Designing Survey Questionnaires (II): What Types of Survey Questions Do Public Officials Not Respond To?

Robert Lipinski, Daniel Rogger and Christian Schuster*

May 6, 2022

Abstract

Surveys of public servants sharply differ in the extent of item non-response: respondents skipping or refusing to respond to questions. Item non-response can affect the legitimacy and quality of public servant survey data. Survey results may be biased, for instance, where 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 non-response we theorize that public servants are less likely to respond to questions that are complex (as they are unable to) or sensitive (as they are unwilling to). We assess this argument using a newly developed coding framework of survey question complexity and sensitivity, which we apply to public service surveys in Guatemala, Romania and the U.S. We find one indicator of complexity – the unfamiliarity of respondents with the subject question – to be the most robust predictor of item non-response across countries. By contrast, other indicators in our framework or machine-coded algorithms of textual complexity do not predict item non-response. Our findings point to the importance of avoiding questions which require public servants to speculate about topics they are less familiar with.

*Lipinski (rlipinski@worldbank.org) and Rogger (drogger@worldbank.org): Development Impact Evaluation Research Group, The World Bank; Schuster (c.schuster@ucl.ac.uk): University College London. We are grateful to XXX for helpful comments. We 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. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Lessons for Practice

- Surveys of public servants typically rely on voluntary responses from public servants. As such, **they may suffer from not only unit non-response – i.e.public servants not responding to surveys at all – but also item non-response - i.e. public servants not responding to particular survey questions.**
- Assessments of three surveys of public servants spanning three continents imply that **item non-response is a significant concern in the public sector**. In some survey modules non-response can be as high as 30%.
- Public servants are typically more educated than the average survey respondent and their daily duties are closely aligned to the task of filling in a questionnaire. As such, **the determinants of non-response in surveys of public servants may be distinct to 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.
- **We find one indicator of complexity – the unfamiliarity of respondents with the subject question – to be the most robust predictor of item non-response across countries.** Surveys of public servants should carefully consider the need for questions which require public servants to speculate about topics they are less familiar with, as they are associated with greater item non-responses.
- In contrast, no other margin of complexity or sensitivity is a particularly acute source of non-response. At least in terms of missing data, **our analysis implies that public officials can handle many aspects of complex and sensitive topics.**
- We compare our manual coding approach to common machine-coded assessments of complexity and find that a manually coded assessment of unfamiliarity outperforms machine-coded variables.

1 Introduction

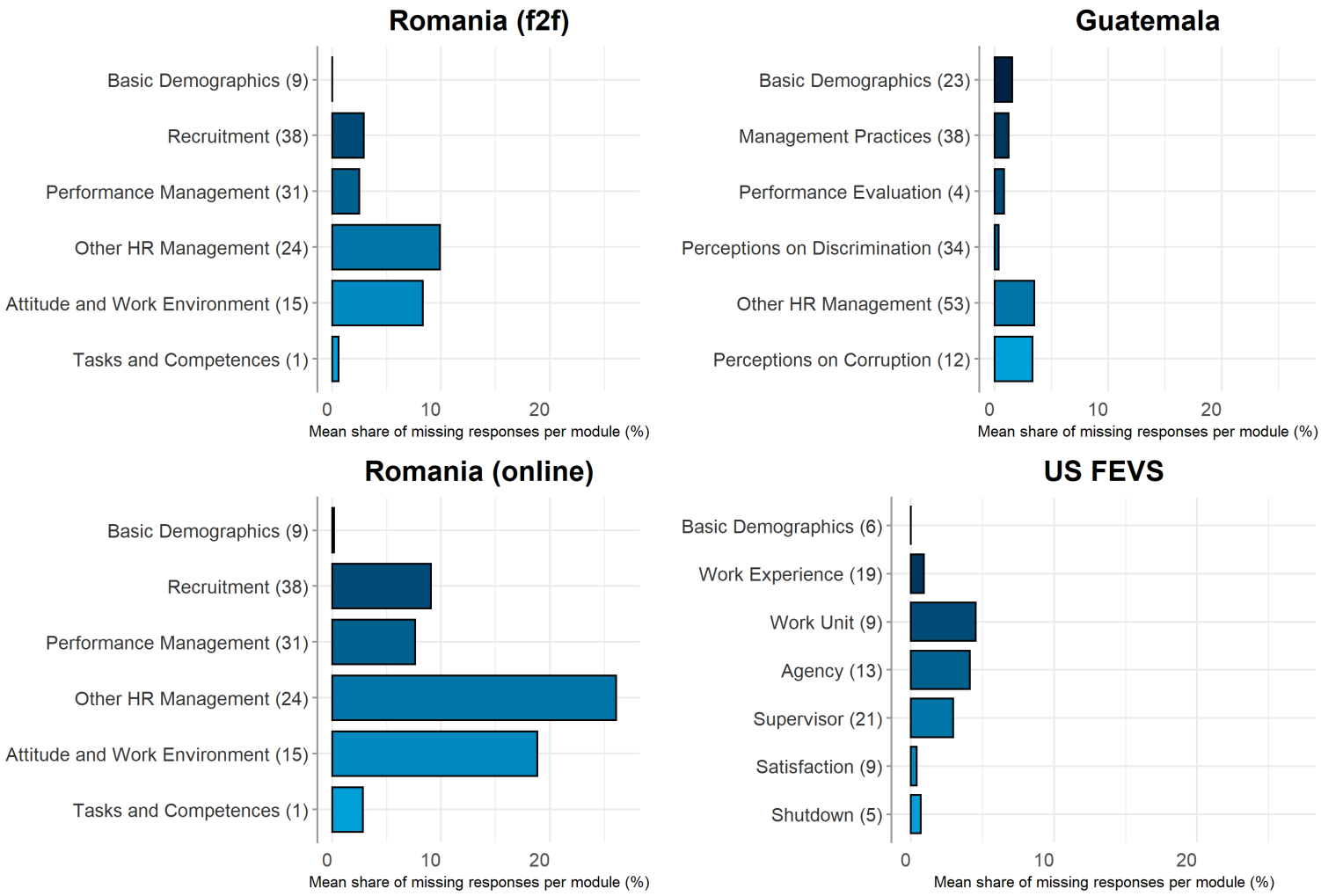
Surveys of public servants typically rely on voluntary responses from public servants. As such, they may suffer from not only unit non-response – i.e. public servants not responding to surveys at all (cf. chapter XXXX) or dropping out of the survey during survey completion (cf. chapter XXXX) – but also item non-response: 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 'Don't know' as their response to questions.

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

Understanding what types of questions public servants tend to respond to and what type of questions enhance item non-response is thus important for survey designers. It provides a basis for designing questions which reduce item non-response and thus enhance public service survey quality and legitimacy. This is important not least as surveying public servants about certain topics - such as satisfaction, motivation or assessment 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 thus the only alternative for valid data collection.

To-date, public service surveys vary in the extent to which their questions yield non-response. As illustrated in figure 2 – which draws on data from public service surveys in Guatemala, Romania and the U.S., which we will use throughout this chapter - item non-response varies across survey modules from almost 0% to up to almost 30% in some settings (and up to 60% for certain individual questions). The figure implies that for certain topics non-response is a substantive concern in public service surveys. The variation we observe across questions also implies that question characteristics determine the likelihood that they will be answered.

Figure 1: Share of Missing Responses by survey module



Numbers in parentheses in y-axis labels indicate the number of questions in each module

Why do public servants respond to some survey questions but not others? This chapter offers a conceptual framework and empirical evidence to better understand this question. Conceptually, we build on the survey methodology literature that has broadly argued for causes of item non-response: question complexity and question sensitivity. Question complexity leads to item non-response as respondents are unable to answer a question, even if willing. This is due to an excessive cognitive burden on one or more steps in the mental process of answering: i) comprehension of the question; ii) information retrieval from memory; iii) information integration; iv) translation to 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 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 judgement does not correspond to available answer categories. All these might also be a larger burden for certain groups of respondents, for example the elderly.

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

While the survey methodology 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 various information, 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

[1] According to Worldwide Bureaucracy Indicators published by the World Bank the share of publicly paid employees with tertiary education across the world is 54.2%, whereas in the private sector it is around half of that - 26.9% (average over 2010-2018 period).

surveys – for instance about employee perceptions 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 assessing empirically item non-response in public employee surveys, rather than simply extrapolating findings about item non-response from household surveys.

This chapter does so by analyzing missing response patterns in three public administration surveys - the U.S. Federal Employee Viewpoint Survey (hereafter U.S. FEVS) and two World Bank surveys of public officials in Romania and Guatemala [2] Our analysis is based on, first, the creation of a coding framework to assess the different elements of complexity and sensitivity of questions; second, the application of this framework to code each of the questions in the aforementioned surveys in terms of complexity and sensitivity; and, third, regressions to assess which of the elements of complexity and sensitivity predict item non-response.

We find that, contrary to literature findings in other contexts, public officials do not appear to shy away from answering questions that are longer, characterized by more complex syntax or requiring more cognitive effort to answer. We also find only limited evidence that question sensitivity is associated with greater item non-response. By contrast, we find robust evidence that one sub-indicator of complexity – the 'unfamiliarity' of topics in questions – is associated with item non-response 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 the 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 functioning of their organization into one response option.

Given the manual nature of our approach to coding the complexity and sensitivity of survey questions, a 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 our core results to the predictive ability of machine-coded measures. We find that our unfamiliarity index continues to be the most effective approach to identifying questions that suffer from non-response in public officials surveys.

The chapter is organized as follows. Section 2 presents an overview of

[2] More information on the surveys used and the reason for their selection is presented in the 'Methodology' section.

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

2 Understanding item non-response: lessons from the survey methodology literature

In essence, the survey methodology literature posits two broad underlying causes of item non-response. Respondents are 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 those two central causes of potential item non-response. To build our coding framework, we discuss the literature on, first, complexity and, then, sensitivity.

2.1 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?' is asking for the same information, but could leave respondents confused about what the question is actually asking for. Although needlessly complicated in this example, some survey questions are longer, more convoluted or otherwise hindering a respondent willing to respond from providing an answer. This is what the literature typically refers to as question complexity (Knäuper et al. 1997, Yand and Tourangeau 2008).

