北京大学中国经济研究中心
China Center for Economic Research
讨论稿系列
Working Paper Series

E2025008 2025-10-20

Design Effects in the Developing Context: Lessons from a Paradoxical List Experiment on Corruption in Egypt

Ashraf Momtaz
Shiyao Liu

Abstract:
In 2014, Egypt launched the National Anti-Corruption Strategy (NACS) as part of constitutional commitments. To evaluate NACS, we conducted a survey to measure Egyptian residents' perceived levels of corruption and their personal experiences with corruption. We addressed social desirability bias, which can distort direct survey responses with a list experiment, the first of its kind in Egypt. Analysis of the paradoxical results from the list experiment revealed a violation of the "no design effect" assumption in the Egyptian context. We propose that the cognitive load caused by the irrelevance of control questions to the survey theme may have led to under-reporting of non-sensitive items, resulting in an overestimation of corruption prevalence. This finding complements prior research on design effects in list experiments: unlike underestimates found in Kenya's list experiment (Kramon and Weghorst, 2019) or bias from inattentive respondents (Blair et al., 2019), Egypt's data showed overestimates, indicating a different error mechanism. We provided practical advice for conducting list experiments in developing contexts.

Keywords: list experiment, corruption, Egypt, design effect

Design Effects in the Developing Context: Lessons from a Paradoxical List Experiment on Corruption in Egypt

Ashraf Momtaz[1]

Shiyao Liu[2] (corresponding author[3])

Abstract

In 2014, Egypt launched the National Anti-Corruption Strategy (NACS) as part of constitutional commitments. To evaluate NACS, we conducted a survey to measure Egyptian residents' perceived levels of corruption and their personal experiences with corruption. We addressed social desirability bias, which can distort direct survey responses with a list experiment, the first of its kind in Egypt. Analysis of the paradoxical results from the list experiment revealed a violation of the "no design effect" assumption in the Egyptian context. We propose that the cognitive load caused by the irrelevance of control questions to the survey theme may have led to under-reporting of non-sensitive items, resulting in an overestimation of corruption prevalence. This finding complements prior research on design effects in list experiments: unlike underestimates found in Kenya's list experiment (Kramon and Weghorst, 2019) or bias from inattentive respondents (Blair et al., 2019), Egypt's data showed overestimates, indicating a different error mechanism. We provided practical advice for conducting list experiments in developing contexts.

Keywords: list experiment, corruption, Egypt, design effect

Introduction

In 2014, Egypt launched a National Anti-Corruption Strategy (NACS) as part of its constitutional commitment. The Egyptian President initiated the NACS to demonstrate the highest-level political leadership's strong commitment to enforce it decisively. As part of a

[1] Institute of South-South Cooperation and Development, Peking University

[2] China Center for Economic Research, Institute of South-South Cooperation and Development, National School of Development, Peking University;

[3] Authorship listed in an alphabetical order. Please address correspondence to shiyaoliu@nsd.pku.edu.cn.

thorough evaluation of the strategy, we conducted an in-person survey to assess (a) the current perceived level of corruption, and (b) personal experiences with corruption among Egyptian residents.

Asking direct questions about respondents' perceived corruption or personal experiences with corruption is less effective because corruption is often seen as a sensitive topic. Respondents are therefore influenced by social desirability bias (Blair et al., 2020; Agerberg, 2022), the tendency to give answers that align with social norms, which can lead to under-reporting of corrupt experiences. Methodologists introduce survey experiments, such as the randomized response technique (Warner, 1965; Tourangeau and Yan, 2007), list experiments (Miller, 1984; Imai, 2011), and conjoint experiments (Hainmuller et al., 2014), to alleviate such concerns. Among these, being logistically more straightforward to implement, the list experiment is one of the most frequently used. Indeed, the World Value Survey included a list experiment in its Wave 7 (2017-2021) questionnaire on political trust in 10 pilot countries, and is expanding it to all sampling countries and regions in Wave 8 (2024-2026).

We, to the best of our knowledge, implemented the first list experiment in Egypt. We randomly assign respondents to two versions of the questionnaires, where all but one question are identical.

Respondents in the control group are asked the following question:

> Here is a list of different situations that you might have encountered or experienced during the past 12 months. Please read the list carefully and indicate how many of these situations you have encountered or experienced. Do not indicate which situations; only how many situations. Please select a number from 0 to 4.
>
> - Spend a vacation at North Coast.
> - Search for a new job.
> - Buy a cell phone.
> - Attend a relative's or friend's wedding.

