Journal of Development Economics 148 (2021) 102582

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Regular Article

Sex, lies, and measurement: Consistency tests for indirect response survey methods☆

Erica Chuang [a], Pascaline Dupas [b,*], Elise Huillery [c], Juliette Seban [d]

[a] Population Council, USA
[b] Stanford University, NBER, CEPR, BREAD and J-PAL, USA
[c] Université Paris Dauphine, PSL Research University, and J-PAL, France
[d] Sciences Po, France

A B S T R A C T

Social scientists seeking to analyze socially sanctioned behaviors or attitudes increasingly rely on indirect response survey methods, meant to veil the answers of individual respondents. We propose simple internal consistency tests for two such methods, the list experiment and the randomized response technique (its Warner and Crosswise variants). We implement these tests in two studies on sexual and reproductive health behavior in Cameroon and Côte d'Ivoire. Non-compliance with instructions among surveyed individuals appears high and not easily char-acterizable. The tests we propose can be easily and cheaply embedded in measurement tools, allowing researchers to at least know whether their data is reliable before using it.

1. Introduction

Understanding attitudes and behaviors is a core objective of economics, and of social science re-search more generally. Data is obviously critical for this endeavor–knowing the prevalence, corre-lates and, ideally, causes and effects, of a given behavior or attitude allows the researcher to develop the appropriate theories and, when applicable, formulate potential solutions.

How do social scientists obtain data? Many studies rely on direct questioning, where an enu-merator directly asks a respondent about a behavior or belief, in order to record and measure an outcome of interest. However, there are some topics about which it is difficult to gain a true response without putting the individual at risk of feeling judged/stigmatized/punished (e.g. sexual behav-ior, political preferences, illegal activity). A comprehensive meta-analysis by Tourangeau and Yan (2007) shows that many respondents misreport when they answer sensitive questions (even when the questions are self-administered). Respondents may be particularly wary of answering truthfully in contexts where trust is limited, where institutions are weak and there is concern that private answers might be shared, where the government is corrupt or venal, or where individual freedoms are limited and there may be concern around anyone official looking coming to one's door. These situations are far from uncommon in lower income countries, especially in rural areas where a lot of survey work takes place. Randomized experiments, now commonplace in the field of Development Economics, may be particularly concerned with the risk that respondents exposed to an interven-tion feel compelled to answer a certain way in order to please the research team; such experimenter demand effect or social desirability bias may be especially important for interventions aiming to change norms or behavior.

Indirect questioning/indirect response (IR) techniques are one proposed solution, meant to veil the answers of individual respondents to enumerators and investigators. Two such methods are the list experiment (LE, otherwise known as unmatched count or item count technique, which provides privacy by embedding the sensitive behavior of interest within several nonsensitive behaviors), and randomized response technique (RRT, which provides privacy by letting the roll of a dice or other randomization device decide, unknown to the interviewer, whether the question is to be answered truthfully) (see Section 2 for detailed descriptions of these techniques). The use and popularity of these

☆ We are grateful to Berk Özler, Patrick Francois, and anonymous referees for detailed comments that have improved the manuscript, as well as to participants at the CEGA Research Retreat for helpful feedback at an early stage. This research makes use of original data collected by Dupas, Huillery and Seban in Cameroon and by Dupas, Victor Orozco, Jonathan Robinson and Miron Tequame in Côte d'Ivoire. We thank IRESCO for their collaboration in Cameroon and the World Bank, Orozco, Robinson, and Tequame for their collaboration in Côte d'Ivoire. Dupas gratefully acknowledges the support of the National Science Foundation (NSF) (award number 1254167) and Huillery gratefully acknowledges the support of the CEPREMAP. All errors are our own.

* Corresponding author.
E-mail addresses: echuang@popcouncil.org (E. Chuang), pdupas@stanford.edu (P. Dupas), elise.huillery@dauphine.psl.eu (E. Huillery), juliette.seban@sciencespo.fr (J. Seban).

https://doi.org/10.1016/j.jdeveco.2020.102582
Received 4 June 2019; Received in revised form 13 October 2020; Accepted 27 October 2020
Available online 11 November 2020

techniques has grown over time, particularly in the last 15 years. Fig. 1 shows the number of studies with original data collection that utilize either technique over time. The growth of use for both of these methods coincides with the adoption of rigorous methodologies for field research in academic subjects including, but not limited to, economics, political science, and psychology.

While these methods are sound in theory, they face potential challenges in practice. Participants may not trust that their privacy is protected–e.g. they may assume that the researchers can retrieve the truth from their answers. As a consequence they may disregard the instructions and give a safe answer to protect themselves. What's more, given that IR techniques are somewhat convoluted, even trusting participants may have a hard time understanding and thus complying with the instructions. Here again, these problems may be particularly acute in lower income countries, where trust and education levels are typically lower.

Assessing how close to the truth the results from IR measurement methods are is difficult due to the very nature of the problem at hand: the underlying true behavior or belief is typically unknown. In this paper, we develop internal consistency tests for both LE and RRT to assess the validity of these methods. The consistency test for LE consists in implementing two List Experiments (i.e. two different sets of control items for the same sensitive behavior of interest) within the sample, and comparing prevalence estimates across the two experiments. If individuals comply with LE, we should find a similar estimated prevalence across the two. The consistency test for RRT, which can be applied to both the original Warner model as well as its popular Crosswise variant, consists in setting the probability that the respondent is asked to answer truthfully to 0.5, for one behavior that is sensitive but not the main outcome of interest. With 50% truthfulness, we should expect exactly 50% of individuals to report engaging in the sensitive behavior, irrespective of the true prevalence. This means that observing a reporting rate that is statistically different from 50% implies non-compliance.

We implement these tests on data from Côte d'Ivoire and Cameroon. We find that both methods fail to pass these internal consistency tests, suggesting that respondent compliance with their protocols is low. We also document that low compliance is unlikely to be due to poor understanding of the protocols, but rather to low levels of trust in the extent to which they veil answers on sensitive items.

Our paper contributes to the sensitive behavior measurement methodology literature. A first strand of this literature has focused on comparing prevalences estimated through IR techniques with those observed through direct questionning (DQ) when both methods are used in the same survey (Coutts and Jann, 2011; Eady, 2017; Chou, 2019). These comparison studies presume that higher prevalence estimates of undesirable behaviors are more valid (the so-called "more is better" assumption) and interpret the gap between prevalence estimates from DQ and IR as a measure of under-reporting in DQ. But what if IR techniques are not themselves accurate due to non-compliance?

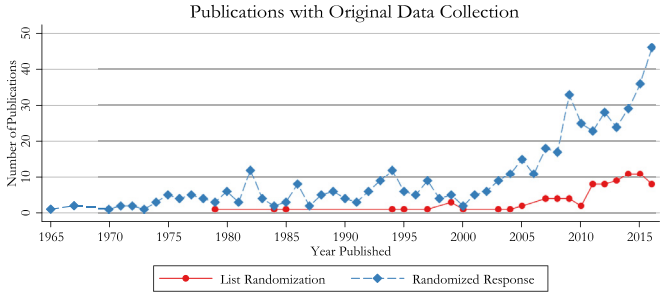

Fig. 1. Importance of indirect response techniques over time.
Source: EBSCO, Web of Science, Scopus, PubMed and ProQuest. Searches were limited to peer-reviewed publications and book chapters, and limited to social science publications (e.g. economics, political science) with original data collection. Articles prior to 1965 for Randomized Response and 1978 for List Randomization were excluded. Searches were conducted in English, though non-English language articles identified through the English language search were included. Searches were conducted from June through October 2017. After the initial search, duplicates were removed to minimize redundancy and articles were screened for relevance to either indirect response method.

A second strand of this literature compares estimates from IR techniques to the "truth" when it is known. Such validation studies are of two types. The first type uses aggregate-level data providing the true population prevalence of the sensitive attribute, e.g. from administrative records. For example, in the context of anti-abortion policies in Mississippi, Rosenfeld et al. (2015) study how LE, RRT, and DQ estimates compare to actual vote outcomes at precinct-level. They find evidence that DQ matches poorly to actual proportion of votes, and that IR methods only partially close the gap. An earlier meta-analysis of six validation studies found that RRT and LE typically provide more accurate prevalence estimates than DQ (Lensvelt-Mulders et al., 2005). One concern with such aggregate-level validation studies is that prevalence estimates may seem valid at the aggregate level even if the IR methods do not actually work well: false positives and false negatives may cancel each other by chance (Höglinger and Jann, 2018).

