

Measuring Corruption using a Bayesian Randomized Response Item Theory Model

October 1, 2025

Corruption carries significant political and economic consequences. There is extensive evidence that corruption can lead to a reduction in the provision of public goods and undermine trust in democratic institutions. Scholars have been attempting to reliably measure the prevalence of corrupt behaviors in the population using surveys with direct question formats. However, these measures are likely biased due to social desirability and non-response biases. Indirect questioning survey techniques have been designed to minimize these biases and elicit truthful answers to sensitive topics and behavior. However, the canonical design of these techniques only allows the measurement of group-level estimates. This paper presents empirical evidence of an extension of the Randomized Response Technique, known as the *Randomized Item Count Response Technique (RIRT)*, to estimate both group- and individual-level corrupt behaviors ($n = 6058$ and $n = 3692$). At the group level, we found prevalence rates ranging from 60% in the case of patronage to 1% in administrative corruption. For individual-level estimates, we found that the distribution of respondents' underlying traits for engaging in corrupt behavior follows a power-law distribution. We implemented several approaches to identify inattentive study participants, demonstrating that our results are robust once these respondents are excluded.

1 Introduction

Measuring corruption is inherently a measurement problem, as it is challenging to observe and quantify due to its secretive and illegal nature. International efforts in measuring corruption have largely relied on a combination of composite indicators, which are partly derived from survey data, as well as purely survey-based measures. Prominent examples include the Corruption Perception Index and the World Bank’s Control of Corruption Index, the Global Corruption Barometer, and the International Crime Victimization Survey. These instruments typically seek to generate experience-based prevalence rates by asking respondents directly about the behavior of interest.

Although this approach is relatively straightforward, the reliability of the estimates derived from this questioning approach hinges on respondents’ willingness to answer and provide truthful answers. The evidence so far is that survey measures using direct questions are prone to be biased due to many reasons such as social desirability bias, fear of retribution, or even legal consequences (Blair, Imai and Lyall, 2014, Yusaku Horiuchi Horiuchi, Markovich and Yamamoto, 2021, Singh and Tir, 2022, Smallpage et al., 2022). Several survey techniques for sensitive questions (SST) have been developed to mitigate these biases by increasing the confidentiality of respondents’ answers. These methods enable researchers to estimate prevalence rates for the targeted behavior only at the group level either by introducing noise to individual responses or by aggregation.

One of the key limitations of these techniques is that these methods enable researchers to estimate prevalence rates for the sensitive behavior at the group level, yet at the expense of losing information at the individual level. This trade-off renders the conventional approach relatively inefficient, as it limits the ability of researchers and policymakers to gain insights into more deeply rooted patterns of individual corrupt behavior. Furthermore, conventional indirect techniques only allow for measuring experience-based indicators of the corrupt behavior of interest, but do not capture the underlying latent trait that drives such behavior.

This paper contributes to overcoming a key limitation of indirect questioning techniques by pro-

viding empirical evidence on an extension of the Randomized Response Technique that integrates Item Response Theory (Fox, 1997, 2005). This extension not only yields prevalence rates of corrupt behavior at the group level, but also provides individual-level estimates of respondents' latent disposition toward corruption as well as their likelihood of engaging in the behaviors in question.

Measuring individual-level estimates of latent traits and propensity to engage in corrupt behavior has important implications for our understanding of this subject. A common strategy for examining heterogeneity is to estimate the prevalence of corrupt behavior across groups. While informative, this approach is limited to differences between predefined groups and fails to capture the full spectrum of variation in prevalence estimates across individuals. To uncover this richness within a population, it is necessary to estimate propensities to engage in the behavior in question at the individual level.

This technique provides a more robust alternative to the conventional single-question approach, offering a more stable and comprehensive assessment of citizens' underlying, unobservable trait of engaging in corrupt behavior. Latent scores can provide a more direct proxy for actual conduct than attitudinal indicators. Individual behavioral propensities can have significant implications for policy, when widespread, can translate into systemic practices that impose substantial economic costs by undermining investment, and erode institutional capacity.

Several observational and experimental studies require individual-level scores either as an outcome or a covariate. On one end, capturing individual-level latent scores avoids the problems of the ecological fallacy, which involves making claims based on higher-level estimates, while also avoiding downstream bias due to measurement error in the final estimation.

Finally, this approach has important implications for assessing the reliability of survey techniques. By estimating individuals' likelihood of engaging in corruption, it enables validation against direct individual-level survey responses or ground-truth measures on a subject-by-subject basis, rather than relying solely on aggregate comparisons.

Given the multivariate framework of the RIRT, our paper expands the scope of analysis in the public sector, ranging from well-documented practices such as bribes and vote-buying to more specific

and understudied ones such as influence peddling, sextortion, and document tampering.

We addressed the concerns raised by an extensive number of studies regarding confidence, attentiveness, and procedural adherence regarding SSM (John et al., 2018, Ibbett, Jones and St John, 2021), following multiple designed-based and estimation-corrections estimates.

We provide evidence of this technique from two online surveys conducted in Chile ($n = 6,044$ and $n = 3,721$), where we comprehensively measured the prevalence rates of 11 corrupt behaviors identified in local governments. At the group level, our results show that patronage and the exploitation of political connections are more prevalent than other forms of corruption, such as sexual solicitation from public officials and giving or soliciting bribes. As for individual-level estimates of respondents' latent traits, we observed respondents in our sample yield low scores in their underlying propensity to engage in corrupt behavior. Most of our sample's latent trait are clustered around zero, but a few subjects obtained considerably higher scores.

This paper aims to contribute several research strands that examine survey methods and aim to measure corruption. The first main contribution is in survey methods for sensitive topics, as it provides empirical evidence on an extension of the randomized response technique that incorporates Item Response Theory. It also contributes to the research that seeks to measure corruption, where we provide population-level and individual estimates of a comprehensive number of corrupt behaviors. To the best of our knowledge, this is the first paper that provides subject-level propensity estimates of engaging to corrupt behaviors using an indirect questioning technique. A third empirical contribution is to the literature investigating survey inattentiveness. We conducted several novel strategies to identify inattentive respondents and provided bias-corrected estimates at the group and individual levels.

The paper continues as follows. Section 2 provides a summary of the different approaches used to measure experience-based corruption. Section 3 introduces the Randomized Response and the Randomized Item Response Technique, outlining the corruption item included and the sampling used in this survey. Section 4 discussed the estimation approaches for the RIRT design. Then, in Section 5 reports and elaborates on group and individual-level estimates, followed by a number of robustness

checks to identify how sensitive our results are to inattentiveness and procedural adherence. Finally, Section 6 summarizes the findings of this study and its limitations.

2 Measuring Corruption: Challenges and Approaches

Scholars and policymakers have attempted multiple approaches to measuring experienced-based forms of corruption. These approaches mostly comprised surveys, official records, and qualitative interviews. Some of these surveys are representative of the population or targeted of public officials, business executives. These surveys

Measures from official records stem from data provided by institutions such as National Statistics, public procurement units, financial regulations and tax authorities, customs borders, law enforcement and audit agencies, and the judicial body. Most of these institutions have been geared to identify corrupt behavior such as suspicious financial activities, smuggling, tax evasion, fraudulent financial activity, irregularities in the public sector, and misuse of public funds. In this paper, we focused in different forms of corruption that citizens face once they engage in with their local government.

There is also a growing body of literature that links different institutional frameworks to both the prevalence and type of corruption (Dawson et al., 2024). This literature builds on the theories that decentralized entities are more prone to moral hazard problems, particularly in situations where local governments have considerable discretion or are exempt from regulations in various policies, such as hiring processes or public procurement.

Most measurement efforts aim to capture forms of corruption related specific forms of corruption such as bribe and vote-buying, but there is little there is a lack of policies targeted low curb pretty/administrative corruption such as bribes solicitation, vote-buying, document tampering, procurement fraud, regulatory bypass, resource misuse, information leaking. Furthermore our motivation to measure corruption at the local level is empirically motivated.

Corruption in Chile Chile stands out within the south American region as one of the countries with lowest perception of corruption based on the Corruption Perception Index (2020, Transparency International), yielding a corruption perception index close to industrialized countries such as France and Portugal. Similarly, OECD Anti-Corruption and Integrity Outlook (2024) provides a similar picture, but more nuanced faced. Even there are institutional arrangements, municipalities and their subsidiaries public entities ("corporations") are excepted from many of these rules. mid-level bureaucrats, procurement procedures are limited. Furthermore, most of the d

The AmericasBarometer has provided a long survey evidence of corruption in the population, and the Council for Transparency have provided measured for most of these metrics refer to institutional arrangements that. Despite the overall assessment citizens' beliefs of the prevalence of corrupt behavior remain high.

3 Randomized Item Response Technique, Sampling and Survey Design

Randomized Response Technique The RR method was developed by Warner (1965), and its most basic level introduces random noise that would help delink study participants' responses to a sensitive question. In practice, RR designs introduce noise using an instrument such as a die or a coin. Respondents are asked to throw a die or flip a coin, and the outcome from this device determines which question the respondent has to answer or which forced answer the respondent has to give. However, the researcher knows the probability distribution that describes the likelihood of all possible outcomes induced by the instrument. This known probability distribution then allows for estimating the ratio of affirmative answers. The RR design provides complete confidentiality, as neither the interviewer nor the researchers know the outcome of the device. Furthermore, in its basic design, we yield group-level

estimates without disclosing which respondents engage in the sensitive behavior.¹

The evidence of the effectiveness of RR technique in minimizing social desirability and non-response bias is overall positive. Lensvelt-Mulders et al. (2005) concluded from their meta-analysis that this technique was effective in yielding robust prevalence estimates of sensitive topics once compared to the direct questioning format.² Similarly, Höglinger, Jann and Diekmann (2016) conducted a study comparing four different RR variants to direct questioning by looking at student misconduct in Universities in the Netherlands. The authors found evidence that all RR designs outperformed the direct questioning version in at least three of the five sensitive questions included in the study. This mounting evidence has buttressed the RR design as a reliable and efficient technique to obtain truthful answers from respondents.

Indeed, the evidence of empirical validation of RRs against official records is small yet promising. Rosenfeld, Imai and Shapiro (2016) found that RR performed better than other techniques in estimating the proportion of people voting 'No' in the Mississippi anti-abortion referendum in 2011.³ The authors also noted that the RR technique produced robust estimates with significantly lower variance than the other indirect questioning techniques.⁴

The crosswise design stands out among all the RR variants due to its statistical efficiency and higher procedural adherence. Blair, Imai and Zhou 2015 extensively analyze different RR designs and their performance by looking at their main statistical properties. The authors strongly support the crosswise design [say more]. Similarly, Sagoe et al. (2021) conducted a meta-analysis that included 35 studies using the crosswise design. The author found that this variant outperforms the direct questioning technique based on the 'more is better' validation criteria. The author also reported that

¹It is important to clarify that there is no deception in the RR designs. However, researchers can use some information provided by the respondent, such as their month of birth, to produce an unrelated question; for example, the unrelated question could be: "I was born in the three months of the year".

²Their meta-analysis comprised two meta-analyses: one meta-analysis included only six individual validation studies, whereas the second meta-analysis contained 32 comparative validation studies.

³In their study, they compared the estimates from the RR design to a direct questioning format and a list and endorsement experiments.

⁴The authors validated their results at the individual level using official records, which is state-of-the-art in terms of validation approach among these survey methods.

the crosswise design is particularly effective when measuring the prevalence of susceptible issues. Overall, the existing evidence provides a solid ground for using the method to elicit genuine answers.

Randomized Item Response Theory. RIRT combines RR with Item Response Theory (IRT) (Fox, Veen and Klotzke, 2019, Fox, 2005, Fox and Wyrick, 2008). This extension allows the measurement of underlying attitudes or abilities as a function of an individual’s responses to a set of closely related questions. Thus, rather than asking just one or a few sensitive questions, respondents must answer a set of closely but mutually exclusive forms of corrupt behavior. This collection of questions allows for estimating individuals’ single unobservable trait of interest (Wu, Wood and Stevenson, 2019).

Using RIRT over conventional RR designs and other indirect questioning techniques has two advantages. Firstly, and most importantly, this technique yields individual-level estimates of the lurking trait of interest (Fox, 2002, 2005). In contrast, all conventional indirect questioning techniques only allow for estimating the prevalence of sensitive behavior at the population or group level. Secondly, RIRT allows for hierarchical analysis where researchers can investigate group differences of the sensitive traits at sub-levels to the level at which the data is gathered. Both features, in theory, would contribute to providing precise estimates of both population, sub-group, and individual-level estimates of the behavior in question. Fox, Avetisyan and van der Palen (2013) validated this technique, finding that RIRT effectively identified smokers from non-smokers. The authors compared individual estimates of respondents’ smoking traits versus post-survey expired-air CO test results that revealed respondents’ true smoking status.

We can formalize how the RR design incorporates IRT. In the standard RR design, a respondent will answer affirmatively ($Y_{ijk} = 1$) to an item k is a function of π_{ijk} . This probability is calculated based on each respondent’s latent trait. In this study, the underlying trait of interest is the individual propensity to engage in a particular type of corrupt behavior. In the IRT literature, the θ_i parameter corresponds to the individual unobservable trait. We can further assume that individuals are nested in j groups, where individuals in some clusters are more likely to answer positively than others (if that is

the case, this would be denoted as θ_{ij}). In the RR set-up, given that we introduce some random noise, the probability of observing a positive answer ($Y_{ijk} = 1$) is the product of π_{ijk} and the probability of also answering affirmatively to the question related to the random device. In this case, the researchers know the probability of the random device, which we can denote as q . Putting all of these parameters together, we can formalize the probability of answering positively in the RR set up to an item k as follows:

$$P(Y_{ijk} = 1) = p\pi_{ijk} + (1 - p)(1 - \pi_{ijk}) \quad (1)$$

In RIRT, the probability of answering positively also depends on each item's attributes. Thus, π_{ijk} is a function of α_k and β_k parameters. α_k represents the discrimination power of an item k . This parameter defines how item k differentiates individuals based on their underlying attribute(s). An item with high discrimination power gives a larger probability of answering positively to respondents with a greater underlying trait. Instead, it provides a lesser probability to those with a lower underlying trait. Conversely, an item with low discrimination power provides roughly the same probability levels to answering positively to questions with different levels of respondent's underlying traits. β_k is the difficulty parameter and determines the percentage of respondents answering favourably to an item. Thus, an item with a higher difficulty level yields a lower chance of positively answering a question. In contrast, an item with a lower difficulty level will provide a higher probability of what of answering affirmatively. If we include these two parameters into the response function, the likelihood of answering affirmatively to an item is as follows.

$$P(\tilde{Y}_{ijk} = 1 | \theta_{ij}, \alpha_k, \beta_k) = \Phi(\alpha_k \theta_{ij} - \beta_k) \quad (2)$$

In this logistic model, the chance of answering positively to the item k is a function of the cumulative probability distribution function Φ and the parameters mentioned before. We can then integrate this into the RR's set up, where π_{ijk} is equal to $P(\tilde{Y}_{ijk} = 1 | \theta_{ij}, \alpha_k, b_k)$:

$$\pi_{ijk} = P(\tilde{Y}_{ijk} = 1 | \theta_{ij}, \alpha_k, \beta_k) = \Phi(\alpha_k \theta_{ij} - \beta_k) \quad (3)$$

Combining Equations 1 and 3, we formally incorporate IRT into the RR design:

$$P(Y_{ijk} = 1) = p\Phi(\alpha_k \theta_{ij} - \beta_k) + (1 - p)(1 - \Phi(\alpha_k \theta_{ij} - \beta_k)) \quad (4)$$

As researchers know the probability distribution of the random device, we can then replace the value of p with its chance of answering affirmatively to the unsensitive device question. For example, in the case of a die, the probability of responding positively is equal to $1/6$, thus $p = 1/6$ and $1 - p = 5/6$; therefore, Equation 4 is equivalent to:

$$P(Y_{ijk} = 1) = \frac{1}{6}\Phi(\alpha_k \theta_{ij} - \beta_k) + \frac{5}{6}(1 - \Phi(\alpha_k \theta_{ij} - \beta_k)) \quad (5)$$

Thus, in addition to estimating the auxiliary parameters α_k , and β_k , we will be able to estimate respondents' latent trait θ_{ij} , which is the parameter of interest for this study.

