Is it Time to Put a Moratorium on List Experiments for Domestic Violence Elicitation?[1]

Andreas Kotsadam (Ragnar Frisch Centre for Economic Research)
Mette Løvgren (NOVA, Oslo Metropolitan University)

Abstract

Using data from over 24,000 respondents in the Norwegian Crime Victimization Survey, we conducted a double list experiment to measure domestic violence (DV). Both list experiments revealed a statistically significant decrease in reporting when including a sensitive DV item. This clear violation of the "no design effects" assumption is not only explained by floor effects. One possibility is that the results indicate a "fleeing" behavior whereby respondents try to avoid association with DV. Combined with the inherent power limitations of list experiments in many contexts, these results underscore the need for caution in employing list experiments to measure DV, even in large samples.

[1] Forthcoming in Sociological Methods & Research. This paper uses data from the Norwegian Crime Victimization Survey funded by the Norwegian Ministry of Justice. All opinions are our own. Anonymized data and replication files are available at the Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CLQSJK).

1 Introduction

Domestic violence (DV) imposes significant costs on society, with one in three women worldwide affected by it within their lifetime (Heise and Kotsadam, 2015; Fearon and Hoeffler, 2014). To accurately assess and mitigate this issue, precise measurement of its prevalence is essential. However, reliance on official sources, such as police reports or hospital visits, is highly inadequate. This is due to the low proportion of abuse victims who report to these sources (Palermo et al., 2014). As a result, researchers frequently turn to survey data. Yet, measuring DV and other sensitive behaviors through surveys are often presumed to be challenging due to stigmatization, fear, and sensitivity bias. Consequently, researchers often employ indirect measurement techniques, with list experiments being a particularly popular strategy.

List experiments, also known as the "item count technique", are conducted by randomly splitting the sample into two groups. The individuals are then asked to count the number of true statements on a list that either includes a sensitive statement or not. Comparing the number of statements reported as true across the two groups yields a measure of the variable of interest without any individual having revealed their own status. By also asking a question about the sensitive statement directly to the respondents researchers can assess the degree of underreporting by comparing the results when using the two different ways of asking. By analyzing differences in responses to direct questions and the list experiment, researchers further hope to investigate the degree of sensitivity bias in the population. While an ever increasing share of papers on DV in sociology, economics, political science, and public health use list experiments, the results are often mixed, and the jury is still out on whether and when list experiments should be used.

A major limitation of list experiments in most settings is low power. List experiments for elicitation of experiences are much noisier than direct measures by design and often severely underpowered in practice. Blair et al. (2020) show, for

instance, that list experiments often require 14 times as many observations to produce as precise prevalence estimates as direct questions. The power to detect sensitivity bias is well below 80 percent in most published studies, which typically have far fewer than 2000 observations. While there are ways of increasing power, such as including double lists and negatively correlated items, most papers have such small samples that the quest of finding meaningful sensitivity bias seems futile.

By embedding list experiments and state of the art measures of DV in the self-administered online Norwegian Crime Victimization Survey, we are able to conduct a double list experiment on a sample of over 24,000 individuals. Using direct questions we find that 2.4 percent of the respondents were abused by their partners in Norway in 2020. In both of our list experiments (where one was before and one after the direct questions about abuse) we find that the number of items reported are statistically significantly lower for individuals with the extra sensitive item added to the list. This shows that the list experiment failed completely, as it is not possible to experience negative amounts of DV.

Assuming that randomization of participants to the various lists works, there are two assumptions that need to be fulfilled for the difference in average counts to be attributed to the key sensitive item. These assumptions are commonly labeled “no liars” and “no design effects” (Blair and Imai 2012). The “no liars” assumption requires that individuals getting the list with the sensitive item answer truthfully about the sensitive item. The “no design effects” assumption requires respondents not to change their answers to non-sensitive items depending on whether the key item appears in the list.

Our result with negative list estimates show that these assumptions do not hold. Furthermore, using the statistical test for the no-design-effect assumption suggested by Blair and Imai (2012) we find that the assumption does not hold for any of our two lists.

Negative point estimates from list experiments have been called “fleeing behavior” (Chuang et al., 2021; Hinsley et al., 2019) and Gilligan et al. (2024) explicitly refer to respondents "fleeing" to avoid being identified with an item

relating to DV. Since the estimates are negative, it must be that individuals change their responses to the non-sensitive items. This therefore represents a particular type of violation of the no design effect assumption. The fleeing notion builds on the idea that respondents do not provide truthful answers even for non-sensitive items, given a fear that researchers will infer information about their behavior linked to sensitive items (Chuang et al., 2021). There may, of course, be other reasons why the assumptions fail in our setting, such as non-strategic misreporting. One particularly important factor discussed in the literature is artificial deflation (Kiewiet et al., 2014). Since lists are longer when the sensitive item is included, there have been concerns that some respondents may systematically report more true statements for longer lists. This can lead to an overestimate of rare items. Artificial deflation occurs when people automatically respond with somewhat higher numbers to longer lists, even though the true number is higher still. Importantly, this would still produce positive list experiment estimates, and not negative ones as we find. When we conduct list experiments in a sample where all respondents report DV on a direct question we do find positive estimates. However, for this sample we only identify around half of the cases of DV. One explanation for this result may be artificial deflation as DV by definition is very prevalent in this sample.

While our results of negative list experiment effects may seem surprising they are actually not that uncommon. While some studies find higher reporting of DV using list experiments (Bulte and Lensink, 2019; Cullen, 2023; Lépine et al., 2020; Traunmüller et al., 2019), several other studies using list experiments find similar prevalence as in direct questions (Agüero and Frisancho, 2022; Kotsadam and Villanger, 2022; Krebs et al., 2011). Using several list experiments in Ethiopia, Gilligan et al. (2024) find lower reporting than in direct questions and sometimes negative estimates as well. In the broader health literature using list experiments, it is quite common to find that list experiments do not yield more accurate reporting (e.g. Chuang et al. (2021); Haber et al. (2018)). We further suspect that publication bias and file drawer problems give a distorted view of the published literature. It is likely that studies finding negative

estimates may simply not report the results as the experiment "did not work".

The relative advantage of list experiments versus direct questions depends on how much sensitivity bias there is, which in turn depends on how sensitive the question is, in what format the question is asked, and in what context (Blair et al. 2020). Tourangeau et al. (2000) identify three psychological dimensions that may lead to underreporting in surveys: social (un)desirability, intrusiveness, and disclosure risk. The psychological dimensions are likely to interact with the mode of data collection. In modes where an interviewer is present, like face-to-face interviews, having list experiments may be more important by ensuring respondent anonymity and by allowing respondents to conceal sensitive answers. In contrast, self-administered methods like online surveys provide inherent anonymity, which might reduce the advantage of list experiments (Li and Van den Noortgate, 2022).

