Appendix A Research design

A.1 Case selection

In this paper, we probe our assumptions about what the sensitive response is in a sensitive survey question. Fundamentally, our assumptions about what the sensitive response is in a question about vote buying depends on assumptions about the prevailing social norms. Norms and attitudes towards vote buying are likely to vary across different contexts. When forming expectations about sensitivity bias, we therefore also need to consider the particularities of our empirical case.

We explore sensitivity bias in vote buying in a representative sample of the adult Nigerian population surveyed after the 2019 elections in the country. Nigeria represents a suitable case to re-evaluate some of the prevailing assumptions in the quantitative literature on vote buying. On the one hand, vote buying is, as in many other countries, illegal in Nigeria. This underscores the likelihood that fear of sanctions is real source of sensitivity bias. Recent laws, reportedly enforced during the 2019 elections, explicitly sought to curtail the incidence of vote buying by restricting the ability of party agents to monitor vote-choice at the day of polling (Obe, 2019). The public discourse in the 2019 electoral environment thus brought attention to its unlawful and undemocratic aspects.

On the other hand, Nigeria represents a case where vote buying is deeply intertwined with the nature of electoral politics. Vote buying has been documented throughout the country's shifts in political power, military coups and democratic transitions (Olaniyan, 2020). Since Nigeria's return to multiparty elections in 1999, vote buying has been regular feature in every single Nigerian election (Olaniyan, 2020). According to several observers it even increased in significance in the 2019 elections as improvements in electoral technology made other tools of electoral manipulation unfeasible (Onuoha and Ojo, 2018; Olaniyan, 2020; Obe, 2019). The prevalence of vote buying in the 2019 elections was acknowledged both by representatives of official electoral management bodies, domestic and international observers (Okakwu, 2019; EU Election Observation Mission, 2019) and scholars (Obe, 2019). Existing qualitative accounts also indicate that vote buying and vote selling has become a widely accepted norm in Nigeria (Sakariyau et al., 2015). Obe (2019, p.113) remarks on the peculiar honesty of those engaged in vote-selling in Nigeria, with many voters in the 2019 elections *expecting* money to show up and openly asking candidates what they were willing to pay.

Nigeria thus represents a case where the connotations surrounding vote buying is sufficiently ambiguous that our statistical exploration of sensitivity bias in vote buying questions is meaningful. While Nigeria arguably represents a most-likely case for finding that vote buying is not stigmatized, there are many other countries where vote buying is similarly ubiquitous, such as Ghana, Kenya and Benin, to mention a few in Africa.

A.2 Sampling

For our empirical analysis, we draw on an original, nationally representative survey of 2400 Nigerian citizens. The survey was fielded by the Nigerian research agency Practical Sampling International (PSI) between March 23 and April 16 2019, just after the conclusions of the Nigerian general elections on February 23 (for president and national assembly) and March 9 (for governor and state assembly).

The target population was adult Nigerian males and females, aged eighteen years and above. Respondents were selected using a clustered, stratified, multi-stage random selection procedure across all of Nigeria's 37 states (including the Federal Capital Territory). Within each state, local government areas were stratified based on urban and rural status, and then selected based on probability proportional to size random sampling. To this end, we used the social population projections for 2016, based on the 2006 Nigerian census. The respondents were selected randomly across 132 urban and 168 rural primary sampling units (PSUs), with a 50/50 quota for men and women. The survey was conducted face-to-face, using tablets.

A.3 Ethical considerations

The survey and list experiment were approved by the *[anonymized ERB]* on 2018-11-07 (ID no *anonymized*). In designing and implementing the survey we took a number of steps to respect the integrity and autonomy of survey respondents and our local research partners and ensure minimal harm.

In conducting the survey, we worked closely with Practical Sampling International (PSI), a Nigeria-based public opinion research agency with vast experience from conducting large-scale public opinion surveys. PSI has a team of highly experienced, local enumerators. Jointly, PSI and the research team also carefully trained and prepared the enumerators for the fieldwork. A reference group with PSI staff and two case experts, further helped us to screen the survey questionnaire, including the list experiment, to assure that they did not include questions that were offensive or would in other way harm respondents.

All interviews are based on informed consent. The enumerators informed respondents about the purpose of the study, the random nature of the selection procedure, that their identity would not be revealed, that the survey was conducted by the research agency on behalf of *[anonymized university]*, and that they could withdraw from their voluntary participation at any time during the interview. The enumerators only proceeded once they had received explicit verbal consent. We offered no compensation to respondents.

We protected respondent confidentiality at the individual level in our survey implementation, but also at the community level by not revealing any information in our replication material that could link individuals to small geographical areas.

To make sure our survey did not intervene in the political process, we waited to field our survey more than one month after the announcement of electoral results.

Appendix B Main results

B.1 Summary statistics

Table B.1 shows summary statistics for the variables in our analysis. The data set includes 2398 observations. We discarded 2 observations which were missing a response to the list experiment; this is necessary for multivariate analysis. Survey questions coded "don't know", "refused" or which were missing a response were coded as missing. INR indicates item-non-response rates.