Respondents in the treatment group are asked the identical question, except that a sensitive item on corruption experience had been added to the list.

> Here is a list of different situations that you might have encountered or experienced during the past 12 months. Please read the list carefully and indicate how many of these situations you have encountered or experienced. Do not indicate which situations; only how many situations. Please select a number from 0 to 5.
>
> - Spend a vacation at North Coast.
> - Search for a new job.
> - Buy a cell phone.
> - Attend a relative's or friend's wedding.

- Being asked to pay a bribe to a public official.

Under the assumptions of (1) randomized assignments, (2) no design effects, and (3) no liars, the difference in means of the reported count between the treatment group responses and the control group responses points to the proportion of all respondents who agree with the added sensitive item (Imai, 2011; Blair and Imai, 2012).

The assumption of randomized assignment requires respondents to be randomly assigned to answer either of the two versions of the questions. It ensures that respondents in the treatment and control groups are otherwise identical. As a result, any differences in means of the reported count should be attributed solely to the different versions of the question asked.

The no design effect assumption states that respondents give the same count for non-sensitive items whether they are asked the treatment or the control version of the question. Therefore, differences in the means of the reported counts are caused by the additional (sensitive) item included in the treatment version. The no liars assumption presumes that respondents report an honest count in the treatment group. Consequently, the differences in means can be seen as the proportion of respondents who agree with the additional (sensitive) item.

The randomized assignment assumption typically holds by design, because empirical researchers randomize respondents into different versions of the questionnaire. The no liars assumption, though untestable, is of less concern. A violation of the no liars assumption indicates that respondents may still under-report the total count of statements they agree with, possibly due to the social desirability bias, despite being in the treatment group. In our setting, this means that respondents who have been shown the treatment version of the question may still be reluctant to include the statement in the total count if they have been asked to pay a bribe. However, the level of under-reporting in the list experiment should still be lower than that in direct questioning. Thus, despite the possible failure of the no liars assumption, the list experiment still alleviates, though not necessarily completely removes, the social desirability bias for the sensitive item.

We focus on the no design effect assumption in this paper. We document the fact that the no design effect assumption does NOT hold in our enumeration of the list experiment in Egypt. We further present suggestive evidence that the cognitive load induced by the control version of the question likely causes the design effect. Since the control version does not include the sensitive item, respondents may be confused and thus be distracted when they suddenly encounter a question about daily life that seems unrelated to the survey's central theme. Such a distraction may cause them to under-report the number of statements they agree with for non-sensitive terms, which leads to the violation of the no design effect assumption.

Our study complements previous studies on the violation of design effects in list experiments. Kramon and Weghorst (2019) demonstrate, using data from Kenya, that less

numerate and less educated respondents are more prone to comprehension and reporting errors, due to the length and complexity of the list experiments. They call a list experiment a breakdown when it produces a lower estimate compared with direct questioning for the proportion of respondents who agree with the sensitive statement. Our data show the opposite breakdown, where the estimation from the list experiment exceeds its theoretical maximum. The different direction indicates a different error mechanism: Kramon and Weghorst (2019) attribute the misreporting to the general complexity of the list experiment, applicable to both those in the treatment group and the control group. As discussed later, we focus on the one-sided misreporting, where respondents in the control group, but not those in the treatment group, face a higher cognitive load.

Meanwhile, Blair et al. (2019) and Agerberg and Tannenberg (2021) posit that inattentive respondents who provide random answers to questions may naturally report a higher count for the treated version of the question. As a result, the difference in means estimator may thus be contaminated, a result of the failed no design effect assumption. However, the magnitude of bias observed in our dataset is larger than the measurement error caused by inattentive respondents can account for.

