Corrupted Estimates? Response Bias in Citizen Surveys on Corruption

Supplementary Materials

Mattias Agerberg*

May 29, 2020

*Department of Political Science, University of Gothenburg, Sweden.
Contact: mattias.agerberg@gu.se.

Appendix

Outcome questions for the PB experiment

I asked three different outcome questions to test the PB hypothesis. In addition I also included a standard question on economic perceptions as a benchmark. First, I adopted the following question (*corruption increase*), used for instance in the Global Corruption Barometer: *In your opinion, over the last year, has the level of corruption in this country increased, decreased, or stayed the same?* The respondent was given five answer alternatives ranging from 'increased a lot' to 'decreased a lot'. Second, I asked a commonly used question about the absolute level of political corruption (*corruption in politics*): *In your opinion, about how many politicians in Romania are involved in corruption?* The question has five answer alternatives ranging from 'almost none' to 'almost all'. This question is, for example, asked in several waves of the ISSP survey. Third, I asked how worried respondents are about the consequences of corruption (*corruption worry*): *In general, how worried are you about the consequences of corruption for the Romanian society?* This is a question similar to the questions asked in Peiffer (2018). The question taps into how concerned a respondent is about the consequences of corruption, and hence also how important the issue of corruption is for the respondent. Four possible answer alternatives were given to the question: 'not worried at all', 'a little worried', 'somewhat worried', 'very worried'. Finally, as a point of comparison, I also included a standard question about economic perceptions (*economy worse*) (see Evans and Andersen (2006)): *In your opinion, over the last year, would you say that Romania's economy has got stronger, weaker, or stayed the same?* The five answer alternatives range from 'got a lot weaker' to 'got a lot stronger'. With this design I am able to compare the treatment effect on the corruption questions with the (well-established; see the review above) political bias-effect on the economy question. All outcome questions were coded so that high values indicate 'bad' outcomes; increased corruption, worsened economy, high political corruption, and high worry about corruption.

Survey details and descriptive statistics

The final survey was reviewed and translated to Romanian by the professional translation company *Language Connect*[1]. The sample for the study was recruited by Lucid, an online marketplace for survey respondents. Different providers redirect participants to Lucid, which then redirects subjects to purchasers (typically market research firms or, increasingly, academic researchers). Survey takers are compensated in cash, gift cards, or reward points. Based on background information on subjects in the marketplace, Lucid then constructs a demographically targeted sample using a combination of quota sampling and screening questions. For the study at hand, nationally representative quotas on gender, age, and region were used. The age quota had to be relaxed somewhat to reach the target number of completes. Descriptive statistics for the sample used in the study are presented in Table 3.

Table 3: Sample characteristics

Statistic	Mean	St. Dev.	Min	Max	N
Female respondent	0.50	0.50	0.00	1.00	3,016
City with over 200,000 inhabitants	0.48	0.50	0	1	3,027
University education	0.56	0.50	0	1	3,027
Age	35.60	11.75	15	110	3,027
Household income (Lei per month)	4,500	4,340	0.00	726,468.00	2,941
Persons in household	4.16	1.26	1	7	3,027

Note: Some extreme (probably miscoded) outliers were excluded from the 'Income' variable. For the Income variable, the median household income is reported, and the interquartile range (IQR) for the variable is reported in the 'St. Dev.' column.

The sample shows good representativeness with regard to the quota variables. Gender is balanced in the sample. The share of urban respondents (0.48) matches well with the population share of people living in cities (0.54 according to statistics aggregated by Worldometer[2]). As a consequence of the fact that the age quota had to be relaxed the sample is slightly younger than the Romanian population (35.6 vs 41.6 years). The sample also shows relatively good representativeness for household income, which varies around 4000 Lei in the population. As is often the case in online samples, the respondents are on average more educated than the general population. This means that the heterogeneous effects for the list experiment with regard to education (presented in the appendix) should be interpreted with caution.

Figure 5 presents a breakdown of party preferences in the sample. While the PSD are

[1] `https://www.languageconnect.net/`

[2] `https://www.worldometers.info/demographics/romania-demographics/`

under-represented (as noted in the main text), the sample contains a good representation of all the major political parties in Romanian politics.

Finally, Table 4 presents descriptive statistics for the main outcome variables used in the analyses.