The second type of validation studies is done at the individual-level. Since having "true" individual-level data on the behavior of interest is typically impossible, such validation studies are rare. They typically test compliance with RRT on respondents pre-selected on the behavior of interest (e.g. from administrative records), so that true prevalence is known to be 100%. Böckenholt et al. (2009) do this in the Netherlands, in the context of non-adherence to conditions required to receive social security benefits. They find that RRT does not eliminate response biases completely due to self-protective and non-self incriminating responses, which leads to underestimation of the true incidence of non-adherence. Preisendörfer and Wolter, 2014 compare RRT with DQ on whether one has ever been convicted for subjects who had been convicted under criminal law in Germany. They find that RRT does not improve accuracy over DQ.

A creative way to conduct a validation study with any available sample is to introduce a known zero-prevalence item (i.e. "ever received a donated organ"), as proposed by Höglinger and Diekmann (2017). They use this method to test the internal validity of RRT and estimate 8% of false positives. Other validation studies have been done using online platforms, where true behavior can be measured unbeknownst to the respondent. Höglinger and Jann (2018) use online dice games to test DQ and RRT against actual cheating behavior at the aggregate and individual level. They find that none of the RRT variants they tried produced overall more valid measurements than DQ. The reason is that RRT induces a substantial number of false positives, likely due to confusion with RRT instructions. Importantly, the misclassification issue cannot be identified with aggregate data because false positives and false negatives may cancel out. This result also invalidates the "more is better assumption."

The literature to date thus suggests that IR techniques are not bulletproof: IR techniques may lead to either under- or over-estimation of the true prevalence. What's more, anticipating the likely accuracy of IR measurements in one's context is difficult. Existing meta-studies have not provided clear guidelines on when IR is likely to be accurate. This means that it is crucial to check for compliance with these techniques within a study sample. In this paper, we propose low-cost tests for the internal consistency of these techniques. These tests can be easily incorporated by researchers during implementation, and our finding that IR methods failed in the two contexts we consider suggest that researchers ought to do so: absent these tests, researchers could wrongly assume that the IR technique they used provided meaningful estimates. This paper is thus closest to the literature testing the internal consistency of IR techniques.

The internal consistency tests proposed to date include that of Blair and Imai (2012), who develop statistical methods to detect and adjust for certain types of violation of LE's validity assumptions, in particular that responses to control items do not change with the addition of a sensitive item to the list (*no design effect*). Aronow et al. (2015) propose a test using the combination of LE and DQ: people who admit the behavior in DQ

should all include the sensitive item in LE response, resulting in an estimated prevalence of the behavior in LE of 100% in this group. Both types of test generate necessary but not sufficient conditions for internal consistency.

For RRT, a method called "cheating detection model", first proposed by Clark and Desharnais (1998) and later expanded by Moshagen and Musch (2012), Ostapczuk et al. (2009), and Moshagen et al. (2012), is to split the sample in two and assign them different randomization probabilities.

This allows a test that the two prevalence estimates obtained from the subsamples are similar, and if not, with further assumptions, one can estimate the proportion of non-compliers. This test comes at the cost of efficiency, however, since estimating a difference in means on two subsamples reduces estimate precision relative to only estimating a single mean on the full sample. More recently, Heck et al. (2018) proposed an extension of the cheating detection model which does not estimate the proportion of non-compliers but allows for indicating whether prevalence estimates can be trusted with no loss in efficiency.

Our paper contributes to this literature by proposing new internal consistency tests for both RRT and LE. Our objective is *not* to provide tools to measure the extent to which non-compliance occurred nor provide a prevalence estimate that corrects for non-compliance, which requires fairly strong assumptions. Nevertheless, the proposed tests allow for generally evaluating whether preva-lence estimates can be trusted or not. Moreover, contrary to most existing consistency tests, they do not reduce estimation efficiency, nor rely on fairly strong assumptions about the reasons why respondents do not comply. Assumption-free, efficiency-preserving tests to evaluate the consistency of IR techniques can help ensure empirical studies of sensitive behaviors are accurate.

The paper is organized as follows. Section 2 describes, for each technique considered, the way the technique works and the design of the internal consistency tests we developed. Section 3 presents the results of the test for LE with data from Côte d'Ivoire, and Section 4 presents the results of the test for RRT with data from Cameroon. Section 5 concludes.

2. Indirect response techniques and proposed tests

2.1. List experiment (LE)

The List experiment (LE), also known as unmatched count or item count technique, was introduced by Raghavarao and Federer (1979) and Miller (1984). In LE, a sample is randomized into two groups, A and B – one which receives a list of non-targeted, non-sensitive ("baseline") statements (e.g. I1, I2, I3, I4), the other which receives the same list of baseline statements plus one extra sensitive statement (S). S is the object of interest: the researcher wants to gauge the prevalence of S behaviors. The respondent provides the enumerator the number of statements that are true without indicating how true any one statement is, and the difference between the means of the sensitive versus non-sensitive list recipients is the estimated prevalence of the behavior the researcher is trying to measure. A well-known challenge in using this method is the risk of "floor" or "ceiling" effects: extreme value responses (declaring having done none or all of the behaviors) perfectly reveal the sensitive trait. This means that to make sure no respondent has an incentive to lie, the choice of non-sensitive statements must be done carefully so that everyone has done at least one non-sensitive behavior but no respondent has done all of them.

Recently, LE has been utilized in studies on the misallocation of loans to small and medium enterprises in Peru and the Philippines (Karlan and Zinman, 2011), illegal migration in Mexico, Morocco, Ethiopia, and the United States (Mckenzie and Siegel, 2013), or the degree of racial hatred in the United States (Imai, 2011). These studies include the targeted sensitive question in the list provided to only one group – either group A or group B (single list experiment).

The test we propose and implement consists in including the sensitive behavior in two lists.

E.g. as above, group A is given the baseline list (I1, I2, I3, and I4) while group B is given the augmented list (I1, I2, I3, S and I4) (set 1), but in addition, group A is given the augmented list (I5, S, I6, I7, and I8) and group B the baseline list (I5, I6, I7, and I8) (set 2). Hence the prevalence of the behavior S can be backed out in two ways: by computing the difference-in-means $B - A$ in set 1 and $A - B$ in set 2. Since groups A and B are randomly drawn from the same population, the expected value of the prevalence of S is equal in both groups. If individuals comply with LE, we should thus find similar estimated prevalences using both sets.

The idea of having two lists is not new: Droitcour et al. (1991) proposed the Double List Experiment (DLE) as a way to generate two difference-in-means estimators that can be averaged, and demonstrate that using DLE reduces the variance compared to the single list experiment. Glynn (2013) expands on this idea and show that the variance of the estimator can be reduced further if there is a positive correlation between the number of true items in the two baseline lists. These papers do not suggest using DLE to test respondent compliance, however. Our innovation is to compare two fairly precisely estimated difference-in-means obtained through DLE, rather than average them. We argue that DLE constitutes another diagnostic test for compliance to LE instructions. As soon as DLE provides two statistically distinguishable estimates, there must be either design effects (i.e. responses to control items change with the addition of a sensitive item to the list), ceiling effects, or floor effects on at least one of the list. Passing the test is thus a necessary condition to trust the prevalence estimates. It is however not sufficient: estimates may still be biased if there are no design, floor or ceiling effects but people lie about the sensitive items in the same way in both lists. This can happen if some respondents act 'as if' they did not have the behavior even though they did engage in it and perfectly comply with LE instructions under this fake status. Since list 1 and list 2 are randomized, the proportion of liars in the two lists would be the same in expectation, resulting in the same prevalence estimates but both biased. As a consequence, if the data pass the DLE test, researchers can have a higher level of confidence in the validity of the prevalence estimate than absent such a test, though not 100%.

We also study the extent to which the accuracy of LE depends on the sensitivity of the baseline statements. It is well understood that the choice of baseline items in the list matters for the accuracy of the technique. According to Tsuchiya et al. (2007), baseline items should not be too uncommon (i.e. close to a "0" mean) or common (i.e. mean is close to the maximum number of items) in order for a respondent to feel protected. Glynn (2013) shows that the variance of the estimator will be minimized when there is negative correlation among the items on the baseline list. It is much less understood whether the topic and sensitivity of the baseline items also influence the results of the technique. We thus vary whether the baseline items are innocuous, or whether they include items that are related to the targeted statements either in topic and/or by being somewhat sensitive statements themselves. This allows us to test whether prevalence estimate increases when the targeted statement does not stand out. Under the assumption that a higher prevalence estimate means better compliance (which may not always be true due to false positives), this test tells us whether less innocuous baseline items make respondents possibly less suspicious of the researcher's intent.