Egyptian Context and Sampling of Respondents

Egypt is situated in North Africa and the Middle East, with an estimated population of 114 million and a GDP per capita of US$3,457 in 2023, according to the World Development Indicators (World Bank, 2025). Administrative corruption remains a persistent problem in Egypt, significantly impacting the country's economic growth. Transparency International's (TI) 2023 Corruption Perceptions Index ranks Egypt 108th out of 180 countries and regions, with a score of 35/100 (Transparency International, 2023). For comparison, the United States was ranked 24th, Malaysia 57th, Morocco 97th, Ukraine 104th, and Mozambique 145th. Among 54 African countries listed, Egypt is ranked 23rd. In 2018, 84.8% of the surveyed Egyptian residents gave a rating of 7 or above on a 10-point scale for the level of corruption in Egypt in the World Value Survey (Haerpfer et al, 2022). Notably, 45.7% indeed gave Egypt a 10 out of 10 (“there is abundant corruption in my country”). Data from the Afrobarometer report that among survey Egyptian residents who had been in contact with a public school, a public clinic or hospital, a government office for documents/licenses, water, sanitation, or electric services, the police, or the court, 52% ever paid a bribe for these services in 2015 (Afrobarometer, 2017).

Enterprise surveys conducted by the World Bank in 2020 revealed a similar pattern: although only 5 percent of firms reported experiencing at least one bribe request, 41 percent stated that they had been asked or expected to give gifts or informal payments to obtain a construction permit (World Bank, 2022). Corruption was listed as the third most common obstacle to business in Egypt, after only tax rates and political stability.

Table 1 Summary of Demographics and Balance Table

	Mean	Mean	N	N	t-statistic	p-value

	Control	Treatment	Control	Treatment		
Female=1	0.328	0.364	250	250	-0.845	0.399
Age (20-29)	0.212	0.228	250	250	-0.431	0.667
Age (30-39)	0.256	0.296	250	250	-0.999	0.318
Age (40-49)	0.236	0.212	250	250	0.643	0.521
Age (50-59)	0.148	0.132	250	250	0.515	0.607
Age (60-69)	0.116	0.096	250	250	0.725	0.469
Age (>=70)	0.032	0.036	250	250	-0.246	0.806
Edu (Pre-University)	0.236	0.244	250	250	-0.209	0.835
Edu (Technical Edu)	0.600	0.636	250	250	-0.827	0.408
Edu (University and above)	0.164	0.12	250	250	1.409	0.159
Monthly Income (~$US 21~104)	0.172	0.168	250	250	0.119	0.905
Monthly Income (~$US 104~208)	0.384	0.456	250	250	-1.632	0.103
Monthly Income (~$US 208~312)	0.312	0.244	250	250	1.698	0.09
Monthly Income (>=$US 312~)	0.132	0.132	250	250	0	1
City (Cairo)	0.404	0.400	250	250	0.091	0.928
City (Alexandria)	0.320	0.280	250	250	0.975	0.33
City (Suez)	0.160	0.160	250	250	0	1
City (Asyut)	0.116	0.160	250	250	-1.426	0.154

Note: the null hypothesis for the t-test is the equality between the treatment and the control means

We conducted our survey through face-to-face interviews using paper forms. The survey was administered in Arabic. Table 1 summarizes the demographic information of our respondents. Our sample includes 500 respondents, with 250 randomly assigned to the treatment version of the survey and the remaining 250 to the control version. By design, our study covers Egypt's four largest cities: Cairo, Alexandria, Suez, and Asyut. To ensure respondent diversity, we intentionally select participants from the government, business community, civil society, legal sectors, and ordinary citizens. Our sample also spans different ages, genders, educational backgrounds, and income levels. Although the sample itself is not a probability sample and therefore not nationally representative, the diverse backgrounds of the respondents provide valuable insights into how residents in Egypt might respond to the list experiment. Additionally, we note that the list experiment is itself an experiment, ensuring the internal validity of the study by design.

Further, we conduct a series of t-tests against the null hypotheses that the means of demographic characteristics of the treatment group are equal to those of the control group.

The t-tests fail to reject the null hypotheses, which indicates a good pre-treatment balance between the treatment and control groups. It also demonstrates that the randomized assignment assumption is likely satisfied in our dataset.

Survey Design

Figure 1 shows our survey flow. We ask respondents about their perceptions of the corruption in Egypt, and then present the list experiment. In the first component, we mainly adopt the direct questioning, where respondents are asked directly about their perceptions on the level of corruption in Egypt, their personal experience of witnessing or being engaged in corruption in the past 12 months before the survey, to what extent they consider extra payment is necessary and/or acceptable for public services, and whether an intermediary is involved to obtain public services. We further asked the respondents to evaluate the prevalence of different types of corruption, such as bribery, nepotism, abuse of power, and embezzlement, in public sectors, as well as to rate the level of corruption in different sectors in Egypt. For the list experiment, we print two versions of questionnaires in advance, and randomly distribute them to the respondents to ensure the random assignment.