Table B.1 Summary statistics

Variable	N	Mean	Sd	Min	Max	INR
List experiment treatment	2398	0.5	0.5	0	1	0
List experiment count	2398	2.4	1.19	0	5	0
Direct measure	2224	0.51	0.5	0	1	0.07
Male	2398	0.5	0.5	0	1	0
Age	2391	32.63	10.93	18	85	0
College education	2394	0.14	0.35	0	1	0
Employed	2381	0.6	0.49	0	1	0.01
Registered voter	2388	0.79	0.4	0	1	0

Table B.2 shows the raw scores of the list experiment for the control and treatment groups. We note that a high proportion of respondents in the treatment group (10.9%) have indicated the highest number of items, thus willingly revealing to the interviewer that they have experienced vote buying. Apart from this, we do not notice any noteworthy deviations from what would be expected in the distribution of answers in the control and treatment groups.

Table B.2 Observed responses in the list experiment for the control and treatment groups

	Control group		Treatment group	
#LE items	Frequency	Proportion	Frequency	Proportion
0	11	0.9%	11	0.9%
1	328	27.4%	291	24.3%
2	379	31.6%	323	26.9%
3	303	25.3%	304	25.4%
4	178	14.8%	139	11.6%
5			131	10.9%
Total	1199		1199	

B.2 Test of priming effects

Our direct measure of vote buying was presented after the list experiment. In order to ensure that respondents' answers to the direct question were not

primed by the sensitive item in the list experiment, we compared the estimated proportions of vote buying according to the direct question between the treatment and control group in the list experiment. Figure B.1 below shows the estimated proportions and confidence intervals for the two groups. No difference in the estimated proportion of vote buying can be detected from this figure. A t-test for difference in mean confirm this, with a mean difference of 0.005, t-value of 0.25, and a p-value of 0.8. Thus, there is no evidence to support priming effects from the list experiment.

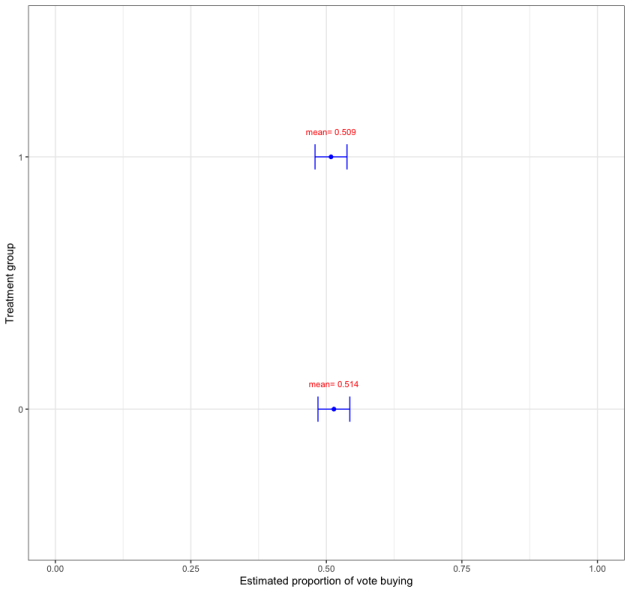

Fig. B.1 Estimated proportion of vote buying in the direct question conditional on the list experiment treatment (1) and control groups (0)

B.3 No design effects

In addition to our main exploration of design effects, we also implement a formal statistical test for design effects as outlined in Blair and Imai (2012). As shown in Table B.3 , none of the estimated proportions are negative and we therefore fail to reject the null hypothesis of no design effects with a Bonferroni-corrected p-value of 1.

There are, however, limitations to the no design effects statistical test (Blair and Imai, 2012, p. 64-5). One such limitation is that the test may fail to detect violations unless the design effects are very large. The test could also fail if design effects are positive among some respondents and negative among others. Since it possible that the formal test yields a false negative, Blair and Imai (2012) warn that a failure to reject the null should not be taken as evidence to support the assumption of no design effects.

Table B.3 No design effects

Population proportions	Estimated proportion	Standard error
$\pi(Y_i(0) = 0, Z_i = 1)$	0.00	0.00
$\pi(Y_i(0) = 1, Z_i = 1)$	0.03	0.02
$\pi(Y_i(0) = 2, Z_i = 1)$	0.08	0.02
$\pi(Y_i(0) = 3, Z_i = 1)$	0.08	0.02
$\pi(Y_i(0) = 4, Z_i = 1)$	0.11	0.01
$\pi(Y_i(0) = 0, Z_i = 0)$	0.01	0.00
$\pi(Y_i(0) = 1, Z_i = 0)$	0.24	0.01
$\pi(Y_i(0) = 2, Z_i = 0)$	0.24	0.02
$\pi(Y_i(0) = 3, Z_i = 0)$	0.18	0.02
$\pi(Y_i(0) = 4, Z_i = 0)$	0.04	0.01
Bonferroni corrected p-value	1.00	

B.4 No liars

We have argued that assumptions of polarity determine which respondent types are considered floor- and ceiling liars.

To make this point clearer, we reproduce the illustration of floor and ceiling effects in Blair and Imai (2012, p. 66) with a modification. Following Blair and Imai (2012), Table B.4 shows possible respondent types in the treatment group for each observed count of list experiment items. Respondent types are characterised here by their latent count of control items Y_i (where Y ranges from 0 to J=5), and their latent answer to the sensitive item Z_i (coded 1 for yes, 0 for no). For each observed answer to the list experiment, there are several possible respondent types. For example, respondents who answer 4 in the list experiment could have experienced 3 control items and also vote buying (3,1) or 4 control items and not vote buying (4,0). The table also indicates additional respondent types who are "floor"- and "ceiling-liars", using bold face.[16] For example, among the respondents who answer 4 list experiment items, there are some respondents who truthfully have experienced all 5 items but prefer not to reveal vote buying while forfeiting confidentiality (this is situation i outlined in the main text). So in Table B.4 , respondent type (4,1) for $Y_i(1)$=4 is considered a ceiling-liar.