2.2. Randomized response (RRT)

Randomized Response technique (RRT) was first proposed by Warner (1965). In this method, a surveyor gives respondents a list of binary questions. The respondent is then given an instrument, for example a six-sided die, and instructed to tell the truth for the question(s) given if the die lands on a particular side, such as six–otherwise lie. In order to preserve anonymity of responses, the survey should be implemented such that the surveyor cannot see or learn which side the respondent landed on. As long as the probability p that the respondent is asked to be truthful (e.g., $p = 1/6$ for a six-sided die) is different from 50%, and assuming that people comply 100% with the protocol, it is possible to back-out the true

prevalence s of the sensitive behavior as follows: the share r of individuals who report engaging in the behavior will be the sum of those that truly did it and those that did not do it but were told to lie:

$$r = ps + (1-p)(1-s) \quad (1)$$

Thus the true prevalence of behavior among the sample is

$$s = \frac{r + p - 1}{2p - 1} \quad (2)$$

where r is the share of "yes" answers to the particular question.

This is the original RRT model developed by Warner (1965) (also known as Mirrored Question Design). Other versions of this technique include Forced Response, Unrelated Question, Disguised Response, and Crosswise designs (Blair et al., 2015). Among them, the Crosswise design has retained much attention. It makes use of a nonsensitive question whose aggregate prevalence is known (e.g., whether the respondent is born between August and December) that is assumed to be independent of the sensitive item (e.g., whether the respondent is a drug user). For each sensitive item, the crosswise design asks respondents to choose whether the yes/no answer to the sensitive item is (Aronow et al., 2015) identical to or (Blair and Imai, 2012) different from the yes/no answer to the paired non-sensitive item. This design is particularly attractive because it does not require a randomization device, it is easy to understand, it avoids self-incrimination, and it does not offer any clear self-protective strategy. Interestingly, our internal consistency test also applies to this design since equation (Aronow et al., 2015) works the same for the Crosswise design, with s the true prevalence of the behavior, r the share of "same" answers, and p being the prevalence of the nonsensitive question.

Many studies have used RRT in order to measure behaviors such as substance abuse among adolescents in the United States (Fisher et al., 1992), illegal fishing of red abalone in Northern California (Gavin et al., 2009), abortion rates in the United States (Greenberg et al., 1971), xenophobia and anti-semitism in Germany (Krumpal, 2012), and illegal resource use in Uganda (Solomon et al., 2007). None of these studies include a test for whether respondents comply.

Equation (Aronow et al., 2015) above suggests a very simple test: if the probability p is set to 0.5, we should observe $r = 0.5$ for any true prevalence of behavior s. Thus implementing RRT with $p = 0.5$ (e.g. a coin flip or three values of a die in the Warner model, or whether the respondent is born between July and December in the Crosswise model) for at least one question allows the researcher to test whether respondents are, on average, complying with the RRT instruction, and therefore trust the estimated prevalence obtained with $p \neq 0.5$. In order to avoid loss in efficiency, the test can be done on the whole sample by adding one sensitive item in addition to the other items of interest. Ideally the additional item would be a sensitive question that is not measuring a key outcome of interest for the study in question, yet close enough in sensitivity with the subject of interest so that compliance behavior should be similar across questions. If the other RRT questions use a die, the compliance-testing RRT question should do so as well but give 3 values for which respondents are instructed to tell the truth so that the conditions between this question and the other ones are as close as possible. Our recommendation is that researchers who use RRT incorporate in their survey design a test with $p = 0.5$ for at least one sensitive item. This may be done at the pilot stage (if the pilot has a large enough sample), in order to determine whether the method is worth using in the context of interest or not.

While observing $r = 0.5$ when $p = 0.5$ is necessary to assert that participants complied with the protocol for the test case, it does not necessarily imply compliance with the protocol for targeted items. First, respondents may behave differently when p is equal to 0.5 if they are able to discern that with $p = 0.5$ the researcher cannot recover anything. This seems unlikely though as statistical literacy is generally low, especially in lower income contexts, so it is unlikely that respondents, when given RRT instructions, compute what p means for inference. Indeed, as shown by a study of Soeken and Macready (1982) with undergraduate students, there is no effect of the randomiza-tion probability on respondents' perceived protection, except for extreme values of p ($p > 0.91$). Therefore, compliance to RRT rules obtained with $p = 0.5$ can arguably generalize to other values of p, except extreme ones.

Second, estimates may still be biased if some people in the sample acts 'as if' they did not have the behavior even though they did engage in it and comply with RRT rules under this fake status, i.e. responding "NO" when assigned to the truth and "YES" when assigned to lie. In this very specific case, our test will be passed (we would observe $r = 0.5$) but the estimated prevalence from RRT with $p \neq 0.5$ would be underestimated. This means that if RRT data pass the $p = 0.5$ test, researchers can have a higher level of confidence in the validity of the prevalence estimate than otherwise, though not 100%.

3. Application: list experiment in côte d'Ivoire

3.1. Sample and data

2017 individuals from four of Côte d'Ivoire's ten regions were surveyed between December 2012 and May 2013. The survey was conducted in French through enumerator-led paper questionnaires and was focused on understanding coping mechanisms during the post-electoral crisis of 2009–2010, which caused the death of over 3000 civilians and the displacement of more than half a million people in just over 5 months.[1]

One question of interest was whether the conflict spurred entry into commercial sex. To this end, the targeted sensitive items included in the list experiment module focused primarily on sexual behavior. 12 sensitive items were targeted – 8 of which were common to everyone and 4 were tailored to be gender-specific. The list of targeted sensitive items considered can be found in Table 1.

To enable the test described above in section 2.1, 24 lists of 4 baseline items were created, which were split into two sets of 12 lists. Respondents had 24 lists to go through, and were asked to record the number of items on each list that they agreed with. Respondents were randomized into two groups, A and B. Group A received a survey in which the 12 targeted sensitive items were embedded in the first 12 lists (Set 1), whereas Group B received a survey in which the 12 targeted sensitive items were added to lists 13–24 (Set 2). The baseline items for both sets are shown in Table A1. Figure A1 is a visual example of how the two versions of the survey compared (for space reasons, it reproduces only lists 1 and 13 embedding the first targeted sensitive item, so one-twelfth of the survey). As recommended by Glynn (2013), in both lists, we included baseline items that were negatively correlated in order to reduce floor and ceiling effects and improve precision.

Often, lists are analyzed by having baseline statements that are completely innocuous aside from the singular sensitive statement that the researchers want to test for. One concern with this is that it makes the targeted statement embedded in the list very obvious. This may undermine the LE technique if this makes the intent of the question salient to respondents, who may suspect that researchers may have a way to back out their answer to the specific sensitive item (even if in truth they do not). This is a different type of "design effect" than that discussed in Blair and Imai (2012). They consider the case in which the inclusion of the sensitive, targeted item affects some respondents' answers to baseline items, whereas we hypothesize that the inclusion of sensitive items among the baseline items may influence whether the respondents give truthful answers to the targeted item.

In our application in Côte d'Ivoire, various degrees of non-innocuous

[1] The survey was designed by Pascaline Dupas, Victor Orozco, Jonathan Robinson and Miron Tequame and data was collected with funding from the World Bank.