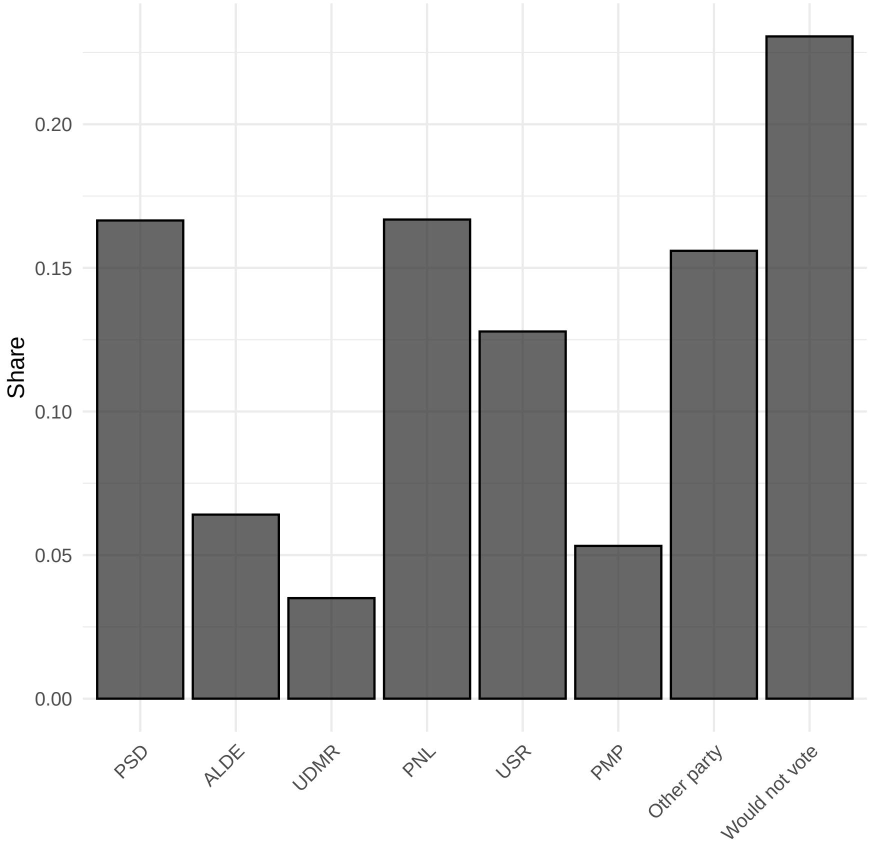

Figure 5: Vote intentions: “What political party would you vote for if the national parliamentary election were today?”

Table 4: Respondent attitudes

Statistic	Mean	St. Dev.	Min	Max	N
Government attitude (5=favoring)	2.09	1.14	1	5	3,027
Government supporter (attitude+party)	0.14	0.34	0	1	3,027
Political interest (4=high)	2.27	0.93	1.00	4.00	2,969
Corruption increase	4.04	1.12	1.00	5.00	1,492
Worse economy	3.82	1.20	1.00	5.00	1,510
Corruption in politics	4.12	1.08	1.00	5.00	1,546
Corruption worry	3.34	0.94	1.00	4.00	1,506
Direct bribe question (1=yes)	0.19	0.39	0.00	1.00	1,516

Power analysis

To decide how many respondents I needed in the final survey (the total number of completes) I conducted two basic power-analyses with simulated data. First, I simulated answers to the list experiment using a list with 4 control items.[3] The simulated responses to the sensitive item (the fifth item in the treatment group) were a random draw from a Bernoulli distribution with $p = 0.3$. In this case I assumed that the answer to the direct sensitive question was a random draw from a Bernoulli distribution with $p = 0.2$. That is, I assumed that the 'true' sensitivity bias was 0.1, or 10 percentage points. For comparison, this is half of the size of the amount of sensitivity bias uncovered in Gonzalez-Ocantos et al. (2012), in their study on vote buying. I then simulated 5000 data sets and calculated the share of the times the null hypothesis of the estimate for the sensitive item being ≤ 0.2 was rejected (using the difference-in-means estimator). I then repeated this process for different numbers of 'respondents'. Figure 1 shows the estimated power (the % of the times the null hypothesis was rejected) for different n.[4]

Next, I simulated data for the political bias hypothesis. I assumed an equal share of two types of respondents; government supporters and 'others'. I then randomly assigned all respondents to either the control group (no political prime) or the treatment group (political prime). The setup simulated a 5-category ordinal outcome variable with a mean value of about 3.55 for the group labeled as 'others' (including both the treatment and the control group). The standard deviation was assumed to be about 1.3 for all groups. 'Government supporters' in the control group were drawn from a random ordinal variable with a mean value of about 3.05, and 'government supporters' in the treatment group were drawn from a random ordinal variable with a mean value of about 2.6. The standard deviation for the latter two groups was about 1.3. This setup thus assumes a baseline difference of about -0.5 between 'government supporters' and 'others' in the control group, and an interaction effect of about -0.45. The power analysis focused on the interaction term since estimation of this term demands the most power. For each data set I estimated equation (1) and tested if δ was negative and statistically significant. Figure 2 shows the estimated power for different

[3] To mimic the actual survey the control items were simulated to be negatively correlated (with ρ between -0.2 to -0.5).

[4] This procedure thus compares the list estimate to a fixed estimate (0.2) for the direct question. A more reasonable simulation strategy is to also view the answers to the direct question as a random variable (a random draw from a Bernoulli distribution with $p = 0.2$). I therefore conducted an additional post hoc simulation, with the same setup as above, but with the direct question modeled as a random variable. For each simulation, I used the procedure in Blair and Imai (2012) to estimate the amount of 'social desirability bias', and whether this quantity was statistically different from 0. Given this setup, I estimated the power for 3000 respondents (a number close to the final sample). The procedure yielded a power estimate of 89.5%.

