CHAPTER 22

Designing Survey Questionnaires

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

Robert Lipinski, Daniel Rogger, and Christian Schuster

SUMMARY

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

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

ANALYTICS IN PRACTICE

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

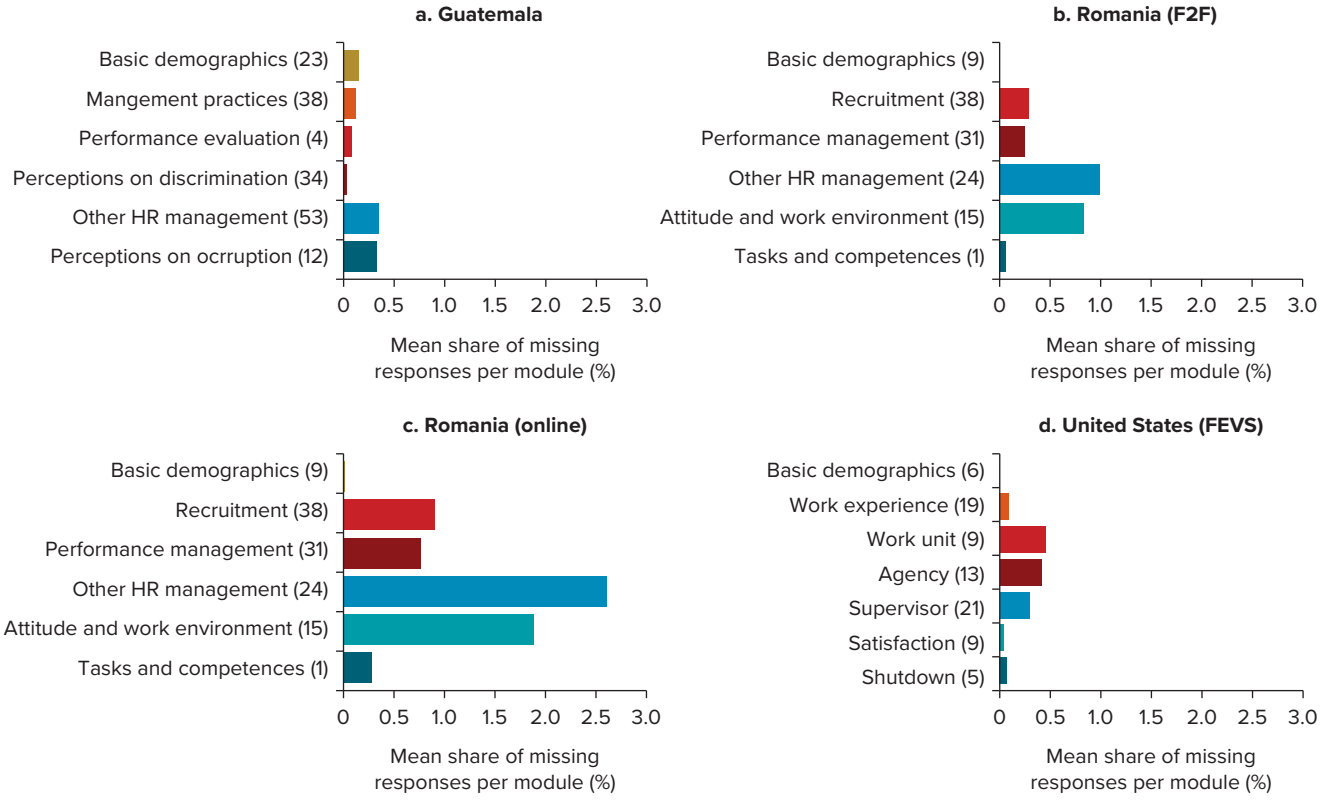
INTRODUCTION

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

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

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

FIGURE 22.1 Share of Missing Responses, by Survey Module



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

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

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

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

several factors, making it more difficult to render a judgment, or because a respondent's judgment does not correspond to the available answer categories. This burden might also be larger for certain groups of respondents—for example, the elderly.