Table 1
List Randomization in Cote d'Ivoire: Estimated Prevalence by set and Comparison with Direct Questioning (DQ).

i	*j*	Statement	Female						Male					
			^{P}set1 (A_i - B_i)	^{P}set2(B_j - A_j)	P-value ^{P}set1$=^{P}$set2	DQ	P-value ^{P}set1$=$DQ	P-value ^{P}set2$=$DQ	^{P}set1 (A_i - B_i)	^{P}set2(B_j - A_j)	P-value ^{P}set1$=^{P}$set2	DQ	P-value ^{P}set1$=$DQ	P-value ^{P}set2$=$DQ
1	13	I once had unprotected sex with someone the same day I met that person.	0.083 *(0.061)*	0.037 *(0.055)*	0.456				0.096 *(0.062)*	0.099 *(0.061)*	0.969			
2	14	I once had sex with a partner whom I had known for less than 1 month and who gave me money.	−0.028 *(0.057)*	0.291 *(0.057)*	0.000				0.114 *(0.066)*	0.172 *(0.060)*	0.382			
3	15	In the last 12 months, I have had concurrent sexual partners	0.070 *(0.060)*	0.167 *(0.058)*	0.108				0.279 *(0.073)*	0.160 *(0.060)*	0.103			
4	16	The last time I had sex, it was with a non-marital/non-cohabitant partner and I did not use condom.	0.024 *(0.051)*	0.231 *(0.061)*	0.000				0.099 *(0.057)*	0.282 *(0.069)*	0.001			
5	17	I have had a sexual partner who supported me financially.	0.064 *(0.054)*	0.224 *(0.065)*	0.003				0.046 *(0.056)*	0.138 *(0.069)*	0.098			
6	18	In the last 12 months, I had a partner whom I suspected to be HIV positive.	−0.083 *(0.052)*	0.119 *(0.064)*	0.000				0.020 *(0.060)*	0.100 *(0.068)*	0.180			
7	19	I had at least two non-cohabitant or non-marital partners in the last 12 months.	0.089 *(0.049)*	0.177 *(0.053)*	0.075	0.028 *(0.005)*	0.224	0.005	0.122 *(0.054)*	0.273 *(0.060)*	0.005	0.108 *(0.010)*	0.790	0.006
8	2	W: I had at least one abortion. M: At least one of my partners had an abortion.	0.274 *(0.068)*	0.230 *(0.068)*	0.517	0.417 *(0.015)*	0.035	0.006	0.173 *(0.076)*	0.273 *(0.075)*	0.193	0.257 *(0.014)*	0.272	0.836
9	21	During the 2010–11 post-electoral economic crisis, I had sex with partners who gave me money.	−0.156 *(0.059)*	0.149 *(0.065)*	0.000				−0.064 *(0.064)*	−0.013 *(0.064)*	0.430			
10	22	W: I had at least one abortion with a traditional healer. M: At least one of my partners had an abortion with a traditional healer.	0.016 *(0.059)*	0.091 *(0.059)*	0.210				0.026 *(0.064)*	0.224 *(0.064)*	0.002			
11	23	W: I had at least one abortion before the age of 18. M: At least one of my partners had an abortion before of 18.	−0.003 *(0.06)*	−0.001 *(0.064)*	0.976				0.033 *(0.061)*	0.078 *(0.068)*	0.467			
12	24	W: I had an abortion in the last two years due to the crisis. M: I have a sexual partner who had an abortion in the last two years due to the crisis.	−0.001 *(0.059)*	0.030 *(0.065)*	0.596				−0.002 *(0.063)*	0.056 *(0.069)*	0.356			

Note: See Figure A1 for illustration of methodology. Each LE prevalence is estimated by comparing average responses between 999 respondents (531 female) randomized into group A and 1012 respondents (526 female) randomized into group B. Standard errors in italics. The order for statements 308/320 and 309/321 in Set B were switched in the endline questionnaire - we switch them back to the order that corresponds with Set 1 for this table. DQ prevalence available for two questions only, asked *after* the LE module had been administered.

statements were incorpo-rated into Set 2 as a method of hiding the statements that the researchers were truly interested in measuring. The non-innocuous statements range from potentially socially undesirable/risky be-haviors that the respondent herself has engaged in (e.g. "I have discussed with friends who think abortion should be legal"), socially undesirable behaviors individuals or groups of individuals in the respondent's community may have engaged in ("Many women have abortion even though it is illegal"), behaviors that straddle the line between socially undesirable and acceptable (e.g. "I met men who go out with very young women."), and other questions that are about sexual and repro-ductive behavior. In Set 2, there are 34 sexual and reproductive health-related baseline items out of 48 total statements (see Table A1 for the complete list). This allows us to compare performance between lists that contain completely innocuous statements and lists that contain non-innocuous statements that would obscure the true behavior that the researchers are interested in. Note that Set 2 always came after Set 1 so the performance comparison relies on the assumption that the ordering of the sets does not affect the resulting prevalence estimates. This assumption may not hold if respondents respond differently to later sets of statements than earlier ones due to, for instance, tiredness.

For two of the sensitive items measured through LE, we also asked the questions directly (DQ). The DQ module included many questions not included in LE. It was administered after the LE module – with a long household roster administered in-between, as a way to distract respondents and reduce the chance that they remembered the specific behaviors listed in LE.

3.2. Results

We present the main results in Table 1, separately for females and males. We show the estimated prevalence using Set 1, the estimated prevalence using Set 2, and a p-value for a test that they are equal. We also show the DQ estimates. We find important discrepancies between the two LE sets: for women, the evidence from set 1 would suggest non-zero prevalence for 3 out of 12 risky behaviors, while Set 2 would suggest non-zero prevalence for 8 out of the same risky behaviors. We can reject with 90% confidence that the estimated prevalence is equal across the two sets in 6 of the 12 behaviors. The gap is somewhat smaller but still large for men: we can reject equality of estimated prevalence between the two sets for 4 of the 12 behaviors.

These results are not driven by imbalances between groups A and B.[2] An alternative explanation for the gaps between the results from the two sets could be the difference in the type of baseline statements that were included in the lists. Recall that the Sets varied in the sensitivity of the baseline items. Specifically, baseline items in Set 2 included somewhat sensitive statements. Did Set 2 work better or worse due to this? In principle the sensitivity and topic of baseline items in the list should not matter as long as the incidence of these behaviors is balanced across groups A and B. But if the inclusion of these items increased the variance in means for baseline items, it could have contributed to adding noise to the estimated prevalence. We show the variance in baseline items in Table A3 and find that, if anything, variance is lower among sensitive baseline items than non-sensitive items. Sensitive baseline items thus produce more precise prevalence estimates than non-sensitive baseline items. Table A3 also shows the frequency of extreme responses, i.e. where baseline items are either too uncommon (where responses are reported as "0", or never engaged in any of the baseline behaviors) or too common ("4", or engaged in all behaviors). We find that extreme "0" responses are relatively rare overall, though more common under Set 1. On the other hand, extreme "4" responses are more common under Set 2, which creates the issue that respondents of this type who engaged into the targeted behavior have to reveal their behavior or lie.

Alternatively, if the targeted sensitive items were too obvious to spot in Set 1, this may have cued to the respondents that attention was being given to a specific behavior, which could have led to lower compliance – not knowing how but assuming the researchers would be able to back out their behavior on the sensitive item, they may have shied away from providing answers on even the non-sensitive items truthfully. This could explain why we see in almost all cases lower estimated prevalence under Set 1 compared to Set 2. It could also explain why we at times find statistically significantly negative prevalence estimates under Set 1, something which is obviously impossible and difficult to explain otherwise. This finding suggests that camouflaging the behavior of inter-est among other sensitive statements may provide more accurate prevalence estimates than using completely innocuous baseline items. However, it is worth noting that this interpretation requires no false positives, which holds only when non-compliance is driven by self-protective strategy. If non-compliance is rather due to confusion with instructions, cognitive load, or carelessness, then higher prevalence estimates may not be more accurate.

Comparing LE estimates to DQ (Table 1), we do not find clear evidence that LE is systematically less biased than DQ. Assuming that DQ suffers from under-reporting (not over-reporting), as is commonly the case, we find that DQ estimates for one behavior (abortion) are substantially larger than LE estimates, suggesting that LE fails to get closer to the truth than DQ. For the other behavior (having multiple partners), DQ estimates are lower than LE estimates. Our finding that DQ estimates may be larger than LE estimates shows that non-compliance is not always driven by self-protective strategy. In fact, participants who would agree to admit abortion in DQ record fewer items when abortion is included than when it is not (both with Set 1 and Set 2 baseline items). This may come from a lack of attention, or confusion due to excessive cognitive load.

Altogether, these results suggest that there is no guarantee that LE yields less biased estimates than DQ, and our recommendation is that researchers should do DLE whenever possible to test compliance, as well as include DQ. Our findings also suggest that it may be best to hide the sensitive item of interest among several lists of sensitive baseline items. Researchers can use more than two lists of sensitive baseline items to show how prevalence estimates vary among equally sensitive lists.