Complexity is a multi-dimensional concept. While its definition is contested, it can perhaps best be conceptualized as a set of hurdles that a respondent can encounter on the mental pathway she has to go through from the moment of being presented with a question to providing an answer (Tourangeau and Rasinski 1988). Or as Knäuper et al. (1997:181) phrased it: "question answering involves a series of cognitive tasks that respondents have to resolve to provide high-quality data". Those tasks might be objectively more or less difficult, but the effort they require might also depend on respondent's characteristics. To go back to the example used at the be-

ginning of this section - the more complex version of the age question would likely pose relatively less trouble to a native-English astrophysicist than to someone for whom English is a second language and who has never learnt about physics.

The literature on cognitive psychology most commonly refers to four steps a respondent goes through in the answering process, as outlined by Tourangeau (1984), Tourangeau and Rasinski (1988) and Tourangeau, Rips and Rasinski (2000). Those are: i) question comprehension; ii) retrieval of necessary information from memory; iii) integrating retrieved information into a judgment or estimate; iv) translating thus formed judgment into an appropriate response. Depending on the type and format of a question those steps might vary in length and cognitive difficulty. For example for a question about age, the information is easily retrieved from the memory, but might require some mapping process if the response is not to be numerical but rather match pre-defined age bands.

In the first step, a respondent has 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 a respondent (Knäuper et al. 1997) - an effort she might or might not be capable or willing to perform.

It is less obvious whether questions that are longer have 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 required cognitive effort required on the part of the respondent. On the other hand, a long question might be simply convoluted, it might touch on too many topics, or it can be difficult to remember in full when providing the answer, thus increasing difficulties for the respondent (Knäuper et al. 1997; Holbrook, Cho and Johnson 2006). Other features of a question, like the number of propositions and logical operators (e.g. '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). The cognitive difficulties in comprehension might also depend on individual working memory capacity, as research by Just and Carpenter (1992) shows that working memory is a key element of both information storage and computations necessary for language comprehension.

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

Apart from the temporal reference frame, attitudes that refer to direct, more recent, as well as 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 therefore surveys should 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 difficulty of integration. Consider this example of three different questions 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 get your first public sector job?**

1 - Very important; 2 - Somewhat important; 3 - Neither important nor unimportant; 4 - Somewhat Important; 5 - Very Important; 6 - Don't Know

3) **Please rank the following criteria in order of importance they**

had for having obtained 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 only requires a respondent to make a binary choice on importance of personal connections. The second version requires a more fine-grained evaluation - not only whether the personal connections were important, but also how strongly so. In the third version, the respondent not only has to judge the importance of personal connections but also four others considerations and evaluate them against each other. Clearly, this last approach requires greatest cognitive effort from the respondent.

Much work in psychology has been conducted to determine how people formulate their judgements from available information. According to Anderson's Information Integration Theory (Anderson 1971), when people formulate a judgement they gather all available pieces on information, assigning value and weight to each of them, before summing them up to form a final judgement. Another view, developed mainly in the works of Tversky and Kahneman, is that people tend to use a range of heuristic methods to arrive at judgements, like using only the readily available instances and examples, resemblance to a prototype, or anchoring based on initial information (Tversky and Kahneman 1974; Tourangeau 1984). A combination of these views have 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 is mapping of the answer into response options available (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 respondent has to simultaneously remember the question, the answer options and map her formed judgement 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 on frequency or Likert-like scale, or if they contain vague words and complex phrases themselves. Second, response formats that are hard to understand or with an ambiguous set of possible responses might compound mapping difficulties. Whereas the role of multitasking as an obstacle depends mainly on the respondent, problems with the response format are usually due to

a faulty questionnaire design. To ease the process of translating 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, complexity in survey questions 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 interviewer for clarifications (Holbrook, Cho and Johnson 2006). Some effects of question complexity, however, are more consequential. In particular, complexity can invite 'acquiescence bias' or *satisficing*, where respondents tend to agree with a complex statement, regardless of their true position, in order to avoid cognitive overload (Lenski and Leggett 1960; Knäuper et al. 1997; Krosnick, 1991). 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 'Don't Know' option. This last option is a particular example of strong satisficing because it requires no cognitive effort whatsoever. [3]

In short, the survey methodology literature suggests that complex survey questions enhance the cognitive effort required along the mental process of answering a question, and may thus lead to satisficing, including item non-response. The empirical literature that complements the theoretical considerations outlined here found supporting evidence that each of these answering stages can increase non-response. For example Knauper et al. (1997) found that respondents answer 'Don't Know' more often to questions that, among the others, contain ambiguous terms, require retrospective reports or quantity reports. The inclusion of those more complex question characteristics raised item non-response in their study by between 0.5 and 7.7 percentage points (and as expected more so for individuals with lower cognitive ability), which is substantial considering that in most sub-groups total share of 'Don't Know' responses stayed well below 10%.

[3] Apart from the degree complexity, which is a stable feature of a question, the likelihood of engaging in satisficing also depends on respondent's characteristics that might increase or decrease her cognitive capacity (e.g. age, education, tiredness) and on her willingness to answer the question. Contextual variables, like to pace in which interviewer conducts an interview or a time pressure (e.g. having only a 20-minute slot dedicated to take a survey), can also impact the degree to which respondent is willing and able to engage in high cognitive effort (Lessler, Tourangeau and Salter 1989; Fazio and Roskos-Ewoldse 2005).

2.2 Sensitivity

Irrespective of how complicated a question may be, the extent to which it requests personally sensitive information may also impact on non-response. 'How many bribes have you accepted in the last month?' arguably has a 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 – asking about behavior which is typically both morally wrong and illegal – a second source of concern for survey designers.

Unlike with complex questions, when people are asked a sensitive question, 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 try to conceal it and avoid discussing it in public (Lensvelt-Mulders 2008). Typically the concerns lies with topics such as 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 public service, or sexual harassment of employees.

The most commonly used classification of sources of sensitivity was developed by Roger Tourangeau, along with several co-authors (Tourangeau, Rips and Rasinski 2000; Tourangeau and Yan 2007). According to them, sensitivity derives from three primary sources - a question might touch on taboo subject, a truthful answer might violate social norms or it might 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 to be a 'taboo' subject (Tourangeau and Yan 2007). This might be a concern for various topics, from one's sexual orientation to salary level. McNeeley (2012) describes how talking about such topics might lead to distress and uneasiness for respondents (and in some cases also enumerators), and, in case of demographic items, also to a threat of identification.[5] Unlike for 'social desirability bias' and other sources of sen-

[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,

sitivity discussed below, those topics are not problematic because revealing requested information could lead to some type of sanctions. Instead, those topics are perceived as sensitive regardless of a respondent's true position (Tourangeau and Yan 2007; Krumpal 2013) because it is not common to discuss them in public or with strangers, like an enumerator. Therefore, they often lead to item non-response rather than misreporting (Höglinger, Jann, and Diekmann 2016), as respondents simply do not want to discuss the topics at all.

Secondly, but arguably most commonly, this wariness to truthfully answer sensitive questions is explained with reference to 'social desirability bias'. It refers to an inner desire to conform to established social norms in a given circle, be it workplace, family, or society at large. Admitting that one has committed an action violating a common norm, either by doing something considered 'wrong' (e.g. taking a bribe) or failing to do 'good' (e.g. helping a colleague in need), is undesirable (Tourangeau and Yan 2007). That is because if someone was to find out, the violator could be frowned upon, criticized, or shunned. The impact it has on responses might further depend on specific social norms one identifies with and how concerned she is about not violating them. For example, Kim and Kim (2016) found 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' threat to the self-image of a respondent (Lensvelt-Mulders 2008:462). Touching on some sensitive topics might re-float feelings of guilt, embarrassment or shame for having done (or failing to do) something or be stressful to discuss for the respondent in general. Therefore, to avoid negative consequences from others as well as her own conscience, a respondent might prefer not to answer a sensitive question or answer it in a socially 'expected' way. In face-to-face surveys, even those believing in full confidentiality of responses, might want to create positive image of themselves or earn social approval from the enumerator (Krumpal 2013) and thus succumb to social desirability bias.

Psychologists have long debated the precise causes of this bias. Paulhus (1984) suggests it has two parts. One is impression management, i.e. a desire to present oneself in positive light in front of the 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.

---

such as a survey conducted within one's workplace" (McNeeley 2012:380).

Thirdly, 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 such interaction between social stigma and formal sanctions in the case of HIV disclosure laws in U.S. states. This is because the introduction of legal sanctions on some actions might add to an already existing social stigma around them. Alternatively, the threat of social disapproval might be a more undesirable consequence than a formal sanction that is small or unlikely to follow. On the other hand, if social and legal norms are not perfectly matched, for example, to use the example mentioned above, 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 easier to a survey respondent, compared to a case when such practice would be socially unacceptable.

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

In sum, item non-response might increase as a result of increased question sensitivity – though, compared with complexity, this effect might be diluted by respondents who answer sensitive items in a socially desirable way rather than not answering at all (Sakshaug, Yan and Tourengeau 2010). Tourengeau and Yan (2007) for example report that item non-response in

NSFG[6] Cycle 6 Female Questionnaire tends to raise by less than 3 percentage points when comparing qustions with very low sensitivity (education - 0.04% non-response rate; age - 0.39%) compared to high sensitivity items (number of times having sex in the past 4 weeks - 1.37%, number of sexual partners - 3.05%). Only the income question had more noticeable non-response, at 8.15%. And whereas experimental methods aimed to reduce question sensitivity, such as unmatched count technique, do significantly affect the mean estimates obtained, they have far less effect on item non-response (Coutts and Jann 2011)[7], suggesting that biasing rather than avoiding the answer is a more prevalent response for people presented with sensitive question. Comparing the effects of unit (although not the item) non-response and measurement error in reports of voting behaviour 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% to 69.4%.

3 Methodology

3.1 Case selection

We evaluate question complexity and sensitivity and their relationship with item non-response in three 2019 central government-wide public administration surveys in Guatemala, Romania and the U.S.

