Online appendix of "Measuring sensitive outcomes through ballot-bag surveys" (Crépon, Elsayed, and Gazeaud)

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 in the main paper.

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

	(1) All	(2) Men	(3) Women
Direct question **Before**			
Prevalence	0.029	0.068	0.016
	(0.004)	(0.012)	(0.004)
Observations	1700	436	1264
Direct question **After**			
Prevalence	0.032	0.092	0.011
	(0.002)	(0.007)	(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: Differences between direct and ballot-bag estimates persist under conservative comparisons accounting for order effects

	(1) Men	(2) Women
DQ after: direct question after ballot-bag		
Prevalence	0.092	0.011
	(0.007)	(0.002)
Number of respondents	1758	4927
BB: Ballot-bag (no correction for potential order effects)		
Prevalence	0.132	0.021
	(0.008)	(0.002)
Number of respondents	1907	5740
BB-corrected: Ballot-bag (correcting for potential order effects)		
Prevalence	0.125	0.018
	(0.000)	(0.000)
Number of respondents	1940	5804
p-value DQ after = BB	0.000	0.000
p-value DQ after = BB-corrected	0.000	0.000

Notes: This table examines whether order effects explain the higher prevalence estimated via the ballot-bag (BB) than via direct questioning (DQ). Ideally, we would compare the prevalence of BB when it is asked before DQ to the prevalence of DQ when it is asked before BB. However, we are unable to identify which BB surveys have been filled before DQ, and instead consider two alternative, conservative comparisons. First, we derive DQ estimates on the subset of respondents for whom it was asked after BB (the results in Table A.3 indicate that this ordering tends to increase DQ reporting among men, and therefore minimizes the BB-DQ difference). Second, we additionally adjust BB estimates for potential order effects. Specifically, because BB is asked after DQ for 20% of the respondents, we adjust BB prevalence by an order effect of a magnitude similar to that estimated for men in DQ (-0.10 SD), scaled by the 20% share. P-values for these comparisons are reported at the bottom of the table.

Table A.5: 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.6: 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.7: 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 in the main paper for details on how these matrices can be identified for the three possible pairs of method.

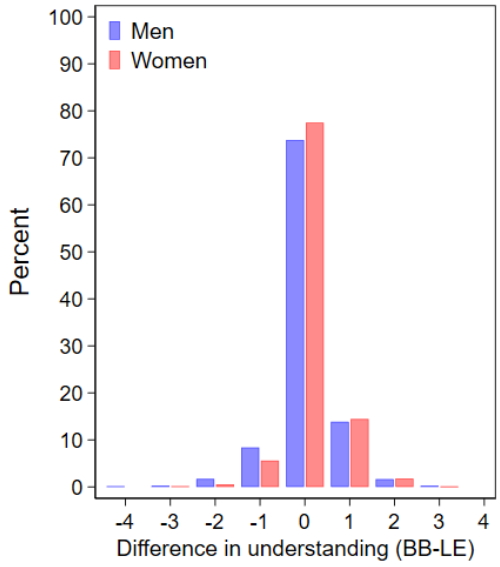
Table A.8: Confusion matrix: list experiment vs. direct questioning

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

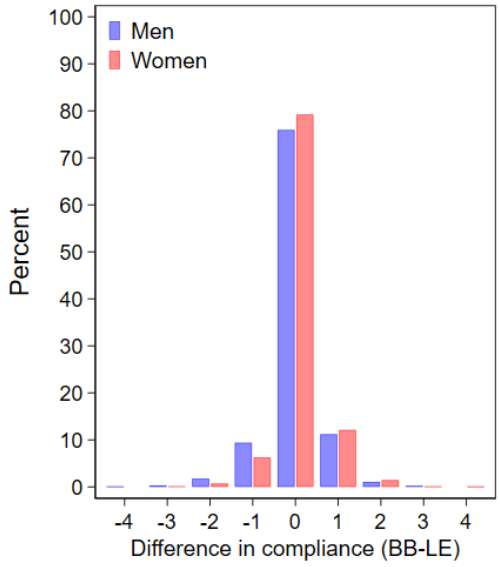
Notes: This table reports estimates from proposition 1 in Appendix A in the main paper. Robust standard errors in parenthesis are clustered at the individual level.

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.6 for more details on the definition of the variables.