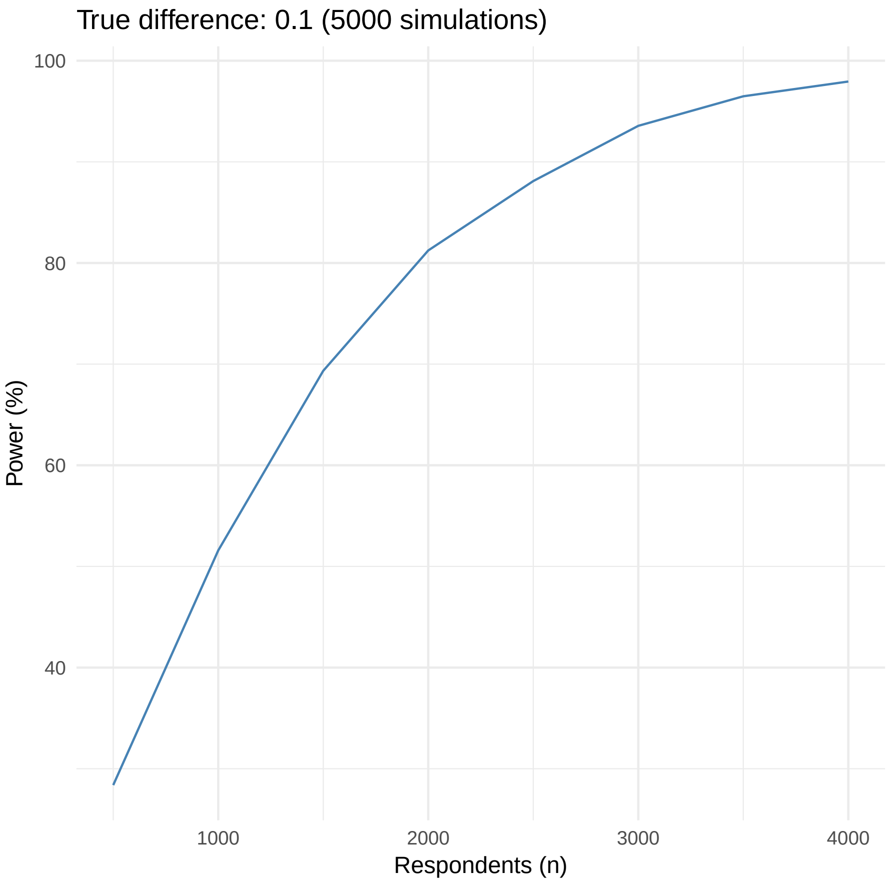

Figure 6: Testing H_2: Estimated power for list experiment (different n).

n based on these simulations.[5]

Because of the randomization scheme described above the experiment has $\frac{1}{4}$ of the respondents in the control group and another $\frac{1}{4}$ of the respondents in the treatment group with regard to each specific corruption/economy question. The effective sample for testing H_1 is hence about half the size of the sample for testing H_2. Based on these two power-calculations I decided that above 2800-2900 respondents would give me enough power to test both hypotheses. This would give me substantial power to detect the main effect in the list experiment (H_2 - over 90% power to detect sensitivity bias of 0.1), and also plenty of room to conduct sub-group analyses (for instance, splitting this sample in half still gives me reasonable power to detect an effect of about 0.1). An effective sample of (at least) 1400 respondents also would give me over 80% power to detect the main effects with regard to

[5] The simulations assumed that the share of government supporters was 0.5. In reality this share was considerably lower (0.14), according to the specific criterion described above. Therefore I also conducted a post hoc power analysis with the share of 'government supporters' set to 0.14, while the rest of the setup was unchanged. For 1500 respondents this yielded a power estimate of 62%. In practice, the estimated effects in the experiment are in general slightly larger and slightly less variable than assumed in this setup.

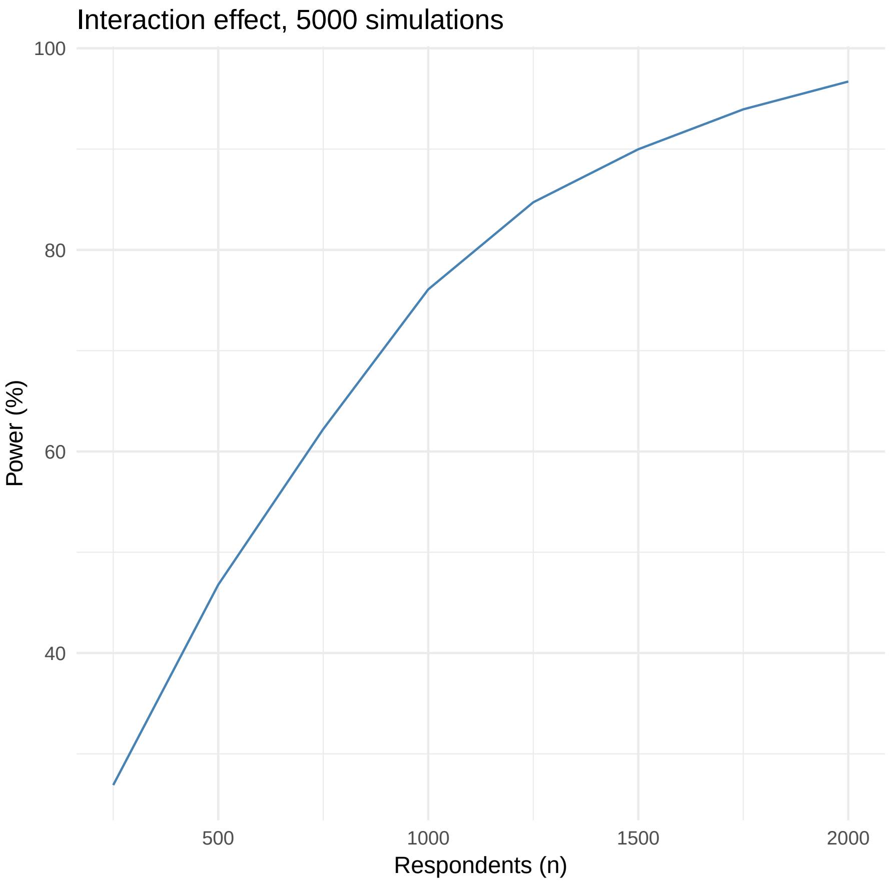

Figure 7: Testing H_1: Estimated power for political bias experiment (different n).

H_1 (given the assumptions stated above). The final sample of over 3000 respondents should thus be considered sufficient to test both hypotheses with high statistical power.

The NLS estimator for list experiments

An important limitation of the difference-in-means estimator is that it does not allow researchers to efficiently estimate multivariate relationships between preferences over the sensitive item and respondents' characteristics. Researchers may apply this estimator to various subsets of the data and compare the results, but such an approach is inefficient and is not applicable when the sample size is small or when many covariates must be incorporated into analysis. To overcome this problem Imai (2011) developed two new multivariate regression estimators that allows the researcher to model the response to the sensitive item as a function of respondent characteristics. The analyses in this study make use of one of these estimators, the NLS estimator.

