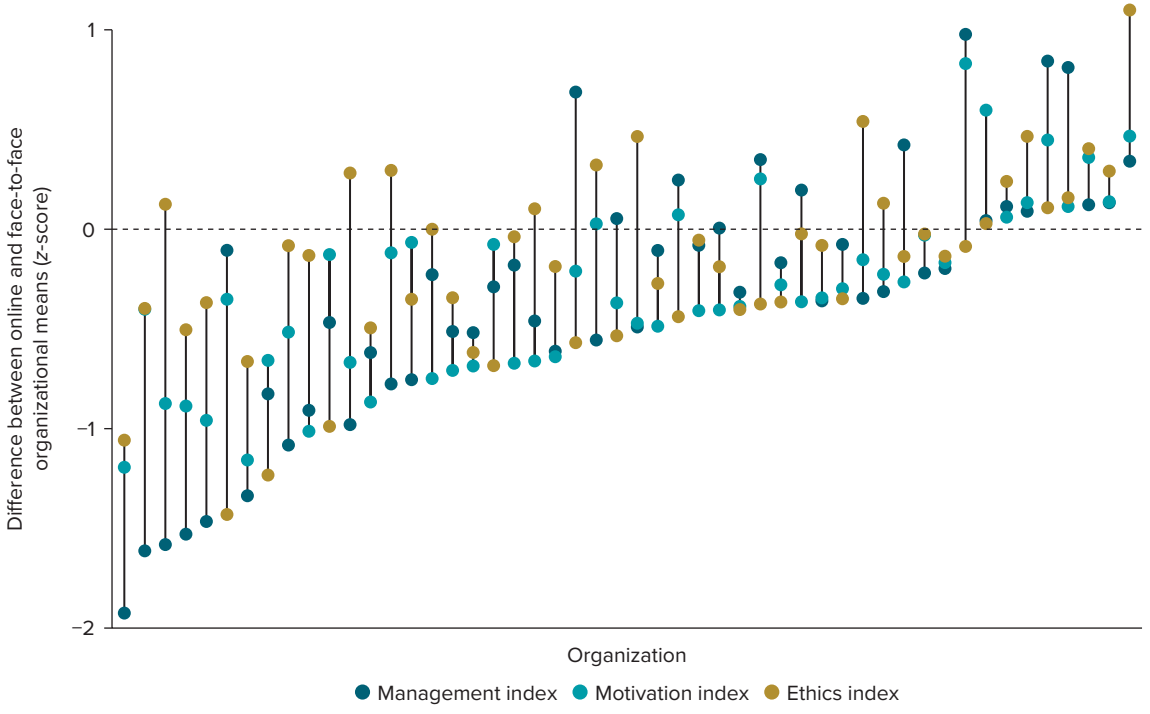
The *average* coefficients at the organizational level are not unlike those at the national level (perhaps naturally, since we are now simply producing a weighted correspondence of the national statistics). The change relative to the national level is the largest for the management index, where the mode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