Sampling. For this study, we conducted three online survey waves using the Nuffield Centre for Experimental Social Sciences subjects panel in Chile. The first wave was conducted in January 2020, where 6,044 respondents completed the survey. This survey included the RIRT items plus additional questions about beliefs about corruption, vote intention, partisanship and demographics.

The second wave was conducted five months after the first wave. The same 6,044 respondents who completed the first wave were invited to participate in the follow-up survey. From the 6,044 study participants that answered the first wave, 3,721 completed the second wave, with a 37% overall attrition rate.

We conducted the third wave by randomly sampling 1,163 respondents who answered at least one of the previous waves. In this round, rather than having all eleven items, we included only three

items, but in a direct questioning format. The prevalence estimates obtained from this wave served as a benchmark to compare against the estimates obtained from the RIRT waves. Although the results from this survey serves to conduct a comparative validation, this is not the primary goal of this paper. The results of this comparative validation are reported in section 5.1 in the Appendix. We complemented this comparative validation analysis by comparing the population-level estimates of bribes obtained from the RIRT to estimates obtained from a representative survey ($n = 2900$) conducted annually in Chile.⁵ that uses a direct questioning format⁶

Within all the RR variants, we chose the crosswise design as it provides a stronger signal of higher levels of confidentiality, higher efficiency, and a moderate level of instruction complexity (Blair, Imai and Zhou, 2015, Hoffmann and Musch, 2016, Höglinger, Jann and Diekmann, 2016, Korndörfer, Krumpal and Schmukle, 2014, Hoffman et al., 2015). Following the design suggestions from Wu, Wood and Stevenson 2019, 11 items were included in the survey, which is a sufficient number of items to achieve lower levels of bias and minimize Root Square Mean Errors.

The items included in the questionnaire were selected based on several existing surveys measuring different forms of corrupt behaviour, such as the Latin American Public Opinion Project (LAPOP, 2017), Chile's Consejo para la Transparencia (de la Transparencia, 2019), and other surveys and studies examining clientelism, electoral violence, and vote-buying (Fergusson, Molina and Riano, 2017, Transparente, 2019). Table 1 presents the 11 dichotomous items that were included in the RIRT survey waves.

To ensure respondents understood the RIRT instructions, we included three test rounds before the respondents answered the sensitive questions in the RIRT format. Participants who answered correctly in the first two rounds were allowed to continue with the sensitive questions. Subjects that responded incorrectly in the first two test rounds were asked to answer the third test before continuing with the RIRT section.

⁵This survey is commissioned the Consejo para la Transparencia, which is an independent non-partisan entity in Chile.

⁶In this survey they only asked about bribes, so we were not able to make additional comparison.

In the second wave, we include four ‘nested’ items reported at the end of Table 1. These questions are designed to flag inattention or poor comprehension, as they identify specific behaviors associated with four core corrupt behaviors. Thus, in theory, nested items should yield lower prevalence estimates relative to their corresponding core item. Finally, we incorporated an *Anchor item* into the second wave of the RIRT, which serves as an additional check of attentiveness. The logic of the item is that if respondents pay attention, we should expect to yield near-zero prevalence estimates for that item, as the prevalence of that type of behavior is very unlikely to occur in the population. We provide more details of these strategies to pin down inattentive respondents in Section 5 of this manuscript.

Table 1: List of items

Core Items		Item name
1	I know personal cases either of friends or relatives who used their personal or political connections to be hired by the municipality	Patronage
2	I used personal or political connections to get better service than others in my municipality	Influence peddling
3	I received money or been offered a special favor, gift or benefit by my municipality to vote in a particular way	Vote-buying
4	I had to give money or goods or gifts to a public official of my municipality so that they could give or award me a contract, service, or benefit.	Bribes
5	A public official of the municipality helped me alter administrative documents or information to obtain a service, benefit, or contract	Forgery
6	A public official allowed me to use public resources of the municipality for personal purposes	Personal
7	A public official of my municipality asked me for sexual favors, either suggestively or openly, in exchange for a benefit or service	Sextortion
8	I pressured municipal officials to skip administrative processes or laws	Lobbying
9	I used personal or political connections or paid public officials of my municipality so they will not oversee an administrative process or project	Compliance
10	A City official threatened me to vote or not vote in a certain way	Electoral violence
11	A City official shared with me privileged information about the municipality	Privilege
Nested Item		
12	I know cases where close relatives such as <i>siblings/partners</i> used their personal or political connections to be hired by the municipality	Patronage
13	I had to give money, gifts or goods to a public official of my municipality so I could obtain a <i>social protection card</i>	Bribes
14	A public official of the municipality has helped me alter administrative documents <i>to obtain my driver's license or a certificate</i>	Forgery
15	Officials of my municipality shared with me privileged information about <i>possible social benefits</i> that the municipality provides	Privilege
Anchor Item		
16	An official from my municipality asked me to pay a bribe to their bank account abroad	

In the crosswise method implemented in this study, respondents were asked to think of a number between one and six and roll a virtual die. Then, they read two statements: a sensitive statement and an unrelated non-sensitive statement. The non-sensitive statement was, "My number is the same as the one on the die". The sensitive statement was one of the eleven corrupt behaviors reported in Table 1. Finally, study participants answered how many of the previous statements were true. Subjects had to choose between option A: *Both or neither of the statements are true* or option B: *Just one of the statements is true*. Figure 1 shows a screenshot of the instructions that respondents saw when they answered the survey.

Figure 1: An example of a randomized response survey item

Please follow the instructions below:


1. Please think of a number between 1 and 6
2. Please throw the die
3. Please read the statements below and answer the question
 - My number is the same as the one in the die
 - I paid a bribe to a city official

How many of the following statements are true?

Please select the appropriate option below

☐ Both or neither of the statements are true

☐ One of the statement is true



Throw the die

Note: This figure shows what respondents observed answered the RIRT survey. Study participants were asked to think of a number between 1 and 6. Then, they were asked to throw a virtual die. They were informed that the researchers did not know what was the outcome of the die. Finally, they had to read the two statements, one insensitive and unrelated statement that asked whether the number they thought was the same obtained from rolling the die. The second statement was about sensitive behavior. The order of the statements was randomized; thus, for some participants, the sensitive statement came up first, and for some respondents, the sensitive statement came up second. Finally, respondents were asked to answer whether the 'Both or neither' statement are true or 'One' of the statement is true.

4 Methods: Explanation, estimation, inference

Using RIRT enables both group-level estimates of sensitive behaviors and individual-level estimates of the underlying sensitive trait. Group-level estimates can be obtained with modified logistic or ordinary least squares models. For individual-level estimates, we rely on a Bayesian estimation framework using MCMC sampling to approximate the posterior distribution of the parameters.

Population-level estimates When using binary responses to measure the prevalence of sensitive issues, we can use an adapted version of a logistic regression that incorporates the randomized response design into the logistic function (Bourke and Moran, 1988, Liu and Zhang, 2017a,b). Following Yu, Tian and Tang (2008)’s formalization of the RR estimation approach, the underlying trait π is a probabilistic measure of the $P(X = 1|Z)$. This probability of answering yes to the sensitive questions is conditional on set of covariates Z that predict engaging to the sensitive trait. In the logistic regression set up, we can derive β , which is unknown parameter of interest (proportion), which we can estimate using maximum likelihood estimation.

$$\pi = \frac{Z' \beta}{1 + Z' \beta} \quad (6)$$

As we do not observe X , which is whether respondent has engaged on the sensitive behavior ($X = 1$), the maximum likelihood for β is as follows:

$$\ln L(\beta|X, Z) = \sum_{i=1}^n [X_i \ln(\pi_i) + (1 - X_i) \ln(1 - \pi_i)] \quad (7)$$

In the crosswise design, we have a response variable R that takes the following values, that is

conditional on the respondents' exposure to the sensitive trait and the result of the unrelated question:

$$R = \begin{cases} 1, & \text{if response is both } (X = 1 \text{ and } Y = 1 \text{ or } X = 0 \text{ and } Y = 0) \\ 0, & \text{if response only one } (X = 1 \text{ and } Y = 0 \text{ or } X = 0 \text{ and } Y = 1) \end{cases} \quad (8)$$

We know the probability of $Pr(Y = 1|Z)$ of answering yes to the unrelated question, which is equal to p . Therefore the probability of $R = 1$ is equal to $\pi p + (1 - \pi)(1 - p)$. Whereas the probability of $R = 0$ is equal to $\pi(1 - p) + (1 - \pi)p$. We then can rewrite the log-likelihood function as follows:

$$\ln L(\beta|R, Z) = \sum_{i=1}^n R_i \cdot \ln[\pi_i p_i + (1 - \pi_i)(1 - p_i)] + (1 - R_i) \cdot \ln[\pi_i(1 - p_i) + (1 - \pi_i)p_i] \quad (9)$$

$$\ln L(\beta|R, Z) = \sum_{i=1}^n R_i \cdot \ln[\pi_i e^{Z'_i \beta} + (1 - \exp^{Z'_i \beta})(1 - p_i)] + (1 - R_i) \cdot \ln[e^{Z'_i \beta}(1 - p_i) + (1 - e^{Z'_i \beta})p_i] \quad (10)$$

We can also estimate the proportion of respondents that have engaged in the sensitive behaviour using a modified version of a linear probability model. In this set up, the probability of engaging of the sensitive behaviour, conditional on a vector of covariates is equal $Pr(X = 1|Z)$. This probability is equal to $Z'\beta$. We can derive β from the following transformation:

$$E(R = 1|Z) = Pr(R = 1|Z) = (Z'\beta)p + (1 - Z'\beta)(1 - p) \quad (11)$$

$$E(R = 1|Z) = (Z'\beta)(2p - 1) + 1 - p \quad (12)$$

The response variable is equal to $\hat{R} = (R + p - 1)/(2p - 1)$, we can replace this function into equation 12:

$$E[(R + p - 1)/(2p - 1)|Z] = \frac{E(R = 1|Z) + p - 1}{2p - 1} = Z'\beta \quad (13)$$

We report the linear probability estimates in Figure 12 in the Appendix. There was a minor coding discrepancy in the probabilities, with negligible consequences for estimation.

Individual Response Probabilities. A key contribution of conducting a RIRT is that it can yield individual-level estimates of the latent trait of interest and probabilities of engaging in the behavior in question. As we pointed out earlier, there is a debate as to whether these latent estimated traits reflect actual behavioral proclivities.

We can formalize the relationship between the observed randomized responses and the latent and true responses, assuming that respondents are clustered in $kj = 1 \dots J$ groups. Thus there are $i = 1, \dots, n_j$ individuals in each group. As before, we can state Y_{ijk} as the individual's latent trait.

$$P(Y_{ijk} = 1) = \phi\Phi(\alpha_k\theta_{ij} - b_k) + \phi(1 - \Phi(\alpha_k\theta_{ij} - b_k)), \text{ where } \phi = \frac{1}{6} \quad (14)$$

We use a Bayesian approach to estimate individuals' underlying trait by conducting a two-parameter RIRT model (Fox, Veen and Klotzke, 2019, Fox, 2005, Fox and Wyrick, 2008). We use two model identification approaches: First, we identify the model by fixing the mean and variance of the ability parameters (θ_{ij})' to zero and one. This minimum identification strategy yields reliable ability parameter estimates. At the same time, the discrimination parameters do not converge well with this model. We believe it is because most sensitive questions show very low population-level prevalence rates (10%).