Imai (2011) develops the NLS estimator to model the response to the list experiment (Y_i) as a function of respondent characteristics. The estimator can be defined as:

$$Y_i = f(\mathbf{x}_i, \gamma) + T_i g(\mathbf{x}_i, \delta) + \epsilon_i \tag{1}$$

Where $\mathbf{x}_i$ is a matrix with respondent covariates, T_i is an indicator variable denoting whether a respondent is in the treatment or control group, and (γ, δ) is a vector of unknown parameters. The model assumes $\mathbf{E}[\epsilon_\mathbf{i}|\mathbf{x}_i, T_i = 0]$. The model thus puts together two potentially nonlinear regression models where $f(\mathbf{x}_i, \gamma)$ represents the conditional expectation of the control items, given the covariates, and $g(\mathbf{x}_i, \delta)$ represents the expected response to the sensitive item, given the covariates. Imai (2011) suggests a two-step procedure to estimate the model where $f(\mathbf{x}_i, \gamma)$ first is fitted to the control group and then $g(\mathbf{x}_i, \delta)$ is fitted to the treatment group using the response variable $Y_i^* = Y_i - f(\mathbf{x}_i, \hat{\gamma})$ where $\hat{\gamma}$ represents the estimate of γ from the first stage.[6] The functional form of the models has to be specified, but Blair and Imai (2012) suggests using logistic regression submodels.[7]

The NLS model is consistent as long as the functional form is correctly specified. In principle, the maximum likelihood (ML) estimator - also developed in Imai (2011) - is a more efficient estimator. However, the ML estimator has shown to be biased in the face of different forms of measurement error that can occur when the list experiment is not perfectly designed and implemented. The NLS estimator, on the other hand, has shown to be quite robust to most such errors (Blair et al. 2019).

Blair et al. (2019) suggests a general specification test, based on the well-known Hausman test, as a formal means of comparing, and deciding between, the ML and NLS estimator.

[6] In the appendix Imai (2011) shows how to obtain heteroscedasticity-robust standard errors for the NLS model.

[7] This would imply that $f(\mathbf{x}_i, \gamma) = J\text{logit}^{-1}(\mathbf{x}_i{}'\gamma)$ and $g(\mathbf{x}_i, \delta) = \text{logit}^{-1}(\mathbf{x}_i{}'\delta)$. See previous section for a description of the logistic regression model.

The idea is that if the underlying modeling assumptions are correct the estimators should yield results that are statistically indistinguishable. In this case the ML estimator will be more efficient. The test takes the following form:

$$(\hat{\theta}_{ML} - \hat{\theta}_{NLS})'(\mathbf{V}(\widehat{\hat{\theta}_{NLS}}) - \mathbf{V}(\widehat{\hat{\theta}_{ML}}))^{-1}(\hat{\theta}_{ML} - \hat{\theta}_{NLS})' \sim \chi^2_{dim(\gamma)+dim(\delta)} \tag{2}$$

where $\hat{\theta}_{NLS} = (\hat{\gamma}_{NLS}, \hat{\delta}_{NLS})$, $\hat{\theta}_{ML} = (\hat{\gamma}_{ML}, \hat{\delta}_{ML})$, and $\mathbf{V}(\widehat{\hat{\theta}_{NLS}})$ and $\mathbf{V}(\widehat{\hat{\theta}_{ML}})$ are their estimated asymptotic variances. The null hypothesis in the test assumes 'correct model specification', in which case the ML estimator should be preferred.

Depending on the exact model specification (which covariates that were included), the test yielded significant results on some occasions, with a p-value of less than 0.05. This suggests that the ML model might not be appropriate to model the data, and that the NLS estimator is the safer option.

Additional analyses and sensitivity checks

Political bias experiment

Additional models. This section contains additional analyses and sensitivity checks for the PB experiment. Table 5 shows results for the basic specification used in the main text, but with covariates included. Table 6 replicates the specification in the main text but uses Ordinal Logistic Regression to estimate the model. The results from these models are clearly in line with the results in the main text but also show that the interaction effect for the *Corruption in politics* outcome is slightly smaller and more variable than for the other outcomes. As a result, this coefficient does not reach conventional level of statistical significance in all models.

Table 7 reports estimates where I instead code *political affiliation* only based on the variable measuring the respondents' attitudes towards the current government (see above). The respondents are coded as either 'opposing', being 'neutral', or 'favoring' the current government.[8] These results show the same pattern as the results above, with neutral respondents being more positive than 'oppose' respondents and 'favoring' respondents being the most positive. The prime also has the strongest effect on respondents favoring the government, followed by neutral respondents. The results from this analysis are in many ways more striking than the results reported above. For instance, for the *corruption change* outcome when comparing respondents in the treatment group favoring the government with respondents opposing the government the total difference is over 2 $((-1.48) + (-0.67))$, e.g. more than two full categories on the 5-point scale.

[8] Where 'opposing' corresponds to category 1-2 on the support variable, 'neutral' corresponds to category 3, and 'favoring' corresponds to category 4-5. The categories were collapsed to make each category sufficiently large.