The addition we make in Table B.4 is to show how what we consider floor- and ceiling-liars differs depending on what we assume is the sensitive response (denoted S_i, coded 1 if 1 if answering affirmatively is assumed to be sensitive and 0 if answering negatively is assumed sensitive). Assuming that answering "no" to vote buying is sensitive, respondents who have experienced all five items should prefer to answer 5; so we would not interpret ceiling effects as an indication of strategic misreporting. Floor effects as defined by Blair and Imai (2012) – respondent type (0,1) for $Y_i(1)$=0 – also does not seem strategic since respondents should prefer to avoid answering 0 if this is the sensitive response. Floor effects as defined by Ahlquist (2018), do however seem strategic, as respondents should prefer to 'get off the floor' since answering 0 amounts to revealing the sensitive response while forfeiting confidentiality.

[16] In our version of this table we indicate the floor effects as defined both by Blair and Imai (2012) and Ahlquist (2018).

Table B.4 Respondent types, under different polarity assumptions

List experiment items	$S_i = 1$	$S_i = 0$
$Y_i(1) = 0$	(0,0) **(0,1)**	(0,0) (0,1)
$Y_i(1) = 1$	**(0,1)** (1,0)	**(0,1)** (1,0)
$Y_i(1) = 3$	(1,1) (3,0)	(1,1) (3,0)
$Y_i(1) = 2$	(1,1) (2,0)	(1,1) (2,0)
$Y_i(1) = 4$	(3,1) (4,0) **(4,1)**	(3,1) (4,0)
$Y_i(1) = 5$	(4,1)	(4,1)

This table shows possible respondent types in the treatment group for each observed count of list experiment experiment items ($Y(1) \in 0, ..., 5$). Respondent types (y,z) are defined by their latent count of control items $Y \in 0, ..., 4$ and their latent answer to the sensitive item $Z \in 0, 1$. Respondent types in bold face are floor or ceiling effect liars, under each assumption of the sensitive response $S \in 0, 1$.

B.5 Maximum Likelihood estimates and model comparison

In the following sections of the appendix we discuss the maximum likelihood estimation for the proportion of respondents with the latent sensitive trait[17] using the standard ML approach proposed by Blair and Imai (2012), as well as a number of alternative maximum likelhiood estimation models proposed by Blair et al. (2019) and Imai (2011). All models are estimated via the `ictreg` function in the `List` package for `R`. With the exception of DiM, all models include age, gender, education, employment and registered voter as covariates. The results for the standard ML estimation approach, used as the baseline comparison model, can be seen in Table B.5 below.

B.5 .1 Floor- and ceiling models

To test the presence of floor and ceiling liars in the data we use two maximum likelihood (ML) estimators proposed by Blair and Imai (2012) which model floor and ceiling effects respectively. These models fit a separate sub-model to the data which models the likelihood of ceiling- and floor liars among respondents. The results, seen in Tables B.6 -B.7 suggests against any substantial presence of ceiling and floor liars, with estimated proportions of both being 0% in the maximum likelihood estimation and 1% or less in the quasi-Bayesian estimation. This should suggest that respondents are not strategically avoiding revealing vote buying in the list experiment.

As the floor- and ceiling liars models are maximum likelihood models which are nested within the regular maximum likelihood model we can test whether or not using these models provide a better fit to the data using a standard likelihood ratio test between the restricted model (the regular ML model) and the unrestricted models (ceiling and floor liars models). The results, shown in Table B.8 , show that the likelihood ratio tests fail to reject the null of

[17] Although we are agnostic as to which is the sensitive response– having participated or not having participated in vote buying– we will adopt the convention of assuming that engaging in the behavior asked about in the survey question is equivalent to holding the latent "sensitive trait". That is, when we refer to respondents "with the sensitive trait" we mean those who truly have participated in vote buying.

Table B.5 Maxmimum Likelihood Model

	Sensitive item	Control items
(Intercept)	-0.89	-0.202
	(0.434)	(0.087)
Age	0.006	0
	(0.011)	(0.002)
Male_d	-0.16	0.109
	(0.254)	(0.051)
Edu_d	0.387	0.026
	(0.375)	(0.074)
Job_d	0.028	0.114
	(0.262)	(0.052)
RegVote_d	0.441	0.349
	(0.321)	(0.061)
Log Likelihood		-3659
AIC		7343
Est. prop. w. sensitive trait		0.410

Note: Standard errors in parentheses. Number of control items is set to 4.