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

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

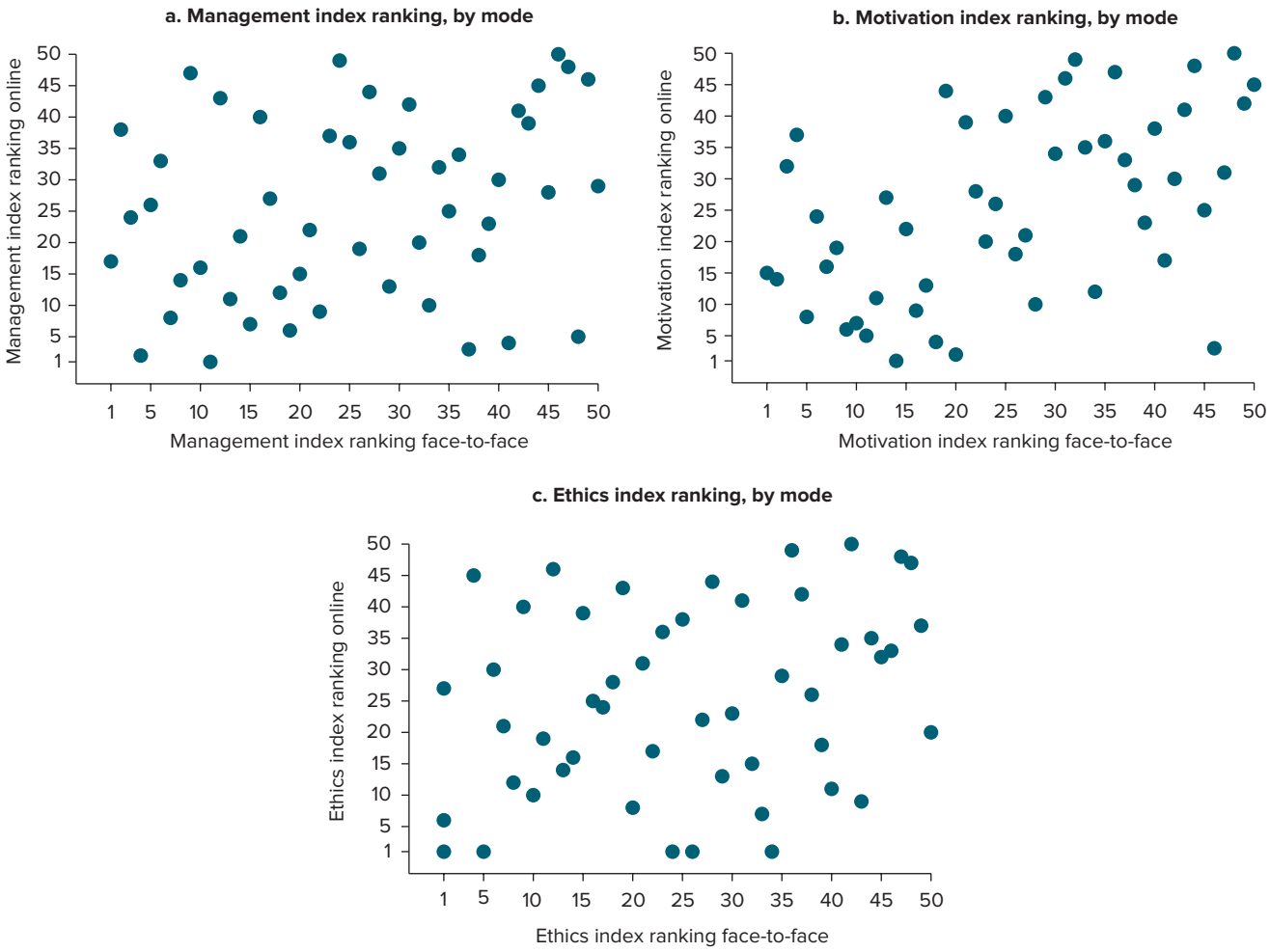
Building on the discussion in chapter 20, these heterogeneous mode effects at the organizational level are of particular concern to policy makers if they intend to present survey results as organizational rankings. Specifically, we find that the rank of a public sector organization (that is, its place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

FIGURE 19.4 Average Modal Difference, by Organization



Source: Original figure for this publication.

FIGURE 19.5 Organization Rankings, by Index and Mode



Source: Original figure for this publication.

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

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

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

Individual-Level Quantities

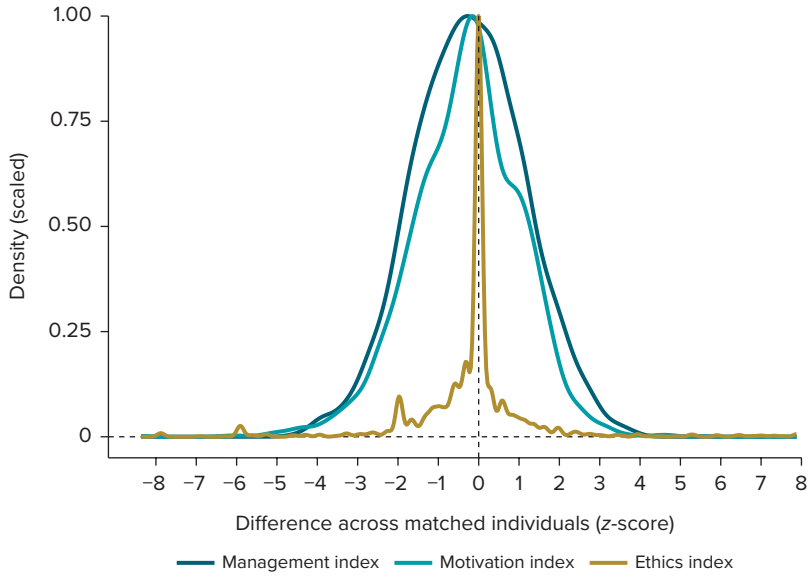
Showing the summary statistics at the individual level, as in panel 3 of table 19.2, requires matching the respondents on their observable characteristics. We use PSM to address the concern of selection bias in who chooses to respond to online surveys (Tourangeau, Conrad, and Couper 2013). PSM employs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