Table 5: The effect of the political prime on perceived corruption, with covariates

	Dependent variable:			
	Corruption increase	Worse economy	Corruption in politics	Corruption worry
	(1)	(2)	(3)	(4)
Government support	-0.86^{***}	-1.16^{***}	-0.68^{***}	-0.39^{***}
	(0.12)	(0.11)	(0.12)	(0.09)
Prime	-0.03	0.15^{*}	-0.02	-0.04
	(0.06)	(0.06)	(0.05)	(0.05)
Gov. support x Prime	-0.59^{***}	-0.46^{**}	-0.31	-0.47^{***}
	(0.17)	(0.15)	(0.18)	(0.13)
Female	0.17^{**}	0.20^{***}	0.25^{***}	0.30^{***}
	(0.05)	(0.06)	(0.05)	(0.04)
City inhabitant	-0.12^{*}	-0.26^{***}	0.20^{***}	0.19^{***}
	(0.05)	(0.06)	(0.05)	(0.04)
University education	-0.21^{***}	-0.15^{**}	0.35^{***}	0.24^{***}
	(0.06)	(0.06)	(0.05)	(0.05)
Top 20% income	-0.14	-0.10	-0.68^{***}	-0.53^{***}
	(0.07)	(0.07)	(0.08)	(0.06)
Age	-0.03^{*}	0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Age^2	0.0003	-0.0001	0.0001	0.0003
	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Constant	4.95^{***}	4.04^{***}	4.29^{***}	3.18^{***}
	(0.24)	(0.22)	(0.20)	(0.22)
Observations	1,486	1,504	1,542	1,500
Adjusted R^2	0.17	0.20	0.19	0.20

Note: All models are estimated using OLS. Robust standard errors in parentheses (HC2). 'Prime' is an indicator variable equal to 1 if a respondent was asked the political questions before a specific corruption/economy question, and 0 otherwise. *p<0.05; **p<0.01; ***p<0.001.

Table 6: The effect of the political prime on perceived corruption, OLR estimates

	Dependent variable:			
	Corruption increase	Worse economy	Corruption in politics	Corruption worry
	(1)	(2)	(3)	(4)
Government support	-1.44^{***}	-1.83^{***}	-1.25^{***}	-0.98^{***}
	(0.19)	(0.19)	(0.19)	(0.19)
Prime	0.01	0.25^{*}	0.10	-0.12
	(0.11)	(0.10)	(0.10)	(0.11)
Gov. support x Prime	-0.98^{***}	-0.81^{**}	-0.46	-0.59^{*}
	(0.28)	(0.28)	(0.27)	(0.26)
Observations	1,492	1,510	1,546	1,506

Note: All models are estimated using ordinal logistic regression. Standard errors in parentheses. 'Prime' is an indicator variable equal to 1 if a respondent was asked the political questions before a specific corruption/economy question, and 0 otherwise. *p<0.05; **p<0.01; ***p<0.001.

Table 7: The effect of the political prime on perceived corruption, oppose/favor government

	Dependent variable:			
	Corruption increase	Worse economy	Corruption in politics	Corruption worry
	(1)	(2)	(3)	(4)
Prime (Government: Oppose)	0.06	0.19^{**}	0.08	−0.03
	(0.05)	(0.06)	(0.07)	(0.06)
Government: Neutral	-0.76^{***}	-0.70^{***}	-0.34^{***}	-0.44^{***}
	(0.09)	(0.09)	(0.09)	(0.08)
Government: Favor	-1.48^{***}	-1.71^{***}	-0.54^{***}	-0.45^{***}
	(0.13)	(0.10)	(0.10)	(0.08)
Government: Neutral x Prime	−0.18	-0.30^{*}	−0.19	−0.18
	(0.12)	(0.13)	(0.13)	(0.12)
Government: Favor x Prime	-0.67^{***}	-0.51^{***}	-0.50^{***}	-0.45^{***}
	(0.17)	(0.15)	(0.15)	(0.13)
Constant	4.43^{***}	4.18^{***}	4.28^{***}	3.57^{***}
	(0.04)	(0.05)	(0.05)	(0.04)
Observations	1,492	1,510	1,546	1,506
Adjusted R^2	0.34	0.33	0.08	0.10

Note: All models are estimated using OLS. Robust standard errors in parentheses (HC2). ‘Prime’ is an indicator variable equal to 1 if a respondent was asked the political questions before a specific corruption/economy question, and 0 otherwise. ‘Opposing’ corresponds to category 1-2 on the government support variable, ‘neutral’ corresponds to category 3, and ‘favoring’ corresponds to category 4-5. *p<0.05; **p<0.01; ***p<0.001.

Removing response-time outliers. It is well known that some respondents might pay less attention while taking a survey and that their responses might be of lower quality. To explore if more inattentive respondents influence the overall results, I classified each respondent according to their completion time by calculating response-time percentiles in the sample (see Alvarez et al. (2019)). I then re-estimated the main models (those included in the main text) and tested different exclusion-criteria.[9] I tried three different criteria to exclude different categories of response-time outliers: excluding the 5% fastest respondents, excluding the 10% fastest respondents, excluding the 5% fastest and the 5% slowest respondents. Table 8 report regression estimates for the PB hypothesis and figure 9 (below) displays results for the list experiment, using these three different exclusion criteria. Overall, the main results are not sensitive to excluding different response-time outliers.

[9] The estimates reported in the main text include *all* respondents - in accordance with the specifications in the pre-analysis plan.

Table 8: Estimates of eq. (1), excluding response-time outliers

	5th pctl < X	10th pctl < X	5th pctl < X < 95th
Corr. increase β_1	-0.92	-0.93	-0.92
	(0.12)	(0.13)	(0.13)
Corr. increase δ	-0.67	-0.66	-0.7
	(0.18)	(0.18)	(0.19)
Worse economy β_1	-1.19	-1.23	-1.23
	(0.11)	(0.12)	(0.12)
Worse economy δ	-0.51	-0.48	-0.43
	(0.16)	(0.16)	(0.17)
Corr. in politics β_1	-0.63	-0.62	-0.68
	(0.11)	(0.12)	(0.12)
Corr. in politics δ	-0.32	-0.38	-0.34
	(0.18)	(0.18)	(0.19)
Corr. worry β_1	-0.38	-0.39	-0.42
	(0.094)	(0.096)	(0.10)
Corr. worry δ	-0.45	-0.44	-0.46
	(0.14)	(0.14)	(0.15)

Note: β_1 and δ refer to coefficient estimates in equation (1). All models estimated using OLS. X denotes response time percentiles. Robust standard errors in parentheses (HC2).

