I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18072

Unbiased and Accurate: Measuring Sensitive Outcomes Through Ballot-Bag Surveys

Bruno Crépon
Ahmed Elsayed
Jules Gazeaud

AUGUST 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18072

Unbiased and Accurate: Measuring Sensitive Outcomes Through Ballot-Bag Surveys

Bruno Crépon
CREST-ENSAE and IZA

Ahmed Elsayed
American University in Cairo, J-PAL, MENA and IZA

Jules Gazeaud
Université Clermont Auvergne, CNRS, IRD and CERDI

AUGUST 2025

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Unbiased and Accurate: Measuring Sensitive Outcomes Through Ballot-Bag Surveys[*]

Prevailing methods for measuring sensitive outcomes confront researchers with an inherent bias-variance trade-off: direct questioning is prone to a sensitivity bias, while indirect methods such as list experiments are substantially less precise. We introduce the ballot-bag, a novel technique that relaxes this trade-off by mitigating bias in direct questioning while improving precision over indirect methods. In a field experiment in Egypt, where direct questions on irregular migration are biased, ballot-bag estimates closely align with those from a list experiment but exhibit significantly lower variance. Consequently, treatment effects are highly significant via the ballot-bag and not via the list experiment.

JEL Classification: C83, C93, J61, O15, D83

Keywords: sensitive behaviors, survey methods, sensitivity bias, field experiments, irregular migration

Corresponding author:
Ahmed Elsayed
American University in Cairo
AUC Avenue, P.O. Box 74
New Cairo 11835
Egypt
E-mail: a_elsayed@aucegypt.edu

[*] We thank Maha Adel, Nada Alaa, Fatma Bakr, Amira El-Mashad, Diaa Elsonny, Nada Esmail, Mahmoud Hussein, Anas Mamdouh, Mohamed Mofreh, Mazen Mohamed, Nada Rostom, Engy Saleh, Hanan Salem, and Maria Sedhom for excellent research assistance, project management, and field coordination. We also thank the management of Prominsing Egypt for their efforts in data collection, including Mohamed Zanati, Asmaa Elfoly, Eslam Abutaleb, Ahmed Allam, Mostafa Abdelreheem, Eman Abdelfattah, and Dai Salah. We are grateful to Simone Bertoli, Elise Huillery, Aurélia Lépine, Francesca Marchetta, Adam Osman, Raúl Sanchez de la Sierra, and participants at various conferences, workshops, and seminars for helpful comments. This research makes use of original data collected with Dina Abdel Fattah and MSMEDA as part of a broader project on irregular migration (AEARCTR-0010604; IRB received from the American University in Cairo). We are grateful for their collaboration. We acknowledge financial support from the European Union's Emergency Trust Fund for Africa and the Egypt Impact Lab. Gazeaud acknowledges support from the Agence Nationale de la Recherche of the French government through the program "France 2030" ANR-16-IDEX-0001.

1 Introduction

A persistent challenge in empirical research is how to obtain truthful answers to sensitive questions, as misreporting can systematically bias estimates and misinform policy. Direct questions on sensitive topics (e.g., drug use, prejudice, corruption, intimate partner violence, irregular migration) often suffer from underreporting due to respondents' concerns over social judgment, potential sanctions, or self-image (Tourangeau and Yan, 2007; Rosenfeld et al., 2016; Chuang et al., 2021). In response, researchers have introduced methods to enhance respondent anonymity and mitigate biases. However, despite decades of innovations, the choice between methods still confronts researchers with a stark bias-variance trade-off (Blair et al., 2020). Indirect methods are substantially less precise, while direct questioning and existing self-administered methods do not fully address respondents' concerns about anonymity and may therefore be biased.

In this paper, we introduce the *ballot-bag*, a self-administered technique designed to provide full anonymity to respondents while achieving a level of precision similar to direct questioning. In the context of in-person surveys conducted by enumerators, respondents privately report their answers to one or more sensitive questions on anonymous paper sheets, which they then fold and place into a transparent bag held by an enumerator. Crucially, respondents are instructed not to include identifying information such as their names on the paper. Once each respondent mixes their paper with those already collected, responses in the ballot-bag cannot be linked to individual respondents or to the rest of the questionnaire.[1] The method is particularly appealing due to its simplicity and because respondents can visually understand how anonymity is provided. The spirit of the method is not entirely new and mimics the technique used by Gallup and other polling organizations in the early days of public opinion polls (Benson, 1941; Turnbull, 1947), where respondents marked their views on sensitive topics on a ballot and placed it in a box labeled "secret"—a similar technique is used by Bishop and Fisher (1995) to measure voter choices in exit polls.[2] We adapt the method for in-person surveys conducted by enumerators, using a low-cost protocol and transparent bags that are easy to transport. This approach may prove particularly useful in

[1] For the first respondents, enumerators explain that papers from other respondents will accumulate in the bag, such that no individual paper can eventually be traced back to them.

[2] The method belongs to the family of self-administered techniques, of which audio computer-assisted self-interviewing (ACASI) is the most commonly used (Chauchard, 2013; Park et al., 2024; Peterman et al., 2024). However, a key limitation of existing self-administered methods is that they conceal answers only from enumerators, not from analysts. In the context of demographic surveys in Nigeria, Valente et al. (2024) show that respondents are less likely to provide sensitive answers when analysts can still observe individual answers.

field experiments which are now commonplace in economics.

We test the ballot-bag in the context of a large-scale field experiment in Egypt examining the impact of active labor market policies on irregular migration to Europe.[3] We compare estimates using the ballot-bag with those obtained through direct questioning and a (double) list experiment. First introduced by Miller (1984), the list experiment (a.k.a., the item count technique or list randomization) has become the indirect method of choice for eliciting truthful responses on sensitive topics across social sciences, including in economics.[4] While other indirect methods exist—e.g., the randomized response technique (Warner, 1965) or the cross-wise technique (Yu et al., 2008)—a common trait of all these methods is that they reduce sensitivity bias by adding noise to the measurement. Blair et al. (2020) show that under typical conditions, list experiments are approximately 14 times noisier than direct questions. McKenzie and Siegel (2013) use a list experiment to estimate irregular migration rates in Ethiopia, Mexico, Morocco, and the Philippines, and find that the 95% confidence intervals typically span 20 percentage points despite sample sizes exceeding 1,000 in each country. The scale of our experiment (N=8,385) offers an ideal setting to conduct highly powered tests for the presence of a sensitivity bias and to assess the potential of the ballot-bag to address it.

Our main results challenge the idea that ensuring respondents' anonymity must come at the expense of precision. Consistent with the presence of a sensitivity bias, we estimate that aspirations for irregular migration are higher using the list experiment than using direct questioning. However, the list experiment is several times less precise than direct questioning, which highlights the existence of a bias-variance trade-off in the setting we study. Interestingly, the ballot-bag appears to similarly address concerns over sensitivity, as its estimates fully align with those from the list experiment. But unlike the list experiment, the ballot-bag achieves a level of precision on par with direct questioning, showing that statistical accuracy need not be sacrificed for anonymity. This gain in precision has major implications when estimating the treatment effect of an intervention to deter irregular migration: while we estimate a significant reduction in men's aspirations for irregular migration using the ballot-bag (p-value<0.01), the list experiment completely fails to detect this effect (p-value=0.66). The

[3] We identify treatment status using paper sheets of different colors.

[4] List experiments have been used to measure topics such as irregular migration (McKenzie and Siegel, 2013), discrimination (Neggers, 2018; Bursztyn et al., 2020; Aksoy et al., 2025; Osman et al., 2025), religiosity (Bryan et al., 2021), corruption (Detkova et al., 2021; Okunogbe and Pouliquen, 2022), gender-based violence (Dhar et al., 2022; Cullen, 2023; Gilligan et al., 2024; Bertelli et al., 2025), child labor (Jouvin, 2024), among others (Karlan and Zinman, 2012; Chen and Yang, 2019; Lépine et al., 2020; Armand et al., 2023). See Section 2.2 for details on the list experiment.

minimum detectable effect size is 0.42 standard deviations using the list experiment and 0.12 standard deviations using the ballot-bag.

We conclude our analysis by comparing the consistency of individual responses across the different methods. Even when two elicitation methods yield identical averages, this does not imply that the same individuals responded affirmatively in both methods. Some respondents might report "no" in the first method but "yes" in the second method, while other respondents do the opposite. To assess the extent of these inconsistencies, we summarize the joint distribution of responses across methods using confusion matrices.[5] In our setting, this exercise is challenging because ballot-bag responses are anonymous and cannot be linked directly to individual answers in the other methods. Nevertheless, we show that the joint distribution is identifiable under minimal assumptions, using auxiliary variables collected on the ballot and also available in the general survey. We find that individual responses from the ballot-bag and the list experiment are consistent for men but not for women—possibly reflecting the lower underlying prevalence of irregular migration or a lack of understanding of the list experiment among women.

The main contribution of our paper is to introduce and test a new method for measuring sensitive topics. Prior research has documented substantial misreporting when respondents are asked sensitive questions directly (Tourangeau and Yan, 2007). Yet, despite decades of innovations to improve the precision and consistency of alternative methods (Miller, 1984; Blair and Imai, 2012; Glynn, 2013; Aronow et al., 2015; Blair et al., 2020; Chuang et al., 2021), prevailing techniques still confront researchers with an inherent bias-variance trade-off. This trade-off is compounded in settings like ours, where the goal is not only to estimate the prevalence of a sensitive outcome but also to assess the effect of an intervention on that outcome. We provide encouraging evidence that the ballot-bag method can protect respondent anonymity without sacrificing statistical accuracy.