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 4 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% (World Bank 2020a). All respondents were surveyed in person by trained enumerators.

The Romania survey 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

[6]National Survey of Family Growth

[7]Some other methods, like randomized response technique, actually lead to higher item non-response, but rather due to convoluted instructions they entail rather than the nature of the question itself

overlapping between the two versions are used in the analyses below. The Romania data was 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 (response rate of 24%) and 3,316 answered the face-to-face (response rate of 92%; for details see World Bank 2020b).

Responding the survey online may increase the sense of comfort and privacy of respondents, thus reducing the perceived threat posed by sensitive questions (McNeeley 2012). On the other hand, online surveys lack an enumerator that could 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 U.S. Federal Viewpoint Survey (FEVS) has been fielded by the Office of Personnel Management biannually since 2004 and annually since 2010 (cf. chapter X). 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, giving a response rate of 42.6%

Our case selection enables us to understand item non-response in public service surveys of countries from across diverse cultures, regions, levels of development and education. Findings about item non-response that travel across all three contexts give us plausible reason to believe that they are generalizable to other surveys of public administrators.

3.2 Coding framework

Understanding item non-response – and whether different dimensions of complexity and sensitivity shape item non-response in public service surveys – requires measuring complexity and sensitivity consistently across and within surveys. To do so, we developed a coding framework that allows us to assign a numerical value that reflects the degree of complexity and sensitivity of every survey question. Our 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.[8]

Our complexity and sensitivity indices comprise several sub-dimensions, as described in Tables 1.1 and 1.2. The complexity index is composed of 10 sub-dimensions, which are conceptually based on the 4-stage mental process of answering a question (cf. Tourangeau and Rasinski (1988); Tourangeau, Rips and Rasinski (2000)) and synthesize the measures proposed by, among the others, Knäuper et al. (1997), Belson (1981), as well as Holbrook, Cho and Johnson (2006). Our sub-dimensions include complexity of syntax, number of sub-questions, presence of a reference frame, and unfamiliarity of the subject.

The Sensitivity Index is constructed using 4 sub-dimensions suggested by the literature: invasion of privacy, socio-emotional threat of disclosure, threat of formal sanctions (Tourangeau, Rips and Rasinski 2000; and 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 we study along the dimensions outlined in Tables 1.1 and 1.2 and score it a value of 0, 1, or 2. The value of 0 was given to questions that do not present a particular sub-dimension of complexity or sensitivity at all. The value of 1 referred to cases in which questions could potentially be viewed as creating problems for respondents on a given sub-dimension, whereas 2 was used for cases where such problems were clearly substantive. The full coding framework is presented in an online appendix.

Three research assistants applied the coding framework to assess the complexity and sensitivity of each of the questions in the three surveys. Each research assistant first coded the values independently and then their scores were compared. In 86.8% 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. We calculate the values of both indices as arithmetic means of the scores across their respective sub-dimensions.

[8]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 non-response.

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 qualification, job-specific skills, knowing someone with political links, having personal connections, etc. 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 a respondent to calculate his/her length of service in the current institution by subtracting his/her starting year from the current year. This is not a complicated calculation, but still one that is likely not performed often and might require some mental effort if a respondent joined long time ago, hesitates if periods like initial internships should be included etc. 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' since the respondent might feel that there might be social consequences of disclosing that information or even formal ones if they did not inform relevant authorities about the bribe-taking behavior.

Table 1.1: Complexity Coding Framework

Sub-dimension	Description	U.S. FEVS	Romania	Guatemala	Aggregate
Comprehension					
Complex Syntax	This component assesses (1) the length of a question, which is measured by the number of characters (n); (2) the complexity of the syntax or the grammatical arrangements of words and phrases, which is determined by the sentence structure. Definitions: (1) simple syntax: simple sentence(s) with three part of speech; (2) moderately difficult syntax: (i) simple sentence(s) with more than 3 part of speech (3) complicated syntax: complex or complex-compound sentences.	1.2 (0.53)	1.09 (0.61)	0.85 (0.65)	1 (0.63)
Vagueness	This component assesses the extent to which the language used in a question is vague, unclear, imprecise, ambiguous (Edwards, Thomas et al. 1997) or open to interpretation. Common terms such as "good" are pre-determined in a list of vague words.	0.54 (0.55)	0.64 (0.5)	0.34 (0.48)	0.48 (0.52)
Reference Category	This component assesses the extent to which necessary frame(s) of reference are available in a question so that respondents understand the question in the way intended.	0.16 (0.48)	0.26 (0.51)	0.17 (0.47)	0.2 (0.48)
Number of Questions	This component measures the number of sub-questions embedded in the question block to which a question belongs. A sub-question must only ask for one issue so a compound sub-question is not counted as one sub-question.	0.21 (0.46)	0.11 (0.33)	0.3 (0.55)	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 closer knowledge of (for instance, their own experience versus their perceptions of experiences of other employees in the organization).	0.35 (0.51)	0.34 (0.53)	0.93 (0.79)	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).	1.04 (0.4)	1 (0.44)	0.91 (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.	0.02 (0.16)	0.07 (0.26)	0.07 (0.29)	0.06 (0.26)
Scope of Information	This component assesses the extent to which answers are derived from information beyond the personal experience of respondent.	0.32 (0.47)	0.46 (0.55)	0.38 (0.52)	0.39 (0.52)
Translation to Answer					
Categories Mismatch	This component assesses the extent to which the available answer options match the true answer to the question.	0.04 (0.25)	0.09 (0.41)	0.06 (0.28)	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 (0)	0.01 (0.09)	0.06 (0.28)	0.03 (0.2)

Notes: The final four columns show mean and standard deviation (in parentheses) of scores for each sub-dimensions and survey.

Table 1.2: Sensitivity Coding Framework

Sub-dimension	Description	U.S. FEVS	Romania	Guatemala	Aggregate
Privacy					
Invasion of Privacy	This sub-indicator measures the extent to which the respondent is 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.	0.04 (0.19)	0.31 (0.6)	0.08 (0.27)	0.14 (0.41)
Informal Sensitivity					
Social-Emotional Threat of Disclosure	This sub-indicator measures the degree to which the respondent may be concerned with the social and/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 sensitive if the respondent's truthful answer departs from socially desirable behaviors or social norms.	0.79 (0.56)	0.7 (0.65)	0.55 (0.51)	0.65 (0.58)
Formal Sensitivity					
Threat of Formal Sanctions	This sub-indicator measures the degree to which the respondent 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.	0.15 (0.5)	0.15 (0.46)	0.26 (0.6)	0.2 (0.54)
Interaction					
Relationship between informal and formal sensitivity	This sub-indicator measures the likelihood that a behavior/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 might be frowned upon in one's social circle, i.e. reporting colleagues taking bribes might be considered as 'snitching', but it might 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)	0.13 (0.49)	0.2 (0.48)	0.2 (0.41)	0.19 (0.45)

Notes: The final four columns show mean and standard deviation (in parentheses) of scores for each sub-dimensions and survey.

3.3 Analysis

To investigate non-response 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. Our regressions take the respondent-question as the unit of observation, meaning that each row corresponds to particular respondent's answer to a given question. We define item non-response as 1) a 'Don't Know' answer, 2) a refusal to answer, and 3) skipping of the question.

We control for individual-level characteristics that might affect non-response, 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 non-response for 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 different cost-benefit calculus when deciding on 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)[9].

We further control for the overall response rate in the government organization/agency to which a respondent belongs. Lower response rate might reflect unobservable characteristics of the organization or its employees that shape item non-response. We also control for the position of a question within a questionnaire (coded as an integer variables starting from 1). 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). In general, we find that men tend to have lower item non-response, whereas non-managers 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.

Our data thus takes a 'long' format, with each row corresponding to a particular respondent's answer to a question, accompanied by the respondent's individual characteristics, the variables pertaining to her organization, the question's complexity, sensitivity, position in the questionnaire and finally if the respondent answered a given question (1) or not (0).

[9]The exclusion of this variable does not change our substantive conclusions

We first look at simple correlations between our key variables of interest and then go on to regress item non-response variable on our indices of complexity and sensitivity as well as their various sub-dimensions in OLS regressions. In order to account for possible correlation of residual errors in our dataset we use multi-way clustering on individual and question-level, which allows us to correctly estimate standard errors and the corresponding significance levels.

4 Results

Descriptively, our independent variables vary across the components of complexity and sensitivity we code. As detailed in Table 1.1 and 1.2, 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 information integration seem the least variable. Among the sensitivity components, socio-emotional threat records both the highest mean score and greatest variation. Invasion of privacy scores the lowest in mean and standard deviation.

To ensure our coding framework meaningfully captures distinct sub-dimensions or components of complexity and sensitivity, we assess correlations between different components of sub-dimensions of complexity and sensitivity. Figure 3 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, as in between syntax complexity and vague wording or between 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 are most often occurring 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 our coding framework and its application.

Next, as presented below, we can observe that item non-response is a challenge across our three surveys, though to a varying extent. As illustrated in Figure 2, in the U.S. FEVS online survey, questions have an average item non-response of 2.4%. This number increases to 5% in the face-to-face public service survey in Guatemala, and 5.9% in the face-to-face public service survey in Romania. In the online version of this survey in Romania in turn,

average item non-response increases to 15.9%.[10]

[10]This is consistent with prior work which associates online surveys with higher non-response than face-to-face surveys (Heerwegh and Loosveldt 2008).