As an additional identification strategy, we anchored two difficulty item parameters that yield the highest and the lowest group-level prevalence rates. While the individual-level ability parameters are consistent across the two models, we obtain more reliable discrimination parameters using the second model. We use a dispersed normal prior ($N(0, 100)$) for the difficulty parameters and a uniform prior

$(U(0, 5))$ for the discrimination parameters. The equations below summarize the estimation strategy

$$Y_{ik} | \alpha_k, \beta_k, \theta_i \sim \text{Bernoulli}(\pi_{ik}) \quad (15)$$

$$\pi_{ik} = \phi \Phi(\alpha_k \theta_i - \beta_k) + \phi(1 - \Phi(\alpha_k \theta_i - \beta_k)) \quad (16)$$

$$\alpha_k \sim \text{Uniform}(0, 5) \quad (17)$$

$$\beta_{(-k_{highest}, -k_{lowest})} \sim N(0, 100) \quad (18)$$

$$\beta_{k_{highest}} = \Phi^{-1}(\pi_{k_{highest}}) \text{ and } \beta_{k_{lowest}} = \Phi^{-1}(\pi_{k_{lowest}}) \quad (19)$$

$$\theta_i \sim N(0, 1) \quad (20)$$

Respondents' probability of being corrupt. We determined corrupt behavior for each subject, when the posterior probability of the corrupt latent score is above a specific threshold. We followed Fox's approach by imposing an arbitrary threshold called 0_c . A positive prediction of the individual hold the true $X_i = 1$, when the model predict a respondent to be corrupt, where the true corrupt status is equal to $D_i = 1$. The posterior probability of a positive diagnosis $X_i = 1$, conditional on the response pattern is given by the following model:

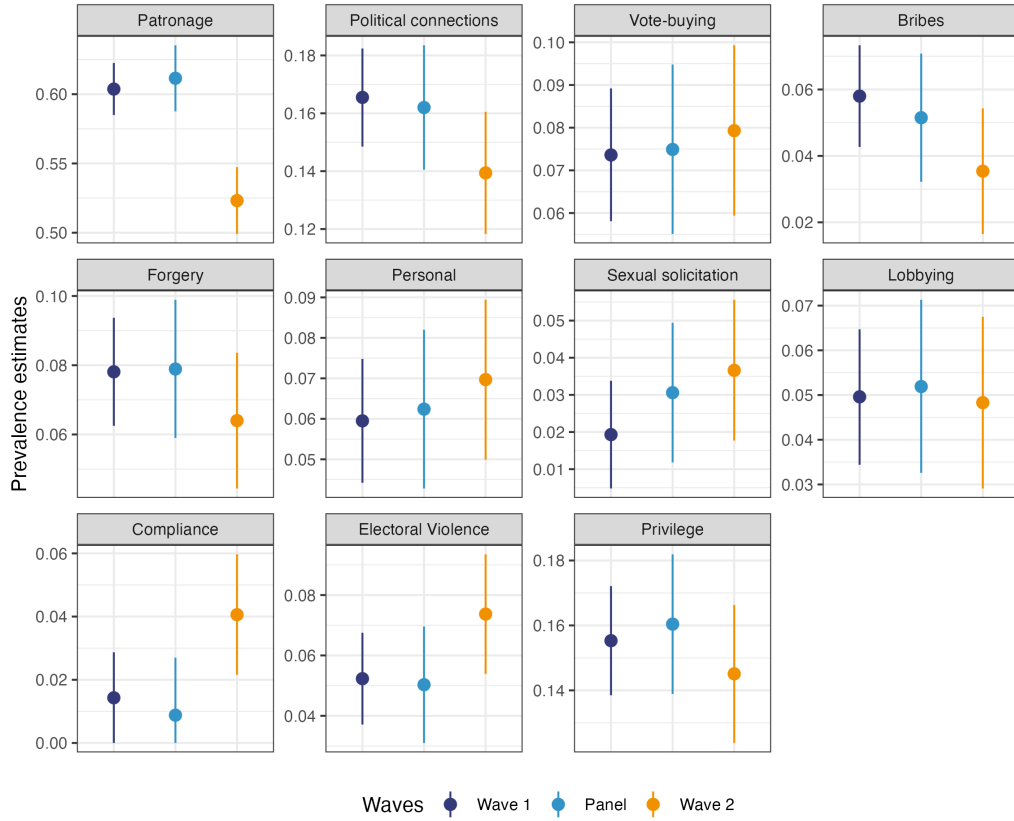
$$p_{ic} = P(X_i = 1 | \mathbf{y}_i, \theta_c) = P(\theta_i > \theta_c | \mathbf{y}, \theta_c) \quad (21)$$

$$= \int_{-\infty}^{\theta_c} p(\theta_i|\mathbf{y})d\theta_i \quad (22)$$

5 Results

Group-level estimates We report the prevalence of each sensitive behavior using equation ???. As presented in Figure 2, there is variation across the different corrupt behaviors, going from 52-60% of cases related to patronage to around 1-4% to paying bribes to reduce administrative compliance. We can also observe that the estimated prevalence of the different corrupt behaviors tends to be relatively similar across both waves. However, there is a significant drop of around 8% in the case of patronage. Table 4 in the appendix reports the *Naive* estimates for both waves, as well as first wave respondents that also took part on the second wave, we refer to this sub-sample in this manuscript as *Panel*. We observe similar estimates between *Panel* and *Wave 2*, where in some items the prevalence estimates are the same to differences of a 9%. We complemented this analysis providing estimates from a linear probability model regression estimates for both waves of the RIRT in Figure 12 in the Appendix. Estimates from the modified OLS tend to be similar across time, except for patronage, which drops substantially on the second wave.

Figure 2: Prevalence estimates of each item - RIRT Waves



Note: This figure report prevalence estimates for all 11 items included both in *Wave 1*, *Panel*, and *Wave 2*. These are the 'Naive' and unweighted estimates. In dark blue, we show estimates for the first wave, in light blue for the *Panel*, and in yellow, the results from *Wave 2*. We can derive from this plot that there is only a statically significant difference for the 'Patronage' item.

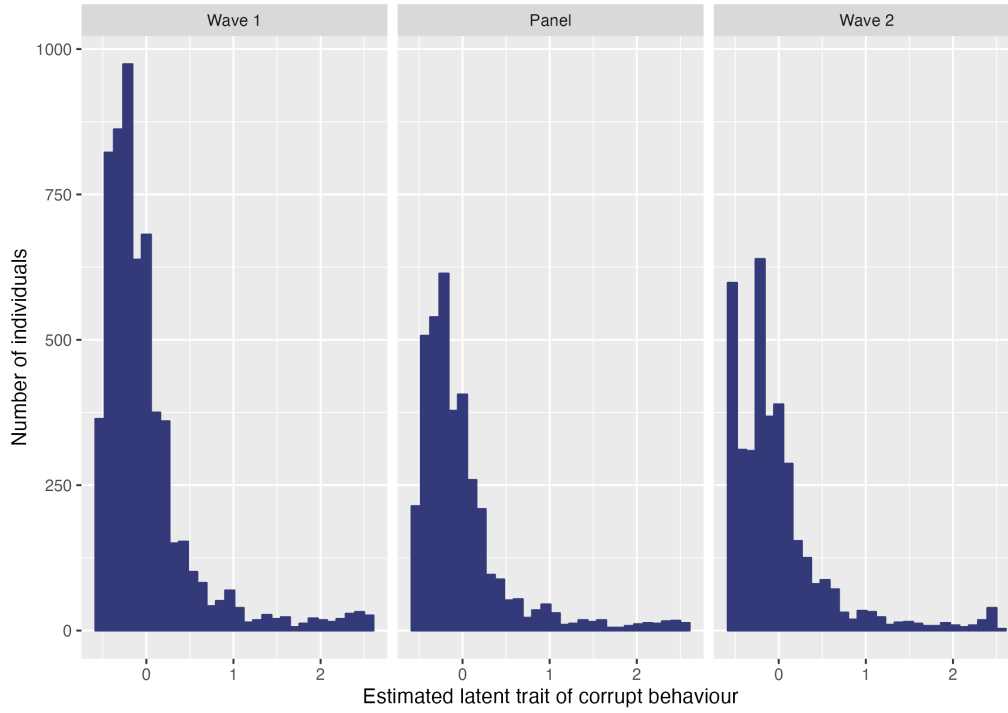
We also find considerable differences between *Wave 1*, and *Panel* and *Wave 2* for the 'Compliance' and 'Electoral Violence' items. For the 'Compliance' item, prevalence rates were around 0.015, whereas for the *Wave 2* estimates around 0.04. Similarly, for the 'Electoral Violence' item, estimates both in the first wave and the *Panel* respondents was around 0.05, but then increased up two per cent in the follow-up wave. We can explain differences between *Wave 1* and *Wave 2* are partly due to temporal changes (assuming there is full instruction compliance and full attentiveness) in prevalence rates of corrupt behaviors. We can rule out that these changes are due to changes in the sampling composition, given that the prevalence estimates obtained from *Panel* are closely pegged to the estimates from the

first wave. Differences on the prevalence rates and their respective p-values are summarized in Table 7.

Given this study was administered using a sample from an opt-in unrepresented respondents' pool, we provide estimates in Figure 2 that include sample weights of the naive estimates. We applied inverse probability weighting (Horvitz and Thompson, 1952) to account for population differences across communes. This weighting strategy shows that the results of the estimates are somewhat consistent with this weighting scheme, as roughly eight of the items yield similar estimates. We extended this analysis using different weights. The results of this additional analysis are documented in Figure 13 in the Appendix.

Individual-level estimates We estimated individual-level latent traits of corrupt behavior for the 6,044 respondents in the first wave. We also repeated the estimation for 3,721 respondents in the second wave and *Panel*. We run the MCMC algorithm with 5,000 burn-ins and 5,000 iterations after the burn-ins. Figure 3 shows the distribution of individual-level RIRT estimates of the latent trait of corrupt behavior (θ_i) in each survey wave. As shown in Figure 3, many corrupt behavior items show low group-level prevalence rates. Therefore, the distribution of individual-level latent trait estimates is right-skewed. In other words, because most corrupt behaviours asked did not occur (or were not admitted by respondents), the majority of respondents' latent corruption scores are low. On the other hand, a small group of respondents who have admitted committing these corrupt behaviors receive extremely high estimates.

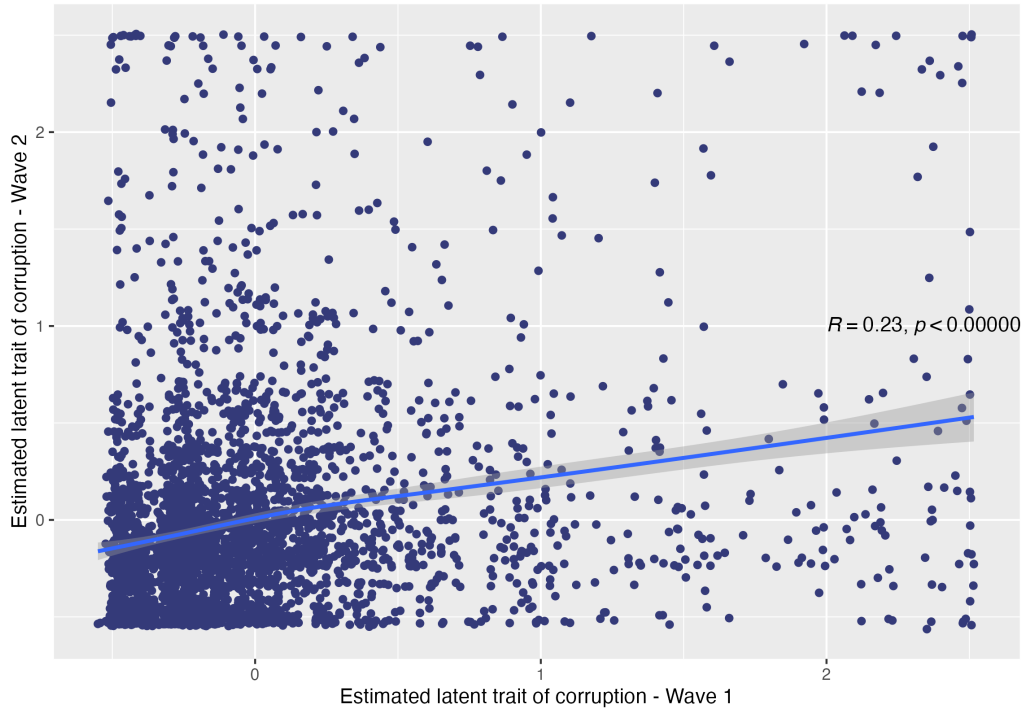
Figure 3: Estimated individual-level latent trait of corrupt behavior



Note: This figure shows the distribution of respondents' latent trait of corrupt behavior for both *Wave 1* and *Wave 2*. This figure illustrates that subjects' corrupt trait distribution is right-skewed. Most observations are clustered around zero in both waves. We can also see outliers that generate large tails for the highest values of these estimates.

Figure 4 shows the relationship between the RIRT estimates in the first and second waves. Although the correlation is not very strong ($R = 0.23$), it is positive. This result would suggest different levels of compliance and attentiveness across respondents. For example, we can explain a subject scoring a lower value on her underlying trait on the second wave for several reasons: 1) The respondent forgot the instances where she was engaged in the corrupt behaviors asked in the surveys. 2) The subject preferred not to disclose one or several corrupt behaviors asked in the second wave. 3) The subject did not comply or understand the instructions in at least one of the waves. In contrast, subjects with higher or relatively similar scores on their underlying trait are perfectly reasonable results. For instance, respondents may or may not have engaged in at least one of the sensitive behaviors asked in the surveys between each wave.

Figure 4: Estimates of individual-level latent trait of corrupt behavior in both waves



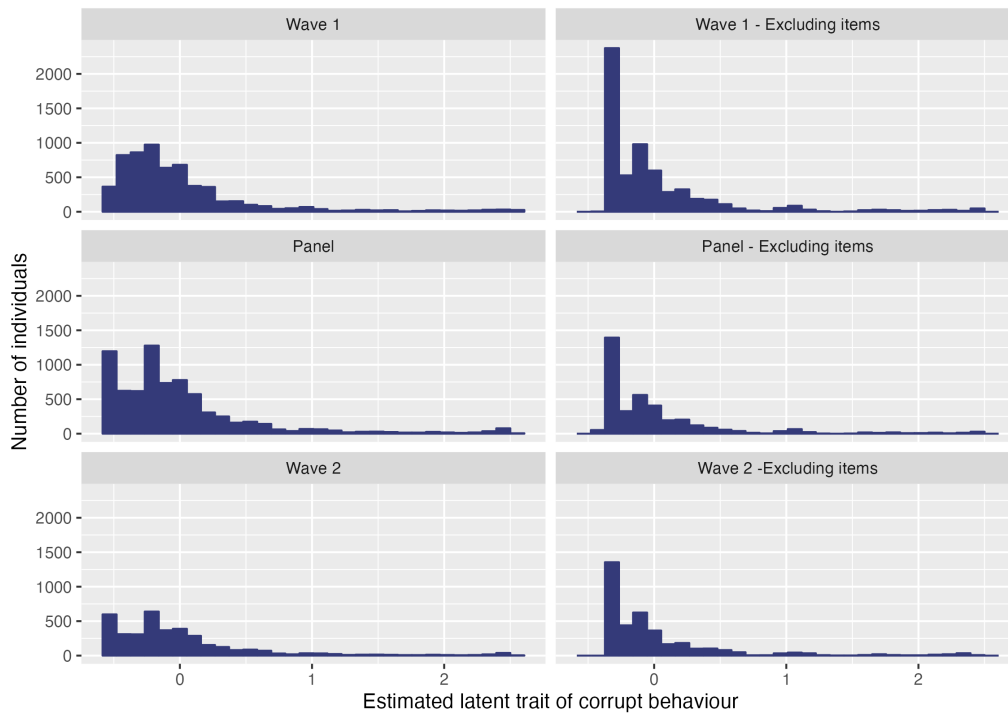
Note: This figure is a scatterplot that reports estimates of respondents' corrupt behavior trait from *Panel* and *Wave 2*. On the x-axis are the estimates from the *Panel*, and on the y-axis are the estimates from *Wave 2*. We also report the coefficient of correlation and add a regression line that shows a positive relationship between these two estimates.