Table B.6 Ceiling liars ML model

	Submodel		
	Sensitive item	Control items	Ceiling liars
(Intercept)	-0.89	-0.202	-0.55
	(0.434)	(0.087)	(1.897)
Male_d	-0.16	0.109	-0.013
	(0.254)	(0.051)	(1.764)
Age	0.006	0	-0.398
	(0.011)	(0.002)	(0.631)
Edu_d	0.387	0.026	-0.003
	(0.375)	(0.074)	(1.767)
Job_d	0.028	0.114	-0.011
	(0.262)	(0.052)	(1.764)
RegVote_d	0.441	0.349	-0.024
	(0.321)	(0.061)	(1.764)
Log Likelihood		-3687	
AIC		7353	
Quasi-Bayesian Liar cond. prob.		0.243	
Quasi-Bayesian Liar pop. prop.		0.010	
ML Liar cond. prob.		0.000	
ML Liar pop. prop.		0.000	
Est. prop. w sensitive trait		0.410	

Note: Standard errors in parentheses. Number of control items set to 4.

equal fit with p-values in excess of 0.84. This implies that we have no evidence to suggest that the floor and ceiling liar models fit the data better than the regular ML model.

Table B.7 Floor liars ML model

	Submodel		
	Sensitive item	Control items	Floor liars
(Intercept)	-0.891	-0.202	-0.465
	(0.434)	(0.087)	(1.859)
Male_d	-0.16	0.109	-0.011
	(0.254)	(0.051)	(1.764)
Age	0.006	0	-0.345
	(0.011)	(0.002)	(0.657)
Edu_d	0.387	0.026	-0.002
	(0.375)	(0.074)	(1.767)
Job_d	0.028	0.114	-0.007
	(0.262)	(0.052)	(1.765)
RegVote_d	0.441	0.349	-0.021
	(0.321)	(0.061)	(1.764)
Log Likelihood		-3687	
AIC		7353	
Quasi-Bayesian Liar cond. prob.		0.295	
Quasi-Bayesian Liar pop. prop.		0.005	
ML Liar cond. prob.		0.000	
ML Liar pop. prop.		0.000	
Est. prop. w sensitive trait		0.410	

Note: Standard errors in parenthesis. Number of control items set to 4

Table B.8 Likelihood ratio tests between the ML model and floor and ceiling liar models

	log.lik0	log.lik1	statistic	df	p
Ceiling liar model	-3659.921	-3658.585	2.673	6	0.849
Floor liar model	-3659.921	-3658.572	2.699	6	0.846

B.5 .2 Uniform and Top-biased error models

To test for the presence of uniform error and top-bias, i.e. respondents who hide their true answer to the list experiment by either picking a random value (uniform error coders) or by picking the maximum answer (top-coders), we use two maximum likelihood estimators proposed by Blair et al. (2019). These models estimate, apart from the effect of the covariates on the sensitive item and the control item, the likelihood of respondents being uniform error coders and top-coders respectively. The results for these models are seen in Tables B.9 -B.10 .

The results from these models highlight a number of interesting patterns. First, looking at the uniform error model, we can see the model does indeed identify a substantial proportion of uniform error coders. Especially interesting here is that the ML model separates the proportion of uniform coders between the treatment and control group, and we see that in the control group there is no evidence of uniform error coders, with an estimated proportion of 0%. In the treatment group, on the other hand, the estimated proportion of uniform error coders is 27.5% which is a substantial proportion. Uniform error is usually

Table B.9 Uniform error model

	Submodel	
	Sensitive item	Control items
(Intercept)	-14.498	-0.189
	(499.812)	(0.09)
Age	0.014	0
	(0.022)	(0.002)
Male_d	0.053	0.105
	(0.55)	(0.053)
Edu_d	0.921	0
	(0.701)	(0.079)
Job_d	0.087	0.142
	(0.587)	(0.054)
RegVote_d	12.731	0.339
	(499.812)	(0.061)
Log Likelihood	-3625	
AIC	7278	
Est. prop. with uniform error (control)	0.000, 95% CI: {0.000,0.000}	
Est. prop. with uniform error (treated)	0.275, 95% CI: {0.210,0.350}	
Est. prop. w sensitive trait	0.206	

Note: Standard errors in parentheses. Number of control items is set to 4.

Table B.10 Top biased error model

	Submodel	
	Sensitive item	Control items
(Intercept)	-14.498	-0.189
	(499.812)	(0.09)
Age	0.014	0
	(0.022)	(0.002)
Male_d	0.053	0.105
	(0.55)	(0.053)
Edu_d	0.921	0
	(0.701)	(0.079)
Job_d	0.087	0.142
	(0.587)	(0.054)
RegVote_d	12.731	0.339
	(499.812)	(0.061)
Log Likelihood	-3552	
AIC	7130	
Est. prop. of top coders	0.086, 95% CI: {0.073,0.101}	
Est. prop. w sensitive trait	0.210	

Note: Standard errors in parentheses. Number of control items is set to 4.

portrayed as nonstrategic, resulting for example from respondents providing random answers in order to "satisfice" (Blair et al., 2019, p. 17). Yet, the fact that uniform error only occurs in the treatment group in our survey suggests that respondents may be strategically choosing a random response as way of concealing their true sensitive response. In other words, the presence of uniform error could indicate sensitivity bias, rather than design failure.

The results for the top-biased error model also show a substantive proportion of respondents as 'top-coders' with an estimated proportion of 8.6%. Blair et al. (2019) argue that top-biased error should be uncommon, since choosing the maximum response is equivalent to forfeiting confidentiality and revealing a socially undesirable behavior. Under the assumption that choosing the maximum response reveals an undesirable behavior, top-biased error would suggest design failure in the form of respondent inattentiveness, misunderstanding or a technical error in the administration or coding of the survey. However, as we have noted in our discussion of floor- and ceiling-liars, choosing the maximum response maybe be strategic under the assumption that doing so reveals a desirable behavior.