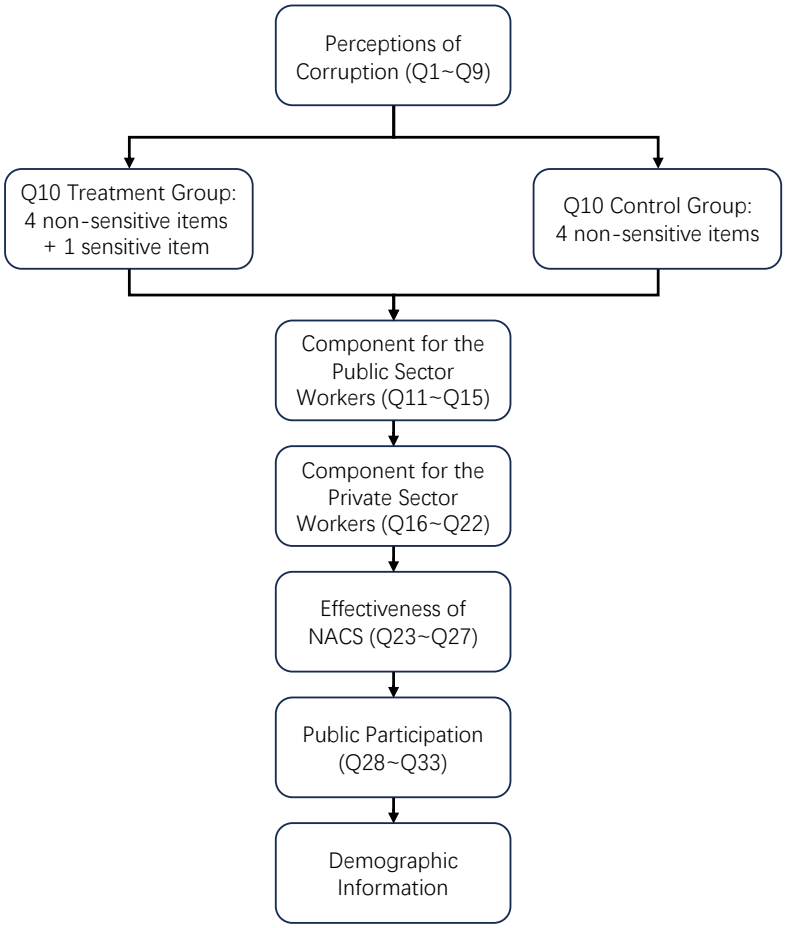

Figure 1 Survey Flow

After the list experiment, we gathered responses based on participants' sectors. For public sector workers, we asked whether their employer had implemented codes of conduct or

training courses, and we also inquired about their perceptions of how many of their colleagues are involved in corruption. For private sector workers, we collected data on the size of their business, their experiences with encounters involving a bribery solicitation by officials, their internal control practices, and which bureaucratic procedure they consider the most corrupt.

The effectiveness of the NACS component asks respondents about their awareness of the program and their evaluation of its effectiveness. The public participation component inquired about the respondent's beliefs and actions in civic engagement for the anti-corruption campaign. Demographic information was collected at the end of the survey.

Paradoxical Results from the List Experiment

To benchmark the results from our survey experiments, we first report respondents' answers regarding their personal experience of witnessing or being involved in corruption in the past 12 months before the survey in Table 2. Direct questions reveal that around 65% to 70% of the surveyed have witnessed or been involved in corruption in Egypt in the past 12 months.

Table 2 Summary from Direct Questions

Question	Proportion of "Yes"	N
In the past 12 months, have you personally witnessed or experienced any form of corruption in Egypt?	71.2%	500
In the past 12 months, have you had to pay extra money or give a gift in exchange for a public service?	64.4%	500