Item Analysis One concern in estimating respondents' underlying trait is selecting items that accurately measure the dimension of interest. Usually, researchers select questions from a pool of validated items with well-known and precisely estimated parameters such as its discrimination and difficulty (Baker and Kim, 2017). However, researchers may want to drop some items in exchange of obtaining more accurate estimates. For example, researchers may want to remove items ex-post because these questions perform poorly in capturing respondents' unobservable trait (Hambleton and Swaminathan, 1985)⁷. In this study, we estimated individuals' underlying trait, dropping the 'Patronage' and 'Sexual solicitation' items. We excluded these items for the following reasons: 1) Both questions may not be capturing the respondent's underlying propensity to engage in corrupt behavior.

⁷For example, Yang et al. (2013) tested the Confusion Assessment Method to identify Delirium (acute confusion), and their suggested dropping items from this instrument to maximize the probability of identifying this illness

In the case of 'Patronage', although this question has been used in surveys that measure patronage (Fornasari et al., 2022), the question opens up the possibility for respondents to claim that they are aware of cases of patronage from third parties. For the case of 'Sexual solicitation', the question's format asks whether respondents have been victims of sexual coercion, but not whether the corruption behavior is not initiated by the respondents. The results of this analysis are in Figure 5 where we see a considerable change in the distribution of subjects' underlying trait in all waves.

Figure 5: Estimates of individual-level latent trait of corrupt behavior in both waves - Excluding items



Note: This figure shows the distribution of respondents' underlying trait of corrupt behavior for *Wave 1*, *Wave 2* and *Panel*. We report the original estimates using all 11 core items in the left-hand column. The distributions in the right column correspond to the distribution of respondents' corrupt trait, excluding the 'Patronage' and 'Sexual solicitation' items.

In both cases, estimates are clustered around zero; we find that the distribution of these estimates is susceptible to change once we drop these two items. We performed a Wilcoxon sign rank test to compare the means of two dependent non-normal distributions. We find statistically significant differences between the two distributions for waves 1 and 2. We do not find statistically significant differences

for *Panel* ($p = 0.037$). More details of the magnitude of individual-level changes are in the Appendix in Table 5.

Addressing Attentiveness and Instruction Adherence One of the shortcomings of the RR designs is that they put a substantial burden on the respondents (Jerke et al., 2021). This additional burden can lead to inattention, lack of cooperation and compliance with the RR instructions. In RIRT, this is particularly important as respondents are required to answer a long list of sensitive items.

A common approach to mitigate the inattentive bias is adding a compliance/attention check RR item whose population-level prevalence estimate is approximately zero (Höglinger and Diekmann, 2017, Atsusaka and Stevenson, 2021). Another common approach is to include a test RR item whose randomization device (e.g., secret number) is known to the researcher. Researchers can use these attention check techniques to screen out non-compliant respondents from the sample.

Atsusaka and Stevenson (2021) proposed another approach where the randomization device is unknown to researchers and the sensitive item is related to the primary research subject. This approach is better at testing compliance because the item is in the same format as the actual RR items. In their design-based attention check question, they include an RR item known to have zero population-level prevalence rate. Equations 24 and 26 provides the bias-corrected estimator and its variance proposed by these authors:

$$P(Y = 1) = \lambda = \{p\pi_{CM} + (1 - p)(1 - \pi_{CM})\}\gamma + \kappa(1 - \gamma) \quad (23)$$

$$B_{CM} \equiv E[\hat{\pi}_{CM}] - \pi = \left(\frac{1}{2} - \frac{1}{2\gamma}\right) \left(\frac{\lambda - \kappa}{p - \frac{1}{2}}\right), \text{ where } \kappa = \frac{1}{2} \quad (24)$$

$$\hat{\pi}_{BC} = \hat{\pi}_{CM} - \hat{B}_{CM} \quad (25)$$

$$Var(\hat{\pi}_{BC}) = V \left[\frac{\hat{\lambda}}{\hat{\lambda}'} \left(\frac{\frac{1}{2} - p'}{2p - 1} \right) \right] \quad (26)$$

γ is the proportion of attentive respondents, κ is the probability with which inattentive respondents pick "both or neither is true." B_{CM} is the bias with respect to the quantity of interest caused by inattentive respondents.

To address this issue, we implemented Atsusaka and Stevenson (2021) attention check technique that helps identifying inattentive respondents. We used the 'Anchor item' reported at the end of Table 1 to correct the bias due to inattentiveness. Respondents were asked the following question: "An official from my municipality asked me to pay a bribe to their bank account abroad"⁸. This item should yield near zero prevalence estimates; as in this context, it is improbable that local government officials would ask citizens to deposit bribes into to their foreign bank accounts.

Table 2: Prevalence estimates - Naive, Weighted and Bias-Corrected

Core Items	Wave 1				Wave 2			
	Whole sample		Panel					
	Unweighted	Weighted	Unweighted	Weighted	Unweighted-Naive	Unweighted-Corrected	Weighted-Naive	Weighted-Corrected
Patronage	0.60	0.59	0.61	0.60	0.52	0.52	0.51	0.51
Political connections	0.17	0.18	0.16	0.18	0.14	0.13	0.15	0.15
Vote-buying	0.07	0.07	0.07	0.07	0.08	0.07	0.08	0.07
Bribes	0.06	0.07	0.05	0.06	0.04	0.03	0.05	0.04
Forgery	0.08	0.09	0.08	0.08	0.06	0.05	0.07	0.07
Personal	0.06	0.05	0.06	0.06	0.07	0.06	0.08	0.08
Sexual solicitation	0.02	0.02	0.03	0.03	0.04	0.03	0.04	0.03
Lobbying	0.05	0.05	0.05	0.06	0.05	0.04	0.05	0.05
Compliance	0.01	0.03	0.01	0.03	0.04	0.03	0.04	0.04
Electoral Violence	0.05	0.06	0.05	0.07	0.07	0.06	0.09	0.08
Privilege	0.16	0.15	0.16	0.15	0.15	0.14	0.15	0.14
Nested - Patronage					0.17	0.17	0.17	0.17
Nested - Bribes					0.05	0.04	0.04	0.04
Nested - Forgery					0.04	0.03	0.04	0.04
Nested - Privilege					0.11	0.10	0.11	0.10
Num.Obs.	6044	6044	3721	3721	3721	3721	3721	3721

Note: This table reports group-level both Naive and Bias corrected estimates of the corrupt behaviors. These estimates are reported as proportions of the population engaged in the sensitive behavior. We report weighted and unweighted estimates. Post-stratification adjusts the sampling and replicate weights so that the joint distribution of a set of post-stratifying variables matches the known population joint distribution. For post-stratification weights we used demographic variables such as age, education attainment, occupation, and gender.

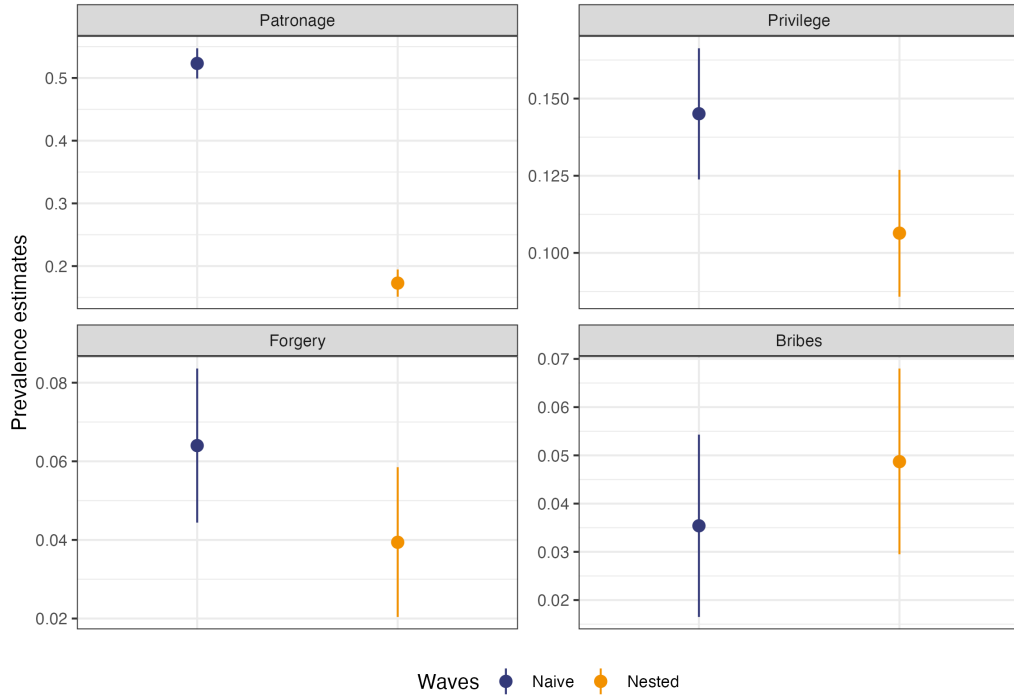
The bias-corrected estimates using the check item are reported in Table 2. Using only the sample

⁸We use their *cWise* package to produce the bias-corrected population-level prevalence rates.

of the second wave, the estimated attentive rate is 97.8%. Because of the high attentive rate, there is little difference between the 'naive' RIRT estimates and the bias-corrected RIRT estimates. Overall, The bias-corrected estimates are lower than the naive estimates by $\sim 1\%$.

Nested Items As we stated before, we included four nested items to evaluate respondents' attentiveness and compliance with the RIRT design. These nested items are a sub-category of the eleven core items included in the survey; therefore, we should expect lower prevalence estimates once we compare them to their respective core items. Figure 6 reports the proportions of both core and nested items of the second wave. We find lower prevalence estimates for three out of the fourth nested items, except for the 'Bribes' item. This evidence implies a reasonable level of diligence, comprehension, and compliance from respondents. Table 6 in the Appendix reports proportions and their respective confidence intervals for unweighted, weighted, naive and bias-corrected estimates.

Figure 6: Comparison of naive RR estimates versus nested Items



Note: This figure shows the point estimates and confidence intervals for the 'Patronage', 'Privilege', 'Forgery' and 'Bribes' items. It provides estimates both from the core and nested items. We can see that the prevalence rate of the core items is higher for three out of the fourth items. Both estimates are from the second wave, given that we only implemented this attention check technique for this wave.

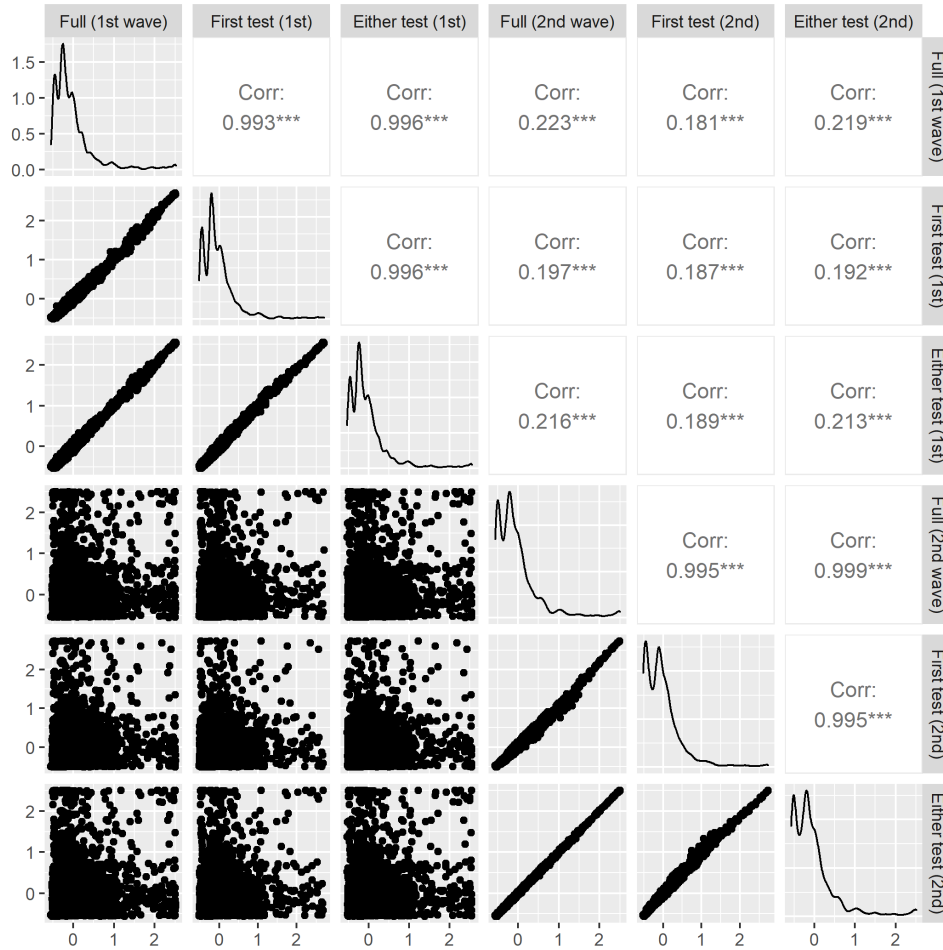
The above strategy to correct inattentiveness bias is only available for population-level estimates. Therefore, we take a different approach to examining the effect of inattentiveness on individual-level estimates. At the beginning of the survey, respondents answered two RIRT compliance test questions, which help me to examine respondents' comprehension of the RR design. If a respondent answers the first and second test questions correctly, the person does not get the third question. If a respondent failed to answer the first or second test question correctly, they had to answer the third test question. We use this design to ensure they understand the RR design well.

Utilizing this information, we conduct two additional RIRT estimations for each wave. First, we re-estimate individual-level outcomes restricted to respondents who correctly answered the initial RIRT compliance test question. In the first wave, 5,217 of 6,044 respondents passed this test, while

in the second wave, 3,263 of 3,721 did so.

Second, we also re-estimate the RIRT model for those who either answered the first question correctly or the follow-up question correctly. In other words, it includes the respondents who answered the initial question incorrectly but answered the follow-up question correctly. 5,753 respondents are in this group for the first wave, and 3,560 respondents are in this group for the second wave.

Figure 7: Comparison of RIRT estimates across subsamples



Note: This figure shows scatterplot and pairwise correlations between individual-level estimates of respondents' sensitive trait from Wave 1 and Wave 2. It breaks down this analysis based on whether respondents correctly answered the test questions conducted in one or both waves. We find a strong positive correlation between the two groups of estimates.