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

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

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

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

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

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

UNDERSTANDING ITEM NONRESPONSE: LESSONS FROM THE SURVEY METHODOLOGY LITERATURE

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

Complexity

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

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

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

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

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

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

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

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

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

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

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

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

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

Much work in psychology has been conducted to determine how people formulate judgments from available information. According to Anderson's (1971) information integration theory, when people formulate a judgment, they gather all available pieces of information, assigning value and weight to each of them, before summing them up to form a final judgment. Another view, developed mainly in the work of Tversky and Kahneman, is that people tend to use a range of heuristic methods to arrive at judgments, like using only readily available instances and examples, using resemblance to a prototype, or anchoring based on

initial information (Tourangeau 1984; Tversky and Kahneman 1974). A combination of these views has been adopted by Zaller (1992) in his "response axiom," which argues that individuals answer survey questions by averaging different considerations, but only those that are immediately salient or accessible to them.

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

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

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

Sensitivity

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

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

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

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

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

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

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

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

Informal and formal sources of sensitivity might also interact with each other. Research by Galletly and Pinkerton (2006) suggests that there is an interaction between social stigma and formal sanctions in the case of HIV disclosure laws in US states. The introduction of legal sanctions for some actions may add to the already existing social stigma around them. Alternatively, the threat of social disapproval may be a more undesirable consequence than a formal sanction that is small or unlikely to follow. Likewise, if social and legal norms are not perfectly matched, admitting to an illegal but socially acceptable practice might be

less difficult for respondents. For example, if the law prohibits hiring one's family members and friends but society generally accepts this practice, then admitting to some degree of nepotism might come more easily to a survey respondent than if this practice were socially unacceptable.

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

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

METHODOLOGY

Case Selection

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

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

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

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

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

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

Coding Framework

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

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

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

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

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

TABLE 22.1 Complexity Coding Framework

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

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

TABLE 22.2 Sensitivity Coding Framework

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

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

Analysis

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

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

BOX 22.1 Applying the Coding Framework: Illustrative Examples from Romania

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

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

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

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

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

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

RESULTS

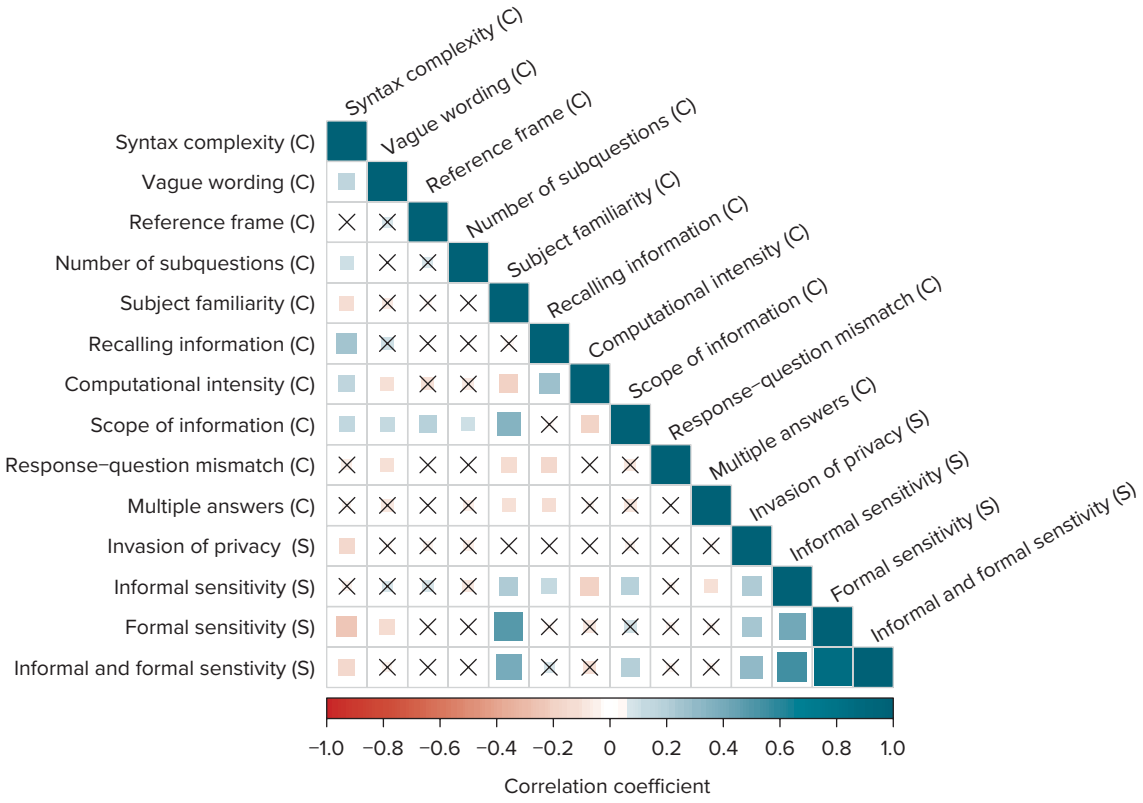
Descriptively, our independent variables vary across the components of complexity and sensitivity we code. As detailed in tables 22.1 and 22.2, among the components of complexity, complexity of syntax and unfamiliarity are the variables with the highest variance. On the other end of the scale, components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