2 Context

We rely on data from a randomized controlled trial (RCT) conducted in Egypt to examine the impact of increased employment on aspirations for irregular migration to Europe (Crépon et al., 2022). We are interested in the answers to the question: "*Do you plan to migrate to Europe without the official papers?*". Given the sensitivity of this question, we use two

[5] A confusion matrix captures the proportion of individuals falling into each possible combination of responses—see Ting (2011) for more details. We focus particularly on false positives and false negatives, using the ballot-bag method as the benchmark.

alternative data collection methods alongside the direct approach: (i) the list experiment, one of the most (if not the most) popular methods in social sciences to veil individual answers to enumerators and researchers, and (ii) the ballot-bag, a novel technique we introduce to address the bias-variance trade-off highlighted in previous work.[6]

2.1 The RCT

Addressing the "root causes" of irregular migration has become a key policy priority in Europe. In 2015, the European Union launched the Emergency Trust Fund for Africa (EUTF) to deter irregular migration flows from 26 origin countries. The program we study is implemented in Egypt by the Micro, Small, and Medium Enterprises Development Agency (MSMEDA). Egypt is an interesting setting for studying irregular migration as it is one of the top countries of origin for irregular migrants arriving to Europe.[7] The program was launched in 2019 under the name "Addressing the Root Causes of Irregular Migration through Employability and Labor-Intensive Works" and received a 28-million-euro grant from the EUTF.

We partnered with MSMEDA to run a field experiment in nine governorates (provinces)[8] to assess the impact of two interventions: (i) a cash-for-work program in which beneficiaries are paid to provide community services; (ii) a training and employment support program in which beneficiaries are provided with training packages and employment services to facilitate access to wage- and self-employment. The objective of these interventions is to reduce irregular migration to Europe by improving the employment prospects of beneficiaries. Both interventions are delivered by local NGOs and target unemployed youth aged 18-35. Interested individuals could apply for one of the interventions and were randomized into treatment and control, with randomization done at the individual level and stratified by NGO and field of training/employment. Our sample for this paper includes the 8,385 individuals who took the follow-up survey.[9] Baseline data collection took place between January 2021 and Octo-

[6] We also used these methods to collect data on the aspirations of other household members for irregular migration, however, we do not focus on this outcome in this paper because of the apparent failure of the list experiment (see Section B.6 in the Online Appendix).

[7] Egypt has ranked among the top ten countries of origin of irregular migrants in each year since 2021. In 2022, Egyptians were the most common nationality among undocumented migrants detected at the EU's external borders with 21,753 undocumented migrants (i.e., 11.5% of the total). See IOM data: https://dtm.iom.int/europe/arrivals.

[8] The nine governorates are: Asyut, Beheira, Dakahlia, Faiyum, Gharbia, Luxor, Minya, Qalyubia, and Sharqia. They were identified by local partners as having high irregular migration rates.

[9] We registered a sample size of 11,733 individuals (AEARCTR-0010604), however, we specified that "*the actual number of observations [would] depend on the number of projects that get canceled and on our*

ber 2022, while follow-up data collection occurred shortly after the end of the interventions, between November 2022 and April 2024 depending on the NGO.

2.2 Three ways to measure the sensitive outcome

Direct Question The direct approach is the traditional method for asking questions in face-to-face surveys. An enumerator simply asks the question "*Do you plan to migrate to Europe without the official papers?*" and records the respondent's answer on a tablet. As with most surveys, we assure respondents that their answers are anonymized prior to data analysis and stored securely. Naturally, respondents who are planning to migrate irregularly to Europe may lack trust in the procedures to manage the data or may not feel comfortable revealing the truth to the enumerators.

Double List Experiment We adopt the method as described by Droitcour et al. (1991) and further developed by Blair and Imai (2012) and Glynn (2013). In a list experiment, respondents are presented with a list of statements that could each be true or false, and they are instructed to only reveal the *number* of statements that are true for them—not which specific statements are true. In the *single* list experiment, respondents are randomized into two groups: a *short list* group, whose list does not contain the sensitive outcome, and a *long list* group, whose list does contain the sensitive statement on top of the exact same statements. The prevalence rate of the sensitive outcome can then be estimated by calculating the difference in the average number of true statements reported by each group. While many applications rely on single list experiments, several papers in the literature highlight the importance of using *double* list experiments to improve the accuracy of the estimates and to test their internal consistency (see e.g., Glynn, 2013; Chuang et al., 2021). The double list experiment involves administering two lists to each respondent, with the sensitive statement randomly included in one. The double list experiment offers two main advantages over the single list experiment. First, by doubling the number of observations per individual, it increases statistical power. Second, it allows checking that the prevalence of the sensitive outcome is consistent across the two lists—a necessary condition for the method to provide reliable estimates (Chuang et al., 2021). Despite these benefits, precision of the

ability to track and find respondents included in the experiment". Prior to the follow-up survey, we excluded from the sample the 467 individuals from the five NGOs whose projects got canceled by MSMEDA due to implementation issues. In addition, 2,365 individuals did not complete the follow-up survey (either because they refused, could not be tracked, or were not sampled during the intensive tracking phase of the survey). Finally, we exclude the 516 individuals in lotteries with only treatment or control observations at follow-up.

double list experiment remains limited, especially when an objective is to assess differences in prevalence rates across groups (Blair et al., 2020). Moreover, the cognitive burden associated with the administration of multiple lists may affect compliance and introduce measurement errors. The specific statements included in our double list experiment are detailed in Table B.1. The sensitive statement is phrased as *"I plan to migrate to Europe without the official papers"*. We detail in Online Appendix B.2 how we selected the non-sensitive statements.

Ballot-Bag The ballot-bag is a new method we developed to provide privacy and anonymity to respondents while addressing issues related to the low accuracy of list experiments. The sensitive question is asked directly during the face-to-face survey ("*Do you plan to migrate to Europe without the official papers?*"), but instead of providing their answers directly to enumerators, respondents report them privately on anonymous sheets of paper. Before responding, participants are told that once they finish, they will fold their paper, place it in a transparent bag held by the enumerator, and mix it with the papers already in the bag. Because respondents can see that there are many papers in the bag, and are instructed not to sign or put their names on the paper, they are provided with some reassurance that their answers are indeed anonymous. Importantly, once a respondent has recorded their answer, there is no way for enumerators or analysts to connect answers in the ballot-bag with the names of the respondents or the rest of the questionnaire. For the first respondents, enumerators are instructed to explain that papers from other respondents will accumulate in the bag such that ultimately it will be impossible to tell which one is theirs.[10] We use papers of different colors to identify treatment and control individuals. However, in other set-ups, this information could also be filled in by the enumerator at the top of the paper, or separate bags could be used for treatment and control. We also include on the sheets of paper basic information on respondents' gender and lottery (randomization strata).[11]

[10] In theory, one could add fake papers to the bags, but this should be done carefully to avoid introducing deception or measurement errors.

[11] Enumerators fill in the fields for gender and lottery at the top of the paper prior to the interview. The lottery information is made as concrete as possible to avoid trust issues. In our setting, randomization into treatment and control was stratified by NGO and field, so we generated unique lottery identifiers from the NGO names and fields (respondents could relate to these identifiers). The lottery information was introduced partway through the data collection, when we realized it was required for rigorous estimation of the treatment effects of the program (62% of the sample was interviewed after we introduced the lottery information). We keep the surveys with missing lottery information and identify them with a specific code.

2.3 Survey structure and implementation

We administer the three methods to each respondent in order to maximize the statistical power of our analysis. However, one concern with this design is that responses in one method may influence responses in the other methods. To address such concerns, we administer the methods in different parts of the survey with extensive blocks of questions in-between. In addition, we randomize the location of the direct question to test for the presence of "order" effects. We chose not to randomize the location of the list experiment and the ballot-bag because testing for order effects using the ballot-bag would require indicating its position on the ballot itself—an abstract information in the eyes of respondents which could lead to trust issues—and because detecting order effects relies on sub-sample analysis, for which list experiments are severely underpowered.

Figure 1 outlines the structure of the survey. The sample is randomly divided into two main groups. Group 1 (20% of the sample) is administered the direct question early in the survey (i.e., *before* the list experiment and the ballot-bag), while Group 2 (80% of the sample) is administered the direct question later in the survey (i.e., *after* the list experiment and the ballot-bag). Group 2 is further divided into two subgroups: Group 2A which includes the sensitive statement in List A, and Group 2B which includes the sensitive statement in List B.[12]

The scripts we use to implement the list experiment and the ballot-bag are reproduced in Online Appendix C. For the list experiment, we include a training list to improve understanding of the method. To assess the quality of the data, we ask enumerators to evaluate respondents' understanding of and compliance with the protocols of the list experiment and of the ballot-bag (using a 5-point Likert scale).[13] For the ballot-bag, field coordinators from J-PAL MENA were responsible for collecting the bags from the survey firm approximately once a month. Double data entry was done by personnel monitored by our research assistants. The scope of fieldwork in our setup was particularly extensive, spanning 18 months across nine governorates spread throughout the country, and involving over one hundred enumerators. This scale posed organizational challenges, resulting in the loss of some bags corresponding to approximately 9% of the observations. In Section 3.1, we discuss how this issue may have influenced our findings and argue that attrition likely occurred at random. Future applications of the method should carefully consider how to minimize bag losses.

[12] We did not divide Group 1 into similar subgroups in order to limit potential bias, as the direct question was asked *just* before the list experiment.

[13] Understanding and compliance appear higher in the ballot-bag than in the list experiment (Table A.5).

2.4 Sample characteristics and balance

Table A.1 presents summary statistics on the respondents and tests whether their characteristics are balanced across the three groups in Figure 1. On average, respondents are 26.7 years old; 74% are female, 57% are married, and 48% report that they engaged in an income-generating activity in the seven days prior to they survey. While we did not collect information on literacy, almost all individuals in our sample have completed primary school (96%), suggesting that most of them are literate. For individuals who struggled with reading in the ballot-bag, enumerators were instructed to explain how the questions are organized on the paper and to read the questions aloud for them while ensuring respondents can reply privately, therefore mimicking human-assisted self-interviewing (Álvarez-Aragón and Champeaux, forthcoming) but with an enhancement in anonymity from analysts. We test the balance between each group across all variables using a multinomial logit specification and find a p-value of 0.80, indicating good balance.