Online Appendix B Relation to pre-registration

The main RCT we study is pre-registered on the AEA registry (Gazeaud et al., 2022). This paper reports only treatment effects on plans to migrate irregularly to Europe. Note that we also specified we would look at treatment effects on many additional outcomes, however we do not focus on these outcomes here because we only used direct questions or the ballot-bag method for these outcomes. Treatment effects for these outcomes will be reported in a separate paper. Below, we summarize the deviations from the pre-registration.

- In our pre-registration, we referred to the ballot-bag approach as the "urne" method.
- The three methods (direct questioning, the double list experiment, and the ballot-bag approach) were also used 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 C.7 in the Online Appendix).
- While this paper relies on a sample of 8,385 individuals, the pre-registered sample was 11,733 individuals. We pre-specified however that "*the actual number of observations [would] depend on the number of projects that get canceled and on our 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.

Online Appendix C More on the List Experiments

C.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: $E(y_{le,i}(A)) - E(y_{le,i}(B))$.[1]

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 $0.5 \times (E(y_{le,i}(A)) + E(y_{le,i}(B))) = E(y_{le,i})$.

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

C.2 Loss of precision in list experiments

The basic idea of list experiments is to add random noise to the measurement of the sensitive item. A loss of precision is therefore expected, implying a higher sampling variance—and thus lower precision—compared with direct elicitation. The magnitude of this loss depends on the design of the list experiment. Suppose the sample is randomly split into two groups of equal size $N/2$, one receiving the short list and one the long list. Under the simplifying assumption that all items are independent, the variance of the prevalence estimator in a simple list experiment with four non-sensitive items is:

$$\text{Var}(\widehat{s}_{LE}) = \frac{2V_S}{N} + \frac{4(V_{A1} + V_{A2} + V_{A3} + V_{A4})}{N}$$

where V_S denotes the variance of the sensitive item and V_{Ak} the variance of the non-sensitive item k in list A, with $k \in (1, ..., 4)$. This is straightforward after noting that $\widehat{s}_{LE} = \overline{S + A1 + A2 + A3 + A4}^{G1} - \overline{A1 + A2 + A3 + A4}^{G2}$, with $G1$ the group with the long list and $G2$ the group with the short list. Given independence and the size $N/2$ of groups $G1$ and $G2$, we directly get $\text{Var}(\widehat{s}_{LE}) = V_S/(N/2) + (V_{A1} + V_{A2} + V_{A3} + V_{A4})/(N/2) + (V_{A1} + V_{A2} + V_{A3} + V_{A4})/(N/2)$, which makes $\widehat{s}_{LE} = 18V/N$ in case all items have the same variance V.

Extending the same logic to a double list experiment including two lists A and B yields

$$\text{Var}(\widehat{s}_{DLE}) = \frac{V_S}{N} + \frac{V_{A1} + V_{A2} + V_{A3} + V_{A4} + V_{B1} + V_{B2} + V_{B3} + V_{B4}}{N}$$

To see this simply note that $\widehat{s}_{DLE} = (\widehat{s}_{LEA} + \widehat{s}_{LEB})/2$ with $\widehat{s}_{LEA}$ and $\widehat{s}_{LEB}$ the single list experiment estimates obtained with respectively list A using $G1$ for the long list and list B using $G2$ for the long list. Thus $\text{Var}(\widehat{s}_{DLE}) = (\text{Var}(\widehat{s}_{LEA}) + \text{Var}(\widehat{s}_{LEB}))/4$. Clearly, the overall variance depends on the variance and covariance structure of the non-sensitive items included in the lists. If, for illustration, all items have identical variance V, then $\text{Var}(\widehat{s}_{DLE}) = 9V/N$, implying a standard error about three times larger than that of a direct question. If the additional items are negatively correlated, however, the efficiency loss could be smaller.

C.3 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[2] 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.

C.4 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 C.1.

C.5 Descriptive statistics

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

[2]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 C.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.

C.6 Consistency check

Table C.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 C.2: Descriptive statistics: Double list experiments

	List A		List B	
	(1) Short list	(2) Long list	(3) Short list	(4) 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) Short list	(6) Long list	(7) Short list	(8) 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

C.7 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 C.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 C.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 C.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 C.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 C.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 D). 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.[3]

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

Table C.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.