List experiment

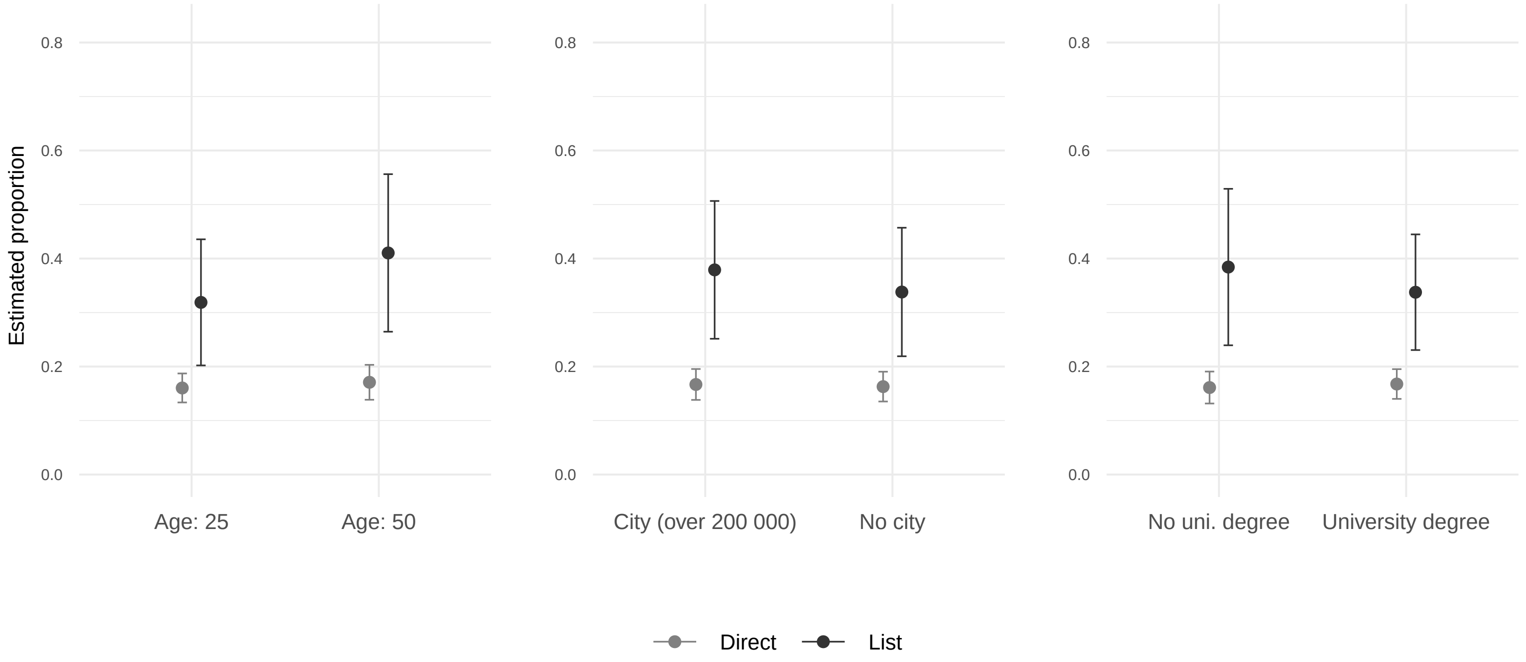

Figure 8: Comparison of direct estimate vs list estimate for sensitive item (being asked to pay a bribe). Different subgroups: Age, City inhabitant, and Education.