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

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

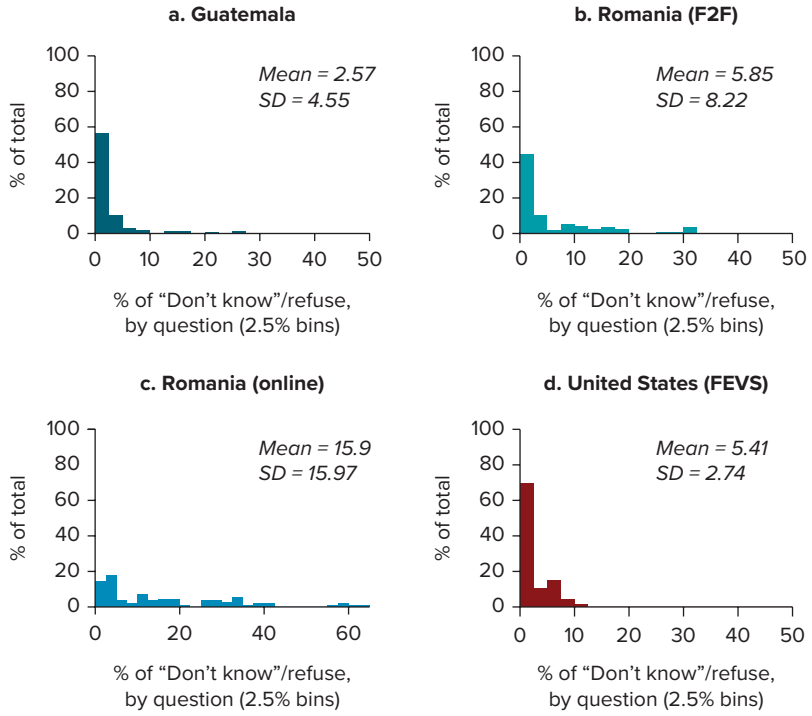
Next, as presented below, we can observe that item nonresponse is a challenge across the three surveys, though to a varying extent. As illustrated in figure 22.3, in the FEVS online survey, questions have an average item nonresponse of 2.4 percent. This number increases to 2.6 percent in the face-to-face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