Apart from wanting to project a positive image to others (i.e., social desirability), sensitivity bias may arise due to self-image concerns or due to fear that others will find out what they have answered. Regarding domestic violence, it is particularly likely that people worry about their partners finding out that they report it. Self-administered online surveys allow respondents to control their environment during the study. In particular, they can chose to respond when the partner is not present. Self-administered questions have previously been shown to yield higher reporting of DV than face-to-face interviews in experiments (Klevens et al., 2012; Ahmad et al., 2009).

While social desirability is not the only reason for sensitivity bias, it is interesting to test how individual level measures of social desirability bias propensity correlate with differential reporting in list experiments. To the best of our knowledge, previous research has not tested this. Using a modified Marlowe-Crowne social desirability scale we find that individuals scoring high on propensity to be affected by social desirability report less abuse, but we do not find that high scores on the Marlowe-Crowne social desirability scale correlate with differential reporting in the list

experiment. We also explore heterogeneity more generally but do not find any consistent patterns. We interpret this as the negative list estimates being a general phenomenon in our setting.

Based on our study, which is the largest list experiment we have seen to date[2], other recent studies, and the inherent low power in list experiments, we conclude that greater caution should be used before using list experiments to measure domestic violence. Direct questioning is likely the best method, at least in online surveys where people likely feel sufficiently anonymous.

2 Data and the list experiments

We conduct the list experiments on a sample of 24,163 individuals that were participating in the Norwegian Crime Survey for 2020 (Løvgren et al., 2022). The survey was conducted in the spring of 2021 and asked many questions about crime victimization the previous calendar year 2020, and also about demographic characteristics. The Norwegian Crime Survey was initiated by its funder, the Ministry of Justice and Public Security. The survey is now repeated yearly.

The questionnaire used in the survey is largely based on those used in victimization surveys in similar countries, such as Sweden and Denmark. It consists of three main parts: background characteristics, concern and feelings of safety, and victimization. The average completion time is 14.8 minutes for respondents with no victimization experience, and about two minutes longer for those who reported at least one incident. Reporting victimization triggers follow-up questions, increasing the average completion time to 15.5 minutes.

[2] For instance, the largest study in the recent meta-analysis by Li and Van den Noortgate (2022) is Krebs et al. (2011) with 5,446 respondents.

A total of 68,000 individuals aged 16 to 84 were randomly drawn from the National Population Register, which contains information on all residents in Norway, and invited to participate in the survey. Individuals born outside Norway were oversampled due to anticipated lower response rates.

The initial invitation to participate was sent by post. To incentivize participation, all invitees were informed that three randomly selected participants would receive gift cards worth a total of 8,000 NOK. Reminders were sent by email or SMS. After three reminders, 24,163 individuals had participated, resulting in a final response rate of 35.56 percent. The final sample is representative of the Norwegian population on observed characteristics such as age, gender, and educational level (Løvgren et al., 2022).

Participants were asked to complete the survey online using a unique identifier. To allow flexibility, progress was saved so that the survey could be completed in multiple sessions. The identifier ensured that only the selected individual could access the questionnaire. However, recognizing that a unique identifier may not fully prevent access by others (e.g., a partner with access to mail or electronic devices), the questionnaire was designed with two key privacy safeguards: (1) each page displayed only one question, and (2) participants could not return to previous questions.

Our main interest lies in measuring domestic violence in 2020. In order to do so we create a dummy variable, *Any partner violence in 2020*, set to one for women who report experiencing any of the following from a partner during 2020: pinching or hair pulling; shoving; slapping; strangling; punching with a fist or an object; being beaten up; another violent attack; or sexual violence, such as unwanted touching or forced vaginal, anal, or oral penetration, either by force or when incapacitated.

We follow best practice in the field in order to maximize disclosure of domestic violence (Ellsberg et al., 2001). In particular, we apply detailed descriptions and ask several questions about different specific abusing acts. These questions follow

internationally validated questions and the sequence of the questions are based on the WHO Violence Against Women Instrument (Ellsberg and Heise, 2002) and the Conflict Tactics Scales (Straus et al., 1982). This approach is shown to reduce underreporting and avoids issues pertaining to understandings of what constitutes abuse (Cools and Kotsadam, 2017; Kishor, 2005; Kotsadam and Villanger, 2022). Since we ask about possibly traumatizing, sensitive, and criminal questions we also offer redress possibilities to the respondents and we give them phone numbers to five different organizations working with abuse victims.

Table 1 shows that 2.4 percent of the respondents report being abused in 2020. This number may seem low but it is substantially higher than the 0.08 percent of the population that reported domestic abuse to the police according to Statistics Norway. The low numbers are very similar to numbers for other years, both in the official statistics of crimes reported to the police and from later versions of the Norwegian Crime Survey (Løvgren et al. 2023). The low numbers are consistent with findings in the literature showing that domestic violence (measured as abuse last year) is much lower in richer countries and in countries with more gender equality (Heise and Kotsadam 2015).

Table 1 Descriptive statistics

	All		List 1		List 2	
	1		2		3	
	Mean	SD	Mean	SD	Mean	SD
Partner violence						
Any partner violence in 2020	0.024	0.154	0.025	0.157	0.023	0.150
Beaten by partner in 2020	0.007	0.084	0.007	0.085	0.007	0.082
List experiment variables						
Number of items on list 1	0.916	0.753	0.877	0.752	0.955	0.752
Number of items on list 2	1.813	0.679	1.832	0.620	1.794	0.731
Treated list 1	0.497	0.500	1.000	0.000	0.000	0.000
Treated list 2	0.503	0.500	0.000	0.000	1.000	0.000
Control variables						
Household experienced a bike theft in 2020	0.065	0.246	0.064	0.244	0.066	0.248
Someone broke into the respondents house in 2020	0.008	0.091	0.008	0.088	0.009	0.093
Health: Very good	0.287	0.452	0.288	0.453	0.285	0.451
Health: Good	0.499	0.500	0.501	0.500	0.497	0.500
Health: Neither good nor bad	0.157	0.364	0.154	0.361	0.159	0.366
Health: Bad	0.053	0.223	0.051	0.221	0.054	0.225
Health: Very bad	0.005	0.072	0.005	0.072	0.005	0.073
Age: Below 33 years	0.326	0.469	0.325	0.469	0.326	0.469
Age: 33-54 years	0.341	0.474	0.343	0.475	0.338	0.473
Age: Above 54 years	0.334	0.472	0.331	0.471	0.336	0.472
Social desirability score	2.135	2.088	2.134	2.082	2.137	2.094
Very high social desirability	0.081	0.273	0.080	0.271	0.082	0.275
High trust in institutions	0.518	0.500	0.520	0.500	0.515	0.500
Female	0.484	0.500	0.483	0.500	0.486	0.500
Immigrant	0.154	0.361	0.153	0.360	0.156	0.362
High education	0.485	0.500	0.479	0.500	0.490	0.500
Disabled	0.126	0.332	0.122	0.328	0.130	0.336
Employed	0.578	0.494	0.583	0.493	0.572	0.495
N	24163		12001		12162	