As shown by Blair and Imai (2012), under the assumptions of randomized assignments, no design effects, and no liars, the difference in means estimator between the average counts reported by the treatment group and the control group identifies the proportion of respondents in agreement with the additional (sensitive) item. We implemented this estimator by the following linear regression:

$$Y_i = \alpha + \beta T_i + \epsilon_i.$$

Column 1 of Table 3 shows our baseline result. Respondents in the control group report, on average, that they have experienced 1.844 out of the 4 non-sensitive activities (vacation at North Coast, search for a new job, purchase of a cell phone, and attendance of a relative's or friend's wedding) within the past 12 months. Meanwhile, respondents in the treatment group report, on average, that they have experienced 3.316 out of the 5 activities (the 4 non-sensitive ones plus the sensitive corruption experience). The difference between the two groups, 1.472, when the assumptions above are satisfied, should be attributed to the additional sensitive item in the survey. Thus, conventionally speaking, it should be interpreted to mean that 147% of respondents had been asked to pay a bribe to a public official in the past 12 months.

Table 3 List Experiment Estimation for Involvement in Corruption

	Dependent variable:			
	Count			
	(1) All	(2) Low Edu	(3) Mid Edu	(4) High Edu
LE (β)	1.472***	1.350***	1.729***	0.822***
	(0.109)	(0.217)	(0.132)	(0.291)
Constant (α)	1.844***	1.831***	1.567***	2.878***
	(0.072)	(0.135)	(0.085)	(0.165)
Observations	500	120	309	71
Adjusted R^2	0.267	0.240	0.355	0.097
p-value from Blair and Imai (2012)	<0.0001	0.0164	<0.0001	1
Note:				*p<0.1; **p<0.05; ***p<0.01
	HC2 heteroskedasticity-robust standard errors in parentheses.			

Obviously, it could not be the case that more than 100% of the respondents have been asked for bribery in the past 12 months. In theory, the proportion should be upper-bounded by 1, while in the data, the estimator gives 1.472. A t-test where the null hypothesis is $\beta < 1$ reports a p-value < 0.0001.

As a result, it must be the case that one of the three stated assumptions is incorrect. The balance of demographic information, as shown in Table 1, across the treatment and control, confirms the randomized assignment assumption. The no liear assumption, if unsatisfied, would generally produce a downward bias, as respondents tend to under-report the count number when they see the sensitive item in the treatment group. Thus, it must be the case that the no design effect assumption fails.

We further conduct a statistical test of two first-order stochastic dominance relationships proposed by Blair and Imai (2012), against the null hypothesis that $E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$, i.e. the average count reported by the treatment group, if they were not assigned to treatment, equals the average count reported by the control group. We implemented the test with the statistical package *list* in R provided by the authors. Results show that the null hypothesis of no design effect can be rejected at a p-value < 0.0001. The strong statistical evidence against the non-existence of the design effect suggests a violation of this assumption.

Possible Reasons for the Design Effect

Measurement error caused by inattentive respondents has previously been identified as a reason for the possible existence of the design effect (Blair et al, 2019). If inattentive respondents are assumed to give answers randomly to all questions, they would, on average, respond with a count of 2 for the control question and 2.5 for the treatment question. Consequently, the mean difference between the treatment and control groups among

inattentive respondents would be 0.5, regardless of whether they agree with the sensitive or non-sensitive items. This means the final difference-in-mean estimator will be a weighted average of the true proportion of respondents agreeing with the sensitive item and 0.5. In fact, the bias equal $p(e-\tau)$, where p is the proportion of non-attentive respondents, e is the difference in means among inattentive respondents, and τ is the difference among attentive respondents. We can demonstrate that the largest possible bias is 1, when p equals e equals 1 and tau equals 0, which is less than the 1.472 estimator we obtain. Therefore, the presence of inattentive respondents cannot fully explain our paradoxical result.

By analyzing the subgroups based on respondents' education levels in Columns (2) to (4) of Table 3, we further investigate the possible reasons behind the design effects. Columns (2) and (3) show the estimated proportion of respondents involved in bribery among those with a pre-university or technical school education. Both estimates exceed 1, indicating the presence of design effects. P-values suggested in Blair and Imai (2012) against the null hypothesis of no design effect are smaller than the usual level of statistical significance, which further confirms the result.

On the other hand, Column (4) shows the estimated proportion involved in bribery among university-level respondents. The result inferred that 82.2% of the university-level respondents were involved in the corruption. The Blair and Imai (2012) test cannot reject the null hypothesis of no design effect. For comparison, the direct question for university-level respondents reveals that 56.34% have personally witnessed or experienced a form of corruption in the past 12 months.