Figure 2: Share of Missing Responses

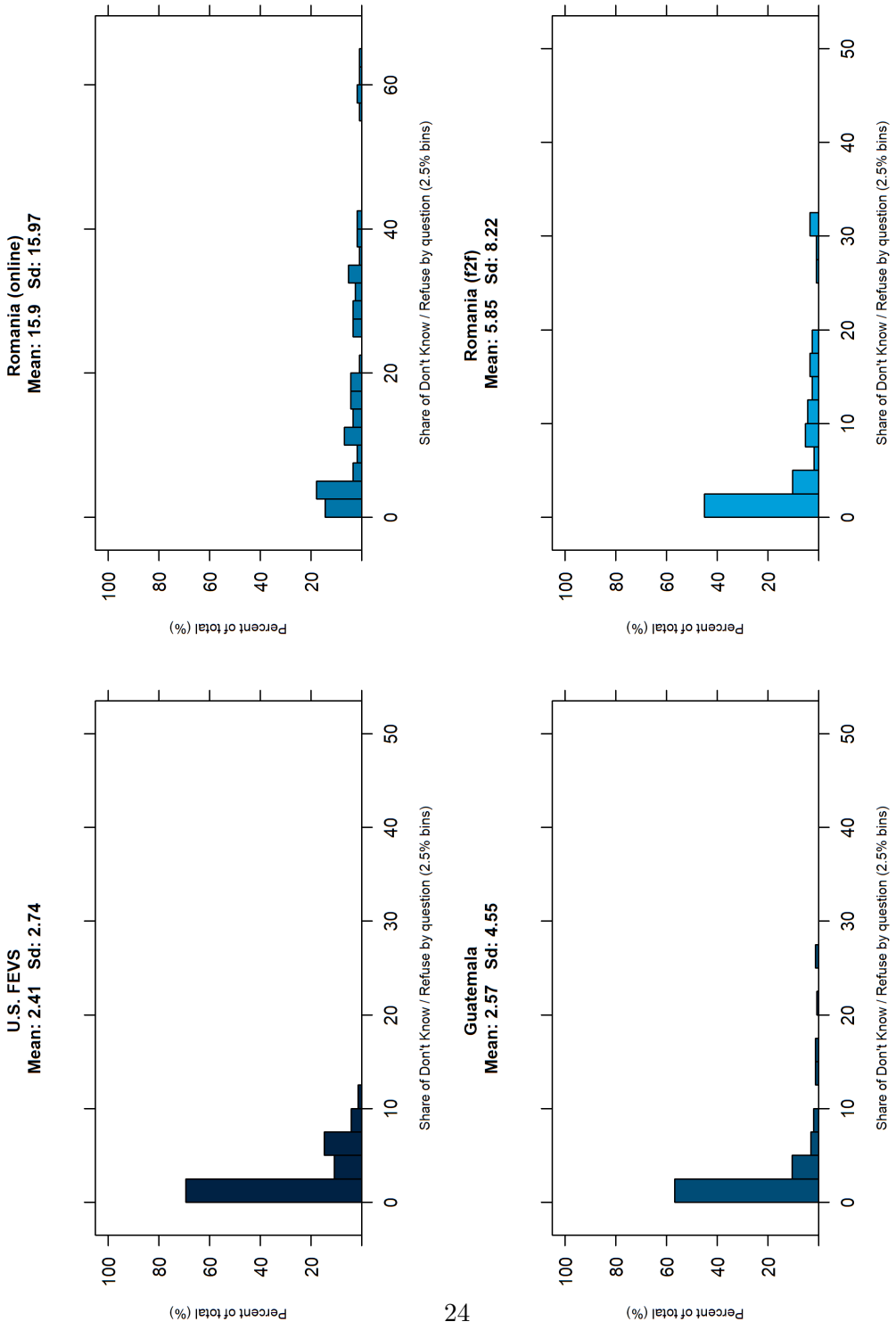
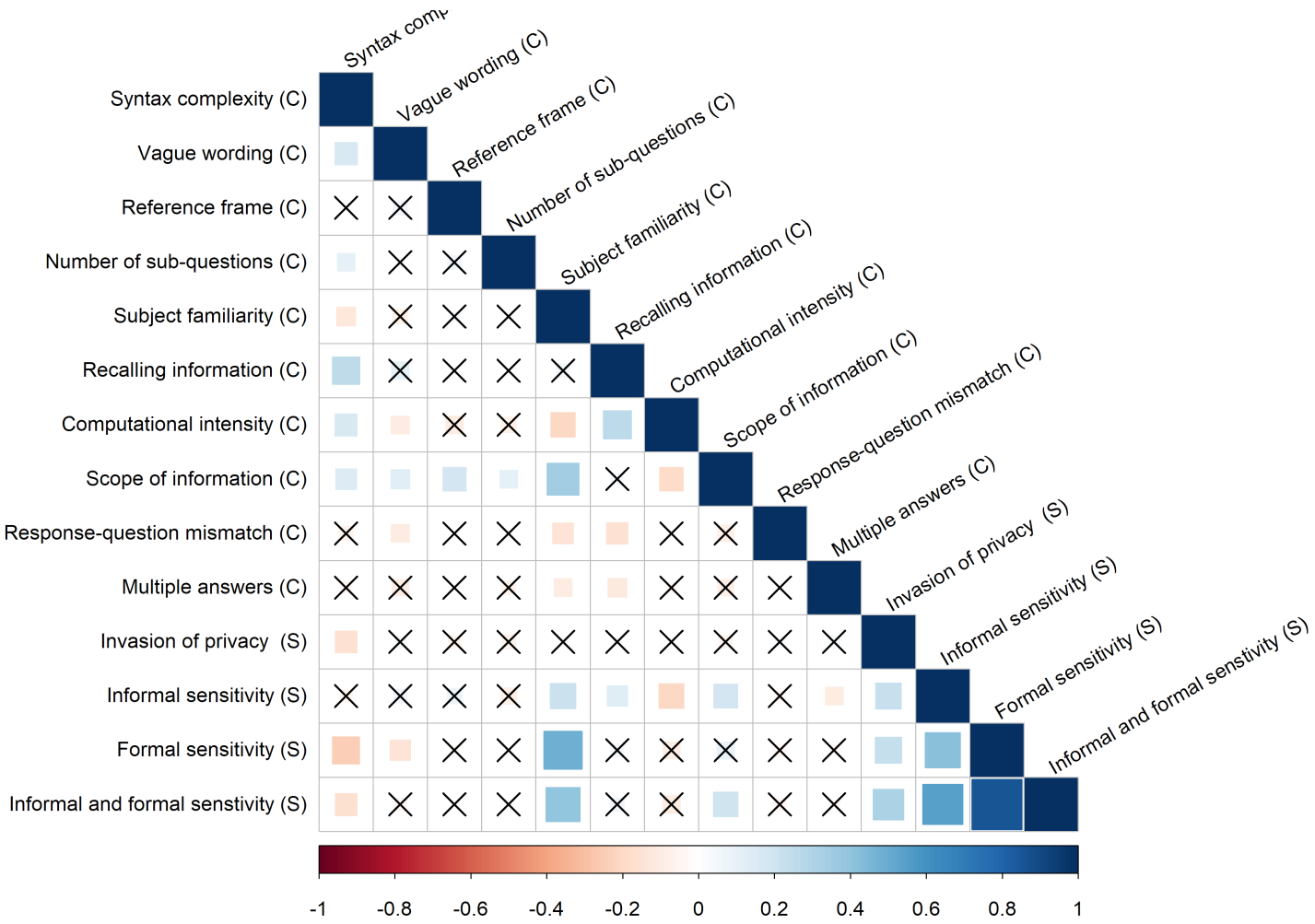
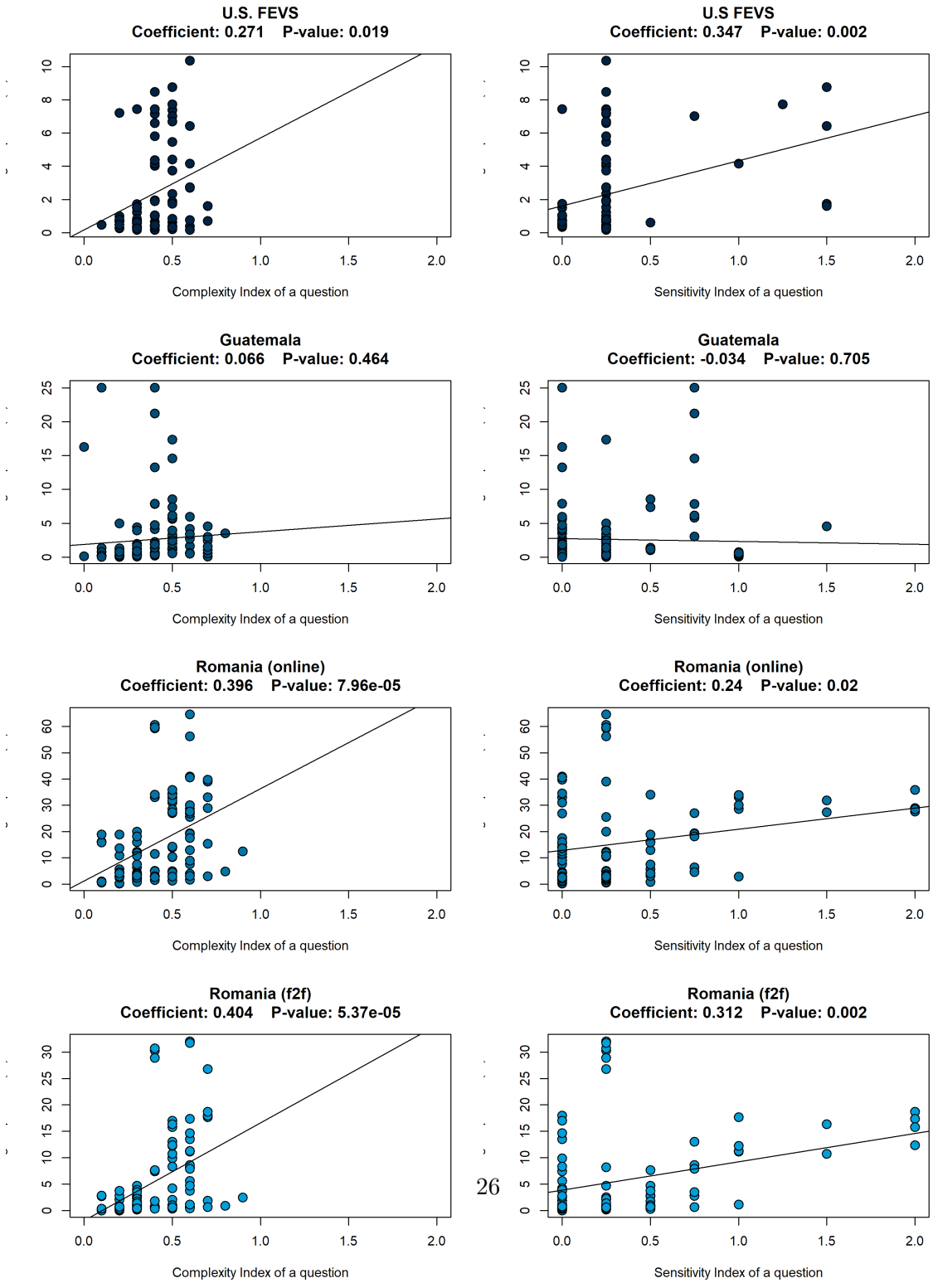


Figure 3: Correlation between sub-dimensions of complexity and sensitivity



Notes: Correlations obtained by pooling together questions across all three surveys (U.S. FEVS, Romania and Guatemala) used in this paper. Crosses mark correlations that are insignificant at 5% level

Figure 4: Relationship between complexity and sensitivity indices and the share of missing responses



To what extent do complexity and sensitivity predict item non-response? Looking, first, at correlations, we find that here is a positive association between complexity and item non-response. Figure 4 presents the correlations separately for each survey. Correlation coefficients range from 0.066 to 0.347 and are significant at the 5% level except for Guatemala. We observe a similar pattern, although slightly weaker in Romania, for correlation between item non-response and sensitivity.