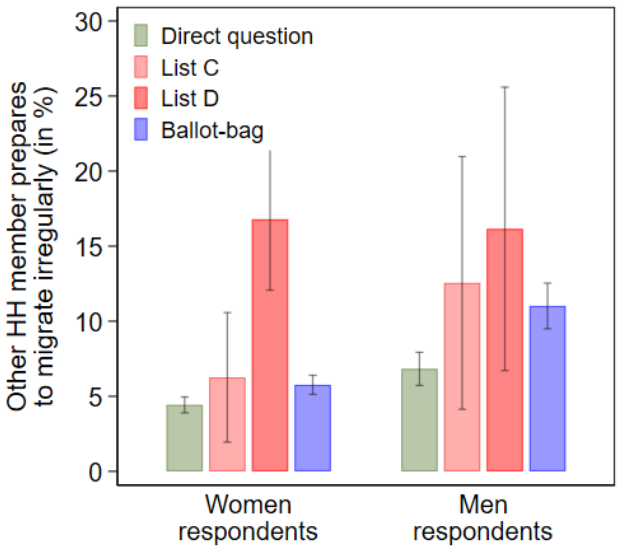
We conjecture that Lists C and/or D may be plagued by another type of design effect whereby the characteristics of the non-sensitive statements alter the responses to the sensitive statement. This effect—akin to a priming effect—may be particularly prevalent for outcomes whose answers are ambiguously 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.[4] 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.

[4]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 C.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.[5] 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 C.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 in the main paper. 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) in the main paper, separately for Lists C and D. All estimates are presented with 95% confidence intervals.

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

To the best of our knowledge, such priming effects have not been considered in previous 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 C.1), which suggests that List C may provide the more reliable estimates.

Online Appendix D Survey scripts

D.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)

D.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)

Online Appendix E Extension for Online Surveys

This appendix reports the results from an extension of the ballot-bag for online settings. The online extension reproduces the core properties of the face-to-face version (separation from the main questionnaire, full anonymity), though without the physical and transparent bag. To test the potential of the online extension, we conducted an online experiment with a sample of approximately 2,000 Egyptian males aged 18-35 residing in Egypt, pre-registered as Gazeaud et al. (2025) on the AEA registry. We assess measurement differences between direct and ballot respondents for three sensitive outcomes as well as for a non-sensitive (placebo) outcome. We find small and non-significant differences between direct and ballot respondents for all outcomes, suggesting that direct respondents already felt comfortable reporting sensitive outcomes truthfully and/or that the online ballot did not provide enough reassurance to reduce misreporting. We conclude with a discussion outlining potential directions for future research in this space.