FIGURE 22.2 Correlation between Subdimensions of Complexity and Sensitivity



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

FIGURE 22.3 Share of Missing Responses



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

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

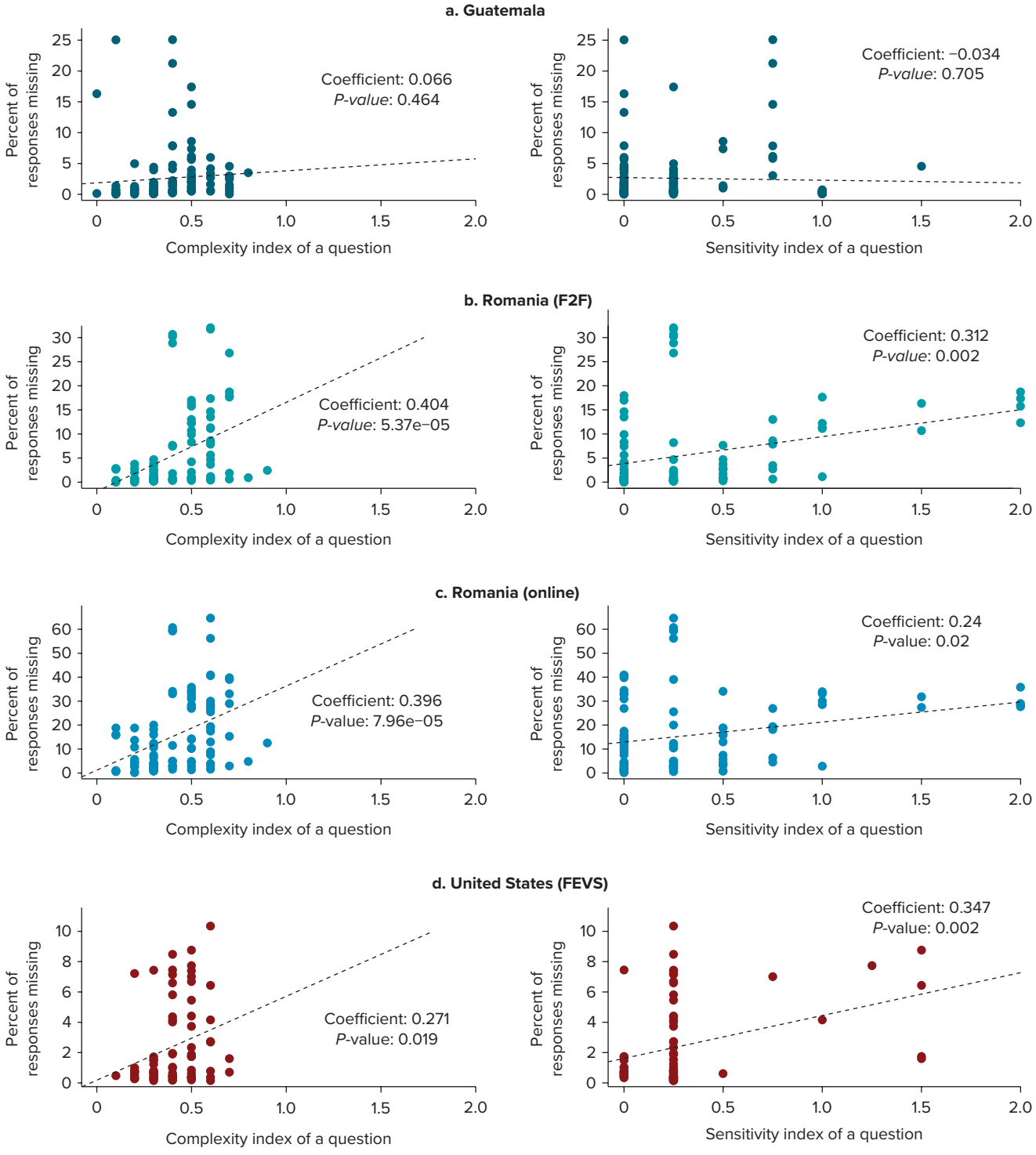
Looking first at correlations, we find that there is a positive association between complexity and item nonresponse. Figure 22.4 presents the correlations separately for each survey. Correlation coefficients range from 0.066 to 0.404 and are significant at the 5 percent level, except for Guatemala. We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

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

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

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

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



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

TABLE 22.3 The Impacts of Complexity and Sensitivity on Item Nonresponse

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

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

TABLE 22.4 Exploratory Factor Analysis

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

(continues on next page)