If we look at the estimated proportion of respondents with the sensitive trait, we can see that this proportion is substantially lower (21%) for these two ML models compared with the regular ML model (41%) and even much lower than the DiM measure (29%). This is especially interesting when we consider that the DiM measure is biased towards overestimation in the face of these error processes (Ahlquist, 2018).

As the uniform and top-biased error models are nested within the regular ML model, we can conduct standard likelihood ratio tests to make pairwise comparisons about whether the uniform and top-biased error models fit the data better than the regular ML model. The results from these LR tests can be seen in Table B.11 below, which shows that the null hypothesis of an equal fit is rejected with $p<0.001$ in favor of the alternative hypothesis that the error models provide a better fit. As the uniform and top-biased error models are not nested in one another we cannot compare the fit between these two models with a standard LR test. Looking at the AIC for the two models, it is however, clear that the top-biased error model provides a better fit for the data, given the number of parameters.[18]

Table B.11 Likelihood ratio tests between the ML model and uniform and top-biased error models

	log.lik0	log.lik1	statistic	df	p
Uniform error model	-3659.921	-3625.157	69.528	2	¡0.001
Top-biased error model	-3659.921	-3552.193	215.457	1	¡0.001

B.5 .3 Unconstrained model

Another possible concern is simply that there is a strong correlation between having the sensitive trait and answering affirmatively to the control items. This is not a violation of list experiment assumptions but it could cause the standard ML estimator to be biased. In these cases it is possible to fit an unconstrained

[18] This result also holds for the BIC which more heavily penalizes the number of parameters in the model.

ML model which allows for separate estimation of the parameters for the control items, depending on if the respondent has or does not have the sensitive trait. The results for this unconstrained model are shown in Table B.12 .

Table B.12 Unconstrained Maximum Likelihood Model

		Submodel	
variable	Sensitive item	Control items0	Control items1
(Intercept)	-3.008	-0.228	4.674
	(0.384)	(0.094)	(1.422)
Age	0.005	0	-0.011
	(0.009)	(0.003)	(0.013)
Male_d	0.517	0.015	-0.851
	(0.218)	(0.054)	(0.59)
Edu_d	1.269	-0.148	-1.663
	(0.355)	(0.104)	(0.554)
Job_d	-0.378	0.214	0.408
	(0.211)	(0.056)	(0.412)
RegVote_d	0.951	0.266	-1.291
	(0.294)	(0.064)	(1.261)
Log Likelihood		-3548	
AIC		7133	
Est. prop. w sensitive trait		0.152	

Note: Standard errors in parentheses. Number of control items is set to 4.

These results show that there is clearly a strong correlation between having the sensitive trait and the answers to the control items. This is most evident in the difference between the intercepts for the submodels for the control items in group 0 (without the sensitive trait) and group 1 (with the sensitive trait). Noteworthy is that the estimated proportion of respondents with the sensitive trait is lowest for this unconstrained ML model with an estimated proportion at 15.2%.

What this means in practice is that the unconstrained ML model suggests that individuals who have participated in vote buying also were more likely to answer in the affirmative to the control items. This may suggest that a large proportion of the 'top-coders' identified in the top-biased error model may in fact be individuals who participated in vote buying.

As the unconstrained model is nested within the regular ML model, we can conduct a likelihood-ratio test to test the null hypothesis of equal fit. The test soundly rejects the null of an equal fit in favor of the alternative that the unconstrained model fits the data better, suggesting that there is indeed a correlation between the sensitive trait and the control items.[19]

[19]Whether the unconstrained model is *strictly* nested within the ML model can be debated. However, if the unconstrained model is not considered to be nested within the standard ML model the conclusion that the unconstrained model fits the data better is still supported by the substantially lower AIC of the unconstrained model. The conclusion also holds if we use the BIC, which more severely penalises the number of parameters.

Table B.13 Likelihood ratio tests between the ML model and unconstrained ML model

	log.lik0	log.lik1	statistic	df	p
Unconstrained ML model	-3659.921	-3548.731	222.38	6	¡0.001

B.5 .4 Model fit comparison

The evidence from the sections above suggest that the standard, constrained, maximum likelihood estimator for the proportion of respondents with the sensitive trait does not fit the data equally well as the error models (section B.4.2) or the unconstrained maximum likelihood estimator (section B.4.3). When comparing these competing models with the standard maximum likelihood estimator, we could use likelihood ratio tests to test whether the more expansive models fit the data better than the regular maximum likelihood estimator. To select the best fitting model we cannot, however, continue to use likelihood ratio tests between the three models which provide a better fit than the standard model, since these three models are not nested in one-another. To compare these models, we instead turn to the estimated Akaike- and Bayesian Information Criteria (AIC and BIC) (Konishi and Kitagawa, 2008) which are based on the log-likelihood of the model, but penalizes the score based on the number of parameters to the model.[20] The AICs and BICs of the tested models can be seen in Table B.14 below.

Table B.14 AIC and BIC comparison

Estimator	log.lik	npar	AIC	BIC
Maximum likelihood	-3659.9	12	7343.8	7413
ML Floor liars	-3658.6	18	7353.1	7457
ML Ceiling liars	-3658.6	18	7353.2	7457
ML top-biased error	-3552.2	13	7130.4	7205.4
ML Uniform error	-3625.2	14	7278.3	7359.1
ML Unconstrained	-3548.7	18	7133.5	7237.3

The results from this analysis shows that the ML top-biased error estimator fits the data best on both AIC and BIC, followed by the unconstrained ML model. As we know that the DiM estimator for the proportion of respondents with the sensitive trait is an overestimate in the presence of top-biased errors, this result should give us some confidence that the DiM estimate in our case is truly an overestimate of the true proportion.