3 Empirical Strategy and Results

3.1 Prevalence of the sensitive outcome across the three methods

For the direct question and the ballot-bag, the share of individuals planning to migrate irregularly to Europe is easily obtained by computing sample averages. The analysis of the list experiment is less straightforward but well-known (see Online Appendix B). In the double list experiment considered here, each respondent is presented with two lists, A and B, and the sensitive statement is randomly included in either list A, list B, or neither.[14] We estimate the prevalence of the sensitive outcome by running the following regression:

$$n_{ij} = \alpha + \delta L_{ij} + \lambda A_{ij} + u_{ij} \tag{1}$$

where n_{ij} is the number of true statements reported by individual i for list j (with $j \in \{A, B\}$), L_{ij} is a dummy variable indicating whether individual i was assigned the long version of list j (which includes the sensitive statement), and A_{ij} is a dummy variable indicating whether list j corresponds to list A. We cluster standard errors at the individual level to account for within-individual data dependence.[15] The parameter of interest is δ,

[14] See Section 2.3 and footnote 12 for more details on the study design and the rationale for including a group of respondents who do not receive the sensitive statement in either list.

[15] In our experiment, individuals are assigned to treatment and control through lotteries at the NGO-field

which estimates the share of individuals who plan to migrate irregularly to Europe. Online Appendix B examines the validity of the double list experiment. Most importantly, the characteristics of respondents receiving the different lists are balanced (Table A.1), and the prevalence estimated is consistent in the two lists (Table B.3), thus supporting the validity of the method.

Figure 2 presents the estimates using each of the three methods for men and women separately. The exact estimates and their standard errors are reported in Table A.2. Three key results emerge. First, the direct question produces estimates that are significantly lower than with the other two methods, indicating that respondents' aspirations for irregular migration are indeed a sensitive topic. Second, the ballot-bag method elicits sensitive responses at rates similar to the list experiment, suggesting that it effectively mitigates concerns over anonymity. For men, the prevalence estimated via the direct question is 8.7%, compared to 13.2% via the ballot-bag and 15.3% via the list experiment. The estimate from direct questioning is statistically different from the other two methods, whereas the ballot-bag and list experiment estimates are indistinguishable.[16] Similar patterns hold for women, though differences are smaller in absolute terms due to the lower overall prevalence of the sensitive outcome (1.2% using the direct question). In relative terms, differences between the direct question and the other two methods remain nonetheless large (+75% for the ballot-bag and +67% for the list experiment).

The third key result from Figure 2, and perhaps the strongest reason to prefer the ballot-bag, is that estimates based on the list experiment are several times more imprecise than those based on the other two approaches. For men, standard errors are 3.4 times larger with the list experiment than with the ballot-bag. This lack of precision is even more pronounced for women, with standard errors 7.8 times larger with the list experiment than with the ballot-bag. In contrast, estimates based on the ballot-bag and the direct question have similar precision. The imprecision of the list experiment has important implications for estimating treatment effects, as we show in the next sub-section.

There are two threats to our interpretation that differences across the three methods

level, identified by the variable Z. To account for variations in the assignment rate across lotteries, we weight observations in all specifications using the formula $T/P(Z) + (1-T)/(1-P(Z))$, where T is a treatment indicator and $P(Z)$ is the proportion of treated individuals in each lottery.

[16] Conducting statistical tests for comparisons involving the ballot-bag is challenging, as responses cannot be matched across methods. However, the visual evidence from Figure 2 strongly suggests that estimates from the ballot-bag are significantly larger than those from the direct question and hardly distinguishable from those based on the list experiment. We show in Appendix B that these tests can be performed under reasonable assumptions and formally confirm these conjectures in Table A.2.

are driven by methodological considerations. First, the results in Figure 2 may be subject to order effects, since the different methods were not randomized across respondents and responding to one method may influence answers in the other methods. As described in Section 2.3, we sought to limit this concern by administering the methods in separate sections of the survey with extensive blocks of questions in-between to "divert" respondents. However, this approach may not fully eliminate order effects and may also introduce variability in survey fatigue across methods (Jeong et al., 2023). To test for such concerns, we randomized the location of the direct question in the survey: for 20% of the respondents the direct question was asked before the list experiment and the ballot-bag, while for the remaining 80% the direct question was asked after the other two methods. Reassuringly, the prevalence estimated via the direct question is similar in the two groups—2.9% and 3.2% respectively (p-value=0.53; Table A.3)—suggesting that order effects are limited.[17] Second, as mentioned earlier, some bags were lost during data collection, raising concerns about selective attrition in the ballot-bag. To account for potential selective attrition, we derive the share of missing ballots at the lottery level, S, and adjust the ballot-bag estimates by weighting observations by $1/(1 - S)$ to give more weight to observations in lotteries with more missing ballots. Results hardly change—estimates decrease slightly from 13.2% to 12.9% for men and from 2.1% to 2.0% for women—suggesting that the loss of bags was close to random (Table A.4).

3.2 Treatment effects across the three methods

In this section, we first outline the straightforward way in which treatment effects can be derived across the three methods. We then demonstrate that the imprecision of the list experiment has major consequences when estimating the effect of an intervention. In the application we consider, the direct question and the ballot-bag capture a significant reduction in aspirations for irregular migration, whereas the list experiment fails to detect this effect.

For both the direct question and the ballot-bag, we estimate treatment effects using a regression of the following form:

$$y_{ij} = \alpha + \beta T_{ij} + \gamma \mathbf{X}_j + u_{ij} \tag{2}$$

[17] For men, there is suggestive evidence that prevalence is higher when the direct question is asked after the other two methods (p-value=0.10). Recall that 80% of the respondents received the direct question after the other two methods. Overall, this suggests that estimates using direct questioning may represent an upper bound relative to a setting in which it had been administered alone, implying that sensitivity bias could be more pronounced than our estimates indicate.

where y_{ij} is the outcome of interest for individual i in lottery j, T_{ij} is a treatment indicator, and $\mathbf{X}_j$ are lottery fixed effects. β captures the average treatment effect of the intervention.

For the list experiment, we extend equation (1) and estimate the following regression:

$$n_{ij} = \alpha + \delta L_{ij} + \lambda A_{ij} + \alpha T_i + \delta' L_{ij} T_i + \lambda' A_{ij} T_i + \gamma \mathbf{X}_j + u_{ij} \tag{3}$$

where δ captures the prevalence in the control group and δ' the average treatment effect of the intervention.

The results are presented in Table 1. For women, the effects are always small and non-significant, which is not surprising as women in Egypt rarely migrate to Europe irregularly. For men, however, the results based on the direct question and the ballot-bag indicate that better employment prospects at origin reduce aspirations for irregular migration. Using the ballot-bag data, we estimate that men assigned to treatment are five percentage points less likely to report plans to migrate irregularly to Europe (p-value<0.01). This corresponds to a 32% decrease relative to the control mean. Consistent with the reporting bias highlighted above, effects are smaller using the direct question (-2.4 percentage points), yet they are still significant at the 5% level and of a similar magnitude in relative terms (-24% compared to the control mean). This suggests that misreports in the direct question follow similar patterns in the treatment and control groups. The most striking result from Table 1 is that the list experiment completely fails to detect this effect (p-value=0.66) because of the imprecision of the estimate. If we consider the minimum detectable effect (MDE) at conventional power (80%) and statistical significance (5%), we estimate an MDE of 0.11 standard deviation (SD) for the direct question, 0.12 SD for the ballot-bag, and 0.42 SD for the list experiment. These figures illustrate that the list experiment is particularly underpowered when the objective is to estimate a treatment effect.

3.3 Confusion matrices

While our evidence so far suggests that direct questioning suffer from a sensitivity bias and that both the ballot-bag and the list experiment are successful to mitigate this bias, important questions remain regarding individual patterns. In particular: Do individuals who report the sensitive outcome in the direct question also report it in the other two methods? And are respondents giving consistent responses in the list experiment and in the ballot-bag?

To address these questions, we turn to the analysis of confusion matrices, which summarize the joint distribution of responses. We focus here on *false negatives* and *false positives*, using the ballot-bag as the benchmark. A false negative occurs when an individual reports the sensitive answer under the ballot-bag but not under the other method. A false positive is the opposite. The identification of confusion matrices involving the ballot-bag is challenging since individual responses cannot be directly matched with those from the other two methods. However, Propositions 2 and 3 in Appendix A establish the conditions for identification, leveraging auxiliary variables collected on the ballot simultaneously with the sensitive question and also available in the general survey (i.e., gender, treatment, and lottery information). These variables should be strong predictors of the sensitive outcome and a standard conditional independence assumption must hold.

Results are shown in Table 2.[18] The first panel compares responses in the ballot-bag and in the direct question. As expected, false negatives in the direct question are widespread—only 44.2% of men and 27.6% of women who reported the sensitive answer in the ballot-bag also reported it in the direct question—whereas false positives remain rare—3.5% among men and 0.7% among women. This evidence confirms that under-reporting in the direct question accounts for both the higher prevalence estimated via the ballot-bag and the larger treatment effect. One possible explanation for the presence of false positives in the direct question is that some individuals may strategically give the sensitive answer to the direct question because they hold the (incorrect) belief that this answer could improve their access to assistance programs. Alternatively, false positives may arise from missing ballots or measurement errors.[19]

The second panel compares responses in the ballot-bag and in the list experiment. For men, we cannot reject the hypothesis that the methods are equivalent. We estimate that 81.1% of men who reported the sensitive answer in the list experiment also reported it in the ballot-bag, whereas only 5.2% of men who did not report the sensitive answer in the list experiment did report it in the ballot-bag. These proportions are imprecisely estimated and we cannot reject the hypothesis that there are no false positives *and* no false negatives (p-value=0.38). For women, the results show a greater divergence and we do reject this hypothesis (p-value=0.06). Strikingly, we estimate that only 22.9% of women who reported the sensitive answer in the ballot-bag also reported it in the list experiment. This large share of false negatives in the list experiment might be due to a lack of understanding of the method and to the low prevalence of the sensitive outcome. As part of our survey instruments,

[18] Estimates of the full confusion matrices are reported in Table A.6.