4. Application: randomized response in Cameroon

4.1. Sample and data

The data for this comes from Dupas et al. (2018), where the full details of the study can be found. A total of 3714 adolescent women were surveyed between January 25 and April 29, 2011. Surveys were in French, and included a module that contained direct questions about sexual behaviors. The questionnaire was administered by an enumerator except for the section on sexual behavior: for a random half of the respondents, this section was self-administered on paper, while for the remainder it was also administered by the enumerator.

The last section of the questionnaire consisted of a Warner model RRT module with four of the same questions as those asked directly. The RRT protocol used a coin flip: respondents were instructed to tell the truth if they got heads, and to lie if they got tails. As discussed in Section 2.2, this means that if participants respect the RRT protocol, the share of respondents answering "yes" to all four RRT questions should be 50%. The four questions were (Aronow et al., 2015): Have you ever had sex without a condom (Blair and Imai, 2012)? Have you ever had a sexual partner over the age of 25 (Blair et al., 2015)? Have you ever had multiple sexual partners in the same period? and (Böckenholt et al., 2009) Have you ever had a sponsor? (meaning, a sexual partner that provides for you financially).

In the survey, direct questions on these sensitive behaviors were

[2] We test this in Appendix Table A2, which shows how the two groups compare in terms of baseline characteristics. The two groups appear balanced, though female participants more so than males, suggesting that the gaps observed in Table 1 are unlikely to be driven by imbalance.

Table 2
RRT with coin flip in Cameroon: Testing that $r = 50\%$.

	Average reported prevalence under RRT conditions (r)			P-value
	(1)	(2)	(3)	Test (2)= (3)
Panel A. By Survey Type	All	In-Person	Self- Administered	
Sex without a condom	0.362	0.340	0.385	0.027
P-value: Test $r = 0.5$	<0.001	<0.001	<0.001	
Sexual partner >25 years old	0.328	0.313	0.345	0.112
P-value: Test $r = 0.5$	<0.001	<0.001	<0.001	
Multiple partners in same period	0.314	0.308	0.320	0.563
P-value: Test $r = 0.5$	<0.001	<0.001	<0.001	
Sponsor/"sugar daddy"	0.337	0.329	0.346	0.388
P-value: Test $r = 0.5$	<0.001	<0.001	<0.001	
Observations	3712	1959	1753	
Panel B. By Direct Question Answer		Yes in DQ	No in DQ	
Sex without a condom		0.580	0.336	<0.001
P-value: Test $r = 0.5$		0.002	<0.001	
Sexual partner >25 years old		0.515	0.317	<0.001
P-value: Test $r = 0.5$		0.669	<0.001	
Multiple partners in same period		0.532	0.309	<0.001
P-value: Test $r = 0.5$		0.572	<0.001	
Sponsor/"sugar daddy"		0.408	0.315	<0.001
P-value: Test $r = 0.5$		<0.001	<0.001	
Panel C. By Treatment Status[a]		Treatment	Control	
Sex without a condom		0.360	0.371	0.755
P-value: Test $r = 0.5$		<0.001	<0.001	
Sexual partner >25 years old		0.322	0.313	0.783
P-value: Test $r = 0.5$		<0.001	<0.001	
Multiple partners in same period		0.311	0.292	0.561
P-value: Test $r = 0.5$		<0.001	<0.001	
Sponsor/"sugar daddy"		0.328	0.327	0.973
P-value: Test $r = 0.5$		<0.001	<0.001	
Observations (Panel C)		2473	367	

Note: T-tests were run to test whether the means of each question are different from 0.5.
[a] We consider the randomized HIV education interventions from Dupas et al. (2018). The number of observations is lower for this panel than for Panels A and B because we focus on the subset of girls directly involved in the experiment (grade 8 girls).

included in a long list of other questions that were asked first, before the RRT section (done at the very end). This may affect response to RRT questions because people who denied the behavior in DQ may want to remain consistent and act 'as if' they do not have the behavior in RRT as well. They may 'comply' to the RRT rules but under this fake status, i.e. responding "no" when assigned to the truth and "yes" when assigned to lie, as if they truly did not engage in the behavior. If some respondents who truly did the behavior behave in this way (i.e. denied in DQ and want to appear consistent across measures by 'complying' to the RTT rules under a fake status), our test will be passed (we would observe $r = 0.5$) but the estimated prevalence from RRT would be underestimated (estimated *s*inferior to true *s*). To avoid this situation, it is thus preferable to randomize across respondents if DQ comes before or after IR. That said, as noted above, this type of non-compliance behavior may happen even in the absence of DQ. DQ may just amplify this phenomenon if respondents who denied in DQ want to appear consistent across measures by adopting this very specific non-compliance behavior. However, it seems more likely that these respondents chose to answer "no" in RRT whatever the outcome of randomization, which our test detects. Therefore if non-compliance to RRT instructions is found substantial in our data, it cannot be driven by the presence of DQ before RRT.

4.2. Results

The prevalence observed under DQ is shown in Table A4 Panel A, separately for in-person (enumerator-led) vs. self-administered surveys.[3] Because the majority of respondents declared not being sexually active, information on most of the questions have been imputed as zero. Interestingly, we see some significant differences in reported prevalence across the two modes of administration, with preva-lence of socially non-desirable behaviors like "Ever had sex without a condom" and "Ever had sexual partners >25 years old" being larger under self-administration (+31% and +38% respec-tively, significant at 5% level). For "Multiple partners in same period", prevalence is 47% larger under self-administration mode than in-person mode but this estimate is less precise (p-value 0.15). Consistent with the previous literature (see Tourangeau and Yan, 2007), these findings suggest that if DQ understates the true prevalence, self-administration slightly reduces the problem relative to in-person administration. However, we do not know to what extent there may remain substantial understatement, or even overstatement, under self-administration compared to the true prevalence. Panel A of Table 2 shows the results of the test of compliance with RRT instructions. While the mean share of respondents answering "yes" should be 0.5 under perfect compliance, we can clearly reject that it is the case for all four questions. This is despite the fact that 99.5% of respondents said they had understood the instructions and could accurately answer the comprehension test question "should you lie or be truthful if your coin flip shows "head"?". The problem is only marginally lesspronounced

[3] Prevalences were typically built from triangulation across several questions related to each relevant sexual behav-ior. "Ever had sex without a condom" DQ answers based on "Used condom in the first intercourse", "Ever had sex without condom in the last 12 months", "Ever had sex without condom with the last partner". "Ever had partner >25" DQ answers based on "Age of first partner", "Oldest partner in last 12 months" and "Age of last partner". "Multiple partners" DQ answers based on "Ever had multiple partners". "Ever had a sponsor" DQ answers based on "Total number of sponsors" and "Had a sponsor in last 12 months.".

among self-administered survey respondents.

To understand patterns of non-compliance, we first look at RRT responses by types of answer in the direct questions. We do this in Panel B of Table 2, where we look at the probability of "yes" and "no" answers in RRT given respondents' answers in DQ.[4] We find that individuals who are OK with admitting to have engaged in the behavior in DQ appear quite likely overall to be compliant with the RRT instructions: we cannot reject that the average prevalence is 50% as expected for two of the four behaviors ("Ever had sexual partner $>$ 25 years old" and "Multiple partners in the same period"), and distance to 50% is much smaller for these individuals than those who answer "No" in DQ for all four behaviors. We note that for one of them, "Ever sex without condom", we find a prevalence greater than 50%, however, which is possibly suggestive of false positives (people over-reporting). Overall, individuals who admitted to have engaged in socially undesirable behaviors in DQ fairly comply to instructions. This indicates that these respondents were not confused nor used a self-protective strategy.