[20]BIC provides a harsher penalty when $log(n) > 2$, i.e. when the number of observations is greater than 7.

B.6 Subgroup comparisons for differences in sensitivity bias polarity

To explore our suspicion that there may be differences in polarity across different sub-groups of respondents we investigated the differences in the estimated sensitivity bias across a range of subgroups. The results from this analysis can be seen in figures B.2 -B.8 . In the sample as a whole we observe over-reporting: a larger proportion of respondents report vote buying in the direct measure compared to in the list experiment. However, there may be sub-groups who instead under-report, in which case the difference between the direct measure and the list experiment measure should be negative for these subgroups. It is important to keep in mind that this sub-group analysis is exploratory. The experiment is neither designed to detect these differences in polarity, nor sufficiently powered to do so. In addition, detecting differing polarities cannot resolve the problems for the maximum likelihood models evaluated in section B.5. We consider the detection of differing polarities as a further problem in the analysis of list experiments which merits careful consideration in future research. Nonetheless, we find it useful to consider sub-group differences in order to better understand the puzzling aggregate results in the experiment.

The results of the sub-group analysis does indeed suggest that there are sub-groups where sensitivity bias is in the opposite direction compared to the majority of respondents. This does particularly seem to be the case for the Yoruba ethnic group which under-report vote buying. The result also holds for the South-West region, where the vast majority (94.8%) of respondents are Yoruba. Similarly, in the Lagos region, the difference in between the two measures is near zero. This may, however, be driven by the fact that a substantial proportion (65.7%) of respondents in Lagos are Yoruba which may indicate that this null difference is not due to vote buying being non-sensitive in Lagos but due to different polarity of the sensitivity bias of the respondents. Looking specifically at the respondents from Lagos (figure B.4), we can see that the Yoruba in Lagos have negative point estimates while other ethnicities in Lagos have positive point estimates. While these results are not statistically significant (the survey is not powered for this), this observation further highlights the importance of investigating the assumption of uniform polarity in list experiments.

There also seems to be an effect on the polarity of the sensitivity bias depending on what actor the respondent believes is sponsor of the survey. Among the small number of respondents (69) who believe that a media organization is the sponsor of the survey, vote buying is vastly under-reported in the direct measure compared to the list experiment difference in means measure. Similarly, for respondents who believe that a political actor is the sponsor of the survey, the point estimate indicate that vote buying is under-reported in the direct measure, although this is not statistically different from zero. That the perceived sponsor of the survey affects the respondents behavior is in line with recent research on the topic (Isani and Schlipphak, 2022; Blair et al., 2020).

Subgroup differences along ethnic groups and perceived survey sponsor support our central findings which suggest that the misreporting we observe is strategic, rather than design failure. Indeed, a core assumption of the "social reference theory" of sensitivity bias is that individuals adjust their responses according to who they believe can access the data and according to the likely consequences (positive or negative) associated with their response (Blair et al., 2020, p. 1299). It follows that one observable indicator of sensitivity bias – as opposed to inattentiveness– is that misreporting should vary by perceived survey sponsor.

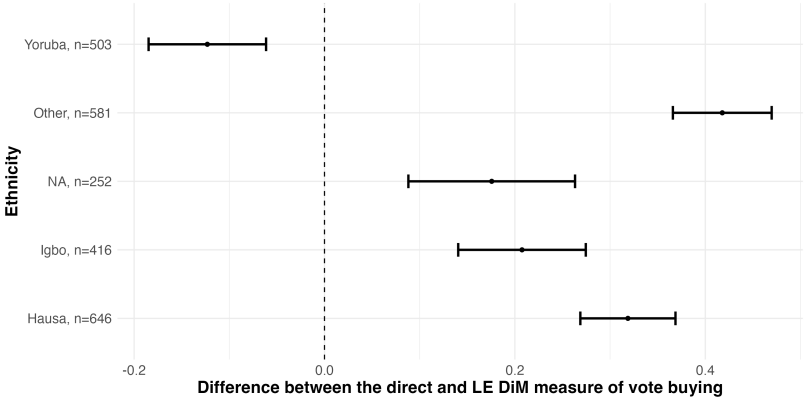

Fig. B.2 Difference between estimated vote buying according to the direct and list experiment difference in means measure across ethnic groups with 95% confidence intervals