[19] We observe similar patterns when comparing responses in the list experiment and in the direct question (Table A.7), suggesting that false positives in the direct method are not solely driven by missing ballots.

we asked enumerators to assess respondents' understanding of the two methods using a 5-point Likert scale. Interestingly, enumerators reported that women's understanding of the protocols was 0.15 SD higher for the ballot-bag than for the list experiment (p-value < 0.01; Appendix Table A.5).[20]

4 Concluding remarks

In this paper, we introduced and tested the ballot-bag, an innovative tool for eliciting truthful responses to sensitive questions in surveys. The approach is designed to address the limitations of direct questioning, which is subject to sensitivity bias, and of indirect techniques such as list experiments, which suffer from high variance. We showed the potential of the method in the context of a large-scale field experiment in Egypt, and conclude that it provides a promising tool to ensure both data reliability and statistical precision. Nonetheless, the method is not without challenges. First, it entails nontrivial logistical demands, and robust protocols are needed to minimize the risks of losing data. Second, although literacy rates have improved globally, illiteracy remains a challenge in many contexts, and the method needs to be adapted for illiterate populations. One option is to explain how questions are organized on the ballot—possibly using pictograms next to each question—and have enumerators read the questions aloud for respondents while ensuring they can reply privately. Third, and more fundamentally, the anonymity that underpins the method's effectiveness also precludes the possibility of linking responses to other parts of the questionnaire, thereby limiting its usefulness in explaining heterogeneity in sensitive behaviors. To limit such issues, one can include on the ballot information on key dimensions of interest. Despite these challenges, the ballot-bag represents a solid addition to the toolkit for measuring sensitive outcomes. Future applications of the method may explore its effectiveness in different populations and across other sensitive topics.

[20] Overall, understanding appears excellent for both men and women, with average scores of 4.34 and 4.50 for the list experiment (the best score is 5). Figure A.1 presents the distribution of within-respondent differences in understanding between the two methods. More than 70% of respondents were reported to have the same level of understanding for both methods.

In-text Appendix

A Confusion matrices

We explain how to estimate confusion matrices for each pair of the methods we consider. For two binary variables A and B, the confusion matrix represents their joint distribution. If the marginal probabilities $P_A = P(A = 1)$ and $P_B = P(B = 1)$ are known, and a conditional probability such as $P(A = 1 \mid B = 1)$ is identified, the full confusion matrix can be recovered.[21]

A.1 List Experiment and Direct Question

The confusion matrix is directly identifiable.

Proposition 1. *In the regression*

$$n_{i,j} = (a_0 + b_0 L_{i,j} + c_0 A_{i,j})(1 - y_{i,dq}) + (a_1 + b_1 L_{i,j} + c_1 A_{i,j}) y_{i,dq} + u_{i,j}$$

- *b_k identifies $P(y_{LE} = 1 \mid y_{DQ} = k)$,*
- *The assumption $\{y_{DQ} = 1\} \subset \{y_{LE} = 1\}$ can be tested as $b_1 = 1$.*

Proof: Follows from the standard interpretation of list experiment coefficients (see Online Appendix B.1). If $\{y_{DQ} = 1\} \subset \{y_{LE} = 1\}$, then $P(y_{LE} = 1 \mid y_{DQ} = 1) = 1$.

A.2 Ballot-Bag and Direct Question

Because individual responses in the ballot-bag and the direct question cannot be matched, identification is granted only under some conditions.

Proposition 2. *Suppose there exists some variables Z collected both on the ballot and in the main questionnaire that satisfy the following conditions:*

1. *$Var_Z(P(y_{BB} \mid Z)) > 0$*
2. *$y_{DQ} \perp Z \mid y_{BB}$*

Then in the regression

$$E(y_{DQ} \mid E(y_{BB} \mid Z)) = b_1 E(y_{BB} \mid Z) + b_0(1 - E(y_{BB} \mid Z))$$

[21] Indeed, $P(A = 1, B = 1) = p_{AB} P_B$, $P(A = 1, B = 0) = p_A - p_{AB} P_B$, $P(A = 0, B = 1) = P_B - p_{AB} P_B$, and $P(A = 0, B = 0) = 1 - P_A - P_B + p_{AB} P_B$.

- *b_k identifies $b_k = P(y_{DQ} = 1 \mid y_{BB} = k)$,*
- *The assumption $\{y_{DQ} = 1\} \subset \{y_{BB} = 1\}$ can be tested as $b_0 = 0$.*

Proof: From $y_{DQ} = y_{DQ}\, y_{BB} + y_{DQ}(1 - y_{BB})$, its expectation conditional on $\{y_{BB}, Z\}$ and the conditional independence assumption, we obtain a linear relation between $E(y_{DQ} \mid Z)$ and $E(y_{BB} \mid Z)$, allowing identification of b_k through variation in $E(y_{BB} \mid Z)$.

A.3 Ballot-Bag and List Experiment

Identification combines the two approaches.

Proposition 3. *Suppose there exists some variables Z collected both on the ballot and in the main questionnaire that satisfy the following conditions:*

1. *$Var_Z(P(y_{BB} \mid Z)) > 0$*
2. *$y_{LE} \perp Z \mid y_{BB}$*

Then in the regression

$$E(n \mid A, L, Z) = a(Z) + c(Z)A + b_1 L \cdot E(y_{BB} \mid Z) + b_0 L \cdot (1 - E(y_{BB} \mid Z))$$

- *$b_k = E(y_{LE} \mid y_{BB} = k)$,*
- *The assumption $y_{LE} \equiv y_{BB}$ can be tested as $b_0 = 0$ and $b_1 = 1$.*

Proof: Starting from Equation 1, we condition on $y_{BB} = k$ and Z. Thanks to the independence assumption, the coefficient $b(Z, k)$ can be interpreted as $\mathbb{E}(y_{LE} \mid y_{BB} = k)$. Identification follows from variation in $E(y_{BB} \mid Z)$.

B Joint distribution of estimates

The covariance between the prevalence estimates $\hat{\theta}_q$ and $\hat{\theta}_r$ from two underlying measures y_q and y_r, with $q, r \in \{DQ, BB, LE\}$, is given by:

$$\text{Cov}(\hat{\theta}_q, \hat{\theta}_r) = \frac{P(y_q = 1, y_r = 1) - P(y_q = 1)P(y_r = 1)}{N}$$

When individual-level observations underlying the two measures can be matched, this covariance can be directly estimated from the data. However, when such matching is not possible—as in cases involving the ballot-bag—this approach cannot be used. Nevertheless, under the assumptions stated in Appendix A, the joint probability can be recovered, allowing the covariance to be computed using the formula above.

References

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*, 2025, *71* (1), 659–677.

Álvarez-Aragón, Pablo and Hugues Champeaux, "Measuring Norms and Enumerator Effects: Survey Method Matters," *The World Bank Economic Review*, forthcoming.

Armand, Alex, Britta Augsburg, Antonella Bancalari, and Maitreesh Ghatak, "Public service delivery, exclusion and externalities: Theory and experimental evidence from India," 2023. IFS Working Paper No. 23/37.

Aronow, Peter M, Alexander Coppock, Forrest W Crawford, and Donald P Green, "Combining list experiment and direct question estimates of sensitive behavior prevalence," *Journal of Survey Statistics and Methodology*, 2015, *3* (1), 43–66.

Benson, Lawrence E, "Studies in secret-ballot technique," *Public Opinion Quarterly*, 1941, *5* (1), 79–82.

Bertelli, Olivia, Thomas Calvo, Emmanuelle Lavallée, Marion Mercier, and Sandrine Mesplé-Somps, "What one thinks, what one says and what one does: Male justifications and practices of gender-based violence in Mali," *Journal of Development Economics*, 2025, p. 103479.

Bishop, George F and Bonnie S Fisher, ""Secret ballots" and self-reports in an exit-poll experiment," *Public Opinion Quarterly*, 1995, *59* (4), 568–588.

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.

_ **and Kosuke Imai**, "Statistical analysis of list experiments," *Political Analysis*, 2012, *20* (1), 47–77.

Bryan, Gharad, James J Choi, and Dean Karlan, "Randomizing religion: the impact of Protestant evangelism on economic outcomes," *The Quarterly Journal of Economics*, 2021, *136* (1), 293–380.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott, "Misperceived social norms: Women working outside the home in Saudi Arabia," *American Economic Review*, 2020, *110* (10), 2997–3029.

Chauchard, Simon, "Using MP3 players in surveys: The impact of a low-tech self-administration mode on reporting of sensitive attitudes," *Public Opinion Quarterly*, 2013, *77* (S1), 220–231.

Chen, Yuyu and David Y Yang, "The impact of media censorship: 1984 or brave new world?," *American Economic Review*, 2019, *109* (6), 2294–2332.

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.

Crépon, Bruno, Ahmed Elsayed, Dina Abdel Fattah, and Jules Gazeaud, "Employment and Irregular Migration: Evidence from Two Randomized Controlled Trials in Egypt," 2022. AEA RCT Registry. https://www.socialscienceregistry.org/trials/10604.

Cullen, Claire, "Method matters: The underreporting of intimate partner violence," *The World Bank Economic Review*, 2023, *37* (1), 49–73.

Detkova, Polina, Andrey Tkachenko, and Andrei Yakovlev, "Gender heterogeneity of bureaucrats in attitude to corruption: Evidence from list experiment," *Journal of Economic Behavior & Organization*, 2021, *189*, 217–233.

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.

Droitcour, Judith, Rachel A Caspar, Michael L Hubbard, Teresa L Parsley, Wendy Visscher, and Trena M Ezzati, "The item count technique as a method of indirect questioning: A review of its development and a case study application," in Paul P Biemer, Robert M Groves, Lars E Lyberg, Nancy A Mathiowetz, and Seymour Sudman, eds., *Measurement Errors in Surveys*, New York, NY: John Wiley & Sons, 1991.