Table 2 presents formal regressions of item non-response on sensitivity and complexity (both separately and jointly) with and without the aforementioned set of controls. We observe evidence from the U.S. and Romania that both complexity and sensitivity increase the probability of survey non-response in surveys of public officials. The results in Guatemala are not significant at the 10% level. The effect sizes are relatively small, with a standard deviation increase in our indices having a 1 percentage point increase in non-response in the U.S. In Romania, one standard deviation increase in complexity is associated with at most 6 percentage point increase in non-response, depending on the specification and mode of enumeration.

Table 2: The Impacts of Complexity and Sensitivity on Item Non-Response

Dependent Variable: Binary Variable Indicating Non-Response
Standard Errors: Multi-Way Clustered by Individual-Question
OLS Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
U.S. FEVS						
Sensitivity	0.009**		0.008*	0.009**		0.008*
	(0.004)		(0.004)	(0.003)		(0.004)
Complexity		0.007*	0.004		0.008*	0.005
		(0.003)	(0.003)		(0.003)	(0.003)
Romania (online)						
Sensitivity	0.039***		0.030***	0.027**		0.017
	(0.007)		(0.009)	(0.010)		(0.010)
Complexity		0.062***	0.056***		0.060***	0.057***
		(0.015)	(0.015)		(0.015)	(0.015)
Romania (face-to-face)						
Sensitivity	0.025***		0.020***	0.020***		0.015**
	(0.004)		(0.004)	(0.006)		(0.006)
Complexity		0.035***	0.030***		0.033***	0.031***
		(0.009)	(0.009)		(0.009)	(0.009)
Guatemala						
Sensitivity	−0.002		−0.002	−0.002		−0.002
	(0.004)		(0.004)	(0.004)		(0.004)
Complexity		0.003	0.003		0.003	0.002
		(0.006)	(0.006)		(0.005)	(0.006)
Controls	No	No	No	Yes	Yes	Yes

Notes: *** denotes significance at 1%, ** at 5%, and * at 10% level. Standard errors are in parentheses and are clustered using multi-way clustering at the level of an individual and a question throughout. All columns report OLS estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered '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 - non-manager; 1 - manager), experience in public administration (0 - more than 10 years; 1 - less than 10 years), job satisfaction (0 - respondent not satisfied with her job; 1 - respondents satisfied with her job; in Guatemala pay satisfaction variable is used instead). The control at the organization-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 sub-section 'Analysis' of the 'Methodology' section.

On average, our indices of complexity and sensitivity thus predict item non-response in some but not all cases. Of course, however, it could be that our indices – which simply average out different potentially relevant sub-components of complexity and sensitivity – are not appropriately aggregated. The various sub-components of complexity and sensitivity might not, as theorized, measure a single underlying dimension. To assess this, we perform exploratory factor analysis across all 14 sub-dimensions pooled together. Indeed, instead of finding that two factors are sufficient to describe the data (as would be expected if the sub-dimensions measured only two dimensions - complexity and sensitivity), we find that at least four factors are needed to properly describe the data in each of our surveys.[11]

The results of the exploratory factor analysis with 4 factors are presented in Table 3. The results suggest that, across countries, sensitivity sub-dimensions load unto a single factor (1st factor). While the scores for the second factor exhibit more variation, two sub-dimensions consistently score highly across countries: 'unfamiliarity' and 'scope of information'. Both those factors measure whether a question asks about personal or at least proximate experiences of a respondent rather than the broader working environment (e.g. the behavior of employees in the organization as a whole). Both thus relate closely to the concept that can be called 'Unamiliarity' (of topic). The remaining two factors vary, in terms of significant sub-dimensions, across countries and thus do not offer a clear conceptual interpretation.

[11] Decision was made based on the p-values of the EFA models. In Guatemala the model was significant at the 5% level only when 5 factors were used, but for cross-country consistency we employ a 4-factor model throughout the analyses

Table 3: Exploratory Factor Analysis

	1st Factor	2nd Factor	3rd Factor	4th Factor
U.S. 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: categories mismatch				
Translation to answer: number of responses				
Sensitivity				
Invasion of privacy				0.547
Socio-emotional threat	0.489	0.438		−0.259
Formal threat of sanctions	0.989			
Informal-formal threat interaction	0.909			

	1st Factor	2nd Factor	3rd Factor	4th Factor
Romania				
Complexity				
Comprehension: complex syntax			0.247	0.566
Comprehension: vagueness				-0.236
Comprehension: refererence 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			
Socio-emotial threat	0.65			
Formal threat of sanctions	0.806	0.314		
Informal-formal threat interaction	0.938	0.321		

	1st Factor	2nd Factor	3rd Factor	4th Factor
Guatemala				
Complexity				
Comprehension: complex syntax	−0.304	0.219		0.476
Comprehension: vagueness				0.508
Comprehension: refererence 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		−0.342		
Translation to answer: number of responses		−0.281		
Sensitivity				
Invasion of privacy	0.259			−0.325
Socio-emotial threat	0.484			
Formal threat of sanctions	0.776	0.241		−0.404
Informal-formal threat interaction	0.964			

Notes: Only loadings with absolute values higher than 0.2 are shown

We next assess whether the four factors from our EFA models – and, in particular, the sensitivity factor (factor 1) and the unfamiliarity factor (factor 2) – predict item non-response (4). We find that factor 1 (sensitivity) does not predict item non-response. By contrast, factor 2 (unfamiliarity) does predict item non-response in two of the three countries (U.S. and Romania) (the 3rd and 4th factors do not display clear patterns)[12].

As our exploratory factor analysis pointed to the sensitivity index as meaningfully reflecting the empirical structure of the sub-dimensions, while the complexity index consisting of 'unfamiliarity' and other complexity items, we next regress item non-response on unfamiliarity, sensitivity and complexity without unfamiliarity (Table 5). We find that unfamiliarity significantly predicts item non-response in the U.S. and Romania. It is also associated with greater item non-response in Guatemala, though this relationship is not significant at the standard significance levels.

The coefficients on the Unfamiliarity index are larger than those on our basic indices in Table 2. A standard deviation increase in our Unfamiliarity index (implying that the questions are *less* familiar) increases non-response by 3 percentage points in the U.S. and by almost 20 percentage points in the online survey in Romania. Relative to the baseline levels of non-response of 2.4% in FEVS, and 5.9% and 15.9% 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 non-response.

[12]In the following regression tables we present results only with our standard set of controls. The results however hold in unconditional regressions as well.

Table 4: Factor Analysis Regression

	U.S. FEVS	Romania (online)	Romania (f2f)	Guatemala
1st Factor	0.006	0.0001	0.006	0.005
	(0.004)	(0.010)	(0.007)	(0.005)
2nd Factor	0.018***	0.095***	0.048***	−0.002
	(0.003)	(0.013)	(0.007)	(0.005)
3rd Factor	0.003	−0.011	−0.009*	−0.002
	(0.003)	(0.006)	(0.003)	(0.003)
4th Factor	−0.002	0.029	0.016*	0.011*
	(0.002)	(0.015)	(0.008)	(0.005)
Controls	Yes	Yes	Yes	Yes
N	667,425	161,793	181,614	378,472
Adjusted R^2	0.015	0.094	0.057	0.005

Notes: *** denotes significance at 1%, ** at 5%, and * at 10% level. Standard errors are in parentheses and are clustered using multi-way clustering at the level of an individual and a question throughout. All columns report OLS estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered 'Don't Know', refused to answer or skipped a particular question. Factors scores obtained from Exploratory Factor Analysis models with 4 factors, as presented in Table 3. Controls at the individual-level include gender (0 - female; 1 - male), education (0 - below university-level; 1 - university-level), occupational status (0 - non-manager; 1 - manager), experience in public administration (0 - more than 10 years; 1 - less than 10 years), job satisfaction (0 - respondent not satisfied with her job; 1 - respondents satisfied with her job; in Guatemala pay satisfaction variable is used instead). The control at the organization-level is response rate (0-1 scale) and at the question-level the position of a question within a questionnaire.