Figure 7 shows the density of each set of individual-level estimates, pairwise scatter plots and

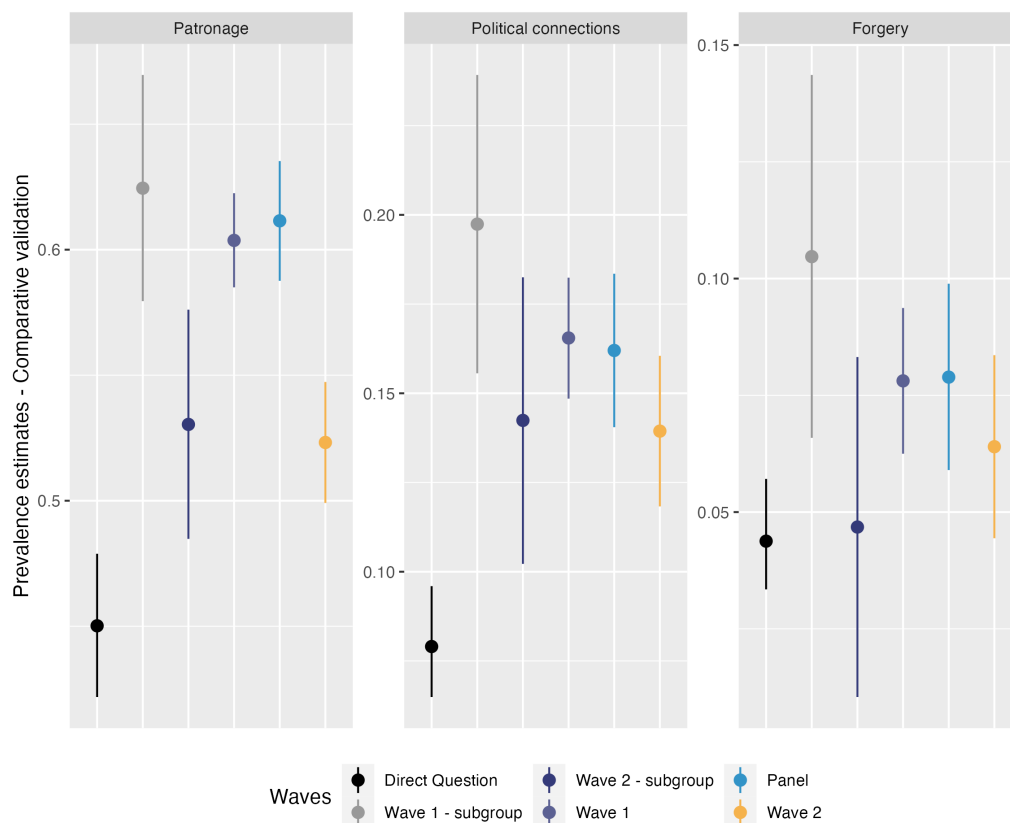
pairwise correlations among the estimates. There are several notable findings. First, there is hardly any difference between the individual-level estimates obtained from the total sample and the estimates from either sub-samples. The scatter plots show that almost all of the estimates align on a straight line, $y = x$, and the pairwise correlations are ~ 0.99 . It indicates that the individual-level estimates are reliable and are not very sensitive to respondents' compliance.

The pairwise comparisons between the two waves are also consistent. The correlations are ~ 0.2 , and the scatter plots also show consistent shape across all pairs. Furthermore, the shape of density plots is also very similar across different estimates. The estimates for the majority are pretty low, indicating that the overall prevalence rate of corrupt behaviors is low. Still, the distribution is skewed to the right, revealing a small number of cases where the estimated latent trait of corruption is high.

5.1 Comparison with direct question format from opt-in and representative samples

As pointed out, we conducted a third survey wave where 1,163 randomly selected respondents from the first wave. In this survey, respondents were asked to answer 3 out of 11 sensitive items in a direct questioning format. Figure 8 compares the results for all three sensitive items, and we can see that the RIRT yields higher prevalence levels. However, in the case of 'Forgery', I did not find a statistically significant difference between the two questioning formats only once we compared these estimates to the second wave. Still, there were substantial differences for the 'Patronage' and 'Political connections'. Given that there are changes in the sample composition between the RIRT waves and the third direct question wave, we provide RIRT estimates from only those respondents who took part in the third wave. We can see that the results still show significant differences in prevalence estimates, with the exception of 'Forgery' item.

Figure 8: Prevalence estimates RIRT vs Direct Questioning

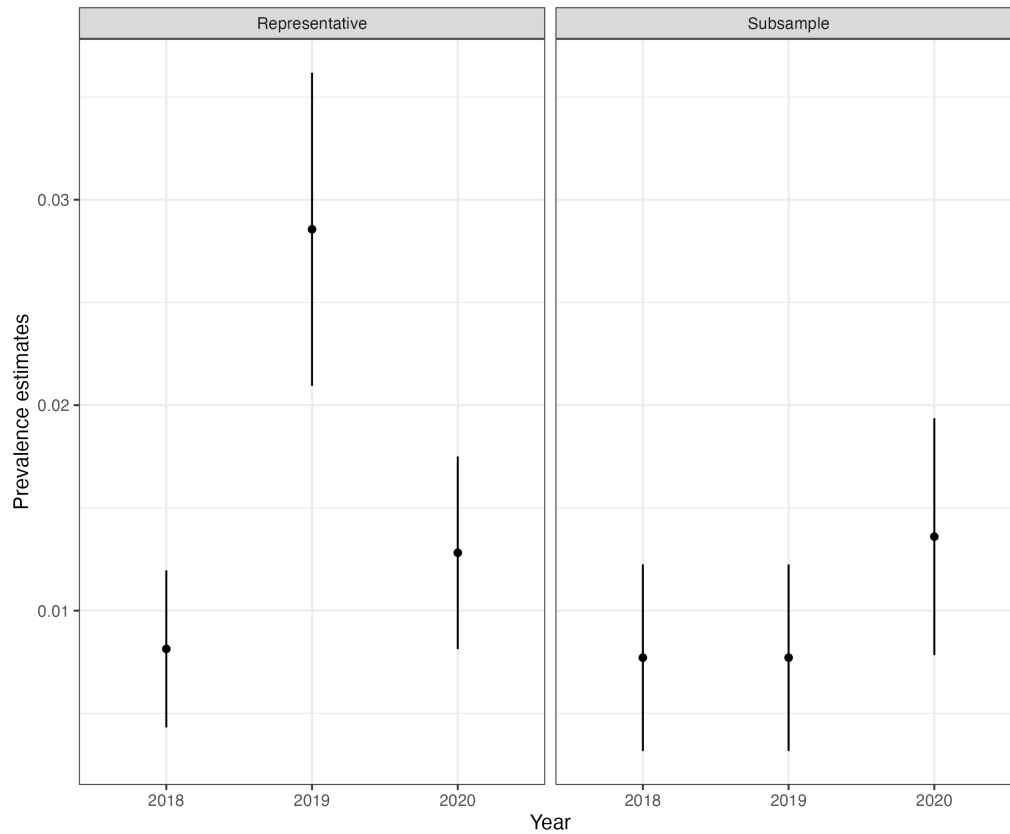


Note: This figure reports estimated prevalence rates from *Wave 1*, *Panel*, *Wave 2*, and the last wave *Wave 3*. *Wave 3* included all items asked in a direct question format. We report the direct question wave in this plot as *Direct Question*. *Wave 1* included only three items: 'Forgery', 'Patronage', and 'Political connections', which corresponded to the three items with the highest prevalence rates on *Wave 1*. All estimates reported are naive and unweighted estimates.

Prevalence estimates from representative surveys We complemented this analysis by looking at the prevalence rates of 'Bribes' from an annual survey commissioned by the *Chilean Consejo para la Transparencia* (para la Transparencia, 2018, 2019, 2020). The Consejo para la Transparencia is an independent, non-partisan governmental organisation that seeks to promote transparency in Chile. This survey is a representative survey containing the same 'Bribe' item question in the RIRT but in a direct question format. The estimates displayed in Figure 9 are significantly lower than the bribe estimates obtained using RIRT. We provide two estimates from this sample: 1) Weighted estimates from the whole sample; 2) Weighted estimates for a subset that only contains respondents from the

same communes included in the RIRT. We find that the estimated proportion of bribes since 2018 has been around 0.008 and 0.01.

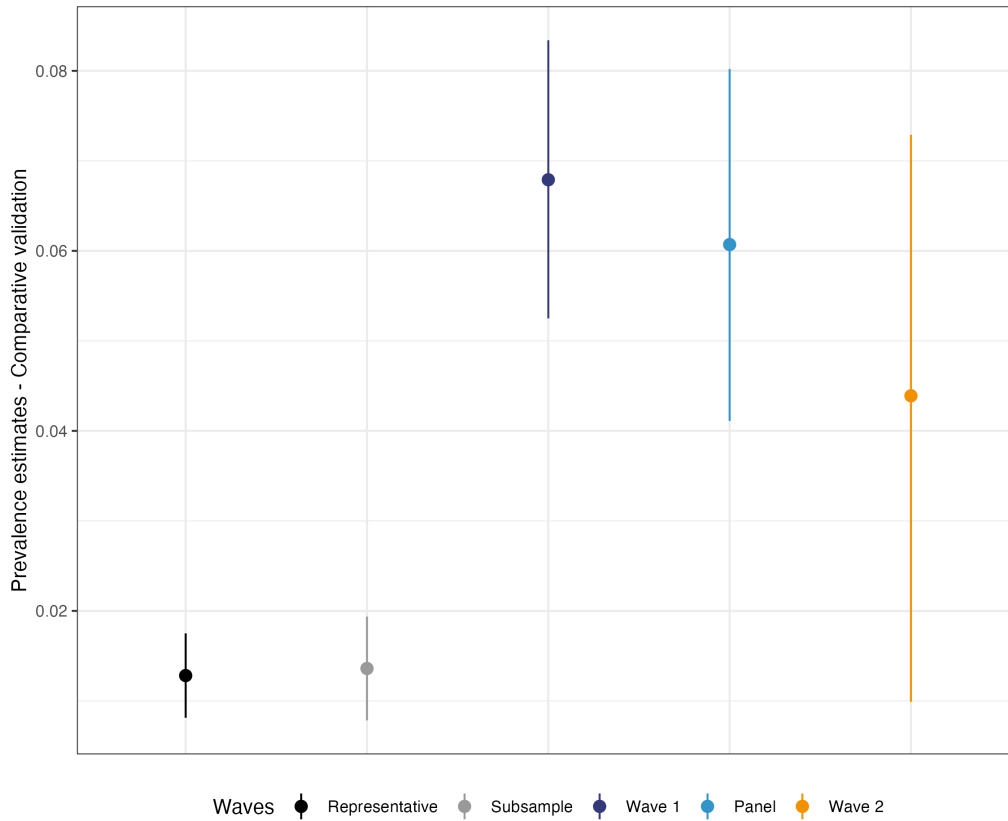
Figure 9: Prevalence estimates - Bribes item



Note: This figure shows the prevalence estimates of the 'Bribe' item from a representative survey conducted annually in Chile. I find that under 1% of respondents claimed to pay a bribe to a local official. Estimates are weighted to be nationally representative. The sample size by year: 2018 (n = 2,860), 2018 (n = 2,850) and 2020 (2,900). For the sub-sample, the sample sizes as follow 2018 (n = 2,100), 2019 (n = 1980) and 2020 (n = 2,170).

From this analysis, we find that prevalence rates from the representative survey are considerably lower than those obtained from RIRT. Figure 10 summarizes the weighted estimates from the representative sample (and subsample) and all waves of the RIRT.

Figure 10: Prevalence estimates - Bribes item RIRT and representative survey



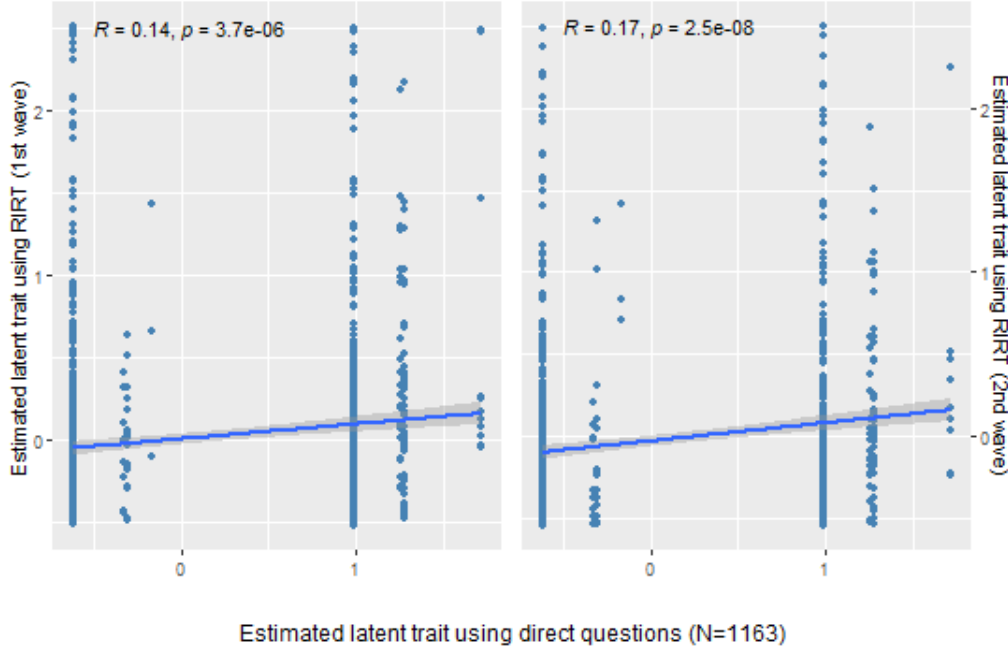
Note: This figure shows the prevalence estimates of the 'Bribe' item from a representative survey conducted annually in Chile. From this analysis, we find that under 1% of respondents claimed to pay a bribe to a local official. Estimates are weighted to be nationally representative. The sample size by year: 2018 (n = 2,860), 2018 (n = 2,850) and 2020 (2,900). For the sub-sample, the sample sizes are as follows: 2018 (n = 2,100), 2019 (n = 1980) and 2020 (n = 2,170). The estimates from the RIRT for *Wave 1* and *Panel* are weighted estimates. *Wave 2* are weighted and bias-corrected estimates.

Individual-level estimates comparison Figure 5 illustrates a comparison of individual-level RIRT estimates to IRT estimates. The IRT estimates are obtained from the three items using direct questioning. I use R's *ltm* package to run the IRT model. The x-axis of Figure 5 is the individual-level estimates of corruption behavior using direct questioning. The y-axes are the individual-level estimates of corruption behaviour using Crosswise RR questioning. The fitted line and Pearson's correlation coefficient on each plot describe the statistical association between each pair of estimates.