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.

Jeong, Dahyeon, Shilpa Aggarwal, Jonathan Robinson, Naresh Kumar, Alan Spearot, and David Sungho Park, "Exhaustive or exhausting? Evidence on respondent fatigue in long surveys," *Journal of Development Economics*, 2023, *161*, 102992.

Jouvin, Marine, "Addressing social desirability bias when measuring child labor use: An application to cocoa farms in Côte d'Ivoire," *The World Bank Economic Review*, 2024, *38* (3), 625–646.

Karlan, Dean S and Jonathan Zinman, "List randomization for sensitive behavior: An application for measuring use of loan proceeds," *Journal of Development Economics*, 2012, *98* (1), 71–75.

Kerwin, Jason, Nada Rostom, and Olivier Sterck, "Striking the Right Balance: Why Standard Balance Tests Over-Reject the Null, and How to Fix It," 2024. IZA Discussion Papers No. 17217.

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.

McKenzie, David and Melissa Siegel, "Eliciting illegal migration rates through list randomization," *Migration Studies*, 2013, *1* (3), 276–291.

Miller, Judith Droitcour, "A new survey technique for studying deviant behavior." PhD dissertation, The George Washington University 1984.

Neggers, Yusuf, "Enfranchising your own? Experimental evidence on bureaucrat diversity and election bias in India," *American Economic Review*, 2018, *108* (6), 1288–1321.

Okunogbe, Oyebola and Victor Pouliquen, "Technology, taxation, and corruption: evidence from the introduction of electronic tax filing," *American Economic Journal: Economic Policy*, 2022, *14* (1), 341–372.

Osman, Adam, Jamin D Speer, and Andrew Weaver, "Discrimination against women in hiring," *Economic Development and Cultural Change*, 2025, *73* (2), 781–809.

Park, David Sungho, Shilpa Aggarwal, Dahyeon Jeong, Naresh Kumar, Jonathan Robinson, and Alan Spearot, "Private but misunderstood? Evidence on measuring intimate partner violence via self-interviewing in rural Liberia and Malawi," *The World Bank Economic Review*, 2024, p. lhae040.

Peterman, Amber, Malick Dione, Agnes Le Port, Justine Briaux, Fatma Lamesse, and Melissa Hidrobo, "Disclosure of violence against women and girls in Senegal," *The World Bank Economic Review*, 2024, p. lhae039.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N Shapiro, "An empirical validation study of popular survey methodologies for sensitive questions," *American Journal of Political Science*, 2016, *60* (3), 783–802.

Ting, Kai Ming, "Confusion matrix," *Encyclopedia of machine learning*, 2011, pp. 209–209.

Tourangeau, Roger and Ting Yan, "Sensitive questions in surveys.," *Psychological bulletin*, 2007, *133* (5), 859.

Tsai, Chi, "Statistical analysis of the item-count technique using Stata," *The Stata Journal*, 2019, *19* (2), 390–434.

Turnbull, William, "Secret vs. Nonsecret Ballots," in "Gauging Public Opinion," Cantril, Hadley and Associates. Princeton, NJ: Princeton University Press, 1947, pp. 77–82.

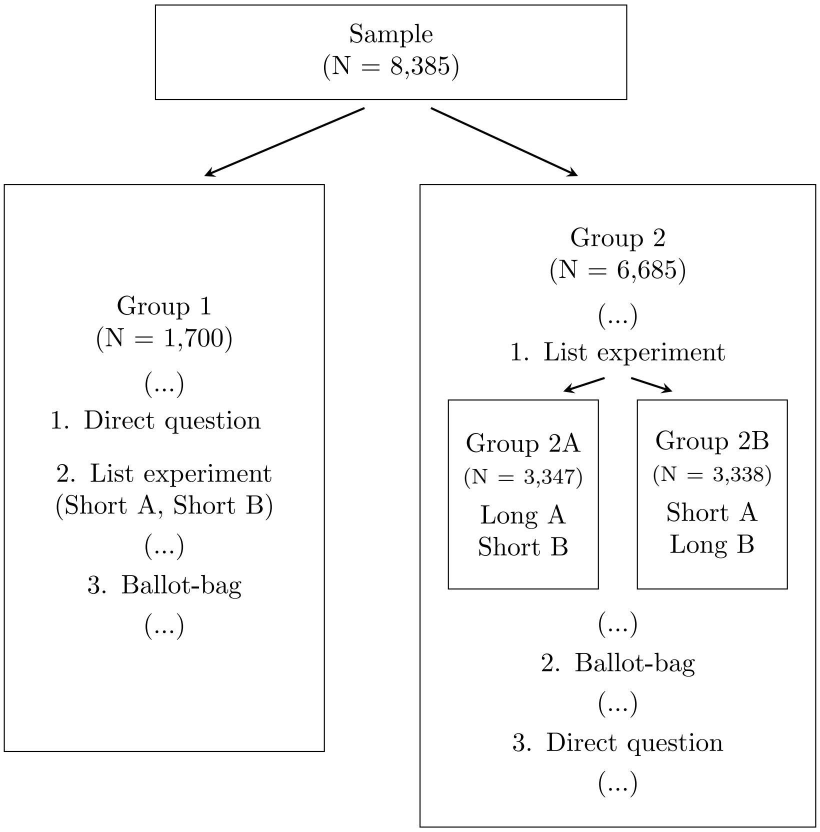
Valente, Christine, Wen Qiang Toh, Inuwa Jalingo, Aurélia Lépine, Áureo de Paula, and Grant Miller, "Are self-reported fertility preferences biased? Evidence from indirect elicitation methods," *Proceedings of the National Academy of Sciences*, 2024, *121* (34), e2407629121.

Warner, Stanley L, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 1965, *60* (309), 63–69.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang, "Two new models for survey sampling with sensitive characteristic: design and analysis," *Metrika*, 2008, *67*, 251–263.

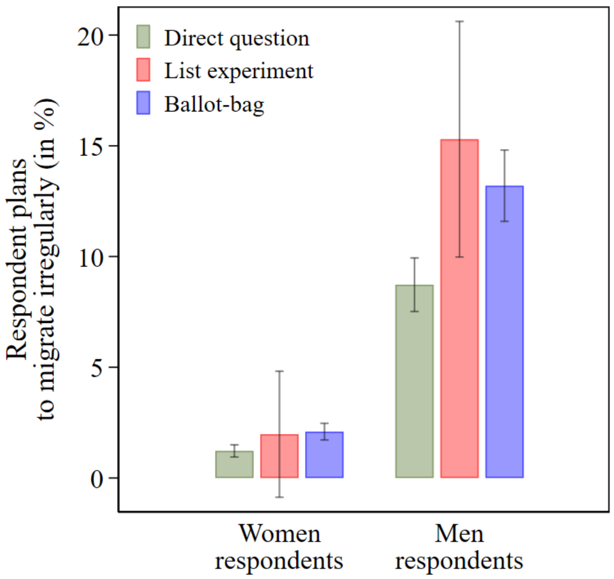
Main Figures and Tables

Figure 1: Survey structure



Notes: This figure presents the different versions of the survey respondents were randomly allocated to. Respondents in Group 1 were administered the direct question *before* the other two methods, whereas respondents in Group 2 were administered the direct question *after* the other two methods. Respondents in Group 2 are further divided into two subgroups: those who were asked the sensitive statement in List A (Group 2A) and those who were asked the sensitive statement in List B (Group 2B).

Figure 2: Prevalence of the sensitive outcome across the three methods



Notes: This figure reports the share of individuals planning to migrate irregularly to Europe using each of the three methods outlined in Section 2.2. For the direct question and the ballot-bag method, we simply report sample averages. For the list experiment, we report estimates of δ using equation (1). All estimates are presented with 95% confidence intervals.

Table 1: Treatment effects on the sensitive outcome across the three methods

	(1) Direct question	(2) Ballot-bag	(3) List experiment
Men respondents			
Treatment	-0.024**	-0.050***	-0.024
	(0.012)	(0.016)	(0.055)
Control mean	0.102	0.157	0.160
Control SD	0.303	0.364	0.366
Number of respondents	2194	1907	2194
Women respondents			
Treatment	0.004	0.005	-0.008
	(0.003)	(0.004)	(0.029)
Control mean	0.010	0.019	0.023
Control SD	0.102	0.136	0.150
Number of respondents	6191	5740	6191

Notes: This table reports the treatment effects of better employment prospects at origin on aspirations for irregular migration to Europe using the RCT described in Section 2. Columns 1 and 2 report estimates of β using equation (2). Column 3 reports estimates of δ' using equation (3). All columns control for lottery fixed effects. Robust standard errors in parenthesis are clustered at the individual level. *** p-value $<$ 0.01, ** p-value $<$ 0.05, * p-value $<$ 0.1.

Table 2: Confusion matrices

	(1) Men	(2) Women
Panel A. Direct Question | Ballot Bags		
$P(y_{DQ} = 1 \vert\, y_{BB} = 0)$	0.035	0.007
	(0.008)	(0.002)
$P(y_{DQ} = 1 \vert\, y_{BB} = 1)$	0.442	0.276
	(0.056)	(0.080)
p-value DQ $\subset$ BB	0.000	0.000
Number of respondents	2194	6191
Panel B. List Experiment | Ballot Bags		
$P(y_{LE} = 1 \vert\, y_{BB} = 0)$	0.052	0.011
	(0.039)	(0.017)
$P(y_{LE} = 1 \vert\, y_{BB} = 1)$	0.811	0.229
	(0.185)	(0.328)
p-value LE $\equiv$ BB	0.383	0.060
Number of respondents	2194	6191

Notes: This table summarizes the joint distribution of responses in the different methods, using the ballot-bag as the benchmark. Panels A and B report estimates from the regressions described in Propositions 2 and 3 respectively (Appendix A). The p-value in Panel A tests the null hypothesis that all respondents who report the sensitive answer in the direct question also report it in the ballot-bag (i.e., there are no false positives in the direct question: $P(y_{DQ} = 1|\, y_{BB} = 0) = 0$). The p-value in Panel B tests the hypothesis that individual responses are consistent in the list experiment and in the ballot-bag (i.e., there are no false positives *and* no false negatives: $P(y_{LE} = 1|\, y_{BB} = 0) = 0$ and $P(y_{LE} = 1|\, y_{BB} = 1) = 1$). Robust standard errors in parenthesis are clustered at the individual level.