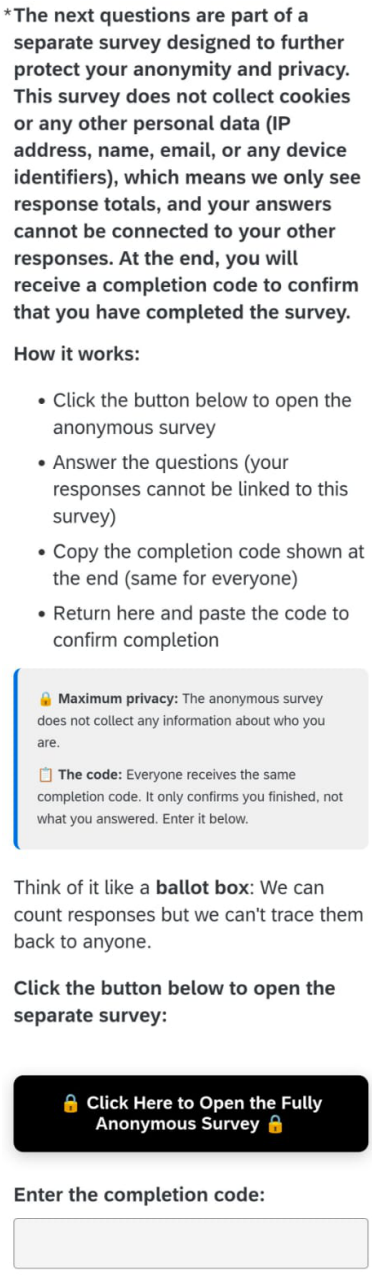
E.1 Online ballot

The protocol is implemented entirely within Qualtrics through two distinct questionnaires. The first is the main questionnaire, which in our case includes basic socio-demographic questions. The second is an *online ballot* designed to enhance anonymity and improve reporting of sensitive outcomes. Respondents access the online ballot by clicking a button in the main survey which opens a separate Qualtrics questionnaire in a new tab. This questionnaire offers full privacy and anonymity (it collects no cookies, IP addresses, device IDs, location information, etc.) and answers cannot be linked to information from the main questionnaire. To ensure that respondents complete the online ballot, they must enter in the main survey a completion code that appears at the end of the ballot. This code does not identify respondents' answers, and we explicitly inform them of this.

We ran several small pilots to test the robustness of the ballot and identified a few challenges, such as how to prevent "ballot-box stuffing" without recording information such as cookies or IP addresses in the ballot, or how to minimize the risk that respondents get lost and fail to return to the main survey after completing the ballot. After a few iterations, we developed a version that was sufficiently robust and could potentially lead to more truthful reporting. The scripts for the key steps (translated from Arabic) are presented in the screenshots below. The Qualtrics *.qls* source code is available from the authors' websites. A key feature is the distinct background and visually salient layout of the

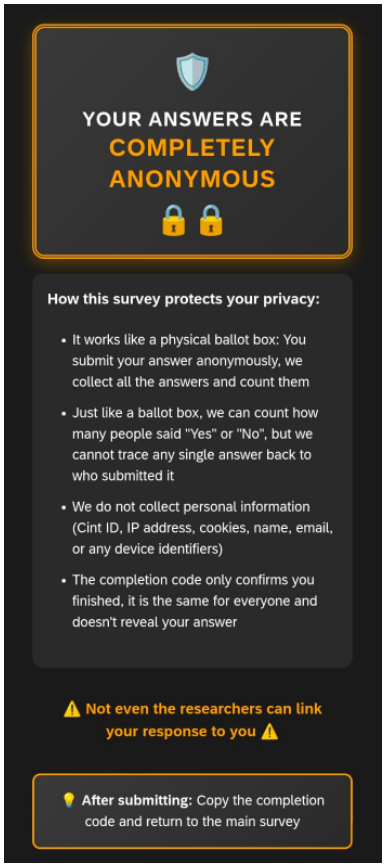
ballot survey, which seeks to reinforce the idea that answers are anonymous.

Step 1: Respondents are introduced to the the ballot procedure and must click on a button to access it. The button is clickable only once to prevent "ballot-box stuffing". Once respondents click, the button turns light grey and respondents cannot use it again (see step 5). This is necessary because the ballot does not collect unique identifiers such as IP addresses or cookies to prevent multiple submissions by the same respondents.



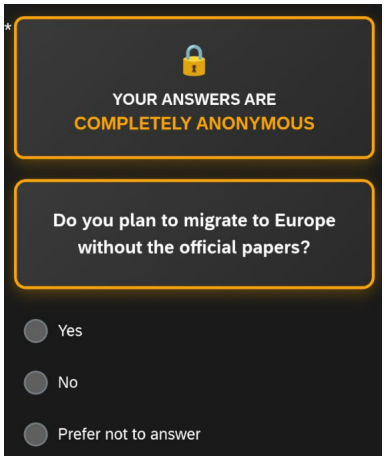
Source: Authors' elaboration

Step 2: The ballot opens in a new tab and respondents are reminded of how it works. The background turns to black and reinforces the idea that answers are completely anonymous.



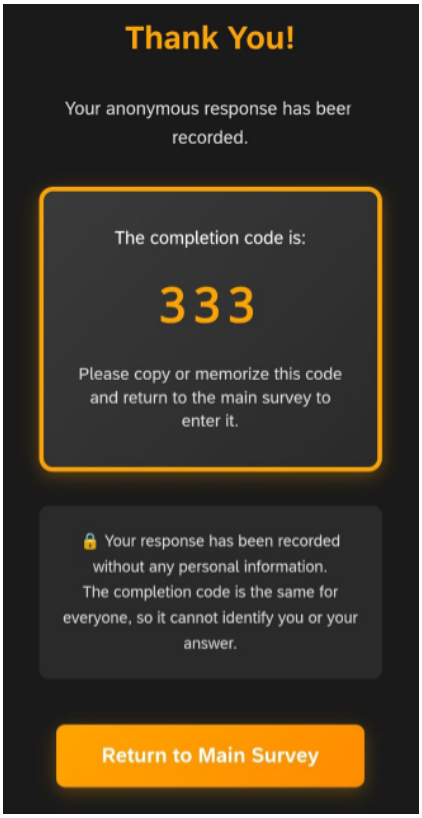
Source: Authors' elaboration

Step 3: Respondents answer the question(s) on the ballot.