The list experiments were introduced at different parts of the survey. The first experiment was included before any questions about abuse were asked and the second experiment was included at the end of the survey. The survey respondents were randomized to be treated in either the first or second list. We are interested in measuring the share of individuals that were beaten by their partner in 2020 and we see in Table 1 that this share is reported to be only 0.7 percent when we ask individuals directly. We include this statement in both of the lists but individuals are only seeing the statement in the list for which they are randomly treated. We see in Table 1 that having the list before or after asking about abuse directly was inconsequential for the direct reporting of abuse. The p-value of the test of difference between any partner

violence for those having the abuse item included in the list before (List 1) and after (List 2) is 0.34.

We also see that the randomization worked in terms of creating balance across the two list samples in general. The full lists are shown below.

List 1 How many of these statements were true for you in 2020?

- I travelled on several vacations
- Me or someone in my household had their bike stolen
- I stayed home the whole year
- *I was beaten by my partner*
- I witnessed a car accident

List 2 How many of these statements were true for you in 2020?

- I was sick a lot
- I had someone break in to my house/flat
- I was in good health
- *I was beaten by my partner*
- I shopped online

The four control questions include four statements that we are not interested in and that are used only to get an average to compare the treated group with. The treatment list includes the same questions and adds the question of interest. The control questions are created to avoid ceiling and floor effects and to include items that are negatively correlated so as to increase power (Glynn, 2013). We also include one question in each list that we know the answer to from other parts of the survey (having a bike being stolen and having someone break into the house). This was done in order

to further increase power by means of controlling for variables that affect the total number of list items reported. In the same vein, we also control for self reported health in some specifications to see if we can further increase power as health is included in list 2.

3 Results

The way list experiments are usually interpreted is best described by an example. Assume that the list control group answers that one of the four statements are true on average and the list treatment group answers that 1.5 of the statements are true on average. As the only difference between the two groups is that the treatment group had an extra question about DV, researchers would conclude that 50 percent of the individuals in the list treatment group had experienced DV.

In Table 2 we show the results of our list experiments. Panel (a) reports results from the two list experiments analyzed jointly. To estimate the share with the sensitive attribute using both list experiments, we first take the difference in means between the long and the short lists separately for List 1 and List 2. These two estimates are then averaged. We compute standard errors following Aksoy et al. (2025), using the standard large-sample formula for a difference in means between two independent samples. We find the same results in least squares estimation specifically designed for a double list experiment using the Stata command "kict ls" (Tsai 2019). We start by using the whole sample. In column 1 we use the full sample and we see that the joint estimate is -0.057 and highly statistically significant. The literal interpretation from this would be that a negative 5.7 percent of the individuals have experienced DV. This is of course factually impossible. We see negative estimates also when restricting the sample to individuals having partners (column 2), which are more likely to have been abused by a partner. The presence of a partner may be endogenous to domestic violence, but since the prevalence of partner violence is non-existent for non-partnered individuals we follow the standard in the DV literature (e.g., Heise and Kotsadam, 2015) and restrict the

sample to partnered individuals in what follows. But we note that the conclusions are the same for both types of samples.

Panel (b) shows the results from two list experiments separately. We estimate these effects using OLS regressions. In columns 1 and 2 we analyze the first list experiment. We see in column 1 that individuals in the control group on average answer that 0.95 of the four statements were true for them in 2020. Individuals in the treatment group, i.e. individuals getting the list with the additional question about DV, answer 0.078 fewer statements on average. Again, the literal interpretation from this would be that a negative 7.8 percent of the individuals have experienced DV. We see negative estimates also when restricting the sample to individuals having partners (column 2).

In columns 3 and 4 we see negative point estimates also for the second list experiment. For this list, control respondents answer that 1.83 of the statements were true for them in 2020. The treatment group, that got the exact same questions plus the additional DV statement, report 0.039 fewer statements as true. Hence, we find negative estimates of DV both when we ask about it before and after the direct questions and for lists with different control questions and different average levels of true statements for the control group. Remember that the samples are practically identical on all other observable characteristics and that they only answer that fewer statements are true when they get the DV statement.

In Table 3 we see that including controls increases the variance explained but the point estimates are very similar. Hence, for both experiments we find that the negative reporting is still visible even when we are able to control for statements on the lists we have the answers to.

Table 2 Main results for the list experiments.

(a) Joint estimation of the list experiments

	(1) Joint	(2) Joint
List Treatment	-0.057*** (0.0058)	-0.062*** (0.0067)
N	23809	17742
Sample	All	All partnered
Controls	No	No

(b) Estimation of the list experiments separately

	(1) List 1	(2) List 1	(3) List 2	(4) List 2
List Treatment	-0.078***	-0.081***	-0.039***	-0.044***
	(0.0097)	(0.011)	(0.0088)	(0.0099)
Control mean	0.95	0.94	1.83	1.84
N	24011	17885	23914	17818
R-squared	0.003	0.003	0.001	0.001
Sample	All	All partnered	All	All partnered
Controls	No	No	No	No