Online appendix

Online Appendix A Additional tables and figures

Table A.1: Balance checks

	(1) N	(2) Mean (sample)	(3) SD (sample)	(4) Group 1 vs. rest of the sample	(5) Group 2A vs. rest of the sample	(6) Group 2B vs. rest of the sample
Respondent characteristics						
Female	8,385	0.738	0.440	0.001	0.003	-0.004
				(0.013)	(0.010)	(0.010)
Age	8,382	26.746	3.905	0.041	-0.007	-0.020
				(0.111)	(0.090)	(0.090)
Primary school completed	8,385	0.956	0.205	-0.005	0.002	0.001
				(0.006)	(0.005)	(0.005)
Married	8,385	0.566	0.496	0.008	-0.016	0.011
				(0.014)	(0.011)	(0.011)
Has children	8,385	0.495	0.500	-0.008	0.002	0.004
				(0.014)	(0.011)	(0.011)
Has an income generating activity	8,385	0.477	0.500	0.016	-0.004	-0.007
				(0.014)	(0.011)	(0.011)
Earnings (in EGP; winsorized 99%)	8,369	858.520	1624.132	40.213	8.623	-35.973
				(46.448)	(37.601)	(36.890)
Has a bank account	8,383	0.115	0.319	0.003	0.009	-0.011
				(0.009)	(0.007)	(0.007)
Household characteristics						
Household size	8,385	4.982	2.220	-0.065	0.027	0.017
				(0.060)	(0.051)	(0.051)
Owns agricultural land	8,379	0.226	0.418	-0.005	0.016	-0.013
				(0.012)	(0.010)	(0.010)
Has livestock	8,385	0.432	0.495	-0.008	0.012	-0.007
				(0.014)	(0.011)	(0.011)
Randomization inference p-value	0.80					

Notes: Column 1 shows the number of non-missing observations in the follow-up survey out of a total of 8,385 observations (1,700 observations for Group 1, 3,347 observations for Group 2A, and 3,338 observations for Group 2B). Columns 2 and 3 show summary statistics for the full sample. Column 4 shows the coefficients from regressing the variables on an indicator for Group 1. Column 5 shows the coefficients from regressing the variables on an indicator for Group 2A. Column 6 shows the coefficients from regressing the variables on an indicator for Group 2B. The last row reports the p-value from a multinomial logit specification that tests balance between each group across all variables. Following the recommendations of Kerwin et al. (2024), we derive the p-value using randomization inference. Robust standard errors are in parenthesis. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table A.2: Prevalence using the different methods

	(1) Women	(2) Men
DQ: Direct question	0.012	0.087
	(0.001)	(0.006)
Number of respondents	6191	2194
LE: List experiment	0.020	0.153
	(0.015)	(0.027)
Number of respondents	6191	2194
BB: Ballot-bag	0.021	0.132
	(0.002)	(0.008)
Number of respondents	5740	1907
p-value: DQ = LE	0.629	0.027
p-value: DQ = BB	0.000	0.000
p-value: BB = DLE	0.941	0.473

Notes: This table reports the estimates from Figure 2. Robust standard errors in parenthesis are clustered at the individual level. The derivation of the p-values involving the ballot-bag is described in Appendix B.

Table A.3: Testing for the presence of order effects

	(1) All	(2) Men	(3) Women
Direct question **Before**			
Prevalence	0.029 (0.004)	0.068 (0.012)	0.016 (0.004)
Observations	1700	436	1264
Direct question **After**			
Prevalence	0.032 (0.002)	0.092 (0.007)	0.011 (0.002)
Observations	6685	1758	4927
p-value Before = After	0.527	0.097	0.222

Notes: This table tests whether asking the direct question before or after the other two methods affects the estimated prevalence. The location of the direct question was randomized as follows: for 20% of the respondents it was asked before the list experiments and the ballot-bag method, while for the remaining 80% it was asked after. Robust standard errors in parenthesis are clustered at the individual level.

Table A.4: Results from the ballot-bags are robust to reweighting observations to account for missing ballots

	(1) Normal weights	(2) More weights to ballots from lotteries with more missings
Men		
Prevalence	0.132	0.129
	(0.008)	(0.009)
Observations	1,907	1,907
Women		
Prevalence	0.021	0.020
	(0.002)	(0.002)
Observations	5,740	5,740

Notes: This table tests whether missing ballots bias the estimates from the ballot-bag method. Column 1 reports the benchmark results from Figure 2. Column 2 presents results where we derive the share of missing ballots per lottery $S = (N_{DQ} - N_{BB})/N_{DQ}$ and reweight observations using the formula $1/(1 - S)$ in order to give more weights to observations in lotteries with more missing ballots. Robust standard errors in parenthesis are clustered at the individual level.

Table A.5: Respondents understand the ballot-bag method better than the list experiment

	Understanding		Compliance	
	(1) Men	(2) Women	(3) Men	(4) Women
Ballot-bag	0.057***	0.113***	0.010	0.077***
	(0.019)	(0.009)	(0.018)	(0.009)
Mean list experiment	4.344	4.503	4.387	4.538
SD list experiment	0.878	0.740	0.840	0.722
Number of respondents	1580	3908	1580	3908

Notes: This table reports respondents' understanding of the list experiment and the ballot-bag method, as well as their compliance with the protocols of each method. For each method and respondent, enumerators assessed understanding and compliance using a 5-point Likert scale (1: very bad; 5: excellent). Understanding: "*How well did the respondent seem to understand the instructions?*" Compliance: "*What is your impression about the seriousness with which the respondent followed the instructions?*" We report the coefficients from regressions of these variables on a dummy indicating whether the method is the ballot-bag method. These questions were introduced partway through the data collection, resulting in a smaller sample size than in other analyses. Robust standard errors in parenthesis are clustered at the individual level. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table A.6: Full confusion matrices

Men

Ballot Bags versus Direct Question

	DQ=1	DQ=0	Total
BB=1	0.058	0.074	0.132
BB=0	0.029	0.839	0.868
Total	0.087	0.913	1.000

Ballot Bags versus List Experiment

	LE=1	LE=0	Total
BB=1	0.107	0.025	0.132
BB=0	0.046	0.822	0.868
Total	0.153	0.847	1.000

List Experiment versus Direct Question

	DQ=1	DQ=0	Total
LE=1	0.056	0.097	0.153
LE=0	0.031	0.816	0.847
Total	0.087	0.913	1.000

Women

Ballot Bags versus Direct Question

	DQ=1	DQ=0	Total
BB=1	0.006	0.015	0.021
BB=0	0.006	0.973	0.979
Total	0.012	0.988	1.000

Ballot Bags versus List Experiment

	LE=1	LE=0	Total
BB=1	0.005	0.016	0.021
BB=0	0.015	0.964	0.979
Total	0.020	0.980	1.000

List Experiment versus Direct Question

	DQ=1	DQ=0	Total
LE=1	0.007	0.013	0.020
LE=0	0.005	0.975	0.980
Total	0.012	0.988	1.000

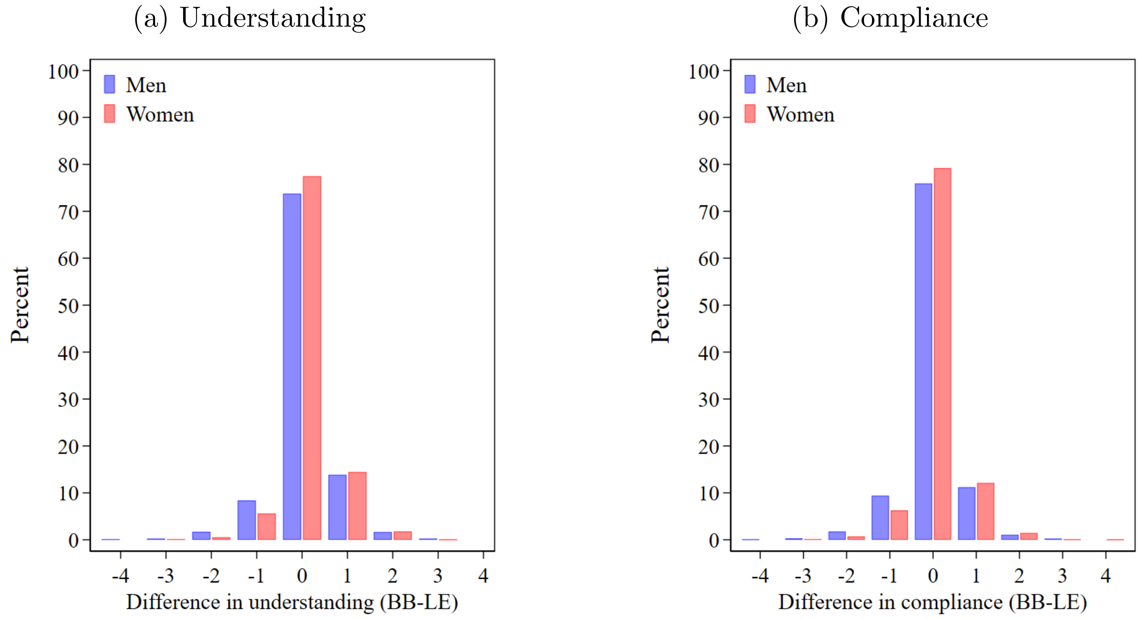
Notes: This table summarizes the joint distribution of responses in the different methods. See Appendix A for details on how these matrices can be identified for the three possible pairs of method.