Table 5: Impact of Sensitivity, Complexity and Unfamiliarity on Non-Response Rate

	U.S. FEVS	Romania (online)	Romania (f2f)	Guatemala
Sensitivity	0.004	−0.005	0.004	−0.005
	(0.004)	(0.011)	(0.007)	(0.005)
Complexity	−0.002	−0.027	−0.011	−0.004
(w/o unfamiliarity sub-dimensions)	(0.003)	(0.015)	(0.008)	(0.005)
Unfamiliarity	0.030***	0.196***	0.097***	0.007
	(0.007)	(0.028)	(0.017)	(0.007)
Controls	Yes	Yes	Yes	Yes
N	667,425	161,793	181,614	378,472
Adjusted R^2	0.012	0.100	0.061	0.001

Notes: *** denotes significance at 1%, ** at 5%, and * at 10% level. Standard errors are in parentheses and are clustered using multi-way clustering at the level of an individual and a question throughout. All columns report OLS estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered '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 'Information retrieval: unfamiliarity' and 'Information integration: scope of information' sub-dimensions. Controls at the individual-level include gender (0 - female; 1 - male), education (0 - below university-level; 1 - university-level), occupational status (0 - non-manager; 1 - manager), experience in public administration (0 - more than 10 years; 1 - less than 10 years), job satisfaction (0 - respondent not satisfied with her job; 1 - respondents satisfied with her job; in Guatemala pay satisfaction variable is used instead). The control at the organization-level is response rate (0-1 scale) and at the question-level the position of a question within a questionnaire.

4.1 The performance of manual versus machine-coded complexity

A potential criticism of our approach is that automated measures of complexity may be as or more effective a means of identifying potential problem questions while requiring far less investment. We therefore turn to assessing the relative effectiveness of our 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 a pre-defined lists of words regarded as easy or difficult. As there are dozens of such indices, with no agreement on which one is 'optimal', we selected nine that are commonly used and calculated their values for each survey question. Correlations among their final scores can be seen in the first column of the Figure 5 below. Given a very high degree of correlation, instead of using all nine scores in a regression, we perform principal component analysis across them and extract the first principal component (which explains between 59% to 66% of the overall variance - see second column in Figure 5 on) to serve as a predictor in our regressions.

Tables 6 and 7 evaluate the predictive power of machine-driven complexity scores. When not controlling for our manually coded indices (sensitivity, unfamiliarity and complexity without unfamiliarity), we find some evidence for an effect of 'machine-coded complexity', with significant positive effects in the U.S. only.

Figure 5: Relationship between machine-coded complexity scores - correlograms and scree plots from principal components analysis

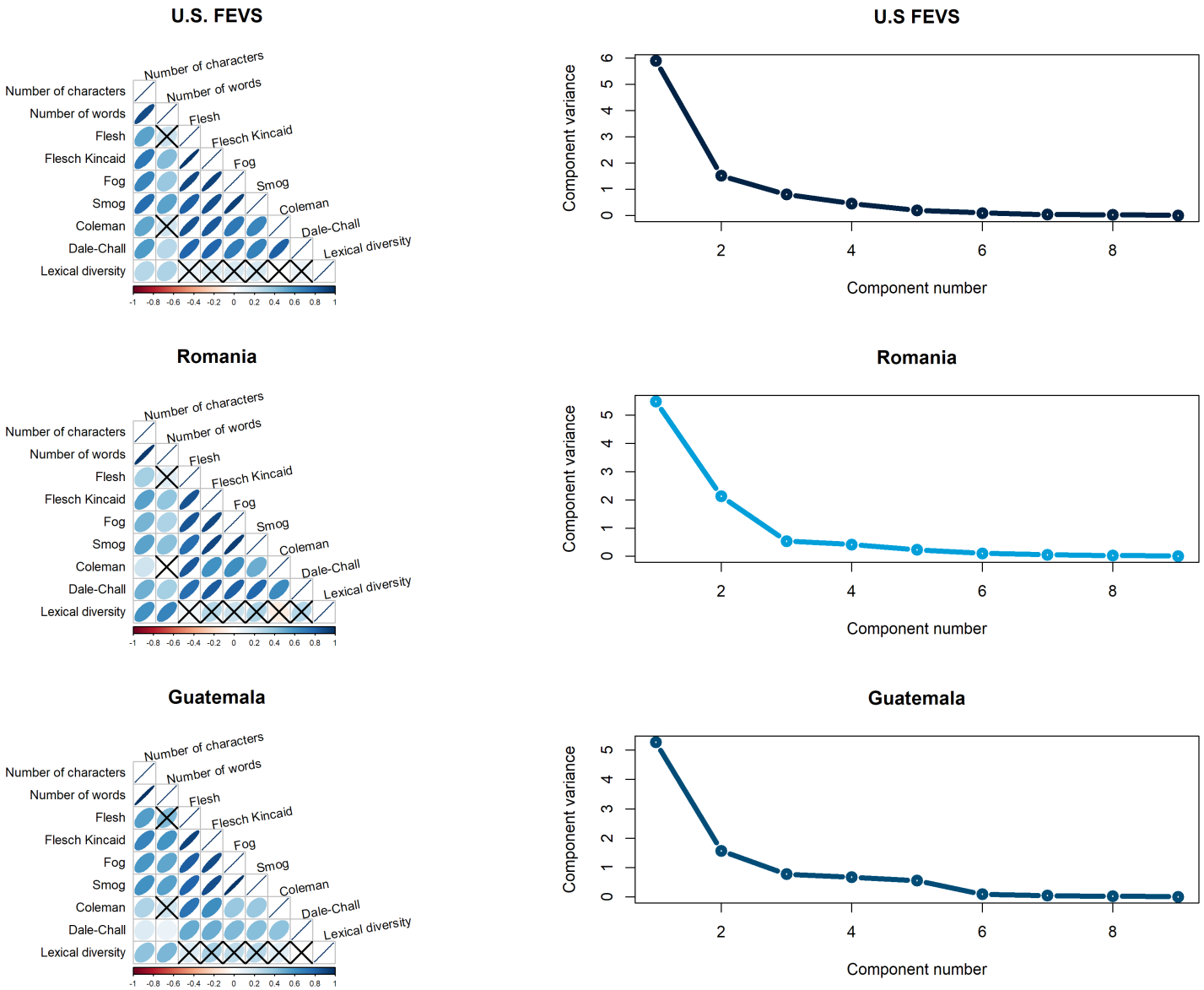


Table 6: Impact of Machine-Coded Complexity on Non-Response Rate

	U.S. FEVS	Romania (online)	Romania (f2f)	Guatemala
Machine-coded complexity	0.010***	0.022	0.009	−0.008
	(0.003)	(0.015)	(0.010)	(0.006)
Controls	Yes	Yes	Yes	Yes
N	667,425	161,793	181,614	378,472
Adjusted R^2	0.007	0.027	0.014	0.002

Notes: *** denotes significance at 1%, ** at 5%, and * at 10% level. Standard errors are in parentheses and are clustered using multi-way clustering at the level of an individual and a question throughout. All columns report OLS estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered 'Don't Know', refused to answer or skipped a particular question. Machine-coded complexity is calculated as the first principal component across 9 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 the Appendix. Controls at the individual-level include gender (0 - female; 1 - male), education (0 - below university-level; 1 - university-level), occupational status (0 - non-manager; 1 - manager), experience in public administration (0 - more than 10 years; 1 - less than 10 years), job satisfaction (0 - respondent not satisfied with her job; 1 - respondents satisfied with her job; in Guatemala, a pay satisfaction variable is used instead). The control at the organization-level is response rate (0-1 scale) and at the question-level the position of a question within a questionnaire.

Once we condition on our indices 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 non-response. Table 7 presents the full regressions. Our measure of how unfamiliar questions are is a significant and positive predictor of item non-response for the U.S. and both modes of the Romania survey. In Guatemala the coefficient on Unfamiliar is positive although smaller in size and just misses the 10% 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.

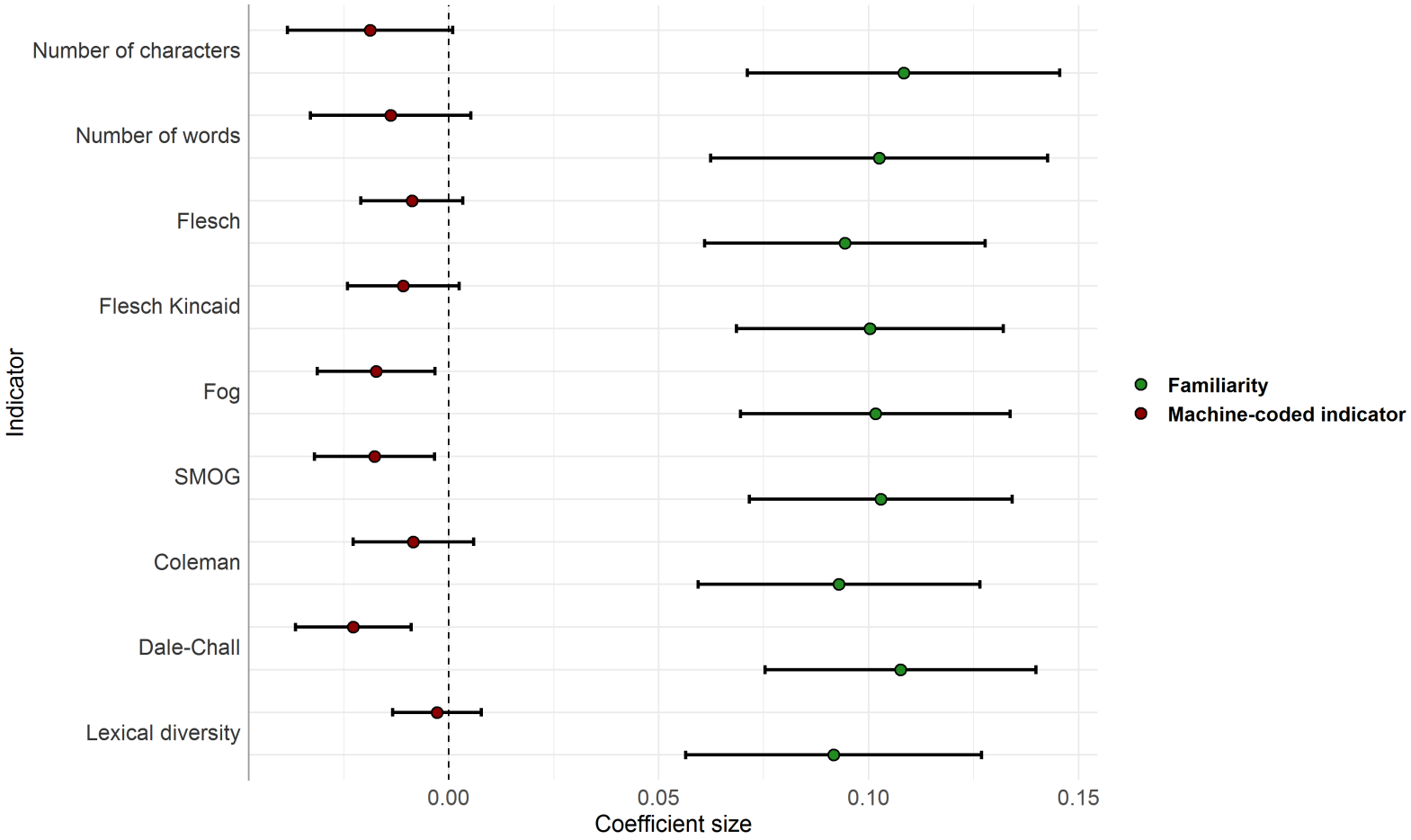
To illustrate the relative predictive power of our framework relative to machine-coded methods, Figure 7 also presents the coefficient sizes of each of individual measure of machine-coded readability and our 'Unfamiliarity' index (see Appendix to get details on each of the individual readability scores presented). For ease of presentation, we present coefficients from the Romania face-to-face survey only, but the patterns are similar throughout. Measuring a lack of familiarity clearly has a far larger predictive ability than any of the syntax-based machine-coded measures.