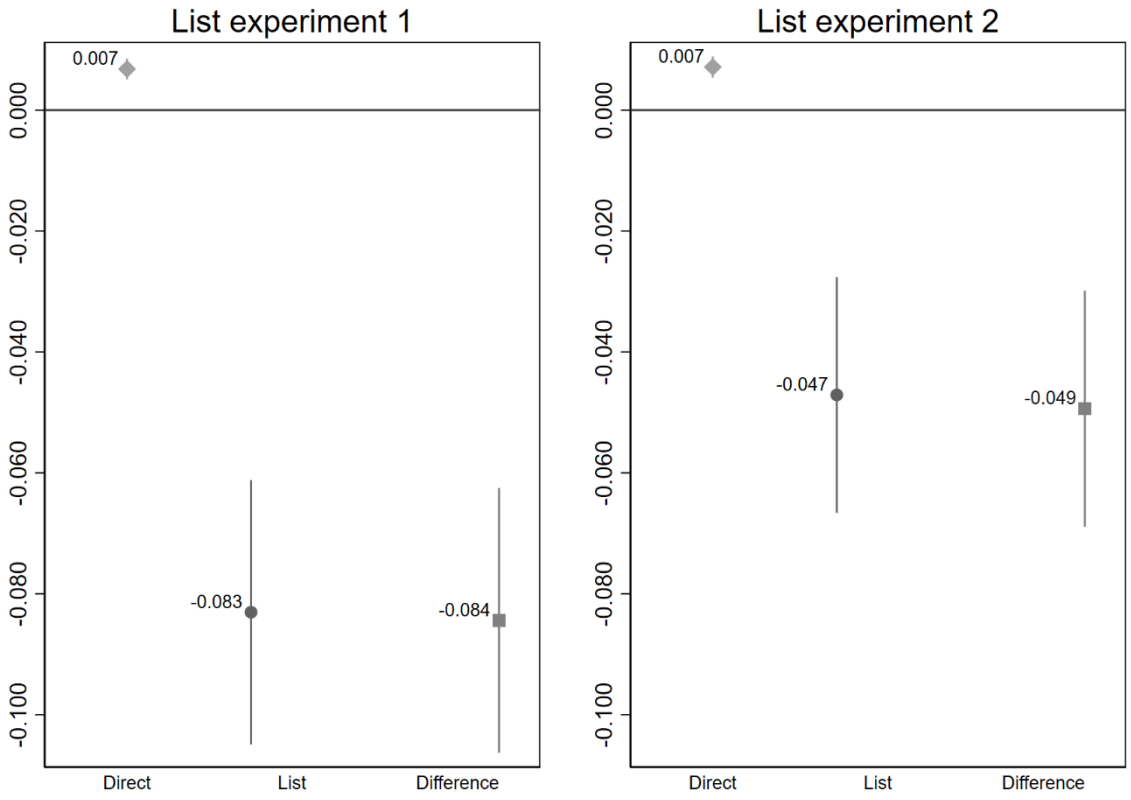
Table 3 Regression results for the list experiments with control variables. Partnered individuals only.

	(1) List 1	(2) List 2	(3) List 1	(4) List 2
List Treatment 1	-0.078*** (0.011)		-0.074*** (0.011)	
List Treatment 2		-0.043*** (0.0097)		-0.042*** (0.0096)
Control mean	0.94	1.84	0.94	1.84
N	17860	17756	17614	17548
R-squared	0.078	0.037	0.087	0.064
Sample	All partnered	All partnered	All partnered	All partnered
Controls	List item	List items	Extensive	Extensive

Notes: The list item in column 1 is a dummy variable for the household having had a bike stolen in 2020 and in column 2 the list items are a dummy variable for having had someone breaking into your house and dummy variables for poor health. The extensive controls include all list item controls, age dummies in three categories, and dummy variables for immigrant, higher education, disabled, and being employed. Robust SE in parentheses.

In Figure 1 we show the estimates when we ask people directly, when we ask using the list experiment, and the difference between the two. The difference is estimated directly in an OLS regression where the dependent variable is replaced by the answer to the direct question for the control group. We see that the difference is highly statistically significant for both list experiments.

Figure 1: Difference between direct questions and list experiments. Partnered individuals only.



Notes: List refers to the estimated prevalence of having partner sometimes hitting in the list experiment. Direct refers to the prevalence when using a direct survey question. Difference refers to the difference between asking in the list experiment minus asking directly. 95 percent confidence intervals are shown.

Table 4 shows the distribution of answers for the treated and control individuals in the two experiments. It seems as if List 1 suffers from floor effects for a relatively large share of the sample as over 25 percent of the sample in the control group answer zero. As 2020 was also a pandemic year it

is possible that a substantial share of the respondents went on precisely one vacation, did not have a bike stolen, and did not witness a car accident. Fortunately this problem is much less severe for the second list where only 3.26 percent in the control group answer zero. Interestingly, we note that more treated individuals answer zero in both lists.

Table 4: Distribution of answers in the list experiments. Partnered individuals only.

	List 1		List 2	
Number of statements reported	Control (%)	Treatment (%)	Control (%)	Treatment (%)
0	25.29	30.31	3.26	4.36
1	59.02	55.95	17.95	22.13
2	10.9	10.01	72.44	66.15
3	4.5	3.37	5.05	5.6
4	0.29	0.23	1.31	0.65
5	–	0.13	–	1.11

To test the key assumption of no design effects in list experiments, we perform a test developed by Blair and Imai (2012). This assumption requires that adding a sensitive item to the list does not change how respondents answer the non-sensitive control items. The test works by examining estimated joint probabilities—the likelihood of observing a given number of responses (R) for individuals in either the control group (S = 0) or treatment group (S = 1). Since probabilities cannot be negative, the presence of any negative estimated joint probabilities suggests a potential violation of the assumption. The "kict deff" test in Stata (Tsai 2019) checks whether these negative values are likely to have occurred by chance. If the null hypothesis (that all joint probabilities are positive) is rejected, it indicates the presence of design effects, meaning the list experiment failed. The test separates the joint probabilities into two sets based on treatment status and evaluates each set independently.

For our first list experiment, the test reveals multiple negative joint probability estimates for the treatment group (S=1), including for Pr(R=0), Pr(R=1), and Pr(R=2),

with all associated p-values below 0.01, indicating these values are statistically significantly below zero. The formal test for design effects confirms this: for the treatment group, the null hypothesis of no design effects is strongly rejected ($P < 0.001$). These results indicate that the list experiment suffers from design effects in the treatment condition, raising concerns about the validity of the estimates produced by the item count technique in our setting. We find very similar effects for the second list experiment. The results show two statistically significant negative joint probability estimates in the treatment group (S=1): Pr(R=0) and Pr(R=1), with p-values below 0.001. Again, formal test for design effects confirms this by rejecting the null for the treatment group ($P < 0.001$). As the tests show that there are negative values at more than point of the distribution for both list experiments, the negative estimates are not only driven by floor effects. To further substantiate this point we note that we get negative estimates in both list experiments even if we exclude everyone answering zero.