Table A.7: Confusion matrix: list experiment vs. direct questioning

	(1) Men	(2) Women
List Experiment | Direct Question		
$P(y_{LE} = 1 \| \, y_{DQ} = 0)$	0.103	0.014
	(0.028)	(0.015)
$P(y_{LE} = 1 \| \, y_{DQ} = 1)$	0.647	0.571
	(0.089)	(0.141)
p-value DQ $\subset$ LE	0.000	0.002
Number of respondents	2194	6191

Notes: This table reports estimates from proposition 1 in Appendix A. Robust standard errors in parenthesis are clustered at the individual level.

Figure A.1: Difference in understanding and compliance between the ballot-bag method and the list experiment


(a) Understanding

(b) Compliance

Notes: For each respondent, we derive the difference in understanding and compliance between the ballot-bag method and the list experiment. These figures show the distribution of within-individual differences, where positive values indicate greater understanding/compliance in the ballot-bag method. See notes to Table A.5 for more details on the definition of the variables.

Online Appendix B More on the List Experiments

B.1 Reminder on list experiments

Let n_s be the number of true statements in the short list and y_{LE} an indicator of whether an individual's response to the sensitive statement is positive. Then, the total number n of true statements that individuals count is simply given by $n = n_s + y_{LE}L$, where L is a random variable indicating whether the individual received the long list. The share of positive responses to the sensitive statement s_{LE} can be estimated using the following regression:

$$n_i = a + s_{LE}L_i + u_i \tag{S.1}$$

When using the double list experiment framework, there are two count variables for each individual: $n_{A,i}$ for the first list experiment and $n_{B,i}$ for the second. There are also variables indicating as before whether the count is from the long list: $L_{A,i}$ and $L_{B,i}$. In our setting, following Figure 1, $L_{A,i} = 1$ only for group 3, and $L_{B,i} = 1$ only for group 2. Let $y_{le,i}(A)$ be the answer reported in the count for list A. Then $n_{A,i} = n^s_{A,i} + y_{le,i}(A)L_{A,i}$, and analogously $n_{B,i} = n^s_{B,i} + y_{le,i}(B)L_{B,i}$, with n^s_A and n^s_B the counts of the non-sensitive statements. We organize the data in a long format with $2N$ observations and $z_i 2 \times 1$ vectors $z'_i = (z_{A,i}, z_{B,i})$ for $z \in \{n, L\}$. Let's consider a last 2×1 variable A_i identifying list A: $A'_i = (1, 0)$. When running the difference-in-differences regression:

$$n_{i,j} = a + bL_{i,j} + cA_{i,j} + \Delta L_{i,j}A_{i,j} + u_{i,j} \quad \text{for } j \in \{A, B\}$$

in this equation, b identifies $E(n_{i,j}|A_{i,j} = 0, L_{i,j} = 1) - E(n_{i,j}|A_{i,j} = 0, L_{i,j} = 0) = E(y_{le,i}(B))$, and Δ identifies $[E(n_{i,j}|A_{i,j} = 1, L_{i,j} = 1) - E(n_{i,j}|A_{i,j} = 1, L_{i,j} = 0)] - [E(n_{i,j}|A_{i,j} = 0, L_{i,j} = 1) - E(n_{i,j}|A_{i,j} = 0, L_{i,j} = 0)]$, which simplifies to $E(y_{le,i}(A)) - E(y_{le,i}(B))$. Thus, in this specification, testing the hypothesis $\Delta = 0$ is a test of the internal consistency of the two list experiments A and B: $E(y_{le,i}(A)) - E(y_{le,i}(B))$.[22]

The final estimate of the prevalence of the sensitive outcome is thus obtained by running the constrained regression:

$$n_{i,j} = a + bL_{i,j} + cA_{i,l} + u_{i,l} \tag{S.2}$$

Indeed, it is possible to show that an analogue to $\hat{b}$ is $0.5 \times ((\overline{n_{A,i}}^{L_{A,i}=1} - \overline{n_{A,i}}^{L_{A,i}=0}) + (\overline{n_{B,i}}^{L_{B,i}=1} - \overline{n_{B,i}}^{L_{B,i}=0}))$, which thus identifies $0.5 \times (E(y_{le,i}(A)) + E(y_{le,i}(B))) = E(y_{le,i})$.

[22] Clearly a identifies $E(n_{i,j}|A_{i,j} = 0, L_{i,j} = 0) = E(n^s_{i,B})$ and c identifies $E(n_{i,j}|A_{i,j} = 1, L_{i,j} = 0) - E(n_{i,j}|A_{i,j} = 0, L_{i,j} = 0) = E(n^s_{i,A} - n^s_{i,B})$.

B.2 Considerations to design the list experiments

We followed three principles to design our lists. First, we sought to include statements that limit the risks of “floor” and “ceiling” effects[23] by including in each list some statements with high prevalence, some statements with low prevalence, and some statements that are negatively correlated (e.g., having flown on a plane versus having never left one’s own country). Second, to improve the precision of the estimates, we sought to include statements that are negatively correlated within each list and positively correlated across lists. Third, we sought to include statements that are well understood/easy to answer (to limit measurement errors), and broadly related to the sensitive outcome (to avoid making the sensitive statement too salient).

Prior to data collection, we conducted a small pilot with 100 respondents to guide us in designing lists that are in line with these principles. We administered a set of possible statements as direct questions and built the lists as follows: (i) we prioritized statements that are easy to understand and answer (after each question, enumerators reported whether the respondent appeared confused by the question or took time to reply—we excluded statements with apparent confusion); (ii) in each list, we aimed to include at least one statement with high prevalence and two statements with low prevalence; (iii) in each list, we aimed to include two statements that are negatively correlated; (iv) across lists, we aimed to include statements that are broadly similar.

B.3 The two double list experiments

We included two double list experiments in our questionnaire: one to measure the migration aspirations of the respondent (Lists A and B); the other to measure the migration aspirations of other household members (Lists C and D). The statements included in the lists are reproduced in Table B.1.

B.4 Descriptive statistics

Descriptive statistics for each list are reported in Table B.2. Reassuringly, few respondents reported the maximum number of statements in the short lists (between 0.9% for List B and

[23] Floor effects correspond to situations where *zero* statements are true, whereas ceiling effects correspond to situations where *all* statements are true. Both types of effects can be problematic because they break the privacy protection of the list experiment—for respondents who report that zero/all statements are true it is possible to infer with certainty the answer to the sensitive statement. However, in our setup, ceiling effects are particularly problematic because they entail that the sensitive statement is true.

Table B.1: Design of the double list experiments

Respondent	
List A	**List B**
1. I have a strong relationship with my cousins	1. I have often been unable to pay my bills
2. I have at least one friends who live in Europe	2. I have lived all my life in Egypt
3. I can afford to buy an apartment in Cairo	3. I already took the plane
4. In general, I think it's easy to find a job in Egypt	4. There are good job opportunities for me in this village
5. I plan to migrate to Europe without the official papers	
Other household members	
List C	**List D**
1. I have been offered at least one job abroad	1. I have already turned down a good job I have been offered
2. In general, I am able to save money	2. I prefer to stay close to my family
3. I have a close relationship with my siblings	3. I have a friend who wants to move to Europe
4. Someone from my family is living in Europe	4. I have already been offered a job in Cairo
5. One (or more) of my household members is preparing to migrate to Europe without the official papers	

4.0% for List D), suggesting that our choice of statements effectively limited ceiling effects. More respondents reported the minimum number of statements in the short lists (between 2.2% for List C and 8.7% for Lists B and D), however, this is not too concerning as reporting zero statements in the long list is not a sensitive answer in our setup.

B.5 Consistency check

Table B.3 reports the results from the consistency check proposed by Chuang et al. (2021). This check verifies that the estimates produced by each list are similar. This is a necessary condition for a double list experiment to provide reliable estimates. For respondents' plans to migrate irregularly to Europe, we find similar estimates for List A and List B: 4.8% and 6.1% respectively (these rates are not statistically different—p-value = 0.65). For other household members, however, we find that List D produces larger estimates than List C: 16.5% against 8.0%. We reject that these rates are statistically similar (p-value = 0.005), suggesting a failure of the second list experiment to estimate the sensitive outcome reliably. We note that the pattern of larger estimates in List D than in List C applies to both men and women, although differences are most pronounced and only significant for women (6.3% for List C against 16.8% for List D—p-value = 0.002).

Table B.2: Descriptive statistics: Double list experiments

	List A		List B	
	(1)	(2)	(3)	(4)
	Short list	Long list	Short list	Long list
Mean	1.661	1.710	1.487	1.548
Standard deviation	0.831	0.908	0.807	0.864
Observations	5,040	3,347	5,047	3,339
Distribution of answers				
0	0.051	0.053	0.087	0.084
1	0.397	0.392	0.442	0.429
2	0.412	0.389	0.374	0.361
3	0.119	0.130	0.087	0.111
4	0.021	0.030	0.009	0.012
5		0.006		0.003
	List C		List D	
	(5)	(6)	(7)	(8)
	Short list	Long list	Short list	Long list
Mean	1.762	1.842	1.551	1.716
Standard deviation	0.836	0.965	0.943	1.061
Observations	4,133	4,169	4,169	4,132
Distribution of answers				
0	0.022	0.025	0.087	0.079
1	0.403	0.398	0.470	0.423
2	0.395	0.358	0.289	0.279
3	0.151	0.158	0.114	0.153
4	0.028	0.048	0.040	0.055
5		0.012		0.011

B.6 What led to the failure of the second list experiment?

In this section, we explore various potential reasons for the failure of the second list experiment (Table B.3). Although our data do not allow us to provide definitive evidence in favor of any single explanation, we suggest that the failure of List C and/or List D may be due to the nature of the sensitive behavior (which pertains to individuals other than the respondent) and to priming effects. We first rule out several alternative explanations and then delve deeper into this hypothesis.