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

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

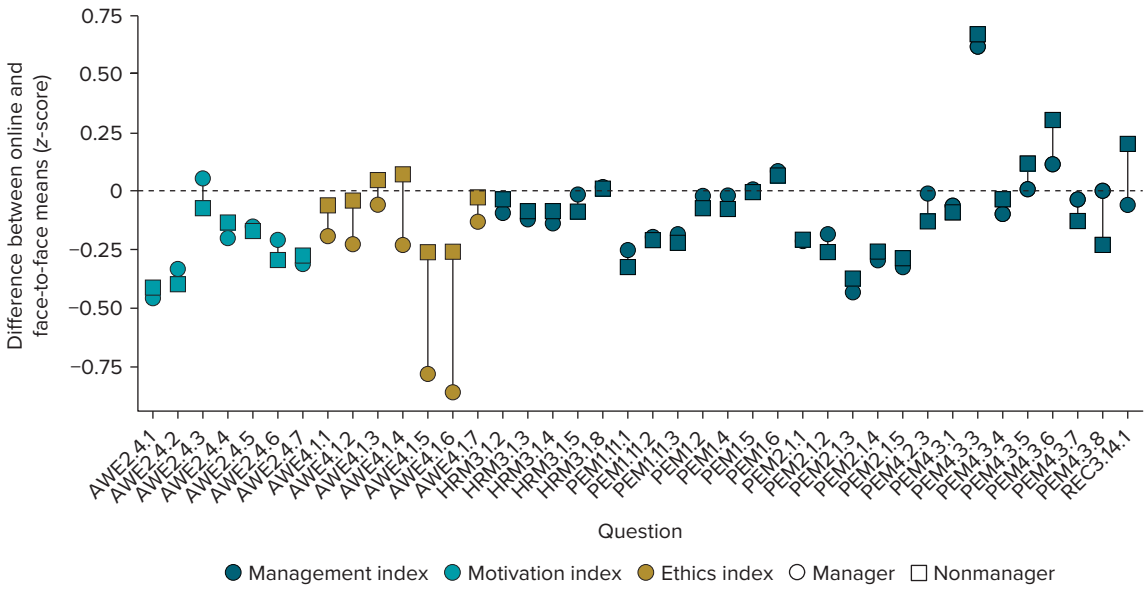
Corresponding to the finding that particular organizations are more sensitive to survey mode effects, it would seem that the distribution of sensitivity across groups of individuals is also important in

FIGURE 19.6 Distribution of Survey Mode Differences across Matched Individuals



Source: Original figure for this publication.

FIGURE 19.7 Average Item Difference, by Managerial Status



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

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

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

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

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

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

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

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

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

THE IMPACT OF COMMON CORRECTIONS

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

How Does the Response Rate Mediate Survey Mode Effects?