TABLE 22.4 Exploratory Factor Analysis *(continued)*

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

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

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

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

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

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

TABLE 22.5 Factor Analysis Regression

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

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

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

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

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

The Performance of Manual versus Machine-Coded Complexity

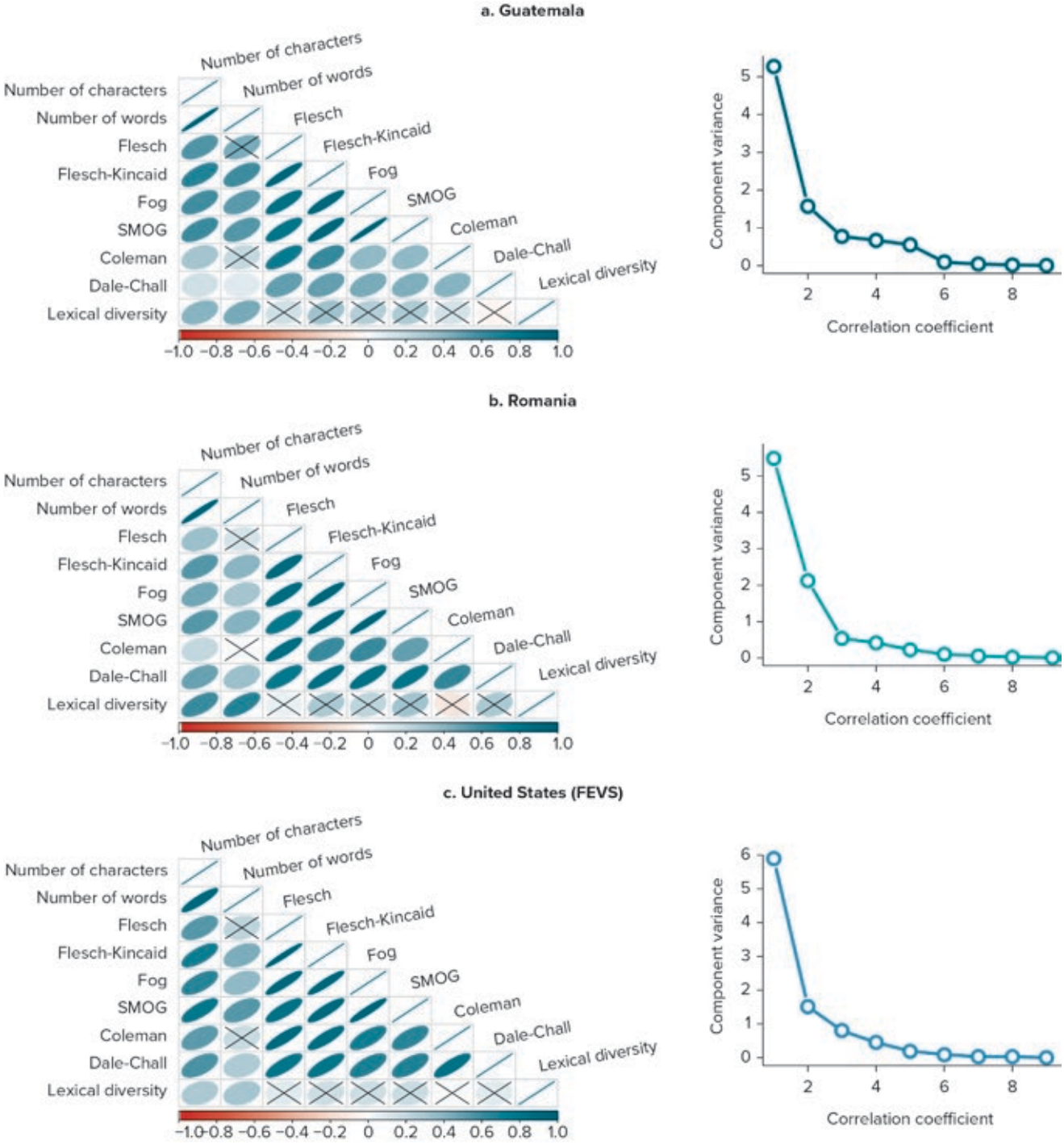
A potential criticism of our approach is that automated measures of complexity may be as effective a means of identifying potential problem questions—or more—while requiring far less investment. We therefore turn to assessing the relative effectiveness of the approach with respect to machine-coded complexity.

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

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

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

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



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

TABLE 22.7 Impact of Machine-Coded Complexity on Nonresponse Rate

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

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

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

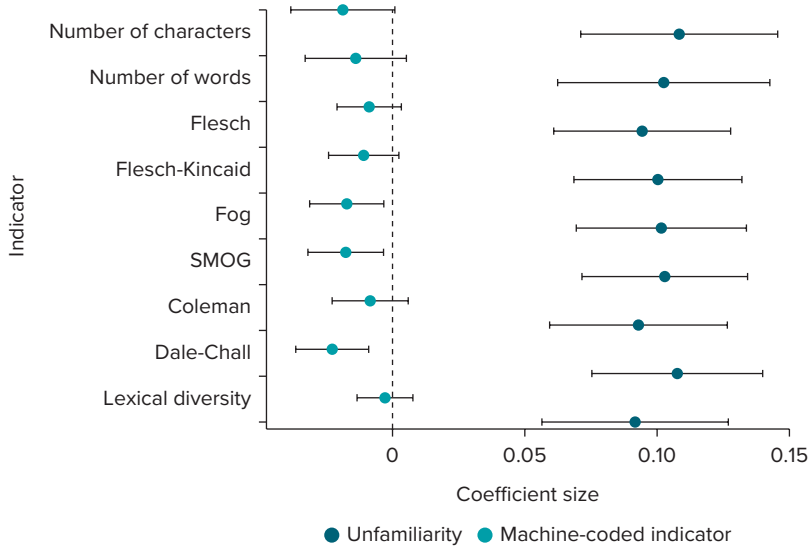
	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
Sensitivity	–0.002 (0.007)	0.002 (0.006)	–0.008 (0.010)	0.003 (0.003)
Complexity	–0.003 (0.005)	–0.011 (0.008)	–0.026 (0.015)	–0.002 (0.003)
Unfamiliarity	0.010 (0.006)	0.109*** (0.016)	0.217*** (0.026)	0.028*** (0.008)
Machine-coded complexity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