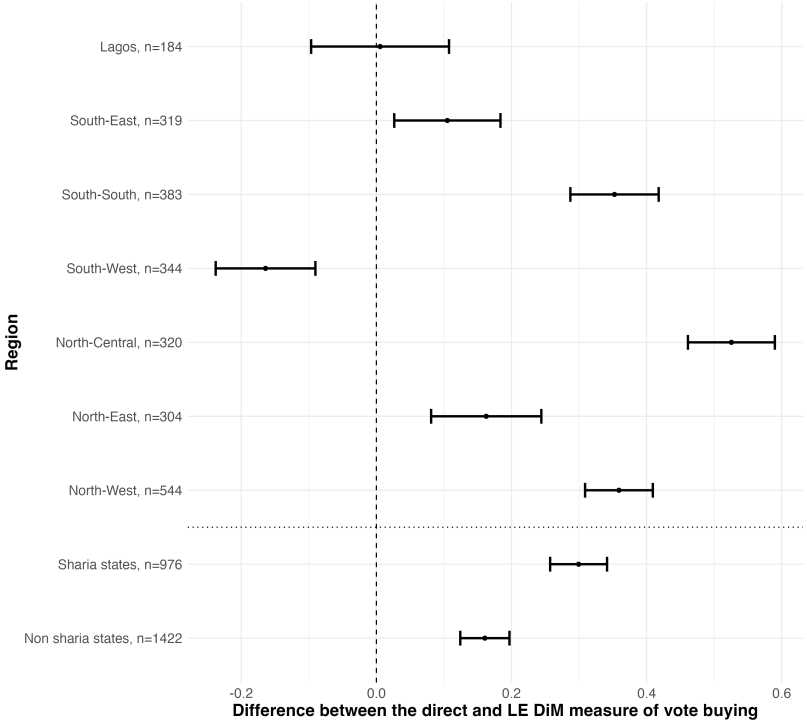

Fig. B.3 Difference between estimated vote buying according to the direct and list experiment difference in means measure across regions with 95% confidence intervals

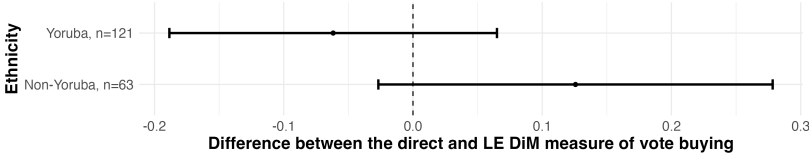

Fig. B.4 Difference between estimated vote buying according to the direct and list experiment difference in means measure across ethnic groups in Lagos with 95% confidence intervals

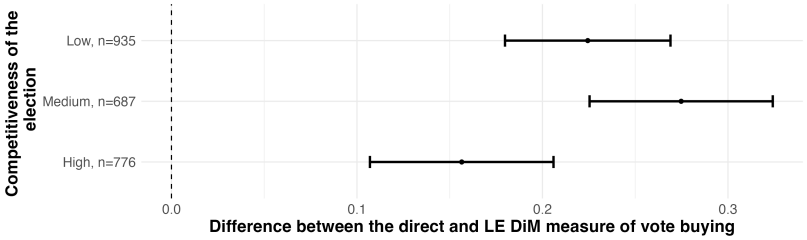

Fig. B.5 Difference between estimated vote buying according to the direct and list experiment difference in means measure across level of competitiveness in the electoral district with 95% confidence intervals

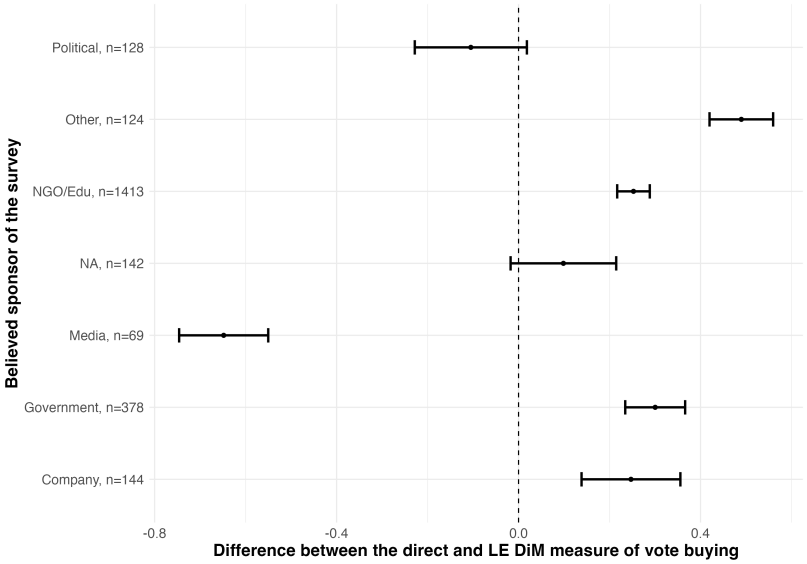

Fig. B.6 Difference between estimated vote buying according to the direct and list experiment difference in means measure across respondent's belief of who sponsored the survey with 95% confidence intervals

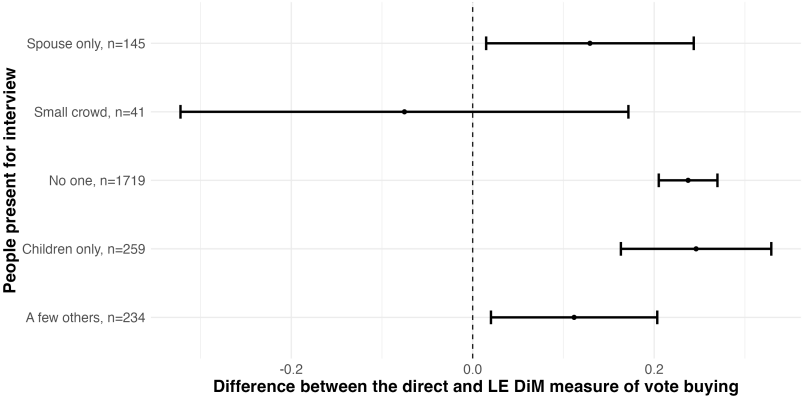

Fig. B.7 Difference between estimated vote buying according to the direct and list experiment difference in means measure across whether other people were present at the time of the interview with 95% confidence intervals