4 Heterogeneity

In this section we conduct a set of heterogeneity analyses in order to try to understand the results. We begin by restricting the sample to individuals we know have reported abuse in 2020. We then continue to investigate differential effects for individuals answering differently on the Marlowe-Crowne social desirability scale. Finally we test for differences for people with varying degrees of trust and for people with different background characteristics.

When we restrict the sample to individuals reporting being hit by their partner when asked directly, the list experiments work better. Table 5 at least shows positive coefficients. While not negative, the estimates are still far from perfect. First of all, the estimates are still different from 1, which is what they should be if everyone reporting DV on the direct question also reports it when asked indirectly. Secondly, the coefficient is only around 0.5. In this sample, where we know everyone has been hit by a partner, we would thus conclude that only 50 percent had been hit by their partners. While negative estimates are avoided, it is not a good sign that the list

experiment only finds reporting for half of the victims.

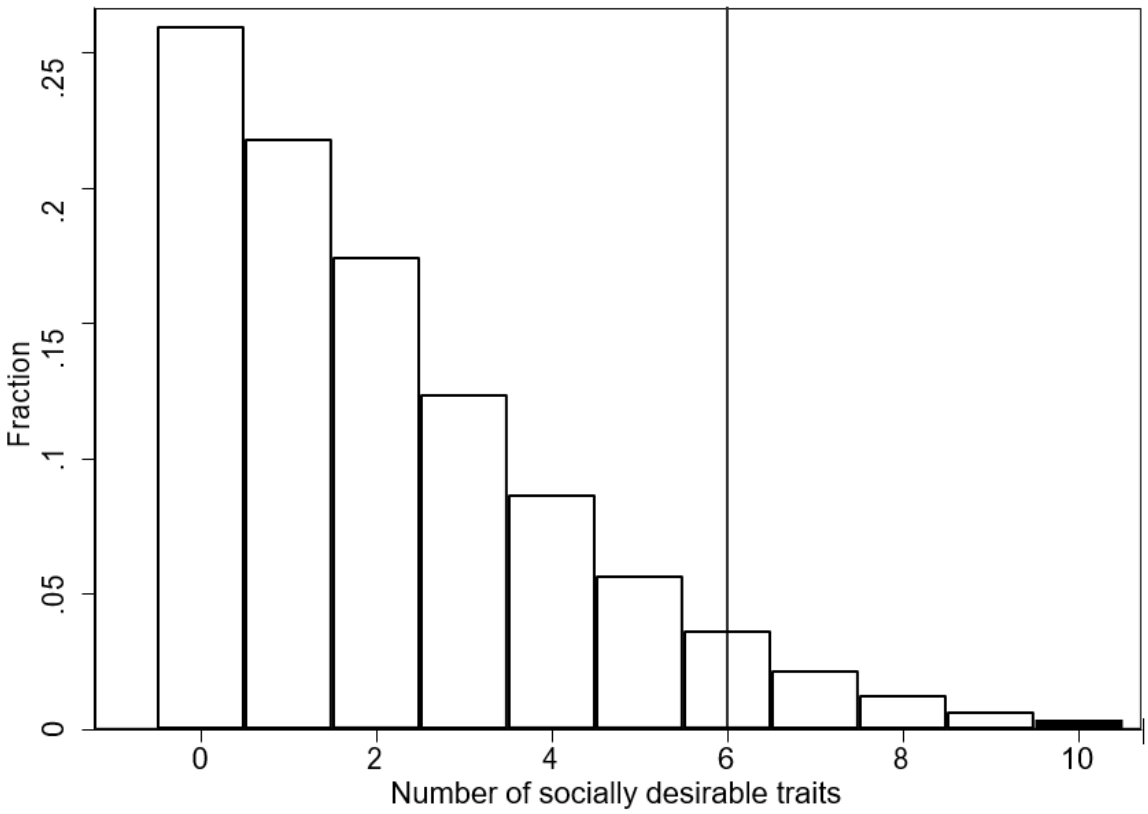
Table 5: Regression results for the list experiments for individuals reporting to have been beaten.

	(1) List 1	(2) List 2
List Treatment	0.53***	
	(0.16)	
List Treatment		0.54***
		(0.15)
Control mean	1.33	2.00
N	124	121
R-squared	0.080	0.099
Sample	Beaten	Beaten
Controls	No	No

Notes: Robust SE in parentheses.

To investigate the role of social desirability for our negative estimates, we use a modified Marlowe-Crowne social desirability scale (Crowne and Marlowe, 1960). We use a validated Norwegian survey module developed by Rudmin (1999) which consists of asking the respondents whether they posses 10 very good traits, such as *always* being a good listener or *always* admitting making a mistake. The traits are sometimes reverse coded to avoid capturing people always answering the same. It is very unlikely that a person actually has several of these saintlike traits (Dhar et al. 2022).

Figure 2 shows the distribution of the traits in our sample. We see that less than 1 percent of the sample possess all 10 traits and 8 percent possess at least 6 of them. We code individuals with more than 5 traits present as having a very high measured propensity to be affected by social desirability, which is indicated by the vertical line in the figure.



Notes: The vertical line is included to show where in the distribution respondents classified as having a very high social desirability bias fall (>5)