In contrast, the patterns of responses are severely skewed away from 50% for those who re-sponded "No" in DQ, suggesting that individuals who report not engaging in the behavior when asked directly are particularly averse to answering "Yes" even if that is a lie. These are the non-compliers that Böckenholt et al. (2009) label "non-pseudo incriminators (NPI)" – respondents who will not give a "self-incriminating" answer, even though they truly have not engaged in the sensitive behavior. Among these "No" answers, there could also be "self--protectors" (also using the Böck-enholt et al. (2009) terminology): respondents who do not trust the RRT procedure to be discrete enough to hide that they truly engaged in the behavior and will therefore implement self-protective (SP) behavior by answering "No" no matter the RRT assignment. These protective responses to RRT may occur because it is difficult to understand why randomization protects confidentiality. It may also result from confusion, although we have no reason to believe that individuals who report not engaging in the behavior in DQ would be more likely to get confused about the instructions than those who admitted to have engaged in the behavior.[5]

Non-compliance with RRT instructions may be particularly problematic in impact evaluations. Indeed, if exposure to a given program reduces the stigma associated with a given behavior, compli-ance with RRT may increase in the treatment group. This could lead to seriously flawed inference. Conversely, if an intervention makes salient what is socially undesirable (e.g. think of smoking pre-vention campaign among teenagers), an intervention may increase non-compliance in RRT. This would also lead to flawed estimates of treatment effects. In Panel C of Table 2, we test whether compliance in RRT is affected by exposure to an intervention. In Cameroon, a subset of classes was randomly selected for HIV education interventions (see Dupas et al., 2018). We look at compliance with RRT separately for girls in the treatment vs. control group. We find significant departure from perfect compliance in both groups, and we cannot reject that non-compliance is comparable across groups. While the interventions considered did not affect compliance levels in this setting, we urge researchers planning to use RRT for primary outcomes in an impact evaluation to incorporate our test in their data collection, since there are clear reasons to think interventions have the potential to influence compliance.

5. Conclusion

This paper provides internal consistency tests for two methods aimed at veiling the answers of individual respondents who may or may not have engaged in a behavior that is not socially desirable: the list experiment and the Warner and Crosswise models of the random response technique. The tests are trivial to include in a survey tool at minimal costs: they may increase the length of the indirect response module by only a couple of minutes.

Data from two settings suggest that both techniques can easily fail at fulfilling validity condi-tions. In the list experiment, we show that the Double List Experiment so far used exclusively to reduce variance can also be very useful to test whether answers vary with the baseline items in the list, invalidating the technique. Regarding random response technique, we find that respondents do not comply to the technique because they are reluctant to report that they engaged in the targeted behavior even if the technique does not allow the interviewer to observe their actual behavior. Only respondents who admit having engaged into the behavior when asked directly appear to somewhat comply, which means that for those individuals direct questioning or random response technique is equivalent. Those individuals who do not admit having engaged into the behavior when asked directly do not comply. Requiring self-incriminating responses from people who behaved according to the acceptable social norm may be the most important challenge for this technique.

Further research is needed to develop new tools to measure socially undesirable behaviors. In the meantime, the simple tests we develop are trivial to include so researchers can at least easily know whether their study population is compliant enough for their indirect response estimates to be meaningful.

Data availability

Data is available online at https://web.stanford.edu/~pdupas.

[4] Included among the "no" answers are individuals who indicated "no" when asked if they ever had sex and were therefore skipped when asked about particular behaviors in DQ. Their "no" answers for the initial question are taken as "no"s for all subsequent questions related to sexual behavior.

[5] While this is not a concern in our case, in RRT studies with a low or high p, respondents may rightfully be concerned that their answer does provide an informative signal about their true behavior. E.g. if a behavior is rare such that s $= 0.10$, and if p $= 0.25$, then under full compliance a respondent who has engaged in the behavior is 3 times as likely to report "yes" than a respondent who has not engaged in the behavior.

Appendix A

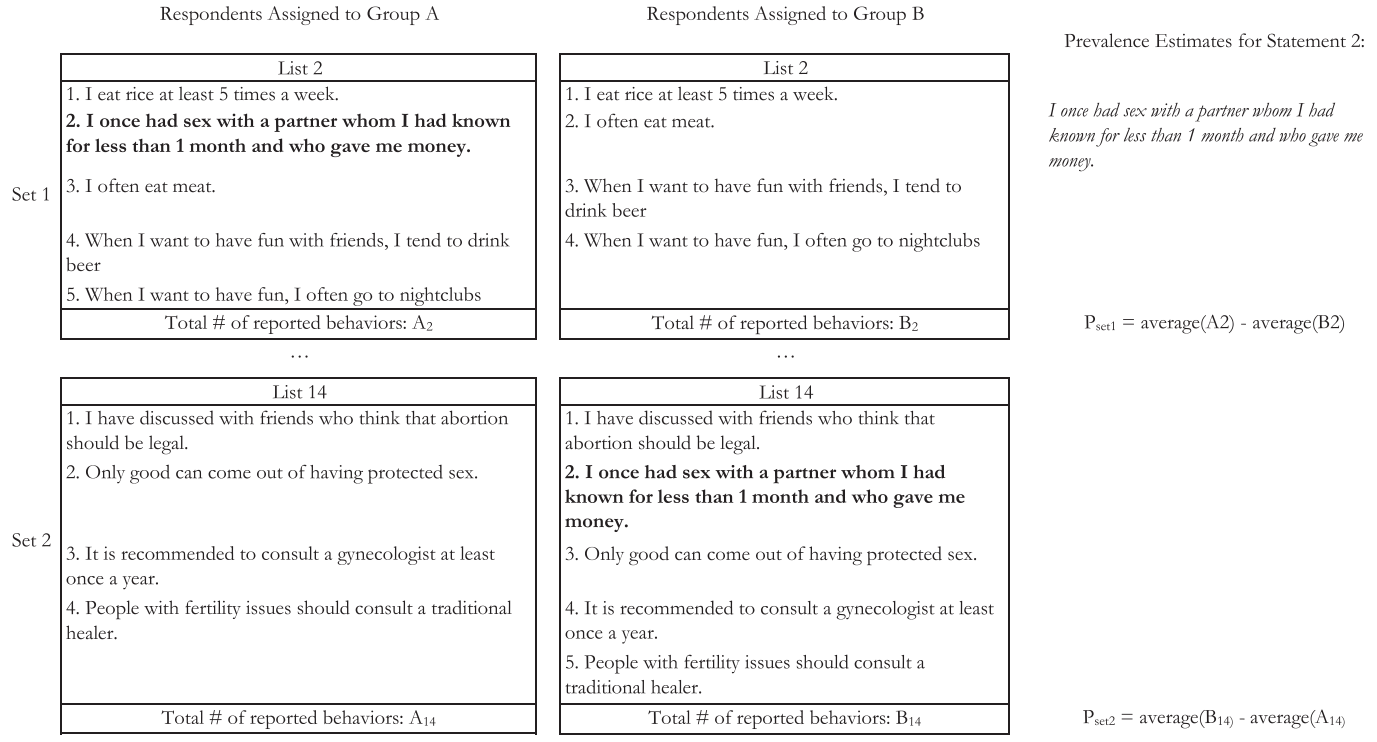

Fig. A1. Illustration of Method Testing Robustness of List Randomization (Example for statement 2)
Notes: We test the sensitivity of LE by comparing the estimated prevalence Pset1 with estimated prevalence Pset2. If the method is not sensitive to baseline statements, these should be equal. We do this statement by statement. These tests are shown in Table 1.

Table A1
List of baseline items in DLE (Cote d'Ivoire)

I	J	Sensitive items	Baseline items: Set 1	Baseline items: Set 2
1	13	I once had unprotected sex with someone the same day I met that person.	I took a private taxi yesterday I own a bank account I took a shared taxi yesterday I am a member of a savings club	I think a woman should not have sex before marriage. I have never attended a wedding for a forced marriage. I met one of my ex's in a maqui/bar. I take time to know my partner well before considering getting married.
2	14	I once had sex with a partner whom I had known for less than 1 month and who gave me money.	I eat rice at least 5 times a week I eat meet often When I want to have fun with friends, I tend to drink beer When I want to have fun, I often go to nightclubs	I have discussed with friends who think that abortion should be legal. Only good can come out of having protected sex. It is recommended to consult a gynecologist at least once a year. People with fertility issues should consult a traditional healer.
3	15	In the last 12 months, I have had concurrent sexual partners	I had to work to pay my school fees I have attended university for at least 2 years I drive a car from time to time I use public mini-buses for my daily commute	People do not use condoms because it reduces sexual pleasure. I always use condoms because they are 100% effective against sexually transmitted diseases. I do not know any woman who has been victim of female genital cutting. I have talked to a woman who thought she needed to go through female genital cutting in order to find a good husband.
4	16	The last time I had sex, it was with a non-marital/non-cohabitant partner and I did not use condom.	I went back home during the holidays I went abroad during the holidays I shop at the open market I shop at the supermarket	I often talk about AIDS with my family. I know a person who got gonorreah. My sexual partners typically visit me at home. I do not need to settle in a long-term relationship for the moment.
5	17	I have had a sexual partner who supported me financially.	Last time I fell sick, I went to see the traditional healer Last time I fell sick, I went to the hospital When I have an infection I buy antibiotics from the drugstore When I get the flu I use traditional medicines to heal myself	At least 50% of women I know cheat on their partner. A woman should know how to cook well for her husband. If I do not want to have sex, I would rather use an excuse than tell my partner. I know women who discovered that their partners cheated on them.
6	18	In the last 12 months, I had a partner whom I suspected to be HIV positive.	I only watch national TV programs I watch satellite TV programs	I have never had sexual relationships with a workmate. I have never received gold from any partner.