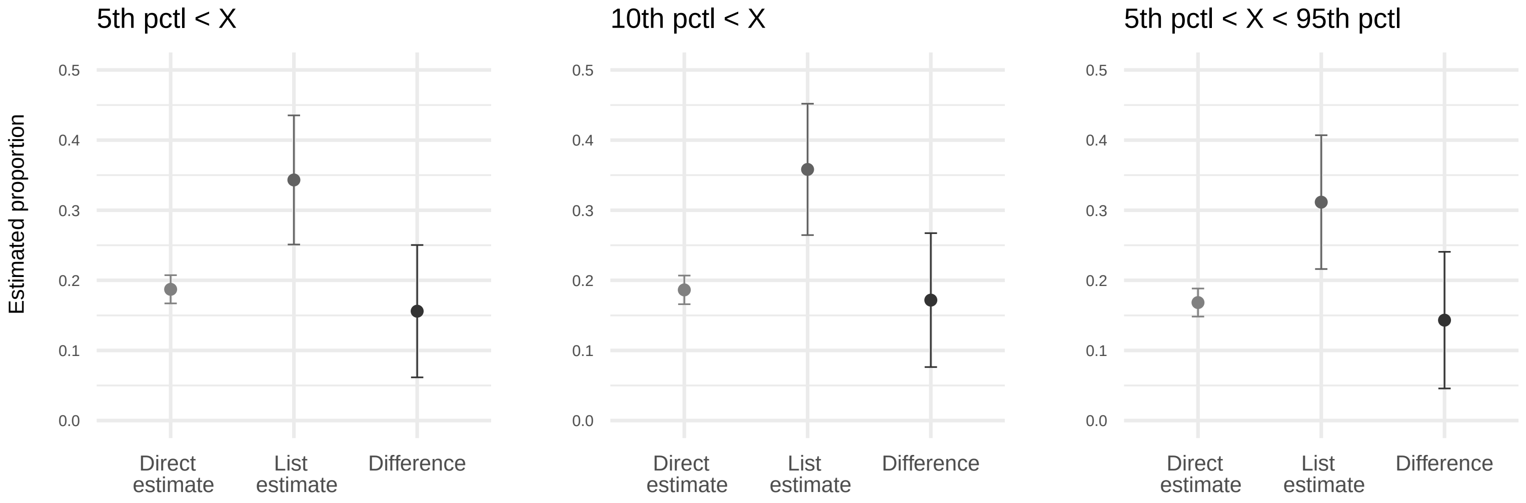

Figure 9: Comparison of direct estimate vs list estimate for sensitive item (being asked to pay a bribe). Removing response-time outliers. X denotes response time percentiles.

Testing for 'design effects' in list experiment. Blair and Imai (2012) proposes a test for detecting *design effects*, e.g. when the inclusion of the sensitive item affects how respondents answer the control items. The proposed test is based on the calculation of the proportion of respondent different respondent types (see above). If one of these proportions would be *negative* this is a violation of the no design effects assumption, and a sign that the list experiment did not work as intended. Formally, the null hypothesis of 'no design effect' can be stated as:

$$H_0 = \begin{cases} Pr(Y_i \leq y | T_i = 0) \geq Pr(Y_i \leq y | T_i = 1) \; \forall \; y = 0, \dots, J-1, \\ Pr(Y_i \leq y | T_i = 1) \geq Pr(Y_i \leq y - 1 | T_i = 0) \; \forall \; y = 1, \dots, J. \end{cases} \tag{3}$$

The alternative hypothesis is that at least one value of y does not satisfy the inequalities described under H_0. Blair and Imai (2012) derives methods to compute p-values for observed proportions under the null hypothesis. Importantly, if *none* of the proportions are estimated to be negative the null hypothesis will not be rejected. The table below shows the estimated distribution of respondent types based on the list experiment in the study at hand.

Table 9: Respondent types, estimated proportions

Respondent type	Est.	s.e.
$Pr(Y_i(0) = 0, Z_i = 1)$	0.007	0.007
$Pr(Y_i(1) = 0, Z_i = 1)$	0.036	0.016
$Pr(Y_i(2) = 0, Z_i = 1)$	0.122	0.018
$Pr(Y_i(3) = 0, Z_i = 1)$	0.076	0.015
$Pr(Y_i(4) = 0, Z_i = 1)$	0.112	0.008
$Pr(Y_i(0) = 0, Z_i = 0)$	0.038	0.005
$Pr(Y_i(1) = 0, Z_i = 0)$	0.192	0.012
$Pr(Y_i(2) = 0, Z_i = 0)$	0.244	0.017
$Pr(Y_i(3) = 0, Z_i = 0)$	0.087	0.017
$Pr(Y_i(4) = 0, Z_i = 0)$	0.086	0.013

Note: The table shows the estimated proportion of respondent types characterized by the total number of affirmative answers to the control questions, Y, and the truthful answer for the sensitive item Z (1 indicates affirmative and 0 represents negative). Standard errors are also provided for each estimated proportion.

As shown in the table, none of the proportions are estimated to be negative, and we can conclude that we do not find evidence of any violations of the 'no design effects' assumption, based on the test.

Question order assumptions and tests

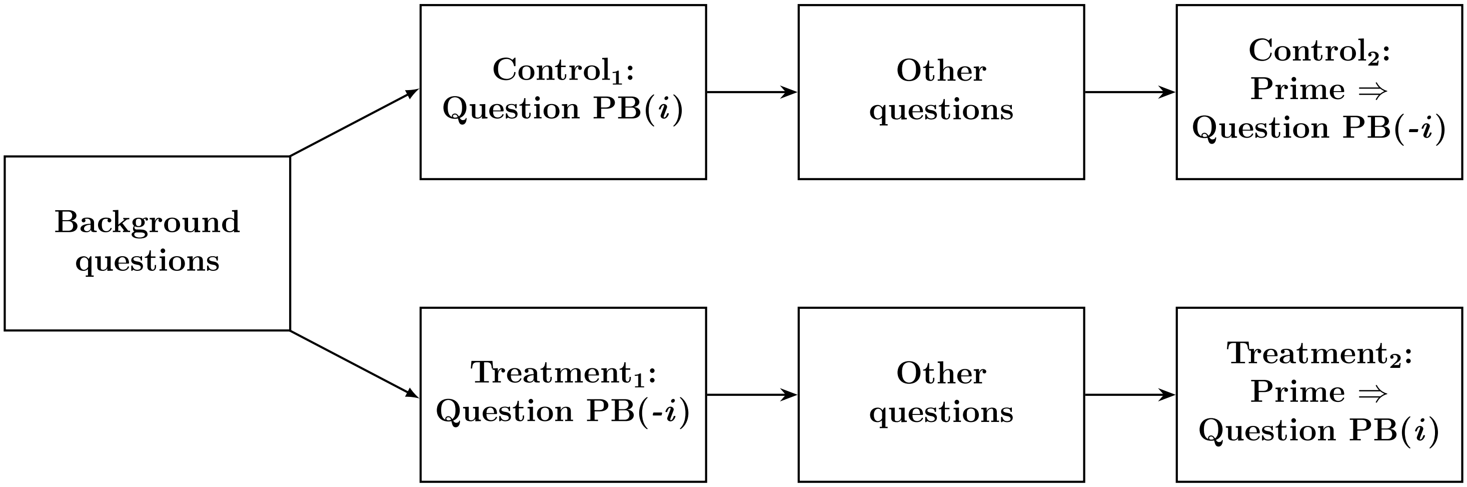
Background questions
Control$_1$: Question PB(i)
Other questions
Control$_2$: Prime ⇒ Question PB(-i)
Treatment$_1$: Question PB(-i)
Other questions
Treatment$_2$: Prime ⇒ Question PB(i)

Figure 10: Basic structure of the PB experiment with regard to PB question i, where i represents a specific question in the set
$PB_i \in \{$*Corruption increase, Economy worse, Corruption in politics, Corruption worry*$\}$.