Table 7: Full model: Impact of Sensitivity, Complexity, Unfamiliarity, Machine-Coded Complexity

	U.S. FEVS	Romania (online)	Romania (f2f)	Guatemala
Sensitivity	0.003	−0.008	0.002	−0.002
	(0.003)	(0.010)	(0.006)	(0.007)
Complexity	−0.002	−0.026	−0.011	−0.003
	(0.003)	(0.015)	(0.008)	(0.005)
Unfamiliarity	0.028***	0.217***	0.109***	0.010
	(0.008)	(0.026)	(0.016)	(0.006)
Machine-coded complexity	0.003	−0.029	−0.017*	−0.009
	(0.003)	(0.015)	(0.008)	(0.009)
Controls	Yes	Yes	Yes	Yes
N	667,425	161,793	181,614	378,472
Adjusted R^2	0.013	0.103	0.064	0.003

Notes: *** denotes significance at 1%, ** at 5%, and * at 10% level. Standard errors are in parentheses and are clustered using multi-way clustering at the level of an individual and a question throughout. All columns report OLS estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered '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. Familarity is calculated as a mean value of 'Information retrieval: unfamiliarity' and 'Information integration: scope of information' sub-components. Machine-coded complexity is calculated as the first principal component across 9 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 the Appendix. Controls at the individual-level include gender (0 - female; 1 - male), education (0 - below university-level; 1 - university-level), occupational status (0 - non-manager; 1 - manager), experience in public administration (0 - more than 10 years; 1 - less than 10 years), job satisfaction (0 - respondent not satisfied with her job; 1 - respondents satisfied with her job; in Guatemala pay satisfaction variable is used instead). The control at the organization-level is response rate (0-1 scale) and at the question-level the position of a question within a questionnaire.

Figure 6: Machine-coded complexity indicators: Romania (f2f)



Values show the size of coefficients of 'machine-coded' complexity indicators when they are entered individually into a regression model with our standard DV and control variables.
The size of the coefficient for 'Familiarity' entered into the same regression is shown for comparison
Error bars indicate 95% confidence intervals.

5 Conclusion

The loss of precision and potential biases introduced by item non-response 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 non-response 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 non-response in surveys of public servants. We apply this framework to government-wide surveys of public officials in Guatemala, Romania and the U.S. As in the existing literature, we find that complexity matters for item non-response. Contrary to much prior work on item non-response however, public servants do not seem to shy away from questions which are complex due to, for instance, syntax, computational intensity or the number of response options (see for example Knäuper et al. 1997). As we had argued in the introduction, this may be as 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 official's about issues with which they have lower familiarity is the feature of question design that is most robustly associated with item non-response. Questions that ask for assessments of public sector organizations as a whole or departments within those, for instance, lead to greater item non-response than questions about public servants themselves.

The implication for survey designers is clear: asking about topics public officials are less familiar with – such as their organization or department, rather than their immediate work environment – is associated with greater item non-response, with concomitant concerns about greater variance and potential biases in estimates. Where data aims to assess practices in larger units or institutions, it would thus be preferable from a non-response perspective to ask respondents about their individual-level experience with the organizational practice and aggregate those.

Our 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.

We compare the predictive ability of our finding to models which in-

clude machine-coded measures of complexity. Our findings underscore the importance of 'manual' assessments by survey designers to assess question complexity. While machine-coded estimates had some predictive power, this was eclipsed by our manual coding approach once that 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 U.S. Our diverse case contexts give us confidence that our findings are generalizable. Probing generalizability should not only extend to testing different country-settings, but also different survey administrators. The noticeable lower item non-response in the FEVS compared to the World Bank-administered surveys might reflect differential levels of trust in the survey administrator itself, for instance. Our framework could equally be employed in employee surveys in private sector companies. One worthwhile area of investigation is to understand whether our findings are unique to public officials or would similarly apply 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 rollout variation in different aspects of these concepts, and in particular unfamiliarity, and assess experimentally whether this leads to improvements in item response rates in their setting.

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

Moreover, our inferences are necessarily limited by the number of surveys (3 countries) and types of questions included. In particular, our surveys contained relatively few highly sensitive questions (see Figure 4), which might partially explain the null results we obtained. It is possible that more discernible patterns on item non-response 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 non-response which highlights a particular aspect of complexity, unfamiliarity, as the fundamental driver of non-response.

Appendix

Definitions of machine-coded complexity scores

The analyses presented above use several measures of complexity of a text that can be put in the form of a computer code and calculated automatically. They are all based on the characteristics of a text syntax, like the number of characters, syllables, words and sentences. Many measure the complexity using school grade or years of education required to easily understand a given text. The 9 measures that were used in the analyses above can were calculated and can interpreted as follows[13]:

1. **Number of characters**

   Sum of all letters and numbers in a text of a question. Punctuation marks, like commas or dots, question marks and blank spaces are all excluded from the count.

2. **Number of words**

   Sum of all words in the question. Numbers are not treated as words. Words connected by a hyphen (e.g. 'machine-coded') are treated as separate words.

3. **Flesch's Reading Ease Score (1948)**

   One of the oldest reading ease scales that are still commonly used. It is based on calculations using an average sentence length, number of syllables and words. Reading ease is measured on 1-100 scale[14], where, contrary to many other indices, higher score indicates greater reading ease, i.e. smaller complexity. Therefore, for consistency, the

[13] All measures were calculated in R using 'stringi' and 'quanteda' packages. Definitions of the measures are taken from the help files attached to these packages and the details can be found at:

- https://www.rdocumentation.org/packages/stringi/versions/1.5.3/topics/stri_stats_latex
- https://www.rdocumentation.org/packages/quanteda/versions/2.1.2/topics/textstat_readability
- https://www.rdocumentation.org/packages/quanteda/versions/1.3.4/topics/textstat_lexdiv

[14] However scores outside this range are possible.

scores were reversed in all the analyses.

Formula:
$206.835 - (1.015 \times ASL) - (84.6 \times \frac{Nsy}{Nw}))$

ASL = Average Sentence Length - number of words divided by the number of sentences
Nsy = Number of syllables
Nw = Number of words

4. **Flesch-Kincaid Readability Score (1975)**

Based on the Flesch Reading Ease Score, the U.S. Navy developed the Flesch-Kincaid scale in 1970's to measure reading ease in a manner that would be more straightforward to interpret. Flesch-Kincaid measures reading ease using U.S. grade level. A score of 8 would therefore mean that the text could be easily understood by someone who has completed at least 8 school grades. It range should therefore be between 0 and 18[15]

Formula:
$0.39 \times ASL + 11.8 \times (\frac{Nsy}{Nw}) - 15.59$

ASL = Average Sentence Length - number of words divided by the number of sentences
Nsy = Number of syllables
Nw = Number of words

5. **Gunning's Fog Index (1952)**

The Gunning's Fog Index is also based on U.S grade level and therefore the scores are intepreted in similar manner to the Flesch-Kincaid Readability Scale described above. This index however also takes into account the length of words, not only sentences, treating words with more than 2 syllables as difficult. The scaling by 100 arises because the original FOG index is based on just a sample of 100 words

[15] Again, exceptions are possible

Formula:
$0.4 \times (ASL + 100 \times (\frac{Ns3}{Nw})$

ASL = Average Sentence Length - number of words divided by the number of sentences
$Ns3$ = The number of words with 3-syllables or more
Nw = Number of words

6. **SMOG (Simple Measure of Gobbledygook) (McLaughlin 1969)**

The SMOG acronym disguises the intended use of the index as a simple measure of gobbledygook, that is a text that is hard to comprehend due to overload of complex and technical language. Similarly to the Gunning's Fog Index it takes into account the presence of long (>2 syllables) words in a text. The scale is based on U.S. grade level.

Formula:
$1.043 \times \sqrt{(\frac{Ns3 \times 30}{Nst})} + 3.1291$

$Ns3$ = The number of words with 3-syllables or more
Nst = Number of sentences

7. **Coleman's Readability Formula 1 (1971)**

One of several versions of Coleman's readability formulas originally created for U.S Office of Education to measure reading complexity of textbooks. The formula used here accounts for the number of words and one-syllable words, on the assumption that one-syllable words are easy to understand. It usually takes scores between 0 and 100, with higher scores indicating less complex text. Therefore, for consistency, the scores were reversed in all the analyses. The scaling by 100 in this Coleman measure arises because it is calculated on a per 100 words basis.

Formula:
$1.29 \times (100 \times \frac{(Ns1)}{Nw}) - 38.45$

$Ns1$ = the number of one-syllable words
Nw = Number of words

8. **New Dale-Chall Readability formula (1995)**

The only formula used here that measures the complexity by comparing words present in a text to a list of pre-defined words judged to be difficult. Higher scores originally mean less complex text. Therefore, for consistency, the scores were reversed in all the analyses.