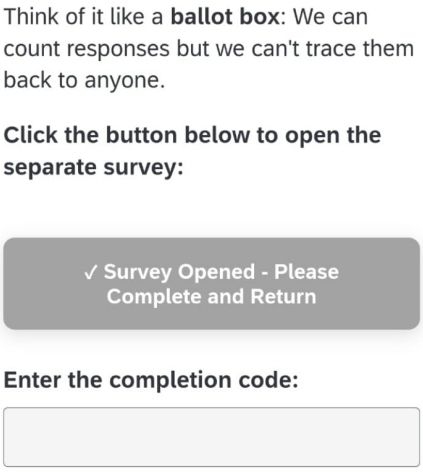
Source: Authors' elaboration

Step 4: At the end of the ballot, respondents receive the completion code and a clickable button to return to the main survey. The completion code serves to confirm that they have completed the ballot and does not identify their answers.



Source: Authors' elaboration

Step 5: In the main questionnaire, respondents have to enter the completion code to proceed until the end of the survey, and cannot click on the ballot button again (to prevent "ballot-box stuffing").



Source: Authors' elaboration

E.2 Experimental design

We use a survey experiment to measure differences between direct and ballot respondents for four outcomes: 1) whether respondents plan to migrate to Europe without the official papers; 2) whether they have ever consumed alcohol, even once; 3) whether they have ever used drugs, even once; and 4) whether they met any of their friends on the day prior to the survey. Outcome 1) is identical to the one used in the in-person version of the ballot. We include outcomes 2) and 3) because they are expected to be sensitive in the Egyptian context and to ensure meaningful analysis in case the true prevalence of outcome 1) is too low in the online sample. Outcome 4) is not expected to be sensitive and therefore serves as a placebo outcome (for this outcome we do not expect any differences between direct and ballot answers).

After completing a basic socio-economic module, respondents are randomly assigned to one of two groups: half answer outcomes 1-4 directly in the main survey and outcomes 2-3 via the online ballot (Group A), while the other half answer outcomes 2-3 directly in the main survey and outcomes 1-4 via the online ballot (Group B).[6] Note that the randomization determines which questions are answered in the main survey versus the ballot, but all respondents eventually complete at least some questions in the ballot. This feature is important because there is a non-trivial number of respondents who do not manage to complete the main survey when the ballot is included, which could create imbalances if a group receives no ballot. In our experiment, out of 187 respondents who started the main survey but did not finish it, 146 (or 78%) stopped on the question with the completion code (i.e., Step 1 in the screenshots above).[7] This is not driven by respondents who access the ballot and cannot return in the main survey or forget the code, as the mismatch between the number of completed main surveys (2,045) and ballot surveys (2,073) is small (28). Nonetheless, this suggests that some respondents may abandon the survey when the script introducing the ballot appears (Step 1) and/or that the ballot effectively screens out very low-quality respondents who do not understand they have to click on the button to proceed and get a completion code.

[6] The order between the direct questions and the ballot is not randomized and all respondents reply to the direct questions first.

[7] To avoid creating imbalances between direct and ballot respondents, we drop all incomplete surveys.

E.3 Recruitment, balance, and sample characteristics

We recruited the sample using the online survey platform CINT. Given the nature of the sensitive outcomes, we focused on Egyptian males aged 18–35, residing in Egypt. We pre-specified that the experiment would run for one week, or until 2,000 eligible individuals complete the survey. This target was reached within the first day of the experiment. Table E.1 presents summary statistics on the respondents and tests whether their characteristics are balanced across the two experimental groups. Respondents are on average 26.3 years old; 99% completed primary school, 40% are married, 84% are urban, and 77% have an income-generating activity. We test the balance between the two groups across all variables using a linear specification and obtain a p-value of 0.30, suggesting that the groups are well balanced.