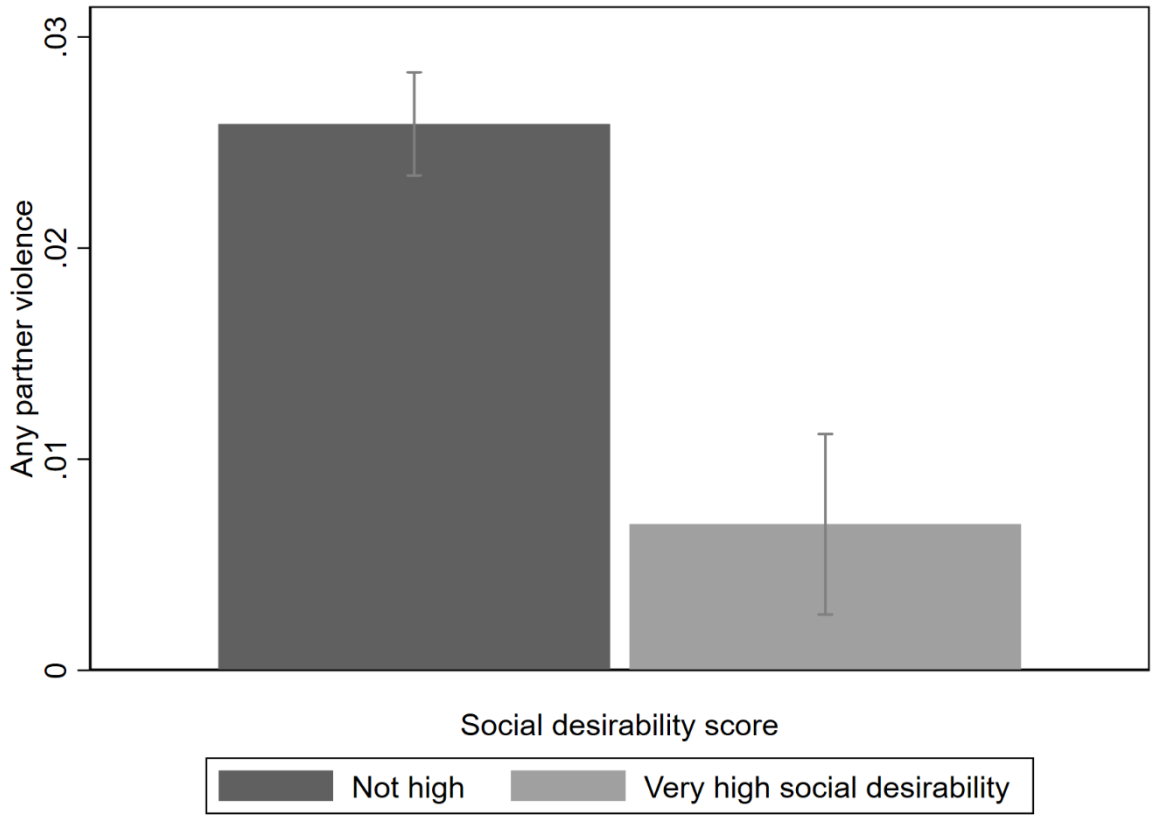
Figure 2: Social desirability index.

We can now investigate reporting behavior in the list experiment for individuals with different degrees of suspected social desirability bias as defined by the index.

We see in Figure 3 that individuals answering to possess at least 6 of the index traits report less DV. This is suggestive of social desirability being important for reporting of sensitive behavior but it is important to note that there may also be other differences between individuals with different degrees of measured propensity to be affected by social desirability.

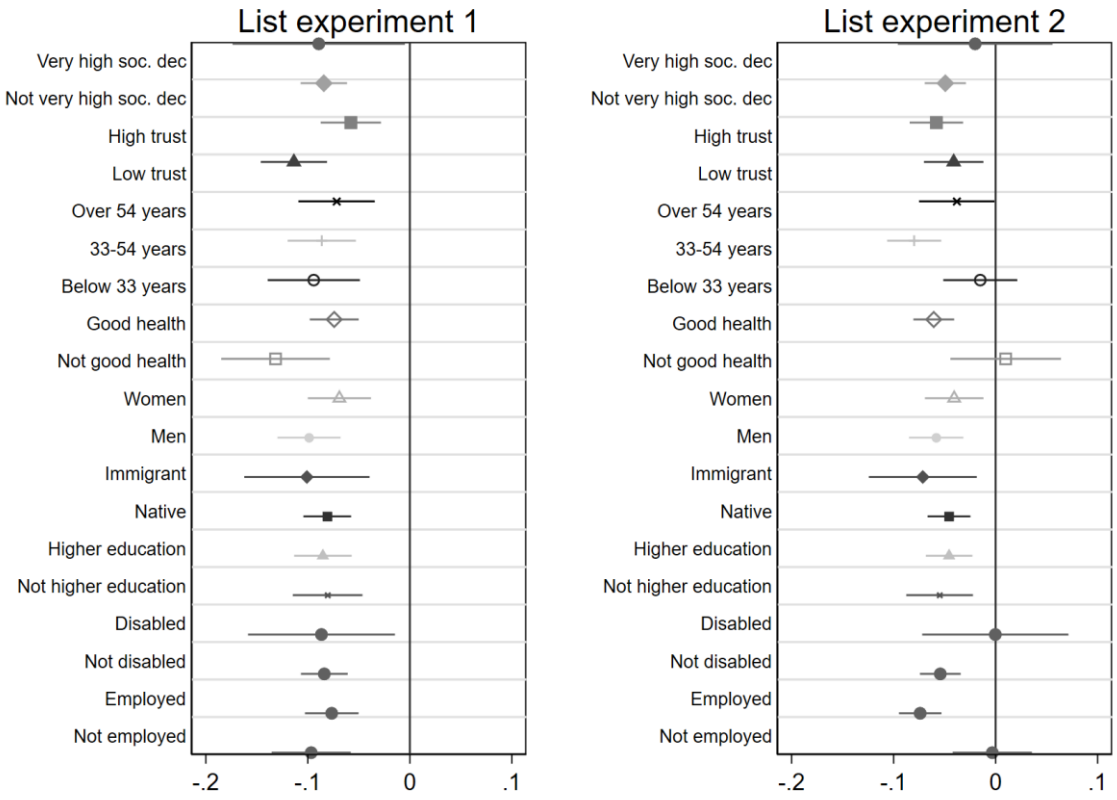
Turning to indirect reporting, do we see that there is a difference between individuals with different measured social desirability bias in reporting then? We see in the first two rows of Figure 4 that there are not large differences for the different groups, but the list experiment effects are imprecisely estimated for the small group scoring above 5 on the index. Most importantly, the results show that also for

individuals with lower social desirability index scores, both list experiments produce negative estimates. In Section B of the online supplement we show results using different cutoffs of the scale and the conclusion remains the same.



Notes: The two bars show the reported prevalence of domestic violence victimization in 2020 by respondents without a very high social desirability score (<6) in blue and with a very high social desirability score (>5) in red. 95 percent confidence intervals are shown.

Figure 3: Reported DV and social desirability index.



Notes: The figure shows estimates on the list experiments for various subgroups. The left part of the figure shows estimates for the first list experiment and the right part for the second list experiment. 95 percent confidence intervals are shown.

Figure 4: List experiment estimates for different subgroups.

The next two rows of Figure 4 shows the list experiment estimates for individuals with a higher trust in institutions or relatively lower trust. One could suspect that not trusting institutions would be correlated with less truthful reporting in surveys. For the first list experiment we here see a difference whereby individuals with low trust have a larger negative estimate. This result is not reproduced in the second list experiment, however. With many sample splits one should be careful not to overinterpret single results and the advantage of the double list experiment is that we can test whether the results replicate. Perhaps more importantly, we note that the negative difference appears for both groups in both experiments.