With the above results, we suggest that the overestimation of the proportion of respondents with bribery experience results from confusion induced by the control version of the survey question. Notably, respondents in the control group began with a series of nine questions about their perceived levels of corruption in Egypt, which should have already prepared their minds for potentially sensitive corruption-related questions. Suddenly, in Question 10, they were presented with a list of non-sensitive activities, such as vacations, jobs, cell phones, and weddings, which seemed unrelated to corruption. This abrupt change in tone and the switch from a yes/no response to reporting a count may have confused the respondents. The confusion may have created cognitive load and distraction, especially for those with lower levels of education. This cognitive load and distraction may have caused less-educated respondents to miscount, or more specifically, to undercount the number of non-sensitive activities they have experienced.

On the other hand, respondents in the treatment condition see a list containing one corruption-related sensitive item along with four non-sensitive items. As a result, they still view this question using the list experiment method, as part of a corruption-related survey, and are therefore not distracted. So they do not tend to undercount the number of non-sensitve activities.

To sum up, the cognitive load required for those who are in the treatment group is less than that required for those in the control group. And this creates the design effect – respondents may change their reported count for non-sensitive items depending on whether they are assigned to the treatment condition or the control condition. Such a design effect may concentrate on the less educated respondents, and less so for the better educated respondents. Yet, a violation of the no design effect assumption, even for a subgroup, is enough to break down the whole assumption.

Since the survey was conducted through paper-based personal interviews, we did not record how long each respondent took to complete it, so we lack additional quantitative data to test our hypothesis further. However, reports from the enumerators, including one of the authors who observed the fieldwork, confirmed that it took them longer to explain the control version of the question to the respondents.

Finally, we note that the choice of non-sensitive items in our control list is commonly used in many other studies, especially those studying corruption, and therefore, the cognitive load induced by the control list may not be unique in our study. Agerberg (2022), in Romania, adopts a control list of four items, including attending a work-related meeting, investing in stock, being unemployed for more than 9 months, and discussing politics with friends or family. The control list of Reisinger et al. (2017) in Russia includes the name of the head of the Central Bank, watching TV every day, owning a cellphone, and whether the pension is too high. A similar control list has been adopted by Tang and Hu (2023) in China. Thus, the cognitive load-induced design effect may exist in studies beyond ours.

Robustness Check

As a robustness check, we repeat the subgroup analyses across all observed demographic covariates in Table 4. Nearly all point estimates for the proportion of respondents who agree with the sensitive item are above 1, the theoretical maximum, which indicates the presence of a design effect.

Table 4 Subgroup Analysis for Different Demographics

	LE Estimator		LE Estimator		LE Estimator
Age (20-29)	1.169*** (0.204)	Gender=Female	0.997*** (0.190)	City (Cairo)	1.578*** (0.167)
Age (30-39)	1.101*** (0.207)	Gender=Male	1.709*** (0.130)	City (Alexandria)	1.555*** (0.208)
Age (40-49)	1.658*** (0.232)	Monthly Income (~$US 21~104)	1.416*** (0.253)	City (Suez)	1.600*** (0.268)
Age (50-59)	1.971*** (0.244)	Monthly Income (~$US 104~208)	1.584*** (0.157)	City (Asyut)	1.048*** (0.251)
Age (60-69)	1.687*** (0.287)	Monthly Income (~$US 208~312)	1.706*** (0.188)		

Age (>=70)	1.722*** (0.554)	Monthly Income (>=$US 312~)	1.000*** (0.315)

Note: *p<0.1; **p<0.05; ***p<0.01
HC2 heteroskedasticity-robust standard errors in parentheses.

Notably, the younger population, female respondents, high-income group, and Assuit residents have a point estimate that is not statistically significantly different from 1. However, this further affirms our theory, as younger individuals and those with higher incomes tend to have higher cognitive abilities. Female respondents, considering the cultural and religious background in Egypt, who agreed to be surveyed, may also belong to the better-educated group. Indeed, among the surveyed females, approximately 21% have a level of education equal to or greater than the university level, while the percentage for the male sample was only about 11%.

Consequently, we conclude that the results of the robustness check do not invalidate our theory about cognitive load-induced design effects.