The failure of the second list experiment is unlikely to be driven by issues with the randomization since the characteristics of individuals assigned to the short lists and long lists are balanced (Table B.4). It is also unlikely to result from ceiling effects as only 2.8% and 4.0% of respondents reported the maximum number of statements in short C and short D (Table B.2). Furthemore, even if ceiling effects were at play, they should reduce the prevalence of the sensitive statement in List D more than in List C, which is not consistent

Table B.3: Consistency check of Chuang et al. (2021)

	(1) All	(2) Men	(3) Women
Respondent			
Double list experiment	0.055	0.153	0.020
	(0.013)	(0.027)	(0.015)
List experiment A	0.048	0.152	0.012
	(0.019)	(0.042)	(0.021)
List experiment B	0.061	0.154	0.027
	(0.019)	(0.038)	(0.021)
p-value A = B	0.647	0.986	0.627
Observations	16773	4389	12384
Other household members			
Double list experiment	0.122	0.144	0.115
	(0.012)	(0.024)	(0.013)
List experiment C	0.080	0.126	0.063
	(0.020)	(0.043)	(0.022)
List experiment D	0.165	0.162	0.168
	(0.022)	(0.048)	(0.024)
p-value C = D	0.005	0.591	0.002
Observations	16603	4361	12242

with what we observe—List D produces larger estimates than List C (Table B.3). It is unclear how floor effects would lead to the failure of List C and/or List D since zero is not a sensitive answer in our setup.

Another possible explanation for the failure of a list experiment is the presence of design effects. Design effects arise when respondents alter their answers to the non-sensitive statements when the sensitive statement is included. We consider this explanation unlikely in our setup because the sensitive statement was placed at the end of the lists and respondents used marbles to 'materialize' their answers. Specifically, respondents were instructed to put both hands behind their backs, start with all marbles in their right hand, transfer one marble to their left hand each time a statement was true, and, at the end, reveal the number of true statements by showing how many marbles are in their left hand (see the detailed protocol in Online Appendix C). In this setup, changing answers to the non-sensitive statements would require transferring marbles in the opposite direction at the end (i.e., when the sensitive statement is introduced). We consider this implausible. To check for the presence of design effects, we implement Blair and Imai (2012)'s statistical test and confirm that there is no evidence to suggest that design effects are at play in our setup.[24]

[24] Blair and Imai (2012) suggest that for each $n \in \{0, \ldots, N\}$, where N is the total number of statements

Table B.4: Balance checks (Lists C and D)

	(1) N	(2) Mean (short C long D)	(3) SD (short C long D)	(4) long C short D vs. short C long D
Respondent characteristics				
Female	8300	0.739	0.439	0.004
				(0.010)
Age	8297	26.747	3.947	0.011
				(0.089)
Primary school completed	8300	0.957	0.203	0.000
				(0.005)
Married	8300	0.557	0.497	-0.016
				(0.011)
Has children	8300	0.488	0.500	-0.012
				(0.011)
Has an income generating activity	8300	0.483	0.500	0.012
				(0.011)
Earnings (in EGP; winsorized 99%)	8285	877.756	1644.367	33.462
				(36.831)
Has a bank account	8298	0.118	0.323	0.005
				(0.007)
Household characteristics				
Household size	8300	4.966	2.162	-0.032
				(0.050)
Owns agricultural land	8294	0.236	0.425	0.018*
				(0.009)
Has livestock	8300	0.433	0.496	0.003
				(0.011)
Randomization inference p-value	0.66			

Notes: The sensitive statement was randomly included in List C or in List D. Column 4 shows the coefficients from regressing the variables on an indicator for including the sensitive statement in List D. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

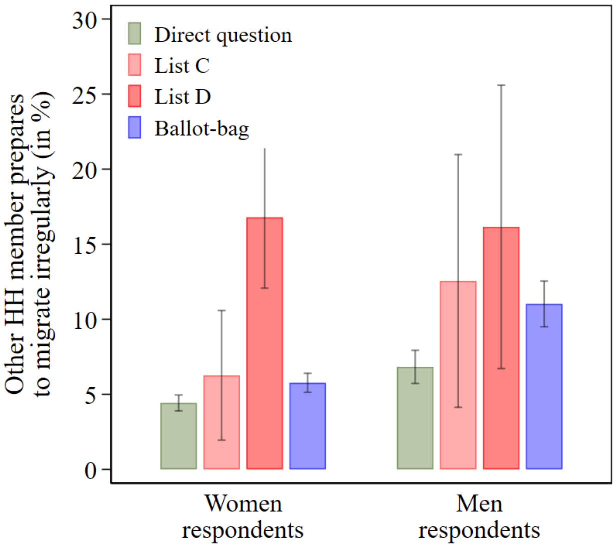
We conjecture that Lists C and/or D may be plagued by another type of design effect whereby the characteristics of the non-sensitive statements alter the responses to the sensitive statement. This effect—akin to a priming effect—may be particularly prevalent for outcomes whose answers are ambiguously defined, such as the outcome covered by Lists C and D. Indeed, Lists C and D inquire about the aspirations for irregular migration of other household members, which may be difficult for respondents to assess.[25] The fact that the test of Chuang

in a given list, the fraction of respondents reporting an answer superior or equal to n should be larger or equal for long list respondents relative to short list respondents. The authors propose a statistical test that estimates the probabilities of all possible responses and checks for non-sensical values (i.e., negative probabilities or probabilities above one). Using the Stata command *kict deff* (Tsai, 2019), we find that only one of the estimated probabilities is below zero and none are above one. Reassuringly, the negative probability is not statistically significant.

[25] In contrast, Lists A and B inquire about the respondents' own aspirations for irregular migration which are likely easier for them to evaluate.

et al. (2021) fails only in the sub-sample of women is consistent with this hypothesis. In Egypt, women are often excluded from important household decisions such as the migration of male household members, and they may therefore particularly struggle to assess whether any of their household members prepares to migrate irregularly to Europe. Looking at the statements included in Lists C and D (Table B.1), we speculate that the fourth statement—placed just before the sensitive statement—may have produced asymmetric priming effects in the two lists. Indeed, the fourth statement in List C ("*Someone from my family is living in Europe*") may prime respondents to answer negatively to the sensitive statement because few respondents in our sample have family members in Europe which may hamper migration.[26] In contrast, the fourth statement in List D ("*I have already been offered a job in Cairo*") may prime respondents to answer positively to the sensitive statement because few respondents in our sample—especially in the women sub-sample—have been offered a job in Cairo and a lack of job opportunities in the home country is generally considered as a push factor for migration.

Figure B.1: List C (but not list D) is consistent with the ballot-bag



Notes: This figure reports the share of other household members preparing to migrate irregularly to Europe using each of the three methods outlined in Section 2.2. For the direct question and the ballot-bag method, we simply report sample averages. For the list experiment, we report estimates of δ using equation (1), separately for Lists C and D. All estimates are presented with 95% confidence intervals.

To the best of our knowledge, such priming effects have not been considered in previous

[26] Decades of migration research highlight the importance of the network at destination for migration decisions.

work on list experiments. Developing statistical tests to detect and quantify these effects represents an interesting avenue for future research. In our setup, we are unable to demonstrate that priming effects are at the source of the failure of Chuang et al. (2021)'s consistency check, and thus cannot determine whether List C or List D provides more reliable estimates. However, it is worth noting that estimates from the ballot-bag are consistent only with the estimates from List C and not with those of List D (Figure B.1), which suggests that List C may provide the more reliable estimates.

Online Appendix C Survey scripts

C.1 List experiment

Now we will play a small game. In the first round, we will try the game so you understand it well. Then we will start.

I will read four statements. I will then ask you how many of these statements are true. You should not tell me which specific statements are true but only the number of statements that are true.

I will give you four marbles and you have to hold them in your right hand. Keep both of your hands behind your back. For each of the statements, if it is true, please transfer one marble from your right hand to your left hand behind you. If it is not true, please do not transfer a marble. I will not be aware, and please do not inform me. At the end, I would like to know the total number of statements that are true. This number should correspond to the number of marbles you have in your left hand. I will now read the statements.

1. I like to listen Om Kalthoom music
2. I met the American president before
3. I like to eat Molokheya
4. I like to listen Shaabi music

How many marbles do you have in your left hand? Is this the number of statements that are true for you?

I will now give you an additional marble. I will read other statements and similarly ask you to keep both hands behind your back and start with all the marbles in your right hand. Please transfer one marble from your right hand to your left hand each time a statement is true for you.

[Enumerators proceed with the lists]

Questions for the enumerators:

- How well did the respondent seem to understand the instructions and the statements that you read? (Very poorly, Poorly, Average, Well, Very well)
- What is your impression about the seriousness with which the respondent followed the instructions? (Very bad, Bad, Average, Good, Very good)

C.2 Ballot Bag

[Note for the enumerator at the beginning of the survey: Before starting with the survey, please take the paper of the colour ${paper_color} and fill in the information about the gender of the respondent and ${ngo_abbreviation}]

[Note for the enumerator at the beginning of the ballot-bag section: Please pick the paper with the color ${paper_color} that you prepared and give the instructions to the respondent.]

Instructions: Now I will give you a paper. This paper contains basic information such as your gender and the NGO. There are also questions I will ask you to answer privately after we read them together. The responses are anonymous so please do not sign the paper or put your name on it. Once you'll have answered the questions please let me know and I will ask you to fold your paper, to put it in this bag, and to mix it with the other papers. Note that all the papers that you can see in this bag are from other respondents but are otherwise exactly the same. This means that once you'll have mixed them together, I or anyone will not be able to identify which one is yours. Do you have any questions? Let's now read the questions.

[Note for the enumerator: Once the respondent is done ask her to fold the paper, to put it in the bag, and to mix it with the other papers.]

Questions for the enumerators:

- Did the respondent agree to fill the paper?
- Did the respondent put the paper in the bag?
- What is your impression about the seriousness with which the respondent followed the instructions? (Very bad, Bad, Average, Good, Very good)
- How well did the respondent seem to understand the instructions and the questions that you read together? (Very poorly, Poorly, Average, Well, Very well)