Formula:
$64 - (0.95 \times 100 \times \frac{Nwd}{Nw} - (0.69 \times ASL))$

ASL = Average Sentence Length - number of words divided by the number of sentences
Nw = number of words
Nwd = number of "difficult" words not matching the Dale-Chall list of "familiar" words

9. **Lexical Diversity**

It calculates the lexical diversity of documents using a variety of indices. The scores are measured on 0-1 scale, where higher values indicate less complex questions. Therefore, for consistency, the scores were reversed in all the analyses. Details can be found at:
www.rdocumentation.org/packages/quanteda/versions/1.3.4/topics/textstat_lexdiv

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 Research 48*(2), 263-295

Barton, A. (1958) Asking the embarassing question. *Public Opinion Quarterly 22*(1), 67-68

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), 269-279

Berman, J., H., McCombs, R. Boruch (1977) Notes on the Contamination Method. Two Small Experiments in Assuring Confidentiality of Responses. *Sociological Methods Research 6*(1), 45-62

Blair, E., S. Sudman, N. Bradburn, and C. Stocking (1977) How to ask questions about drinking and sex: Response effects in measuring consumer behavior. *Journal of Marketing Research 14*(3), 316-321

Bradburn, N., S. Sudman, E. Blair and C. Stocking (1978). Question threat and response bias. *Public Opinion Quarterly, 42*(2), 221-234.

Bryson, A., J. Forth, and L. Stokes (2017). Does employees' subjective well-being affect workplace performance?. Human Relations 70(8), 1017-1037

Catania, J., D. Binson, J. Canchola, L. Pollack, W. Hauck, and T. Coates (1996). Effects of Interviewer Gender, Interviewer Choice, and Item Wording on Responses to Questions Concerning Sexual Behavior. *Public Opinion Quarterly 60*:345–75

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 & Research 40*(1), 169-193

DePaulo, B., S. Kirkendol„ D. Kashy, M. Wyer, and J. Epstein (1996) Lying in everyday life. *Journal of Personality and Social Psychology 70*(5), 979–995

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

De Leeuw, E., J. Hox, and M. Huisman (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics 19*(2), 153-176

Faaß, T, L. Kaczmirek, and A. Lenzner (2008). Psycholinguistic Determinants of Question Difficulty: A Web Experiment. *7th International Conference on Social Science Methodology*

Fazio, R. (1986). How Do Attitudes Guide Behavior?. In: R. Sorrentiono and E. Higgins (eds.) *Handbook of Motivation and Cognition: Foundations Social Behaviour*, Chapter 8 (pp. 204-243) Guilford Press, NY: New York

Fazio, R. (1989). Attitude Accessibility, Attitude-Behavior Consistency, and the Strength of the Object-Evaluation Association. *Journal of Experimenal Social Psychology 18*(4), 339-357

Fazio, R., and D. Roskos-Ewoldsen (2005). Acting as We Feel: When and How Attitudes Guide Behavior. In: T. Brock, and M. Green (eds.) *Persuasion: Psychological insights and perspectives* (p 41-62). Sage Publications, Inc.

Fisher, R. (1993) Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research 20*(2), 303-315

Fuller, C. (1974). Effect of Anonymity on Return Rate and Response Bias in Mail Surveyu. *Journal of Applied Psychology 59*(3), 292-296

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*, 451-461

Gingerich, D. (2010) "Understanding Off-the-Books Politics: Conducting Inference on the Determinants of Sensitive Behavior with Randomized Response Surveys". *Political Analysis 18*(3), 349-380

Gnambs, T., and K. Kaspar (2015) SDisclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. *Behavour Research Methods 47*, 1237-1259

Gnambs, T., and K. Kaspar (2017) Socially Desirable Responding in Web-Based Questionnaires: A Meta-Analytic Review of the Candor Hypothesis. *Assessment 24*(6), 746-762

Gonzales-Ocantos, E., C. de Jonge, C. Meléndez, J. Osorio, D. Nickerson (2012). Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua. *American Journal of Political Science 56*(1), 202-217

Graesser, A., Z. Cai, M. Louwerse, and F. Daniel (2006). Question Understanding Aid (QUAID). A Web Facility That Tests Question Comprehensibility. *Public Opinion Quarterly 70*(1), 3-222

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin, and B. Puranen et al. (eds.). 2020. World Values Survey: Round Seven - Country-Pooled Datafile. Madrid, Spain Vienna, Austria: JD Systems Institute WVSA Secretariat.

Haziza, D., G. Kuromi (2007). Handling item nonresponse in surveys. *Journal of Case Studies in Business, Industry and Government statistics 1*(2), 102-118

He J., F. van de Vijver, A.D. Espinosa, et al. (2014) Socially Desirable Responding: Enhancement and Denial in 20 Countries. *Cross-Cultural Research 49*(3), 227-249.

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), 836-846

Holbrook, A., Y. Cho, and T. Johnson (2006). The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficul-

ties. *Public Opinion Quarterly 70*(4), 565-595

Höglinger, M., B Jann, 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), 171-187

Joinson, A., A. Woodley, and U-D. Reips (2007). Personalization, authentication and self-disclosure in self-administered Internet surveys *Computers in Human Behaviour 23*, 275-285

Just, M, and Carpenter, P. (1992). A Capacity Theory of Comprehension: Individual Differences in Working Memory. *Psychological Review 99*(1), 122-149

Kaplan, R., E. Yu (2015). Measuring Question Sensitivity. *U.S. Bureau of Labor Statistics. Office of Survey Methods Research.* AAPOR015 (October 2015), 4017-4121

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), 181-199

Krumpal (2013). Determinants of Social Desirability Bias in Sensitive Surveys: a Literature Review. *Quality Quantity 47*, 2025-2047

Lalwani A., S. Shavitt, and J. Johnson (2006) What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology, 90*(1), 165–178

Lenski, G., and J. Leggett (1960). Caste, Class, and Deference in the Resarch Inteview. *American Journal of Sociology 65*(5), 463-467

Lensvelt-Mulders, G. (2008).Surveying Sensitive Topics. In: E. de Leeuw, J. Hox, and D. Dillman (eds.) *International Handbook of Survey Methodology* (p. 461–578). Routledge, NY: New York.

Lessler, J., R. Tourangeau, and W. Salter (1989). Questionnaire Design in the Cognitive Research Laboratory. *Vital and Health Statistics 6*(1) (U.S. Deparment of Health and Human Services, DHHS Publication No. (PHS) 89-1076)

Krosnick, J. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology 5*(3), 213-236

Krosnick, J., and S. Presser (2009). Question and Questionnaire Design. In: J. Wright and P. Marsden (eds.) *Handbook of Survey Research (2nd Edition)*. San Diego, CA: Elsevier Available on: https://web.stanford.edu/dept/communication/faculty/kr [Access: 07/04/2021]

McNeeley, S. (2012). Sensitive Issues in Surveys: Reducing Refusals While Increasing Reliability and Quality of Responses to Sensitive Survey Items. In: L. Gideon (eds.) Handbook of Survey Methodology for the Social Sciences (p. 4377-396). Spriner, NY: New York.

Näher, A-F., I. Krumpal (2011) Asking sensitive questions: the impact of forgiving wording and question context on social desirability bias. *Quality Quantity 46*, 1601-1616

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

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: H. Braun, D. Jackson, and D. Wiley (eds.), The role of constructs in psychological and educational measurement (p. 49–69). Mahwah, NJ:Erlbaum.

Peter, J., and P. Valkenburg (2011). The impact of "forgiving" introductions on the reporting of sensitive behavior in surveys: the role of social desirability response style and developmental status. *Public Opinion Quarterly, 75*(4), 779–787.

Rässler, S., Riphahn, R.T. (2006) Survey Item Non-response and Its Treatment. *Allgemeines Statistisches Arch 90*, 217–232 (2006)

Risko, E., L. Quilty, and J. Oakman (2006) Socially desirable responding on the web: investigating the candor hypothesis. *Journal of Personality Assessment 87*(3), 269-276

Sakshaug, J., Yan, T, and Tourangeau, R. (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), pp. 907-933

Schuster, C., J. Fuenzalida, J. Meyer-Sahling, K.S. Mikkelsen, and n. Titelman (2019). Encuesta Nacional de Funcionarios en Chile. Evidencia para un Servicio Público Más Motivado, Satisfecho, Comprometido y Ético. *Informe preparado para la Dirección Nacional del Servicio Civil, Santiago, Chile, January 2020*

Sudman, S., N. Bradburn, (1982). Asking questions: A practical guide to questionnaire design. *John Wiley Sons. First Edition*

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. Jabine, M. Straf, J.Tanur, and R. Tourangeau (eds.) *ognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines* (pp. 73-100). National Academy Press, Washington, DC.

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., L. Rips, and K. Rasinski (2000). The psychology of survey response. Cambridge, UK:Cambridge University Press.

Tourangeau, R., and T. Yan (2007) Sensitive Questions in Surveys, *Psychological Bulletin 133*(5), 859-883

Tourangeau, R., Groves, R, and Redline, C. (2001) Sensitive Topics and Reluctant Respondents. Demonstrating a Lnk Between Nonresponse Bias and Measurement Error*Public Opinion Quarterl 74*(3), 413-432

Tversky, A., and Kahneman, D. (1974) Judgment under Uncertainly: Heuristics and Biases. *Science 185*(4157), 1124-1131

Uriell, A., and C. and Dudley (2009). Sensitive Topics: Are there modal differences? *Computers in Human Behaviour 25*(1), 76-87

Wimbush, J., D. Dalton (1997) Base Rate for Employee Theft: Convergence of Multiple Methods. *Journal of Applied Psychology 82*(5), 756-763

World Bank (2000). Diagnostic Surveys of Corruption in Romania. U.S, Washington, D.C: World Bank

World Bank (2020a) Final Field Report. Encuesta General de Servidores Públicos y Contratistas del Organismo Ejecutivo y Entitdaded Descentralizadas 2019. Washington, D.C: 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, D.C: 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 University Press, Cambridge, UK

Zelenski, J, S. Murphy, and D. Jenkins (2008). The happy-productive worker thesis revisited. Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being, 9(4), 521–537