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

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

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

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



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

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

CONCLUSION

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

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

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

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

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

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

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

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

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

NOTES

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

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Holbrook, A., Y. Cho, and T. Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." *Public Opinion Quarterly* 70 (4): 565–95.
Just, M., and P. Carpenter. 1992. "A Capacity Theory of Comprehension: Individual Differences in Working Memory." *Psychological Review* 99 (1): 122–49.
Kim, S., and S. Kim. 2016. "Social Desirability Bias in Measuring Public Service Motivation." *International Public Management Journal* 19 (3): 293–319.
Knäuper, B., R. Belli, D. Hill, and R. Herzog. 1997. "Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality." *Journal of Official Statistics* 13 (2): 181–99.
Krosnick, J. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5 (3): 213–36.
Krosnick, J., and S. Presser. 2009. "Question and Questionnaire Design." In *Handbook of Survey Research*, 2nd edition, edited by J. Wright and P. Marsden. San Diego, CA: Elsevier.
Krumpal, I. 2013. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality and Quantity* 47: 2025–47.
Lenski, G., and J. Leggett. 1960. "Caste, Class, and Deference in the Research Interview." *American Journal of Sociology* 65 (5): 463–67.
Lensvelt-Mulders, G. 2008. "Surveying Sensitive Topics." In *International Handbook of Survey Methodology*, edited by E. de Leeuw, J. Hox, and D. Dillman, 461–578. New York: Routledge.
Lessler, J., R. Tourangeau, and W. Salter. 1989. *Questionnaire Design in the Cognitive Research Laboratory*. Vital and Health Statistics 6, Cognitive and Survey Measurement 1. DHHS Publication No. (PHS) 89-1076. Hyattsville, MD: US Department of Health and Human Services.
McNeeley, S. 2012. "Sensitive Issues in Surveys: Reducing Refusals While Increasing Reliability and Quality of Responses to Sensitive Survey Items." In *Handbook of Survey Methodology for the Social Sciences*, edited by L. Gideon, 4377–96. New York: Springer.
OPM (Office of Personnel Management). 2019. *2019 Office of Personnel Management Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: OPM.
Paulhus, D. 1984. "Two-Component Models of Socially Desirable Responding." *Journal of Personality and Social Psychology* 46 (3): 598–609.
Paulhus, D. 2002. "Socially Desirable Responding: The Evolution of a Construct." In *The Role of Constructs in Psychological and Educational Measurement*, edited by H. Braun, D. Jackson, and D. Wiley, 49–69. Mahwah, NJ: Erlbaum.
Rässler, S., and R. T. Riphahn. 2006. "Survey Item Nonresponse and Its Treatment." *Allgemeines Statistisches Arch* 90: 217–32.
Sakshaug, J., T. Yan, and R. Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multimode Survey of Sensitive and Non-sensitive Items." *Public Opinion Quarterly* 74 (5): 907–33.
Tourangeau, R. 1984. "Cognitive Sciences and Survey Methods." In *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, edited by T. Jabine, M. Straf, J. Tanur, and R. Tourangeau, 73–100. Washington, DC: National Academy Press.
Tourangeau, R., R. Groves, and C. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74 (3): 413–32.
Tourangeau, R., and K. A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103 (3): 299–314.
Tourangeau, R., L. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.
Tourangeau, R., and T. Smith. 1996. "Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context." *Public Opinion Quarterly* 60 (2): 275–304.
Tourangeau, R., and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859–83.
Tversky, A., and D. Kahneman. 1974. "Judgment under Uncertainly: Heuristics and Biases." *Science* 185 (4157): 1124–31.
World Bank. 2020a. *Final Field Report: Encuesta General de Servidores Públicos y Contratistas del Organismo Ejecutivo y Entitdaded Descentralizadas 2019*. Washington, DC: World Bank.
World Bank. 2020b. *Selecting the Right Staff and Keeping Them Motivated for a High-Performing Public Administration in Romania: Key Findings from a Public Administration Employee Survey*. Washington, DC: World Bank.
Yan, T., and R. Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times." *Applied Cognitive Psychology* 22 (1): 51–68.
Zaller, J. 1992. *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.