Due to the small number of questions, the IRT estimates' variability is limited. However, I could

still find a positive correlation between the IRT and RIRT estimates, validating the overall consistency of RIRT estimates.

Figure 11: Comparison of individual-level estimates between IRT and RIRT



Note: This figure depicts the relationship of respondents' underlying trait for *Wave 1* and *Wave 2* and subjects' underlying trait computed from the three direct questions used in the *Wave 3*. We observe a positive correlation between these two distributions. One corresponds to respondents that answered 'Yes' to at least one of the sensitive questions and '0' to those who responded 'No' to all questions.

We supplemented the previous analysis by looking at the relationship between respondents' estimated probability of engaging in the corrupt behavior and their answers (obtained in *Wave 1*) to the direct questioning format (obtained in *Wave 3*). We depict this relationship in Figure 15 and Table 9. On the x-axis of Figure 15 is the respondent's answer to the sensitive question in a direct question format, and the y-axis is the estimated probability of individuals engaging in each sensitive behavior, computed based on the respondents' underlying trait and the sensitive questions' difficulty and discrimination parameters.⁹

First, for the patronage and connection questions, the average values of these individual-level

⁹We used the following formula to calculate the estimated probabilities: $P(\tilde{Y}_{ijk} = 1 | \theta_{ij}, \alpha_k, \beta_k) = 1 - \Phi^{-1}(\alpha_k \theta_{ij} - \beta_k)$. The reason for using $1 - \Phi^{-1}$ is because we used the inverse logit function in the estimation process.

estimated probabilities are essentially the same as the population-level estimates, as depicted in Figure 2. This is expected from our modeling approach, because the individual traits are standardized to be centered around zero, equating the difficulty parameters (and their transformed value through a logistic function) to the population-level prevalence estimates. The forgery question's average (0.006) is slightly different from the population-level estimates (0.08), but both are very close to zero. When the population-level estimates are close to zero, the RIRT estimates become less reliable due to the lack of variability in the outcome variable. Therefore, we believe this is an inherent limit of the RIRT method, not necessarily a problem in our estimation process.

Second, we also compare the individuals' estimated probabilities of engaging in these behaviors to the answers to the direct questioning format. We classify the individuals into three categories, 'consistent,' 'inconsistent,' and 'inconclusive.' We define the 'consistent' answers as those whose estimated probability of engaging in the sensitive behavior is greater than or equal to 0.5 and those who answered 'yes' when asked the same question in the direct questioning format. On the other hand, the 'inconsistent' respondents are defined as either their estimated probabilities are less than 0.5 but answered 'yes' to the direct questioning format or their estimated probabilities are greater than or equal to 0.5 but answered 'no' to the direct questioning format. We classify those who chose 'prefer not to say' to the direct questioning format as 'inconclusive,' regardless of their estimated probabilities.

As can be seen in Figure 15 and Table 9 in the Appendix, we find that the two measures align well in most cases (76%: 2,662 consistent cases out of 3,489 answers; three questions per respondent). In other words, the respondents who answered 'yes' are likely to have a higher estimated probability of committing the behavior, according to our RIRT estimation. For the 'Patronage' item, we see many inconsistent classifications, mostly assigning high estimated probabilities to those who answered 'No' to the direct questioning format. This result may occur due to the low discrimination power of this item. In this context, we can also compare the item to the 'Political connections' item, in which most predictions were consistent. The highly inconsistent results for the 'Patronage' item could also

suggest that some of those who answered 'No' to the direct questioning format might have lied.

6 Discussion

The results shows that we can leverage RIRT to estimate individual-level prevalence rates of the propensity of each respondent to engage in the sensitive behavior in question. These estimations can be crucial not only to obtain prevalence rates in a particular population of interest, but they could also play an important role as an individual-level predictor whenever researchers are attention checks and investigating topics related to corruption.

Randomized Response technique and its extensions, such as RIRT, may be a promising method to minimize social desirability bias. However, it is still unclear how RIRT's additional burden on respondents may prone them to comply less with the instructions or to answer incorrectly. To address this, we provided bias-corrected estimates at the population-level and sub-sample analyses at the individual-level. From the evidence gathered of the different strategies to measure attentiveness we find encouraging, but somewhat puzzling findings. In on end, we have that around 14% of respondents failed to understand the instructions in the first round of test included before the surveys¹⁰. Meantime, we found strong evidence of high attentiveness and comprehension obtained from the bias-corrected estimates by incorporating an 'Anchor item'. In the case, of nested items, the results also signal high level of attentiveness, as we find that 3 out the 4 nested item yielded lower level of prevalence rates to their respective core items.

Looking at individual-level estimates, we also find evidence that attentiveness is high as individual-level estimates are strongly correlated to

As indicated before, it is important to include sample weights to calibrate the estimations based on the proportions of different groups or populations. I tested various weighting schemes to make accurate inferences about prevalence rates in the Chile population, and I shown that the results are

¹⁰In both waves.

robust different weighting schemes. One particular that I did not exploit in this essay is that we can use individual-level estimates of corrupt behavior to produce predictions at geographical smaller areas of interest using techniques such as Multilevel Regression and Post-Stratification.

One of most prevalent criticisms of all indirect methods techniques, is the over-reliance on what some call the "more-is-better assumption" approach. Such an approach means that whichever technique yields higher prevalence rates is closer to the actual prevalence. According to ?, most RR studies use the "the more is better" validation strategy to evaluate how the different RR variants perform. They also found that different RR variants can capture around 44% to 71% of the actual levels of prevalence of the sensitive behavior. At the same time, ? found slightly more conservative results, with barely any difference in the proportion of respondents that answered truthfully to a RR question versus a direct question format¹¹. Hence, the results of studies that use the "more is better" testing approach must be critically evaluated and taken with a grain of salt.

The results show that RIRT yields higher prevalence rates at the aggregate level than the direct questioning format, at least for the three items evaluated in the direct questioning wave (third wave). This is in line with previous findings where the RR obtains higher prevalence rates once compared to the direct questioning format. Although these results appear promising, I do not make strong claims on whether the RIRT yields estimates closer to the unobserved true prevalence rate.

Once we examine prevalence rates from the RIRT to rates obtained from a representative sample survey, I also find significant differences between these questioning formats. It is important to stress that several factors may explain this difference—for example, different sample composition to weighting schemes. Furthermore, I compare one item, which limits the possibility of making claims on whether they are systematic differences between these survey formats.

Finally, once we compare the propensity to engage in sensitive behavior to respondents' answers from a direct questioning format survey, I find some evidence that RIRT accurately predicts their

¹¹It is important to state that both studies relied on official records at the individual level to estimate the true prevalence, making their validation approach very robust.

self-reported behavior. We observe that RIRT yields a higher probability to those respondents that answered 'Yes' to the items. However, the distribution of the probability values for respondents who answered 'No' is puzzling. RIRT tends to cluster the probabilities in very high or very low values of the probability distribution for the 'Political connections' and 'Forgery' items. Furthermore, it yields high probabilities values near one for a considerable number of respondents, which raises questions about the accuracy of this technique.

In summary, the findings of this study support the claim that attentiveness and compliance are high for this crosswise design. Furthermore, it shows that prevalence rates using indirect questioning techniques tend to be higher than estimates obtained from direct survey surveys. However, more research is needed to examine the accuracy and reliability of individual-level estimates, specifically in validation studies where prevalence estimates are compared to observed behavior or official records.

References

- Atsusaka, Yuki and Randolph T. Stevenson. 2021. “A Bias-Corrected Estimator for the Crosswise Model with Inattentive Respondents.” *Political Analysis* pp. 1–15.
- Baker, Frank B. and Seock-Ho Kim. 2017. *Test Calibration*. Cham: Springer International Publishing pp. 105–125.
URL: https://doi.org/10.1007/978-3-319-54205-8_7
- Ben, Jann, Julia Jerke and Ivan Krumpal. 2012. “Asking Sensitive Questions using the Crosswise Model: An Experimental Survey Measuring Plagiarism.” *American Association for Public Opinion Research* 76(1):32–49.
- Blair, Graeme, Kosuke Imai and Jason Lyall. 2014. “Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan.” *American Journal of Political Science* 58(4):1043–1063.
URL: <http://www.jstor.org/stable/24363542>
- Blair, Graeme, Kosuke Imai and Yang-Yang Zhou. 2015. “Design and Analysis of the Randomized Response Technique.” *Journal of the American Statistical Association* 110(511):1304–1319.
URL: <https://doi.org/10.1080/01621459.2015.1050028>
- Bourke, Patrick D. and Michael A. Moran. 1988. “Estimating Proportions from Randomized Response Data Using the EM Algorithm.” *Journal of the American Statistical Association* 83(404):964–968.
URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478685>
- de la Transparencia, Consejo. 2019. “Estudio Nacional de Funcionarios Públicos.”.
- Fergusson, Leopoldo, Carlos Molina and Juan Riano. 2017. “I Sell My Vote and So What? A New Database and Evidence from Colombia.”.
- Fornasari, Margherita, Christian Schuster, Kim Sass Mikkelsen, Kerenssa Kay, Kay Mukhtarova, Roger Daniel and Zahid Hasnain. 2022. “Global Survey of Public Servants Remote Work Surveys Data Set.”.
- Fox, J.-P. and Cheryl Wyrick. 2008. “A Mixed Effects Randomized Item Response Model.” *Journal of Educational and Behavioral Statistics* 33(4):389–415.
URL: <https://doi.org/10.3102/1076998607306451>
- Fox, J.-P., M. Avetisyan and J. van der Palen. 2013. “Mixture randomized item-response modeling: a smoking behavior validation study.” *Statistics in Medicine* 32(27):4821–4837.
URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5859>
- Fox, Jean-Paul. 2005. “Randomized Item Response Theory Models.” *Journal of Educational and Behavioral Statistics* 30(2):189–212.

- Fox, Jean-Paul, Duco Veen and Konrad Klotzke. 2019. "Generalized linear mixed models for randomized responses." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 15.
- Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks: Sage.
- Fox, John. 2002. *An R and S-Plus Companion to Applied Regression*. Thousand Oaks: Sage Publications.
- Hambleton, Ronald K. and Hariharan Swaminathan. 1985. *Item Banking*. Dordrecht: Springer Netherlands pp. 255–279.
URL: https://doi.org/10.1007/978-94-017-1988-9_2
- Heck, Daniel W. and Morten Moshagen. 2018. "RRreg: An R Package for Correlation and Regression Analyses of Randomized Response Data." *Journal of Statistical Software* 85(2):1–29.
URL: <https://www.jstatsoft.org/index.php/jss/article/view/v085i02>
- Hoffman, Adrian, Birk Diedenhofen, Bruno Verschuere and Jochen Musch. 2015. "A Strong Validation of the Crosswise Model Using Experimentally-Induced Cheating Behavior." *Experimental Psychology* 62:403–414.
- Hoffmann, Adrian and Jochen Musch. 2016. "Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model." *Behavior Research Methods* 48(3):1032–1046.
URL: <https://doi.org/10.3758/s13428-015-0628-6>
- Höglinger, M., B. Jann and A Diekmann. 2016. "Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model." *Survey Research Methods* 10(3):171–187.
- Höglinger, Marc and Andreas Diekmann. 2017. "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT." *Political Analysis* 25(1):131–137.
- Horvitz, D. G. and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association* 47(260):663–685.
- Hsieh, Shu-Hui and Pier Francesco Perri. 2020. "A Logistic Regression Extension for the Randomized Response Simple and Crossed Models: Theoretical Results and Empirical Evidence." *Sociological Methods & Research* 0(0).
URL: <https://doi.org/10.1177/0049124120914950>
- Ibbett, Harriet, Julia P.G. Jones and Freya A.V. St John. 2021. "Asking sensitive questions in conservation using Randomised Response Techniques." *Biological Conservation* 260:109191.

- Jerke, Julia, David Johann, Heiko Rauhut, Kathrin Thomas and Antonia Velicu. 2021. “Handle with Care: Implementation of the List Experiment and Crosswise Model in a Large-Scale Survey on Academic Misconduct.”.
- John, Leslie K., George Loewenstein, Alessandro Acquisti and Joachim Vosgerau. 2018. “When and why randomized response techniques (fail to) elicit the truth.” *Organizational Behavior and Human Decision Processes* 148:101–123.
- Korndörfer, Martin, Ivar Krumpal and Stefan C. Schmukle. 2014. “Measuring and explaining tax evasion: Improving self-reports using the crosswise model.” *Journal of Economic Psychology* 45:18–32.
URL: <https://www.sciencedirect.com/science/article/pii/S0167487014000609>
- LAPOP. 2017. “Latin American Public Opinion Project.”.
- LAPOP. 2019. “Latin American Public Opinion Project.”.
- LAPOP. 2021. “Latin American Public Opinion Project.”.
- Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden and Cora J. M. Maas. 2005. “Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation.” *Sociological Methods & Research* 33(3):319–348.
URL: <https://doi.org/10.1177/0049124104268664>
- Liu, Haiyan and Zhiyong Zhang. 2017a. “Logistic regression with misclassification in binary outcome variables: a method and software.” *Behaviormetrika* 44(2):447–476.
URL: <https://doi.org/10.1007/s41237-017-0031-y>
- Liu, Haiyan and Zhiyong Zhang. 2017b. “Logistic regression with misclassification in binary outcome variables: a method and software.” *Behaviormetrika* 44(2):447–476.
URL: <https://doi.org/10.1007/s41237-017-0031-y>
- para la Transparencia, Consejo Nacional. 2018. “Estudio Nacional de Transparencia.”.
- para la Transparencia, Consejo Nacional. 2019. “Estudio Nacional de Transparencia.”.
- para la Transparencia, Consejo Nacional. 2020. “Estudio Nacional de Transparencia.”.
- Rosenfeld, Bryn, Kosuke Imai and Jacob N. Shapiro. 2016. “An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions.” *American Journal of Political Science* 60(3):783–802.
- Sagoe, Dominic, Maarten Cruyff, Owen Spendiff, Razieh Chegeni, Olivier de Hon, Martial Saugy, Peter G. M. van der Heijden and Andrea Petróczi. 2021. “Functionality of the Crosswise Model for Assessing Sensitive or Transgressive Behavior: A Systematic Review and Meta-Analysis.” *Frontiers in Psychology* 12:2264.
URL: <https://www.frontiersin.org/article/10.3389/fpsyg.2021.655592>

- Singh, Shane P. and Jaroslav Tir. 2022. "Threat-Inducing Violent Events Exacerbate Social Desirability Bias in Survey Responses." *American Journal of Political Science* n/a(n/a).
URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12615>
- Smallpage, Steven M., Adam M. Enders, Hugo Drochon and Joseph E. Uscinski. 2022. "The impact of social desirability bias on conspiracy belief measurement across cultures." *Political Science Research and Methods* 1(15):1–15.
- Transparente, Chile. 2019. Encuesta Nacional de Integridad. Technical report Chile Transparente.
- van den Hout, Ardo, Peter G.M. van der Heijden and Robert Gilchrist. 2007. "The logistic regression model with response variables subject to randomized response." *Computational Statistics Data Analysis* 51(12):6060–6069.
URL: <https://www.sciencedirect.com/science/article/pii/S0167947306004774>
- Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60(309):63–69. PMID: 12261830.
URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1965.10480775>
- Wu, Ahra, R. Andrew Wood and Randolph T Stevenson. 2019. "A Validation Study of Individual-Level Survey Methodologies for Sensitive Questions."
- Yang, Frances M., Richard N. Jones, Sharon K. Inouye, Douglas Tommet, Paul K. Crane, James L. Rudolph, Long H. Ngo and Edward R. Marcantonio. 2013. "Selecting optimal screening items for delirium: an application of item response theory." *BMC Medical Research Methodology* 13(1):8.
URL: <https://doi.org/10.1186/1471-2288-13-8>
- Yu, Jun-Wu, Guo-Liang Tian and Man-Lai Tang. 2008. "Two new models for survey sampling with sensitive characteristic: design and analysis." *Metrika* 67(3):251–263.
URL: <https://doi.org/10.1007/s00184-007-0131-x>
- Yusaku Horiuchi Horiuchi, Yusaku, Zachary Markovich and Teppei Yamamoto. 2021. "Does Conjoint Analysis Mitigate Social Desirability Bias?" *Political Analysis* 1(15).