(*continued on next page*)

Table A1 (*continued*)

I	J	Sensitive items	Baseline items: Set 1	Baseline items: Set 2
			I go to a cyber-café to check my emails	My partner gave me something precious for my last birthday.
			I check email almost daily	Most of my partners are poorer than me.
7	19	I had at least two non-cohabitant or non-marital partners in the last 12 months.	I invite friends over every weekend	In marriage, how much money a man earns is more important than his age
			I prefer to meet my friends outside the house	If I discovered that my partner cheated on me, I'd give him/her a second chance.
			I went to the movies last week	My mother got married before the age of 15.
			I went to the videoclub last week	I prefer men with education to men with money.
8	2	W: I had at least one abortion.M: At least one of my partners had an abortion.	I exercise regularly	I know someone who is gay.
			I exercise about once a month on average	I put more trust in a man who goes to church regularly.
			I read at least one book last year	I know men who beat their wives out of jealousy.
			Quite often I stop reading novels half-way through	I know men who go out with very young women.
9	21	During the 2010–11 post-electoral economic crisis, I had sex with partners who gave me money.	I don't like to drink coffee in the morning	People often meet sexual partners in bars because of alcohol.
			I often eat fruits during the rainy season	If people want to avoid HIV, they should abstain until marriage.
			I spent 2 out of 3 nights watching TV at home with my family	My partner consumes beer very often.
			I often work at night	Smoking is equally bad for men and women.
10	22	W: I had at least one abortion witha traditional healer M: At least one of my partners had an abortion witha traditional healer.	Last Sunday I went to church and made a donation	When a man cares about a woman, he gives her money.
			I sometimes work on Sundays	A man's money is less important than the love he shows for his partner.
			I know at least one person who has been affected by witchcraft	My first partner was a friend of mine.
			I constantly read newspapers to stay informed	Men should distrust women as much as women should distrust men.
11	23	W: I had at least one abortion before the age of 18. M: At least one of my partners had an abortion before of 18	I never go to bed after 11pm	It is better to avoid looking into a man's cell phone.
			I typically get up after 7am	Trusting others is OK but not trusting them is typically safer.
			I am sometimes late paying rent	Men cheat on their wives because of other women.
			I don't like people who are not on time	It is very common for men to have children on the side.
12	24	W: I had an abortion in the last two years due to the crisis. M: I have a sexual partner who had an abortion in the last two years due to the crisis.	I buy newspapers once in a while	Many women have abortions even though it is illegal.
			I watch sitcoms almost daily	Abortions make women infertile.
			I like to listen to traditional music	It is very common for married people to have extramarital affairs.
			I rarely dress up traditionally	It is ok for a man to have multiple wives as long as he can care for all of them.

Table A2
List Randomization in Cote d'Ivoire: Balance check

	Female			Male		
	Group A	Group B	P-value of diff in means	Group A	Group B	P-value of diff in means
Age (mean)	31.465	31.034	0.414	32.311	33.262	0.099
Age (standard deviation)	(8.641)	(8.465)		(8.212)	(9.546)	
Married	0.262	0.285	0.400	0.232	0.280	0.087
Married/Cohabiting	0.516	0.501	0.625	0.483	0.497	0.663
Has a child	0.896	0.641	0.855	0.560	0.554	0.855
Urban	0.896	0.864	0.452	0.863	0.936	0.109
Lives in Autonomous Region of Abidjan	0.531	0.522	0.768	0.455	0.446	0.777
Years of successfully completed education	11.003	11.580	0.069	12.299	12.000	0.295
Avg HH income						
Income <100,000 FCFA/month	0.373	0.361	0.699	0.380	0.425	0.152
100,000<Income<400,000 FCFA/month	0.473	0.434	0.208	0.466	0.423	0.188
400,000<Income<800,000 FCFA/month	0.066	0.090	0.147	0.071	0.076	0.774
800,000<Income<1,200,000 FCFA/month	0.011	0.015	0.572	0.017	0.012	0.531
200,000<Income<1,500,000 FCFA/month	0.002	0.006	0.311	0.015	0.006	0.182
Income>1,500,000 FCFA/month	0.004	0.002	0.572	0.004	0.008	0.444
Employed in the last 30 days	0.667	0.700	0.248	0.773	0.793	0.434
Ever had sex	0.946	0.932	0.339	0.965	0.926	0.007
Age of first sexual intercourse						
<12 years old	0.002	0.004	0.558	0.031	0.033	0.823
Between 12 and 14 years old	0.065	0.082	0.315	0.098	0.097	0.962
Between 15 and 17 years old	0.420	0.425	0.873	0.353	0.355	0.938
Between 18 and 24 years old	0.424	0.412	0.683	0.427	0.384	0.182
Older than 24 years old	0.035	0.010	0.006	0.057	0.056	0.954
Had sex in the last 3 months	0.666	0.658	0.792	0.682	0.657	0.417
Ever used a condom during sex	0.795	0.827	0.195	0.915	0.886	0.145
Has adopted behavior to avoid HIV	0.709	0.707	0.943	0.786	0.730	0.045
Observations	531	523		466	489	

Table A3
List Experiment in Cote d'Ivoire: Baseline items Variance and Extreme Values

	Set 1 (Group B)					Set 2 (Group A)				
	Number of items reported					Number of items reported				# Obs
	Mean	Std. Dev	0 out of 4	4 out of 4	# Obs	Mean	Std. Dev	0 out of 4	4 out of 4	
Panel A. Females										
1 (13)	1.220	0.977	0.239	0.021	522	1.951	0.901	0.041	0.043	531
2 (14)	1.979	0.892	0.038	0.059	522	2.121	0.873	0.032	0.049	531
3 (15)	1.065	0.902	0.289	0.004	522	1.766	0.896	0.068	0.019	530
4 (16)	1.850	0.804	0.021	0.029	521	1.385	0.966	0.189	0.023	530
5 (17)	1.879	0.756	0.019	0.033	522	2.774	0.996	0.009	0.277	530
6 (18)	1.349	0.827	0.123	0.013	522	1.313	1.010	0.252	0.011	528
7 (19)	0.625	0.707	0.487	0.002	522	1.879	0.800	0.034	0.109	531
8 (20)	1.167	1.054	0.341	0.017	522	2.215	1.024	0.028	0.021	527
9 (21)	1.755	0.920	0.075	0.031	522	2.068	1.005	0.047	0.081	531
10 (22)	1.695	0.901	0.077	0.025	522	1.927	0.886	0.036	0.040	531
11 (23)	1.839	0.966	0.050	0.059	522	2.677	0.974	0.009	0.217	530
12 (24)	1.935	0.893	0.042	0.042	522	2.357	1.038	0.038	0.142	530
Average	*1.530*		*0.160*	*0.027*		*2.036*		*0.068*	*0.081*	
Panel B. Males										
1 (13)	1.125	0.940	0.268	0.016	489	2.032	0.935	0.039	0.062	466
2 (14)	2.100	0.969	0.031	0.100	489	2.172	0.902	0.032	0.054	466
3 (15)	1.254	1.018	0.258	0.027	488	1.944	0.912	0.039	0.043	466
4 (16)	1.785	0.815	0.027	0.027	489	1.708	0.973	0.109	0.028	466
5 (17)	1.896	0.778	0.020	0.031	489	2.798	1.021	0.002	0.305	466
6 (18)	1.530	0.880	0.096	0.025	489	1.661	1.008	0.135	0.030	466
7 (19)	0.945	0.795	0.285	0.008	488	1.755	0.845	0.052	0.021	466
8 (20)	1.681	1.100	0.157	0.055	489	2.249	1.063	0.054	0.133	466
9 (21)	1.924	0.944	0.057	0.039	488	2.230	0.970	0.030	0.088	465
10 (22)	1.881	0.964	0.047	0.051	489	2.017	0.977	0.034	0.077	466
11 (23)	1.969	0.922	0.035	0.057	489	2.683	0.966	0.011	0.220	464
12 (24)	1.957	0.947	0.045	0.061	489	2.461	1.067	0.034	0.180	466
Average	*1.671*		*0.116*	*0.040*		*2.114*		*0.049*	*0.096*	

Notes: This table shows the mean, std. dev., and frequency of extreme responses, i.e. where baseline items are either too uncommon (where responses are reported as "0", or never engaged in any of the baseline behaviors) or too common ("4", or engaged in all behaviors). Extreme values are indicative of floor or ceiling effects, both of which can undermine the validity of the LE method.