In the remaining rows of Figure 4 we explore heterogeneous list treatment effects

along all control variables. In all, we observe very little in terms of heterogeneous treatment effects. Two pairwise differences are statistically significant in the second list, the difference for being disabled or not and being employed or not. But these results only appear for the second list and the estimates are negative for all groups except the disabled in list 2. Again, we find it to be an advantage to have two list experiments which enables a direct replication of the results.

In Section A of the online supplement we also present results where we split the sample by response times and we do not find that the negative effects are driven only by respondents answering very quickly. We also find that the tests of “no design effects” fail similarly for the sample of respondents spending more than 20 minutes on the survey.

In all, we conclude that the negative list results seem to be universal in our sample.

5 Conclusion

In a comprehensive analysis involving over 24,000 individuals through the Norwegian Crime Victimization Survey, we conducted a double list experiment to measure domestic violence (DV). The incidence of partner abuse reported in 2020 was below 3 percent when asked directly. However, the list experiments revealed a statistically significant decrease in reporting when a sensitive DV item was included. This is a clear violation of the "no design effects" assumption. One potential reason for the design effect assumption not to hold may be idea that the respondents "flee" to not be associated with domestic violence, a phenomenon that has been labelled “fleeing behavior” in the literature (Gilligan et al. 2024). It may, for instance, be that respondents react negatively to questions that make them feel as if we are trying to trick them.

We found no significant variations in reporting among individuals with different

levels on a social desirability index and we find that the list experiment yields negative estimates in all different subgroups we tested, and those results replicate across both experiments in our large dataset.

Despite the assumption that list experiments might provide a more secure environment for disclosing sensitive information like DV, our study, which is the largest of its kind that we know of, demonstrates that respondents are actually less forthcoming in such setups. The notion of list experiments serving as a shield for truthful reporting or as a "statistical truth serum" (Glynn, 2013) is challenged by our results. Given these findings, alongside the variable results in the literature and the inherently low power of list experiments, we advocate for a more cautious application of this method in DV research. Contrary to being a panacea for eliciting honest responses, list experiments may not offer the protective disclosure environment they are designed to create. This is perhaps especially true in online surveys, where participants may feel sufficiently protected anyway (Krebs et al., 2011; Li and Van den Noortgate, 2022). In all, our conclusion is that direct detailed questions are the best way to elicit domestic violence victimization.

Pre-registration, replication material, and conflict of interest:

This study was not preregistered.

The data and code to replicate the analyses are available in the Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CLQSJK).

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ahmad, Farah, Sheilah Hogg-Johnson, Donna E. Stewart, Harvey A. Skinner, Richard H. Glazier, and Wendy Levinson. "Computer-assisted screening for intimate partner violence and control: a randomized trial." *Annals of internal medicine* 151, no. 2 (2009): 93-102.

Agüero, Jorge M and Veronica Frisancho, "Measuring violence against women with experimental methods," Economic Development and Cultural Change, 2022, 70 (4), 1565–1590.

Aksoy, Billur, Christopher S. Carpenter, and Dario Sansone. "Understanding labor market discrimination against transgender people: Evidence from a double list experiment and a survey." Management Science 71.1 (2025): 659-677.

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, 2020, 114 (4), 1297–1315.

Blair, G., and K. Imai. 2012. Statistical analysis of list experiments. Political Analysis 20: 47–77.

Bulte, Erwin and Robert Lensink, "Women's empowerment and domestic abuse: Experimental evidence from Vietnam," European economic review, 2019, 115, 172–191.

Chuang, Erica, Pascaline Dupas, Elise Huillery, and Juliette Seban, "Sex, lies, and measurement: Consistency tests for indirect response survey methods," Journal of Development Economics, 2021, 148, 102582.

Cools, Sara and Andreas Kotsadam, "Resources and Intimate Partner Violence in Sub-Saharan Africa," World Development, 2017, 95, 211 – 230.

Crowne, Douglas P and David Marlowe, "A new scale of social desirability independent of psychopathology.," Journal of consulting psychology, 1960, 24 (4),

349.

Cullen, Claire, “Method matters: the underreporting of intimate Partner violence,” The World Bank Economic Review, 2023, 37 (1), 49–73.

Dhar, Diva, Tarun Jain, and Seema Jayachandran, “Reshaping adolescents’ gender attitudes: Evidence from a school-based experiment in India,” American Economic Review, 2022, 112 (3), 899–927.

Ellsberg, Mary and Lori Heise, “Bearing witness: ethics in domestic violence research,” The Lancet, 2002, 359 (9317), 1599–1604.

Ellsberg, Mary, Lori Heise, Rodolfo Pena, Sonia Agurto, and Anna Winkvist, “Researching domestic violence against women: methodological and ethical considerations,” Studies in family planning, 2001, 32 (1), 1–16.

Fearon, James and Anke Hoeffler, “Benefits and costs of the conflict and violence targets for the post-2015 development agenda,” Conflict and violence assessment paper, Copenhagen Consensus Center, 2014.

Gilligan, Daniel O, Melissa Hidrobo, Jessica Leight, and Heleene Tambet, “Using a list experiment to measure intimate partner violence: Cautionary evidence from Ethiopia,” Applied Economics Letters, 2024, pp. 1–7.

Glynn, Adam N, “What can we learn with statistical truth serum? Design and analysis of the list experiment,” Public Opinion Quarterly, 2013, 77 (S1), 159–172.

Haber, Noah, Guy Harling, Jessica Cohen, Tinofa Mutevedzi, Frank Tanser, Dickman Gareta, Kobus Herbst, Deenan Pillay, Till Bärnighausen, and Günther Fink, “List randomization for eliciting HIV status and sexual behaviors in rural KwaZulu-Natal, South Africa: a randomized experiment using known true values for validation,” BMC medical research methodology, 2018, 18, 1–12.

Heise, Lori and Andreas Kotsadam, “Cross-national and multilevel correlates of partner violence: an analysis of data from population-based surveys,” The Lancet

Global Health, 2015, 3 (6), e332–e340.

Hinsley, Amy, Aidan Keane, Freya AV St. John, Harriet Ibbett, and Ana Nuno, “Asking sensitive questions using the unmatched count technique: Applications and guidelines for conservation,” Methods in Ecology and Evolution, 2019, 10 (3), 308-319.