Appendix

Descriptive

Table 3 reports the means of relevant variables for all three waves. I observe that, across all waves, most point estimates remain similar, except income, which drops in the direct questioning wave (DC).

Table 3: Means - all waves

	Survey Wave			
	<i>Wave 1</i>	<i>Panel</i>	<i>Wave 2</i>	<i>Wave 3</i>
	<i>RIRT</i>	<i>RIRT</i>	<i>RIRT</i>	<i>DC</i>
Covariates				
Female	65%	66%	66%	67%
Education (Yrs)	15	15	15	15.2
Age	35.9	35.7	35.7	35.9
Partisanship	4.48	4.46	4.46	4.37
Income (USD)	529	506	506	470
Sample	6050	3723	3723	1163

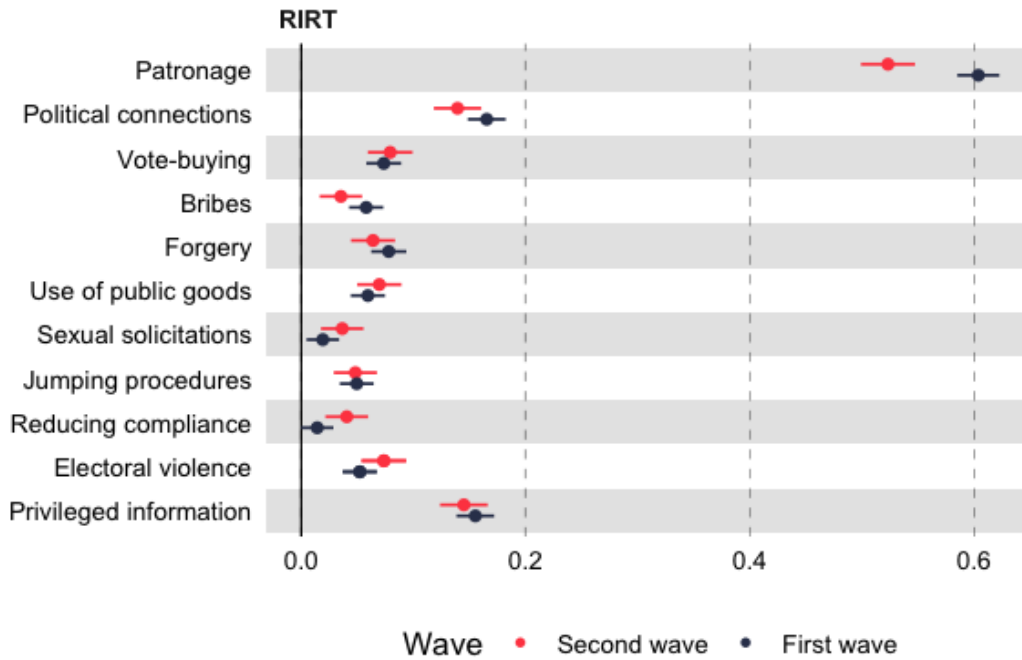
Prevalence estimates

Table 4: Prevalence estimates Waves 1 and 2 and Panel

Items	Wave 1	Panel	Wave 2
Patronage	0.59 (0.010)	0.61 (0.012)	0.52 (0.012)
Political connections	0.18 (0.009)	0.16 (0.011)	0.14 (0.011)
Vote-buying	0.07 (0.008)	0.07 (0.010)	0.08 (0.010)
Bribes	0.07 (0.008)	0.05 (0.010)	0.04 (0.010)
Forgery	0.09 (0.008)	0.08 (0.010)	0.06 (0.010)
Personal	0.05 (0.008)	0.06 (0.010)	0.07 (0.010)
Sexual solicitation	0.02 (0.007)	0.03 (0.010)	0.04 (0.010)
Lobbying	0.05 (0.008)	0.05 (0.010)	0.05 (0.010)
Compliance	0.03 (0.007)	0.01 (0.009)	0.04 (0.010)
Electoral Violence	0.06 (0.008)	0.05 (0.010)	0.07 (0.010)
Privilege	0.15 (0.009)	0.16 (0.011)	0.15 (0.011)

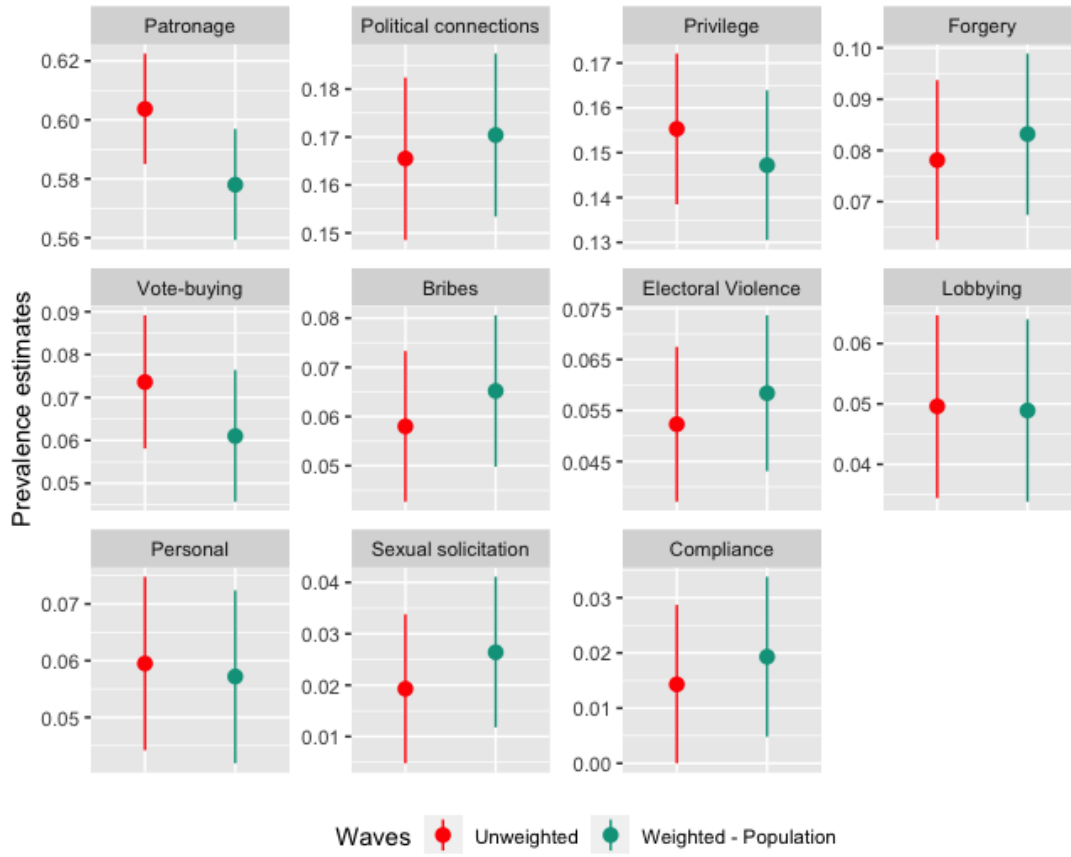
Note: This table reports naive prevalence estimates using Inverse Probability Weights based on population size.

Figure 12: Raw prevalence estimates using uni-variate estimation



Weighting Given that respondents were randomly selected from an unrepresentative opt-in sample, we included sample weights in our estimations using different sampling weights and weighting techniques. Figure 13 provides weighted prevalence estimates once we incorporate population weights, in addition to other demographic parameters such as age, gender, and education. Surprisingly, we observed six items yield higher prevalence estimates once compared to the naive unweighted estimates.

Figure 13: Weighted prevalence estimates



Note: This figure summarizes Naïve prevalence rates from unweighted and weighted estimates. Weighted estimates are weighted using Inverse Probability Weights based on population. We report these estimates for all 11 items. We can identify that prevalence rates somewhat changed once we introduced weights. We observe that most prevalence estimates increased. For example, 'Political connections', 'Bribes', 'Sexual solicitation', 'Forgery' and 'Compliance'.

Table 5: Proportion of individual estimates

Wave	< 0.5 σ	$\geq 0.5\sigma$
Wave 1	0.962	0.038
Panel	-	-
Wave 2	0.923	0.077

Table 6: Difference between core and nested items

Items	Core			Nested			Difference
	Prevalence	Lower C.I	Upper C.I	Prevalence	Lower C.I	Upper C.I	
Unweighted - Naive							
Patronage	0.523	0.499	0.547	0.173	0.151	0.195	0.350***
Bribes	0.035	0.016	0.054	0.049	0.029	0.068	-0.013
Forgery	0.064	0.044	0.084	0.039	0.020	0.059	0.025
Privilege	0.145	0.124	0.166	0.106	0.086	0.127	0.039
Weighted - Bias-corrected							
Patronage	0.512	0.478	0.545	0.167	0.137	0.196	0.345
Forgery	0.070	0.040	0.103	0.035	0.004	0.062	0.034
Privilege	0.144	0.114	0.173	0.103	0.070	0.134	0.041
Bribes	0.044	0.010	0.073	0.039	0.008	0.064	0.005

Note: This table summarizes the prevalence rates for all four core and nested items included in Wave 2. We report the difference between the prevalence rates from core minus nested items. The first set of differences is between unweighted and naive. The second set of differences is between weighted and bias-corrected estimates. We only find differences for the 'Patronage' item *** $p < 0.001$.

Extension on population estimation strategies

In the following section, I provide more details on population-level estimates. I largely borrowed from (Heck and Moshagen, 2018).

For the logistic regression, the expected prevalence of the sensitive behavior is given by the following expression :

$$\pi_i = \text{logit}^{-1}(\mathbf{X}_i\beta) = \frac{1}{1 + \exp(-\mathbf{X}_i\beta)} \quad (27)$$

In this expression, β is the vector of regression coefficients and \mathbf{X}_i is the i -th row of the design matrix \mathbf{X} . In the case of RR designs, the responses are binomially distributed with means given by the expected probabilities λ . λ is equal to the product of a misclassification matrix \mathbf{P} and the π matrix that contains the true probabilities of the sensitive behavior. The entries of P_{ij} of the \mathbf{P} matrix are set by the conditional probabilities to response i (i.e. *Yes* or *no*) to the sensitive item given the true state of a respondent is j (i.e. *Engaged in corrupt behavior* or *Did not engage in corrupt behavior*). In the general case, the likelihood function is the following:

$$f(\mathbf{y}|\beta) = \prod_{i=1}^N \binom{n_i}{y_i} (P_{10}(1 - \pi_i) + P_{11}\pi_i)^{y_i} (P_{00}(1 - \pi_i) + P_{01}\pi_i)^{n_i - y_i} \quad (28)$$

In equation 28, y_i is the number of affirmative responses to the sensitive question, and n_i is the total responses. $P_{01} \dots P_{11}$ are the entries of the misclassification matrix where P_{00} is the conditional probability of answering 'no' to the sensitive question, dependent on that the respondent has not engaged in corrupt behavior. Conversely, P_{11} is the conditional probability of answering affirmatively, conditional on whether the respondent has engaged in corrupt conduct.

Using an expectation-maximization (EM) algorithm, we can retrieve the maximum likelihood estimate of the parameters of interest. This algorithm serves cases with an incomplete-data problem. The EM algorithm works as follows: 1) define the distribution of the Z unobservable variable. 2)

estimate the maximum likelihood with the complete data $L_c(\theta|Y, Z)$ as both X and Y variables were observed. 3) The third step is to set a starting or initial value for the parameter to be estimated θ_0 4) estimate the expectation of the likelihood function.

$$Q(\theta, \theta_0) = E[\ln L_c(\theta|Y, Z, \theta_0)] \quad (29)$$

This algorithm proceeds with the maximization step after obtaining the expectation of the starting value and calculating the current conditional value of the parameter of interest. This is stated formally in equation 30:

$$\max_{\theta} Q(\theta, \theta_1) \quad (30)$$

This algorithm continues in this iterative process of calculating the expected value for a given value of θ and find the value that maximizes the value of this parameter until it converges.

We can extend the modified univariate logistic regression to include predictors of experiencing the sensitive behavior. Ben, Jerke and Krumpal (2012) and van den Hout, van der Heijden and Gilchrist (2007), Hsieh and Perri (2020) formalize this modified version, where $\pi = P(X = 1|Z)$ is the unknown probability of answering 'yes' to the sensitive item, conditional on covariates Z .

Linear estimation. The linear estimation approach for RR can be described as a misclassification scheme that adds random noise to the true responses Heck and Moshagen (2018). The likelihood function of this linear RR regression model contains two elements. 1) The likelihood of the unobservable true states is a function of the mixture distribution of the observed responses and the missclassification matrix \mathbf{P} . 2) n_j^* are the observed frequencies of the $j = 1, \dots, J$ possible RR responses. We can express this likelihood function as follows:

$$\log f_1(\mathbf{y}|\boldsymbol{\pi}) = \sum_{k=1}^J n_j^* \sum_{k=1}^J P_{jk} \pi_k + C \quad (31)$$

The second element is the dependent variable \mathbf{y} (the number of affirmative responses to the sensitive question), which is conditionally normally distributed with standard deviation σ . Then, the expected value is a function of the *true*, *latent* states on the predictors and the observed value for the observed variables. As we do not observe the *true* states, the marginal likelihood of the regression coefficient β has to be computed to reduce the uncertainty induced by the misclassification matrix.