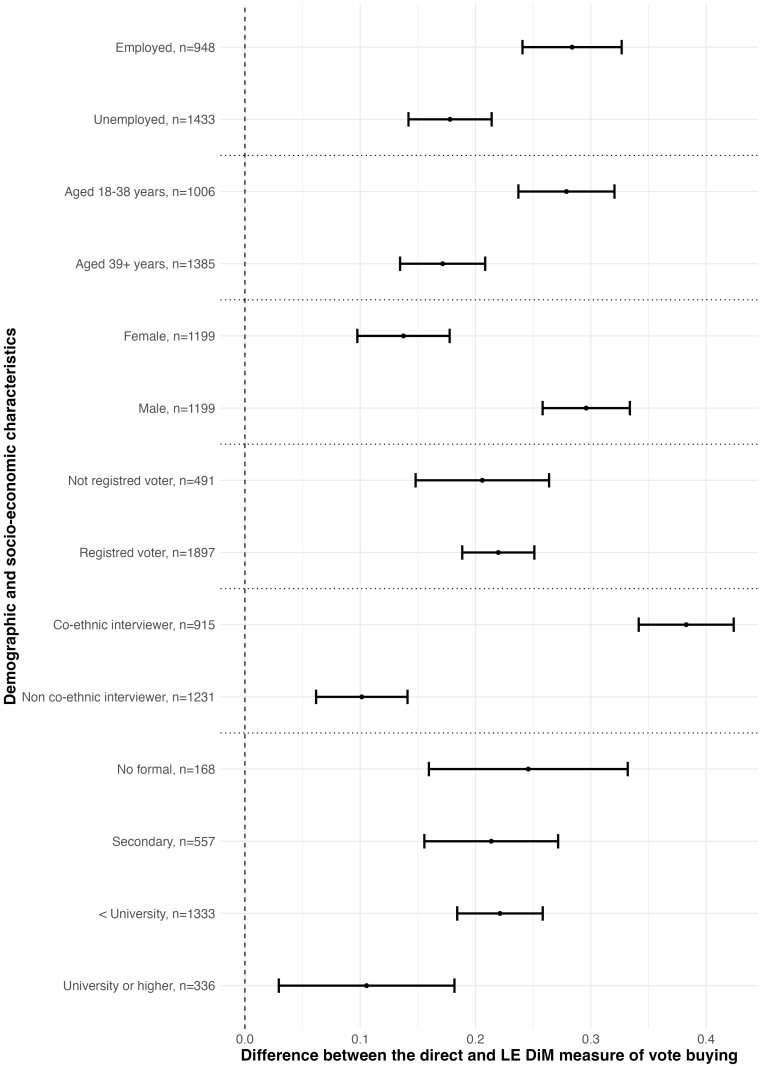

Fig. B.8 Difference between estimated vote buying according to the direct and list experiment difference in means measure across demographic and socio-economic characteristics with 95% confidence intervals

Appendix C Full text of survey questions

Note: The full text of the survey questions used in analysis is shown below. The variable names given in bold preceding each question were not read to respondents. Respondents were randomly assigned to be read either List A or List B in the list experiment.

List experiment:
I am going to read you a list of things that people have told us they experienced during the 2019 election campaign. I would like you to tell me how many of these things you have personally experienced. Please, do not tell me which ones, only HOW MANY. [If you would like me to repeat the list, I will do so.]

List A:

- Politicians put up posters or signs in the area where you live.
- You read the newspaper almost every day to learn about the campaign.
- You met a politician personally to discuss his or her candidacy.
- You discussed the campaign with friends or family.

List B:

- Politicians put up posters or signs in the area where you live.
- You read the newspaper almost every day to learn about the campaign.
- You were offered money from a party or politician to vote in a particular way.
- You met a politician personally to discuss his or her candidacy.
- You discussed the campaign with friends or family.

Direct measure:
Thinking about the 2019 elections, we are interested in whether a party, politician or their representatives tried to persuade you to vote in a particular way. Please let us know if a party or politician did any of the following:

- Provided positive information about their party/candidate
- Promised implementation/removal of a certain policy (ies) after elections
- Offered money during the campaign or on election day
- Offered food or other personal benefits during the campaign or on election day
- Provided negative information about other parties/candidates
- Threatened violence or intimidation during the campaign or on election day
- Threatened violence or intimidation after the election if you did not vote in a particular way

Gender: [Coded by interviewer, options not read]

- Male
- Female

Age:
How old are you?

Education:
What is your highest level of education? [Coded by interviewer from response, options not read]

- No formal schooling
- Informal schooling only (including Koranic schooling)
- Some primary schooling
- Primary school completed
- Intermediate school or Some secondary school / high school
- Secondary school / high school completed
- Post-secondary qualifications, other than university e.g. diploma or degree from a polytechnic or college
- Some university
- University completed
- Post-graduate
- Do not know
- Refused

Employment:
Do you have a job that pays a cash income?

- Yes, full-time
- Yes, part-time
- No
- Don't know
- Refused

Registered voter:
Were you registered to vote in the 2019 elections?

- Yes
- No