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

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

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

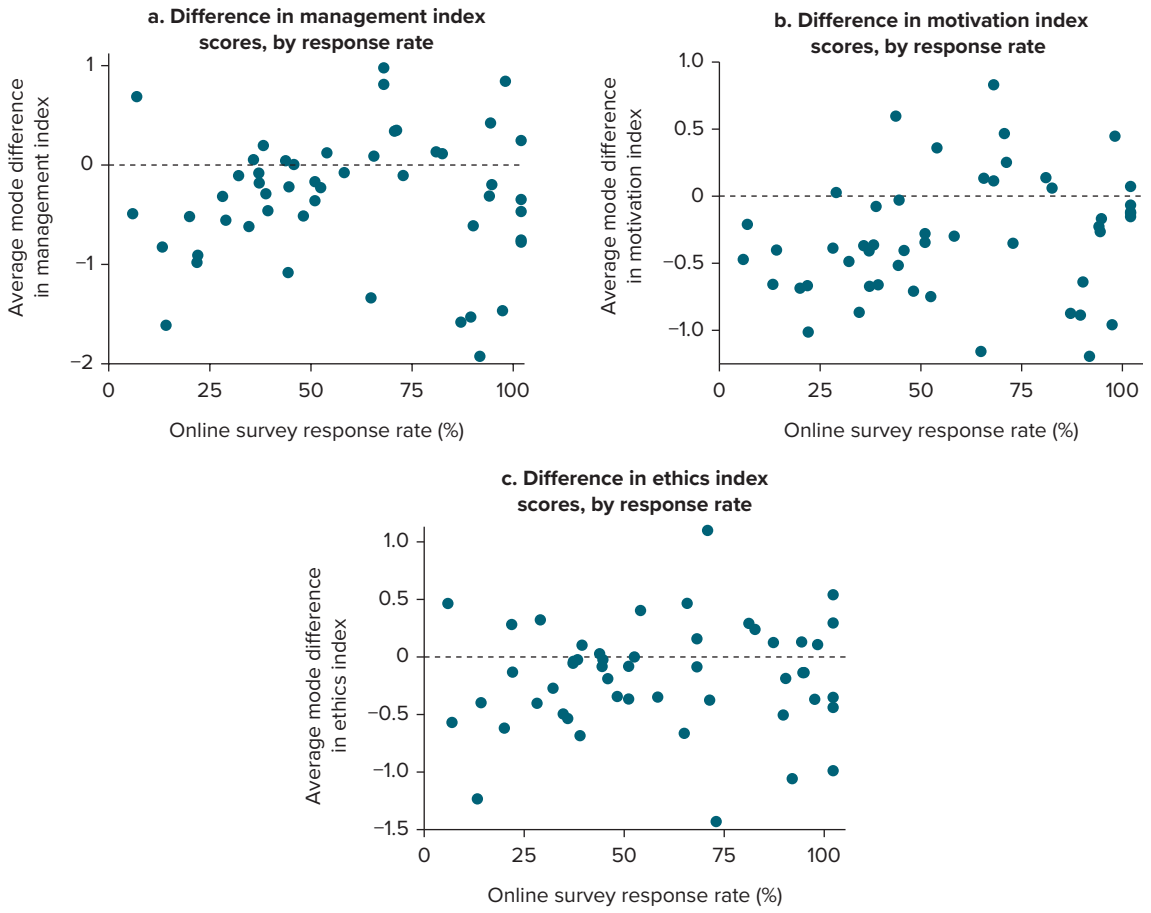
What Is the Impact of Corrections Using Survey Weights?

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

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

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

FIGURE 19.8 Difference in Scores, by Response Rate



Source: Original figure for this publication.

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

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

DISCUSSION

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

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

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

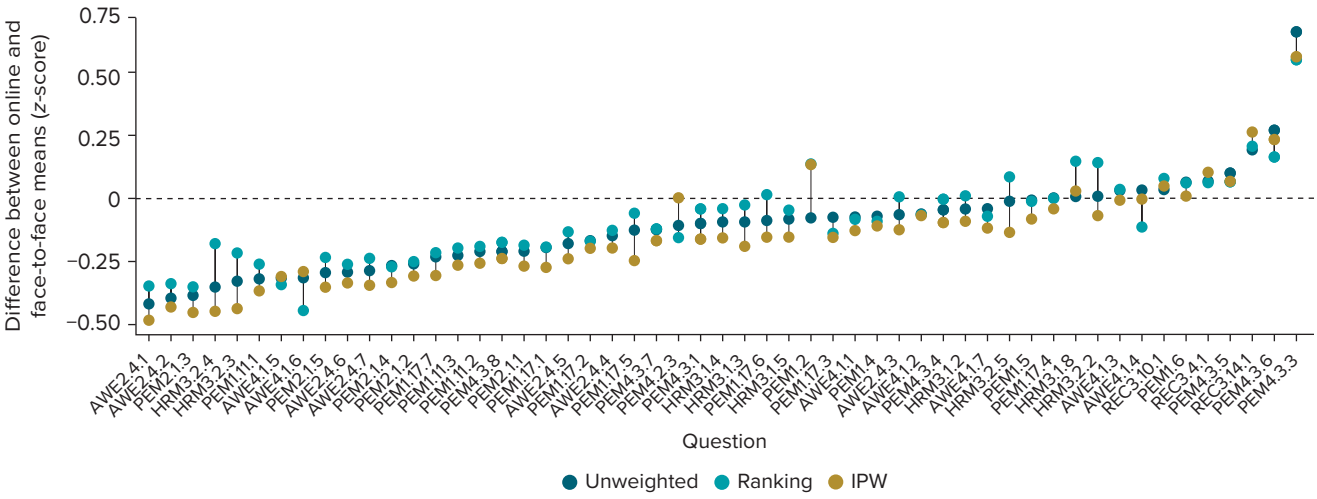
TABLE 19.4 Mean Modal Differences at the National Level, by Weighting Approach

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

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

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

FIGURE 19.9 Average Item Difference, by Sample



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

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

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

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

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

NOTES

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

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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