Kiewiet de Jonge, Chad P., and David W. Nickerson. "Artificial inflation or deflation? Assessing the item count technique in comparative surveys." *Political Behavior* 36 (2014): 659-682.

Kishor, S., “Domestic Violence Measurement in the Demographic and Health Surveys: The History and the Challenges,” Technical Report, United Nations Expert Paper 1-9 2005.

Klevens, Joanne, Laura Sadowski, Romina Kee, William Trick, and Diana Garcia. "Comparison of screening and referral strategies for exposure to partner violence." *Women's Health Issues* 22, no. 1 (2012): 45-52.

Kotsadam, Andreas and Espen Villanger, “Jobs and intimate partner violenceevidence from a field experiment in ethiopia,” Journal of Human Resources, 2022, pp. 0721–11780R2.

Krebs, Christopher P, Christine H Lindquist, Tara D Warner, Bonnie S Fisher, Sandra L Martin, and James M Childers, “Comparing sexual assault prevalence estimates obtained with direct and indirect questioning techniques,” Violence Against Women, 2011, 17 (2), 219–235.

Lépine, Aurélia, Carole Treibich, and Ben d’Exelle, “Nothing but the truth: Consistency and efficiency of the list experiment method for the measurement of sensitive health behaviours,” Social Science & Medicine, 2020, 266, 113326.

Li, Jiayuan and Wim Van den Noortgate, “A Meta-analysis of the relative effectiveness of the item count technique compared to direct questioning,”

Sociological Methods & Research, 2022, 51 (2), 760–799.

Løvgren, Mette, Asle Hagestøl, and Andreas Kotsadam, “Nasjonal trygghetsundersøkelse 2020,” 2022.

Løvgren, Mette, Øivind Skjervheim, Asle Høgestøl, Andreas Kotsadam, Vegar Bjørnshagen, Christer Hyggen, and Stian Lid. "Nasjonal trygghetsundersøkelse 2022." (2023).

Palermo, Tia, Jennifer Bleck, and Amber Peterman, “Tip of the Iceberg: Reporting and Gender-Based Violence in Developing Countries,” American Journal of Epidemiology, 2014, 179 (5), 602–612.

Rudmin, Floyd W, “Norwegian short-form of the Marlowe-Crowne social desirability scale,” Scandinavian Journal of Psychology, 1999, 40 (3), 229–233.

Straus, M.M.A., R.J. Gelles, and S.K. Steinmetz, Behind Closed Doors: Violence in the American Family, Transaction Publishers, 1982.

Tourangeau, Roger, Lance J Rips, and Kenneth Rasinski, “The psychology of survey response,” 2000.

Traunmüller, Richard, Sara Kijewski, and Markus Freitag, “The silent victims of sexual violence during war: Evidence from a list experiment in Sri Lanka,” Journal of Conflict Resolution, 2019, 63 (9), 2015–2042.

Tsai, Chi-lin. "Statistical analysis of the item-count technique using Stata." The Stata Journal 19.2 (2019): 390-434.

Online supplement

Section A: Different time spent on the survey

We investigate whether respondents that spend very little time on the survey are the ones that drive the negative effects. We first split the results by different subgroups for the list experiments in Table A1. In the first column we show the baseline results for the main sample. In column two, we restrict the sample to individuals that answer the survey very quickly, in less than 10 minutes. We see that the negative estimates are slightly larger for both list experiments in panels A and B. When we exclude these individuals in column 3 we still have negative coefficients for the list treatments. Also if we exclude individuals answering less than or equal to 20 minutes, the results are negative and highly statistically significant. Hence, the negative effects are not driven only by respondents answering very quickly. We also find that the tests of "no design effects" fail similarly for the sample of respondents spending more than 20 minutes on the survey.

Table A1: Effects of list experiment by subgroups of different time use

Panel A: List experiment 1

	(1)	(2)	(3)	(4)
	List 1	List 1	List 1	List 1
List Treatment	-0.081***	-0.12***	-0.076***	-0.072***
	(0.011)	(0.028)	(0.012)	(0.021)
Control mean	0.94	0.96	0.94	0.96
N	17885	2386	15479	5993
R-squared	0.003	0.007	0.003	0.002
Sample	All partnered	Less than 10	More than 10	More than 20
Controls	No	No	No	No

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Panel B: List experiment 2

	(1)	(2)	(3)	(4)
	List 2	List 2	List 2	List 2
List Treatment	-0.044***	-0.061***	-0.042***	-0.049**
	(0.0099)	(0.023)	(0.011)	(0.019)
Control mean	1.84	1.90	1.83	1.80
N	17818	2391	15407	5896
R-squared	0.001	0.003	0.001	0.001
Sample	All partnered	Less than 10	More than 10	More than 20
Controls	No	No	No	No

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Section B: Different cutoffs of the social desirability index

In Figure B1 we see the list experiment coefficients for the two list experiments for subsamples defined as having above x on the social desirability score, where x is between 1 to 10. We see negative coefficients throughout but large confidence intervals where we have few observations.

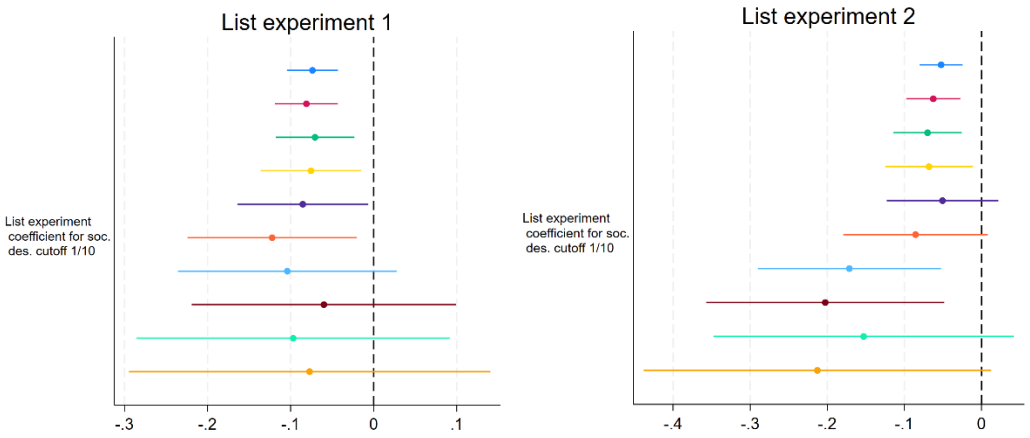


Figure B1: List experiment estimates for different cutoffs of the social desirability index.