The design with the 'political prime', where some respondents answered the corruption questions after the political questions, assumes that the fact that the control group answered the corresponding corruption questions a little earlier in the survey did not affect the outcome. The effect of the prime with regard to a specific PB question i is estimated by comparing **Control$_1$** to **Treatment$_2$** (see figure). Formally, we assume that $\mathbf{E}[Y(0)|T = 0] = \mathbf{E}[Y(0)|T = 1]$ - that the *potential outcomes* for untreated observations on average are the same for respondents assigned to the control and treatment group respectively (see Holland 1986). As part of the pilot study I therefore conducted a test of this assumption where respondents were first randomized into the 'corruption in politics' question (see above) and then received the 'economy question' later in the survey, or the other way around (I only used two questions to retain power with the small pilot sample). However, in this version of the survey the randomized political prime was not included. Therefore if the assumptions stated above holds we should not expect the answer to the questions to differ depending on whether the question was given earlier or somewhat later in the survey. I then conducted an independent sample t-test for each question to see if the placement in the survey itself affected the responses. The results showed no significant differences. Corruption in politics, difference ($earlier - later$): $t_{104} = -0.22, p = 0.83$. Economy, difference ($earlier - later$): $t_{104} = -0.065, p = 0.95$. This suggests that the assumption holds up and that any differences observed in the experiment is due to the political prime.

The design also assumes that the assignment to a specific control group, $\mathbf{PB(\mathit{i})}$, influences whether a respondent reports that he or she is a government supporter or his or her view of

the government's performance (both measured later in the survey). To test this assumption, I modeled (1) whether a respondent reported that he or she would vote for a government party and (2) his or her degree of support for the government. Both outcomes were modeled as a function of control group assignment, using a logistic regression model for (1) and an ordinal logistic regression model for (2). I then conducted a LR-test (a χ^2-test), comparing a null model (only including an intercept) to the model with a set of control group-indicators. Neither model showed any evidence that control group assignment affected respondents' view of the government:
(1): $\chi^2(3) = 1.86$ ($p = 0.6$)
(2): $\chi^2(3) = 1.3$ ($p = 0.73$)

Deviations from the pre-analysis plan

All hypotheses and main analyses were specified in the pre-analysis plan. In the process of working with the paper I deviated from the plan on some occasions and several additional robustness checks were added. The deviations and additions are listed below:

- I slightly changed the wording of hypothesis H_{1a}, by changing 'opposition supporters' to 'other respondents'. The change was made because while 'incumbent supporters' are distinctly operationalized (see above), this group is compared to *all other respondents* among which many could be characterized as 'opposition supporters', but where many also might be characterized as 'undecided'. Therefore 'other respondents' seemed to better reflect the actual categorization used in the study. To increase consistency in the manuscript, I have also changed 'incumbent supporters' to 'government supporters'. All operationalizations and codings of the variable are unchanged from the pre-analysis plan. Based on the same reasoning I also crossed out 'opposition supporters' from H_{1b}, which now only refers to 'government supporters'.

- The pre-analysis plan specified the main tests of H_{1a} and H_{1b} as reported in Table 1, and illustrated in Figure 2. The ordinal logistic regression model was specified as a robustness check of these main results. The test for sensitivity bias (H_2) using the difference-in-means estimator for the list experiment and comparing the results to the direct question modeled using logistic regression was also pre-specified. The sub-group analyses based on six covariates were described - and are still described in the text - as 'exploratory'. The three of these that yielded the most interesting results are included in the main text whereas the other three are included in the appendix. All further analyses and sensitivity tests were not pre-specified and are based on comments and suggestions I have received when circulating the paper.

- The pre-analysis plan was submitted to EGAP after a soft-launch with the first 480 responses had been conducted. The remaining responses were collected after submission. This is clearly stated in the pre-analysis plan.

- I added two 'post hoc' power analyses (see above) that better resemble the true distributions of the variables involved in the experiment. All original power analyses are still included.

References

Alvarez, R. Michael et al. (2019). "Paying Attention to Inattentive Survey Respondents". In: *Political Analysis* 27.2, pp. 145–162.

Blair, Graeme and Kosuke Imai (2012). "Statistical Analysis of List Experiments". In: *Political Analysis* 20.1, pp. 47–77.

Blair, Graeme, Winston Chou, and Kosuke Imai (2019). "List Experiments with Measurement Error". In: *Political Analysis* Forthcoming.

Evans, Geoffrey and Robert Andersen (2006). "The Political Conditioning of Economic Perceptions". In: *The Journal of Politics* 68.1, pp. 194–207.

Gonzalez-Ocantos, Ezequiel et al. (2012). "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua". In: *American Journal of Political Science* 56.1, pp. 202–217.

Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960.

Imai, Kosuke (2011). "Multivariate Regression Analysis for the Item Count Technique". In: *Journal of the American Statistical Association* 106.494, pp. 407–416.

Peiffer, Caryn (2018). "Message Received? Experimental Findings on How Messages about Corruption Shape Perceptions". In: *British Journal of Political Science* Forthcoming.