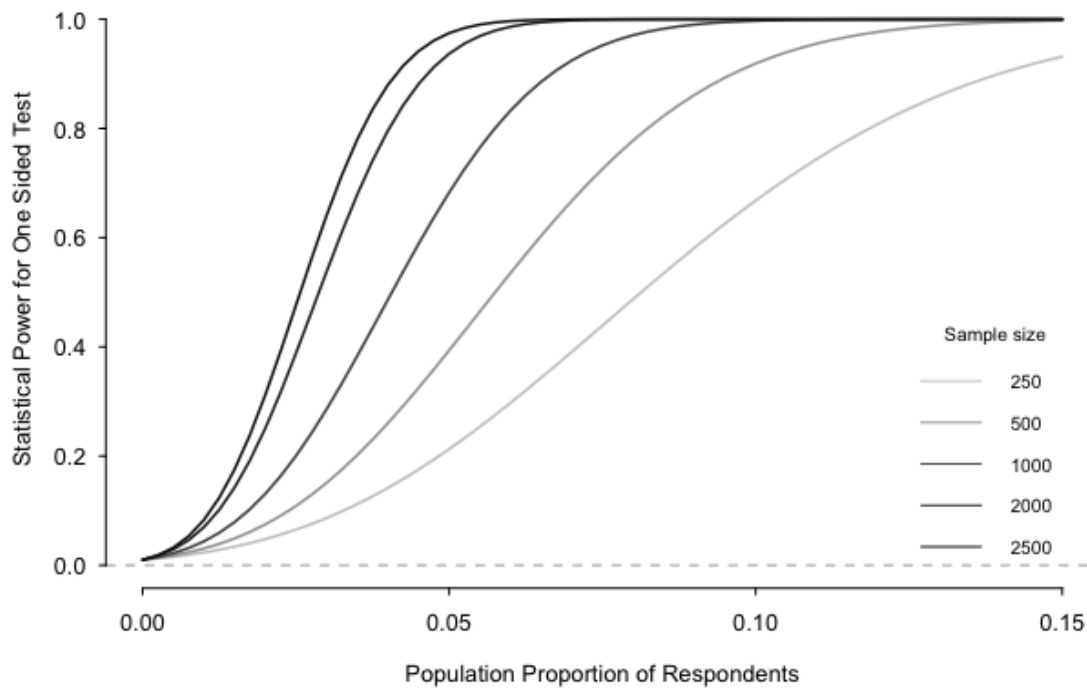
$$\log f_2(\mathbf{y}|\beta, \sigma, \boldsymbol{\pi}) = \sum_{i=1}^N \log \left[\sum_{j=1}^J \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mathbf{X}_i^j}{\sigma}\right)^2\right) \frac{P_{(i)j} \pi_j}{\sum_{k=1}^J P_{(i)k} \pi_k} \right] \quad (32)$$

Combining equations 31 and 32, we can obtain the joint loglikelihood of the whole model. By maximizing this function, we obtain the maximum likelihood of the parameter of interest.

Power Calculations

Power calculations were conducted previous to this study. For population prevalence estimates above 2-3%, all the RIRT waves (first and second waves) are well-powered above 80%. For low-prevalence items, such as sexual solicitations, both the first and second waves would be under-powered.

Figure 14: Power calculations Randomised Response



Research questions and hypotheses

Research questions As we mentioned, researchers should seek to minimize the effects of social desirability bias. Tackling this issue is particularly important for studies that rely on survey-based measurements of a sensitive issue or behavior. Thus, one of the goals of this study is to measure the prevalence of corrupt behavior at the population level using RIRT.

- *RQ 1a: What the prevalence estimates of corrupt behavior at the population level?*

Along with providing population-level estimates of the sensitive trait, we can leverage that RIRT yields individual-level estimates of the sensitive behavior and use these estimates to make predictions of corruption behaviors in small areas. The second research question addresses this point:

- *RQ 1b: What the prevalence estimates of corrupt behavior at the municipal level?*

Most of the literature on RR design has focused on validating this method against other questioning formats. However, no study has investigated whether this technique provides robust estimates. Thus, this question seeks to explore further on this area:

- *RQ 2: Does the RR method provide robust estimates of the prevalence of the sensitive behavior and subjects' underlying trait?*

One major drawback of RR designs is the additional cognitive burden and response times on respondents, which may reduce compliance and attentiveness (Atsusaka and Stevenson, 2021). Thus, the contribution of this paper is to provide new empirical evidence on the magnitude of this bias and to provide bias-corrected estimates of sensitive behavior.

- *RQ 3: Does bias due to inattention lead to significantly larger estimates of the sensitive behaviour?*

Hypotheses Based on recent survey results (LAPOP, 2021, 2019, 2017, para la Transparencia, 2020, 2019, 2018), we should expect to obtain positive prevalence estimates of corrupt behaviors in Chile. We do not have a strong prior about the magnitude of this prevalence estimates, as these estimates vary from one survey to another. Furthermore, there are corrupt behaviors included in this study that have not been measured before¹².

- *H1: RIRT will provide positive prevalence estimates of corrupt behaviors.*

It is reasonable to expect that the lack of comprehension of the RR questioning procedures would yield biased estimates of the prevalence of the behavior of interest (?). According to Atsusaka and Stevenson (2021), naive group-level prevalence estimates that ignore respondents' inattentiveness always overestimate the population prevalence under certain regularity conditions. The conditions are that the population prevalence (π) and the randomization probability (p) are less than 0.5. Because my analysis shows these conditions met, we explore if naive RR estimates are significantly larger than the bias-corrected estimates at the population level—the inattention bias increases as the per cent of inattentive survey respondents increases.

- *H2: Naive estimates of the sensitive behaviors would yield higher prevalence estimates than bias-corrected estimates with inattentive respondents.*

As pointed out before, no studies focus on whether this technique yields robust estimates of the prevalence of the sensitive behaviors in question. However, there is also no systematic evidence or theoretical grounds that would suggest that this method would yield noisy estimates of the behaviors of interest. Thus, my starting point is that this technique should produce robust estimates of the corrupt behaviors at the population and individual levels.

- *H3: RIRT will provide robust estimates of the prevalence of corrupt behaviors.*

¹²In the final version of this survey, we will report the prevalence estimates obtained in other surveys

Table 7: Difference in prevalence weighted naive estimates and their respective p-values

Items	Wave 1 - Panel		Panel - Wave 2		Wave 1 - Wave 2	
	Diff	p-value	Diff	p-value	Diff	p-value
Patronage	-0.011	1.730	0.087	0.000	0.075	0.000
Political connections	-0.002	1.207	0.026	0.002	0.024	0.002
Vote-buying	-0.002	1.253	-0.008	1.818	-0.010	1.931
Bribes	0.007	0.162	0.013	0.011	0.021	0.000
Forgery	0.002	0.783	0.011	0.078	0.013	0.027
Personal	-0.003	1.460	-0.024	2.000	-0.027	2.000
Sexual solicitation	-0.010	1.997	-0.004	1.630	-0.014	2.000
Lobbying	-0.009	1.943	0.014	0.009	0.005	0.308
Compliance	0.000	0.928	-0.014	1.999	-0.014	2.000
Electoral Violence	-0.003	1.403	-0.020	1.999	-0.022	2.000
Privilege	-0.002	1.210	0.008	0.359	0.006	0.452

Note: This table reports the difference in proportions for each of the 11 core items. We only report differences using weighted naive estimates. The three differences documented are: 1) between *Wave 1* and *Panel*, 2) between *Panel* and *Wave 2*, and 3) between *Wave 1* and *Wave 2*.

Randomized response estimate derivation

Owing to a coding quirk in the JavaScript, the probabilities of virtual die rolls were not evenly distributed. The table below shows the probabilities assigned to each die roll outcome. p denotes the probability of the sensitive trait.

Table 8: Probabilities of virtual die rolls

die roll	die probability	selected answer	probability of answering given the die roll
1	0.1	both or neither	$0.1 * p + (1 - 0.1) * (1 - p)$
		either	$0.1 * (1 - p) + (1 - 0.1) * p$
2	0.2	both or neither	$0.2 * p + (1 - 0.2) * (1 - p)$
		either	$0.2 * (1 - p) + (1 - 0.2) * p$
3	0.2	both or neither	$0.2 * p + (1 - 0.2) * (1 - p)$
		either	$0.2 * (1 - p) + (1 - 0.2) * p$
4	0.2	both or neither	$0.2 * p + (1 - 0.2) * (1 - p)$
		either	$0.2 * (1 - p) + (1 - 0.2) * p$
5	0.2	both or neither	$0.2 * p + (1 - 0.2) * (1 - p)$
		either	$0.2 * (1 - p) + (1 - 0.2) * p$
6	0.1	both or neither	$0.1 * p + (1 - 0.1) * (1 - p)$
		either	$0.1 * (1 - p) + (1 - 0.1) * p$

Expectation

Let λ denote the probability of answering "both or neither of the statements are true." Given the above die roll probabilities, λ is defined as below.

$$\begin{aligned}
 \lambda &= 2 * 0.1 * (0.1 * p + (1 - 0.1 * p)) + 4 * 0.2 * (0.2 * p + (1 - 0.2 * p)) \\
 &= \frac{41}{50} - \frac{16}{25} * p
 \end{aligned} \tag{33}$$

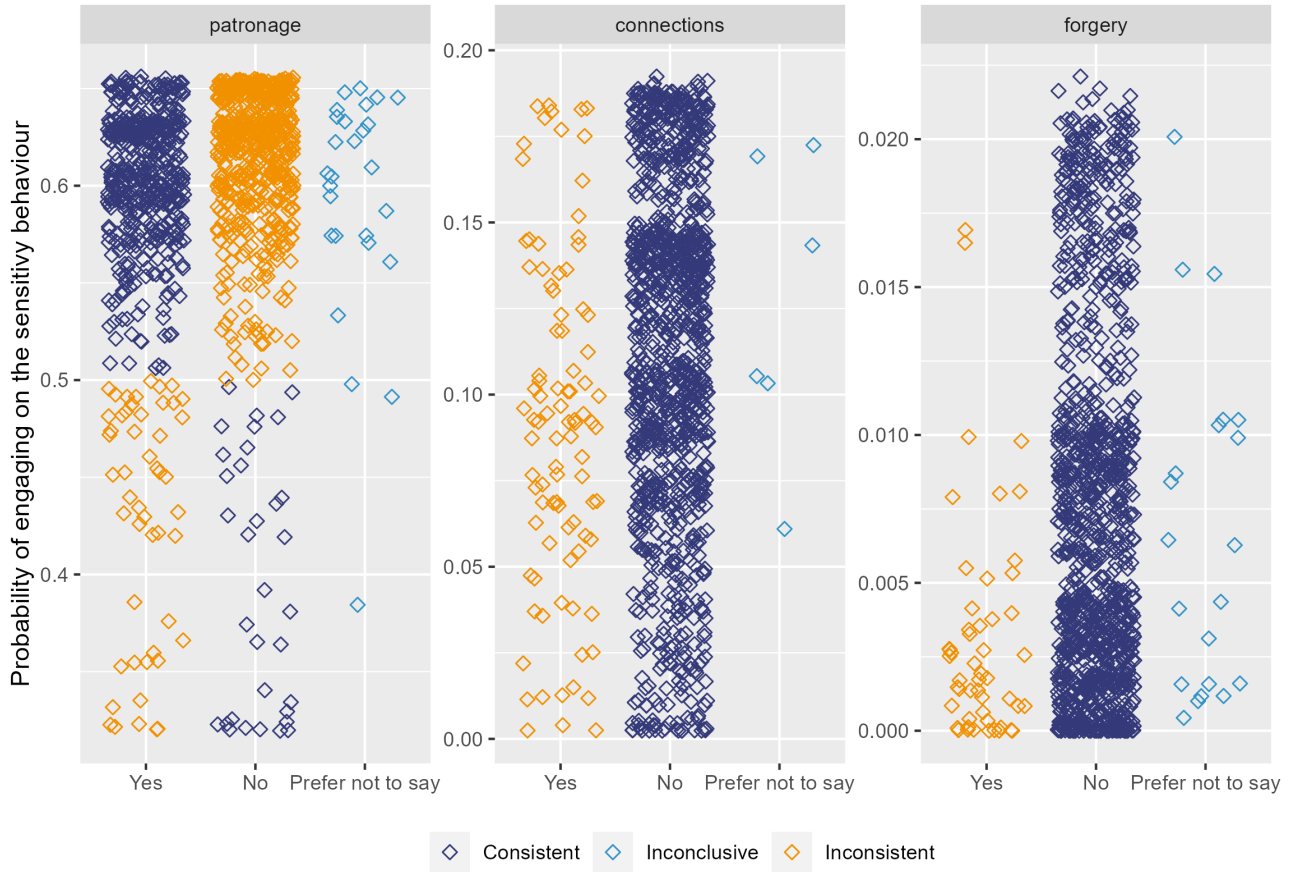
Rearranging the terms, the probability of a sensitive trait is:

$$\begin{aligned}
 p &= 2 * 0.1 * (0.1 * p + (1 - 0.1 * p)) + 4 * 0.2 * (0.2 * p + (1 - 0.2 * p)) \\
 &= \frac{41}{50} - \frac{16}{25} * p
 \end{aligned} \tag{34}$$

λ follows the Bernoulli distribution with a parameter. The expectation of λ is given by

$$\begin{aligned}
E(\lambda) &= E\left(\frac{41}{50} - \frac{16}{25} * p\right) \\
&= \frac{41}{50} - \frac{16}{25} * p
\end{aligned} \tag{35}$$

Figure 15: Comparison of direct question and probability of engaging in the sensitive behavior



Note: This figure shows the distribution of the estimated probabilities of engaging in sensitive behavior for the 'Patronage', 'Political connections', and 'Forgery' items. On the x-axis is whether respondents answered 'Yes', 'No', or 'Prefer not to say' to the same question, but in a direct question format in *Wave 3*. The probability of engaging in the sensitive behavior was computed using respondents' underlying traits, discrimination, and difficulty parameters.

Table 9: Prevalence estimates - Naive, Weighted and Bias-Corrected

Question	Consistency	N
Patronage	Consistent	506
	Inconclusive	27
	Inconsistent	630
Connections	Consistent	1065
	Inconclusive	6
	Inconsistent	92
Forgery	Consistent	1091
	Inconclusive	21
	Inconsistent	51