Table E.1: Online experiment: sample characteristics and balance checks

	(1) N	(2) Sample mean	(3) Sample SD	(4) Group A vs. Group B
Age	2045	26.238	4.839	0.466**
				(0.214)
Primary school completed	2031	0.989	0.106	-0.001
				(0.005)
Married	2017	0.405	0.491	0.026
				(0.022)
Urban	2028	0.836	0.370	0.009
				(0.016)
Has an income generating activity	2011	0.773	0.419	-0.008
				(0.019)
Likelihood to migrate to the Gulf	2045	3.625	1.311	0.009
				(0.058)
Likelihood to migrate to Europe	2045	3.350	1.380	0.031
				(0.061)
Randomization inference p-value	0.30			

Notes: Sample size Group A: 1017; Sample size Group B: 1028. Likelihoods to migrate to the Gulf and Europe are measured using a 5-point Likert scales (1: extremely unlikely, 5: extremely likely). *** p-value $<$ 0.01, ** p-value $<$ 0.05, * p-value $<$ 0.1.

E.4 Results

We find small and non-significant differences between direct and ballot respondents (Table E.2). Among direct respondents, we estimate that 11.2% plan to migrate irregularly to Europe, 13.5% have ever used alcohol, and 10.4% have ever used drugs. Ballot respondents report slightly higher rates for these sensitive outcomes, but the differences are small (0.8pp, 1.3pp, and 0.6pp, respectively) and not statistically significant. For the placebo outcome,

we also estimate non-significant differences: 80.1% of direct respondents report meeting any friends the day prior to the survey, and the ballot effect is a non-significant -1.5pp. This suggests that the lack of ballot effects for sensitive outcomes is unlikely to be spurious. An example of a spurious null ballot effect would occur if respondents pay more attention to the questions in the ballot because of increased saliency of their sensitivity, which could reduce noise and the share of "yes" in ways unrelated to the sensitivity of the outcomes, and in turn mask additional reporting of the sensitive outcome by other respondents.

Table E.2: Online experiment: prevalence estimates do not differ significantly between direct and ballot respondents

	(1) Plans to migrate irregularly to Europe	(2) Ever used alcohol	(3) Ever used drugs	(4) Placebo (met any friends yesterday)
Ballot respondents	0.008	0.013	0.006	-0.015
	(0.014)	(0.015)	(0.014)	(0.018)
Mean (direct respondents)	0.112	0.135	0.104	0.801
Observations	2030	2044	2046	2052

Notes: This table tests for differences between direct and ballot responses for four outcomes (three sensitive and one placebo). Each column shows the coefficient from a regression of the outcome on a dummy variable indicating whether the respondent answered via the ballot method. The number of observations varies due to respondents selecting "prefer not to answer" in some questions and not others, which we code as missing. Robust standard errors are reported in parentheses. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

While these null effects do not allow definitive conclusions about the potential of an online extension of the ballot-bag, they provide useful guidance for more research. They suggest that respondents in the default online setup already felt comfortable reporting sensitive outcomes truthfully, or that our online ballot failed to provide enough reassurance to encourage more truthful reporting among those who under-reported the sensitive outcome. We are unable to distinguish between these two explanations as we lack a benchmark for comparison (such as that provided by a list experiment), and we leave this for future research. We note, however, that sensitivity bias may be less pronounced in online surveys due to the lack of interactions with enumerators, which would support the first explanation. Conversely, the visual features of the in-person ballot-bag could be critical to its effectiveness, which would support the second explanation. More sophisticated versions of the online ballot could incorporate visual elements such as displaying a virtual urn in which each envelope represents a ballot submitted by a previous respondent.

References

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

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.

Gazeaud, Jules, Ahmed Elsayed, and Bruno Crépon, "Measuring sensitive outcomes through online ballots," 2025. Pre-analysis plan. AEA RCT Registry. https://www.socialscienceregistry.org/trials/17301.

_ **, Dina Abdelfattah, Ahmed Elsayed, and Bruno Crépon**, "Employment and Irregular Migration: Evidence from Two Randomized Controlled Trials in Egypt," 2022. Pre-analysis plan. AEA RCT Registry. https://www.socialscienceregistry.org/trials/10604.

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.

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