Concluding Remarks for Studies in China and Beyond

Our paradoxical list experiment results highlight the essential role of the no design effect assumption in list experiments. Specifically, we contend that the extra cognitive load from the control list of non-sensitive items may distract lower-educated respondents and cause an undercount in the control group. In particular, encountering a control list with items that are completely non-sensitive and less related to the survey theme may confuse respondents, especially after they have already been directly asked the question under the theme and are mentally prepared for further ones.

Our results highlight a challenge for researchers working on design for list experiments. Aronow (2015) encourages researchers to ask respondents sensitive questions directly and combine responses from direct questions with list experiment results to improve estimation efficiency. However, including the direct question after the list experiment may introduce post-treatment bias (Montgomery et al., 2018). Conversely, asking the direct question before the list experiment may cause respondents to be influenced by the design effects we observe. So, Practical Advice 1: to ensure the unbiasedness of the result, researchers may have to rely on separate samples for direct questions and list experiments, treating the direct question as a separate arm in the randomized assignment. However, this would undoubtedly cost the efficiency of the estimation.

Second, researchers should carefully consider the composition of the control list, taking into account the potential for violations of the assumption of no design effects. Naturally speaking, items in the control list should be as non-sensitive as possible. Otherwise, researchers may fear that the control list itself is subject to the social desirability bias. As a result, empirical researchers may include items totally unrelated to the sensitive topics they

are working on. For example, in the control list, Tang and Hu (2023) include the name of the head of the Central Bank, watching TV every day, owning a cellphone, and whether the pension is too high, while Li and Meng (2020) include taking public transportation, experience abroad, and receipt of government relief, in their studies on corruption in China.

However, if our theory about cognitive load-induced design effects is correct, such practices could make respondents in the control group vulnerable to these effects, thereby jeopardizing the entire study. Therefore, Practical Advice 2: we recommend that researchers select non-sensitive items under the same survey theme for the control list, ensuring that the cognitive load remains balanced between the treatment and control groups. For example, in Meng et al. (2016), to study when city leaders in China would incorporate citizen suggestions into policymaking, they include local administrative expenditures, influence in attracting foreign investment, and the scope of the migrant population. The World Value Survey (Haerpfer et al., 2022), when assessing the level of trust respondents have in their head of state, includes the names of leaders of other countries in the control list.

Our advice should be read in combination with the existing considerations for the control list, including: (a) the prevention of floor and ceiling effects (Blair and Imai, 2012), which ensures that respondents in the treatment group are not forced to choose the maximum or minimum count, thereby avoiding revealing their answers to the sensitive statement to the enumerators; (b) the negative correlation between non-sensitive items to enhance statistical performance (Glynn, 2013); and (c) an inclusion of a placebo item (Agerberg and Tannenberg, 2021).

Finally, Practical Advice 3: We encourage researchers to adopt good practices that have been shown to reduce the misreporting of respondents in item counts, when possible. Kramon and Weghorst (2019), although discussing a different type of design effect from ours, demonstrate that providing tools to lessen respondents' efforts could improve the accuracy of estimates. They have tested two procedures and found both effective: (a) providing a pen and a piece of paper for respondents to privately mark the items they agree or disagree with, and (b) including a cartoon for each statement. Such practice should also reduce the cognitive load involved in our situation, and thus alleviate the violation of the design effect.

Further, noting that the cognitive load, and thus the misreporting, only happens to the control group. Considering that the control list only includes non-sensitive items, we believe it may help alleviate the design effect if respondents in the control group are asked to provide their item-to-item response rather than a count for the control list, also known as the LISTIT procedure (Corstange, 2009; Blair and Imai, 2012). This approach was not as popular as the canonical version, as initially, the additional complexity in the survey enumeration was mainly for estimation purposes. However, we argue that this approach may carry a previously ignored advantage – to balance the cognitive load between the treatment group and control group, which improves the plausibility of the design effect.

In summary, this paper documents paradoxical results from a list experiment in Egypt, which appears to be the first of its kind as far as we know. The analysis of these results suggests a possible violation of the "no design effect" assumption in the Egyptian context. We present suggestive evidence that the cognitive difficulty caused by the irrelevance of control questions to the survey theme may have led to under-reporting of non-sensitive items, thereby overestimating the prevalence of corruption. This finding complements but is distinct from prior research on design effects in list experiments. We hope this note, along with the practical advice offered, will be helpful, especially given that many failed studies remain unpublished due to publication bias and the file drawer effect (Gelman 2014; Schäfer and Schwarz, 2019).