Table A4
Direct response in Cameroon: Enumerator-led vs. Self-Administered

	In-Person		Self-Administered		P-Value of Diff between Means
	Mean	N. of Obs	Mean	N. of Obs	
Panel A. Share Respondents Who Admitted to a Given Behavior					
Ever had sex	0.314	1961	0.334	1753	0.421
If ever had sex: Ever had sex without a condom	0.290	610	0.353	584	0.032
(Full sample mean with Imputed missing values)*	0.090	1961	0.118	1753	0.036
If ever had sex: Ever had Sexual partner > 25 years old	0.143	615	0.185	585	0.062
(Full sample mean with Imputed missing values)*	0.045	1961	0.062	1753	0.041
If ever had sex: Multiple partners in same period	0.055	615	0.074	585	0.210
(Full sample mean with Imputed missing values)*	0.017	1961	0.025	1753	0.152
If ever had sex: Had Sponsor/"sugar daddy"	0.861	519	0.860	470	0.942
(Full sample mean with Imputed missing values)*	0.228	1961	0.230	1753	0.906
Panel B. Balance on Demographic Characteristics					
Age	15.787	1961	15.921	1753	0.170
	(1.749)		(1.717)		
Married	0.005	1958	0.002	1753	0.282
	(0.068)		(0.048)		
Married/Cohabiting	0.010	1958	0.012	1753	0.566
	(0.098)		(0.109)		
Currently has a child*	0.038	812	0.047	676	0.446
	(0.192)		(0.213)		
Region 2 (Yaoundé)	0.318	2277	0.278	2032	0.455
	(0.466)		(0.448)		
Region 3 (South)	0.109	2277	0.118	2032	0.816
	(0.312)		(0.322)		
Region 4 (West)	0.574	2277	0.604	2032	0.590
	(0.495)		(0.489)		
Currently enrolled in school	0.979	1961	0.969	1753	0.151
	(0.145)		(0.173)		

(*continued on next page*)

Table A4 (*continued*)

	In-Person		Self-Administered		P-Value of Diff between Means
	Mean	N. of Obs	Mean	N. of Obs	
Current class/grade	2.907 (0.673)	1918	2.897 (0.714)	1696	0.714
Mother's highest class/grade reached	2.885 (0.994)	1776	2.709 (0.999)	1571	0.009***
Father's highest class/grade reached	3.425 (1.269)	1571	3.302 (1.407)	1412	0.184

Note: *Imputed zero if respondent was not sexually active or was non-responsive.

References

Aronow, P., Coppock, A., t Crawford, F., Green, D., 2015. Combining list experiment and direct question estimates of sensitive behavior prevalence. Journal of Survey Statistics and Methodology 3, 43–66.

Blair, G., Imai, K., 2012. Statistical analysis of list experiments. Polit. Anal. 20, 47–77.

Blair, G., Imai, K., Zhou, Y., 2015. Design and analysis of the randomized response technique. J. Am. Stat. Assoc. 110 (511), 1304–1319.

Böckenholt, U., Barlas, S., Van Der Heijden, P., 2009. Do randomized-response designs eliminate response biases? An empirical study of non-compliance behavior. J. Appl. Econom. 24 (3), 377–392.

Chou, W., 2019. Lying on Surveys: Methods for List Experiments with Direct Questioning (Working paper).

Clark, S.J., Desharnais, R.A., 1998. Honest answers to embarrassing questions: detecting cheating in the randomized response model. Psychol. Methods 3, 160–168.

Coutts, E., Jann, B., 2011. Sensitive questions in online surveys: experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). Socio. Methods Res. 40 (1), 169–193.

Droitcour, J., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W., Ezzati, T.M., 1991. The item count technique as a method of indirect questioning: a review of its development and a case study application. In: Biemer, P., Groves, R. (Eds.), Measurement Errors In Surveys. John Wiley & Sons, New York, NY.

Dupas, P., Huilllery, E., Seban, J., 2018. Risk information, risk salience, and adolescent sexual behavior: experimental evidence from Cameroon. J. Econ. Behav. Organ. 145 (1), 151–175.

Eady, G., 2017. The statistical analysis of misreporting on sensitive survey questions. Polit. Anal. 25, 241–259.

Fisher, M., Kupferman, L.B., Lesser, M., 1992. Substance use in a school-based clinic population: use of the randomized response technique to estimate prevalence. J. Adolesc. Health 13 (4), 281–285.

Glynn, A.N., 2013. What can we learn with statistical truth serum? Design and Analysis of the List Experiment" *Public Opinion Quarterly* 77, 159–172.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., Horvitz, D.G., 1971. Application of the randomized response technique in obtaining quantitative data. J. Am. Stat. Assoc. 66 (334), 243.

Heck, D., Hoffmann, A., Moshagen, M., 2018. Detecting nonadherence without loss in efficiency: a simple extension of the crosswise model. Behav. Res. Methods 50, 1895–1905.

Höglinger, M., Diekmann, A., 2017. Uncovering a blind spot in sensitive question research: false positives undermine the crosswise-model RRT. Polit. Anal. 25 (1), 131–137.

Höglinger, M., Jann, B., 2018. More is not always better: an experimental individual-level validation of the randomized response technique and the crosswise model. PloS One 13 (8), e0201770.

Imai, K., 2011. Multivariate regression analysis for the item count technique. J. Am. Stat. Assoc. 106 (494), 407–416.

Krumpal, I., 2012. Estimating the prevalence of xenophobia and anti-semitism in Germany: a comparison of randomized response and direct questioning. Soc. Sci. Res. 41 (6), 1387–1403.

Lensvelt-Mulders, G., Hox, J.J., Van Der Heijden, P., Maas, C., 2005. Meta-analysis of randomized response research: thirty-five years of validation. Socio. Methods Res. 33 (3), 319–348.

Mckenzie, D., Siegel, M., 2013. Eliciting illegal migration rates through list randomization. Migration Studies 1 (3), 276–291.

Miller, J.D., 1984. A New Survey Technique for Studying Deviant Behavior. U Microfilms International, Ann Arbor.

Moshagen, M., Musch, J., 2012. Surveying multiple sensitive attributes using an extension of the randomized-response technique. Int. J. Publ. Opin. Res. 24, 508–523.

Moshagen, M., Musch, J., Erdfelder, E., 2012. A stochastic lie detector. Behav. Res. Methods 44, 222–231.

Ostapczuk, M., Moshagen, M., Zhao, Z., Musch, J., 2009. Reducing socially desirable responses in epidemiologic surveys: an extension of the randomized-response technique. Epidemiology 21, 379–382.

Preisendörfer, P., Wolter, F., 2014. Who is telling the truth? A validation study on determinants of response behavior in surveys. Publ. Opin. Q. 78, 126–146.

Raghavarao, D., Federer, W.T., 1979. Block total response as an alternative to the randomized response method in surveys. J. Roy. Stat. Soc. B 41 (1), 40–45.

Rosenfeld, B., Imai, K., Shapiro, J.N., 2015. An empirical validation study of popu- lar survey methodologies for sensitive questions. Am. J. Polit. Sci. 60 (3), 783–802.

Soeken, K.L., Macready, G.B., 1982. "Respondents' perceived protection when using randomized response". Psychol. Bull. 92, 487–489.

Tourangeau, R., Yan, T., 2007. Sensitive questions in surveys. Psychol. Bull. 133, 859–883.

Tsuchiya, T., Hirai, Y., Ono, S., 2007. A study of the properties of the item count technique. Publ. Opin. Q. 71 (2), 253–272.

Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. 60 (309), 63.