References

Afrobarometer. *Afrobarometer Round 6: The Quality of Democracy and Governance in Egypt, 2015*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research (ICPSR), 2017. https://doi.org/10.3886/ICPSR36682.v1.

Agerberg, Mattias. "Corrupted estimates? Response bias in citizen surveys on corruption." *Political Behavior* 44, no. 2 (2022): 653–678.

Agerberg, Mattias, and Marcus Tannenberg. "Dealing with measurement error in list experiments: choosing the right control list design." *Research & Politics* 8, no. 2 (2021): 20531680211013154.

Aronow, P. M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. "Combining list experiment and direct question estimates of sensitive behavior prevalence." *Journal of Survey Statistics and Methodology* 3, no. 1 (2015): 43–66.

Blair, Graeme, Alexander Coppock, and Margaret Moor. "When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments." *American Political Science Review* 114, no. 4 (2020): 1297-1315.

Blair, Graeme, and Kosuke Imai. "Statistical analysis of list experiments." *Political Analysis* 20, no. 1 (2012): 47–77.

Blair, Graeme, Winston Chou, and Kosuke Imai. "List experiments with measurement error." *Political Analysis* 27, no. 4 (2019): 455–480.

Corstange, Daniel. "Sensitive questions, truthful answers? Modeling the list experiment with LISTIT." *Political Analysis* 17, no. 1 (2009): 45–63.

Gelman, Andrew. "Thinking of Doing a List Experiment? Here's a List of Reasons Why You Should Think Again." *Andrew Gelman's Blog* (blog), April 23, 2014.

http://andrewgelman.com/2014/04/23/thinking-listexperiment-heres-list-reasons-think. Accessed October 12, 2025.

Glynn, Adam N. "What can we learn with statistical truth serum? Design and analysis of the list experiment." *Public Opinion Quarterly* 77, no. S1 (2013): 159–172.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. "Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments." *Political Analysis* 22, no. 1 (2014): 1–30.

Haerpfer, C., R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen, eds. *World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0*. Madrid, Spain; Vienna, Austria: JD Systems Institute; WVSA Secretariat, 2022. https://doi.org/10.14281/18241.24.

Imai, Kosuke. "Multivariate regression analysis for the item count technique." *Journal of the American Statistical Association* 106, no. 494 (2011): 407–416.

Kramon, Eric, and Keith Weghorst. "(Mis)measuring sensitive attitudes with the list experiment: solutions to list experiment breakdown in Kenya." *Public Opinion Quarterly* 83, no. S1 (2019): 236–263.

Li, Hui, and Tianguang Meng. "Corruption experience and public perceptions of anti-corruption crackdowns: experimental evidence from China." *Journal of Chinese Political Science* 25, no. 3 (2020): 431–456.

Meng, Tianguang, Jennifer Pan, and Ping Yang. "Conditional receptivity to citizen participation: evidence from a survey experiment in China." *Comparative Political Studies* 50, no. 4 (2017): 399–433.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62, no. 3 (2018): 760–775.

Reisinger, W. M., M. Zaloznaya, and V. L. H. Claypool. "Does everyday corruption affect how Russians view their political leadership?" *Post-Soviet Affairs* 33, no. 4 (2017): 255–275.

Schäfer, Thomas, and Marcus A. Schwarz. "The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases." *Frontiers in Psychology* 10 (2019): 813. https://doi.org/10.3389/fpsyg.2019.00813.

Tang, Wenfang, and Yue Hu. "Detecting grassroots bribery and its sources in China: a survey experimental approach." *Journal of Contemporary China* 32, no. 140 (2023): 207–224.

Tourangeau, R., and T. Yan. "Sensitive questions in surveys." *Psychological Bulletin* 133 (2007): 859–883.

Transparency International. *Corruption Perceptions Index 2023*. Berlin: Transparency International, 2023. https://www.transparency.org/en/cpi/2023.

Warner, S. "Randomized response: a survey technique for eliminating evasive answer bias." *Journal of the American Statistical Association* 60 (1965): 63–69.

World Bank. *World Development Indicators*. Washington, DC: World Bank, 2025. https://data.worldbank.org/indicator.

World Bank. *World Bank Enterprise Survey Egypt 2020*. Washington, DC: World Bank Group, 2022. https://www.enterprisesurveys.org/en/data/exploreeconomies/egypt.