Report on Survey Weights

Maksim Zubok

August 05, 2024

Contents

Introduction

To create weights for all surveys, we are working with the 2020 census data, particularly the cross tabbed gender, age, and university education file here.

We do not include information about region of residence, even though we could do it after harmonising census data with the survey.

Table 1: Population Frame: Census 2020

gender	age_group	university_education	Freq
Man	18-24	BA+	463546
Man	18-24	BA-	5164261
Man	25-34	BA+	2707302
Man	25-34	BA-	7287296
Man	35-44	BA+	3136411

Sample to Population Comparison

To compare our samples to the population, we do two comparisons. First, we compare them to a nationally representative Levada omnibus survey. Second, we compare them to the 2020 Census.

The Table 2 compares the demographic composition of Qualtrics samples collected in August, February, and March to nationally representative Levada Omnibus survey.

Table 2: Comparison of Category Shares by Variable

Variable	Values	Levada	Q. March	Q. February	Q. August	Q. July
Age	18-24	0.08	0.09	0.09	0.07	0.09
Age	25-34	0.17	0.21	0.21	0.22	0.20
Age	35-44	0.23	0.18	0.18	0.28	0.19
Age	45-54	0.15	0.17	0.17	0.21	0.17
Age	55-64	0.18	0.20	0.25	0.10	0.21
Age	65+	0.19	0.15	0.10	0.12	0.15
Gender	Man	0.44	0.45	0.45	0.49	0.45
Gender	Woman	0.56	0.55	0.55	0.51	0.55
Education	BA+	0.28	0.54	0.54	0.57	0.49
Education	BA-	0.72	0.46	0.46	0.43	0.51

The Figure 1 depicts strata shares compared to the 2020 Census.

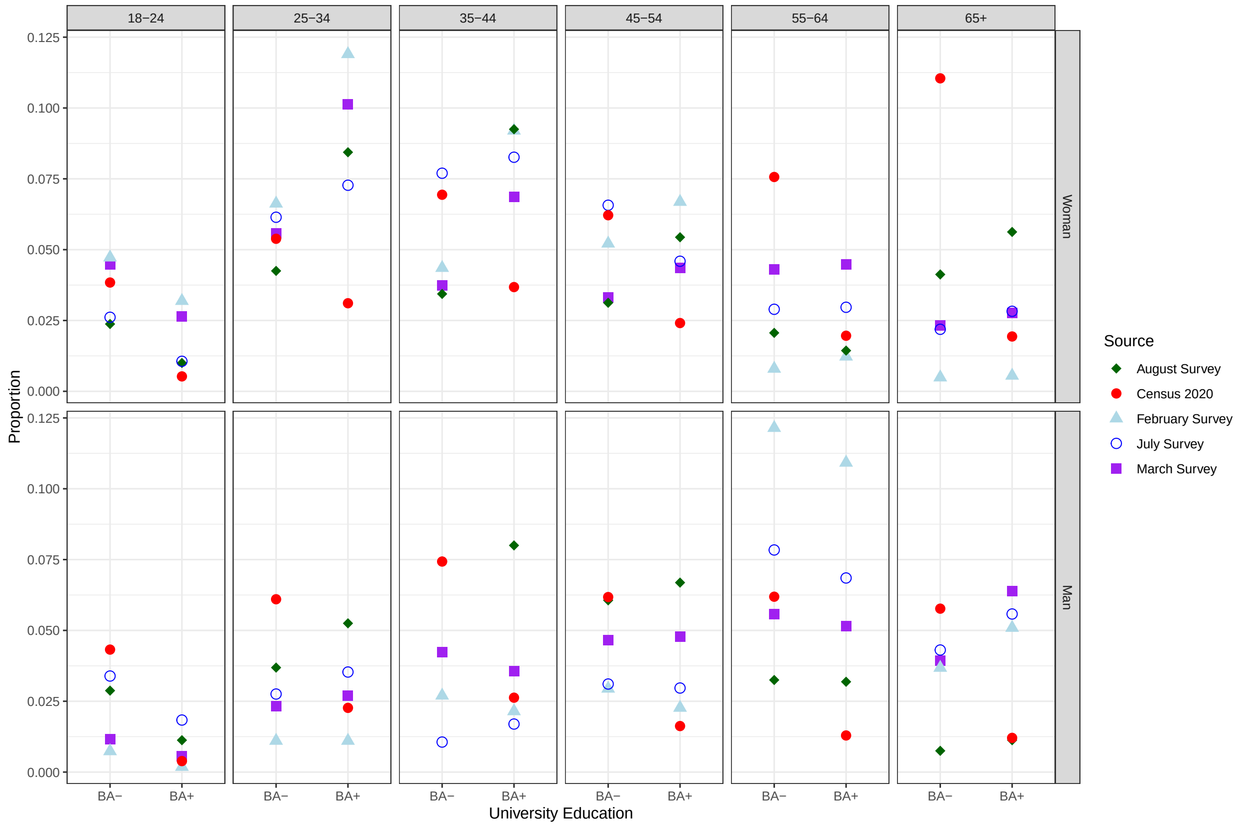


Figure 1: Demographic Comparison on Census 2020

The main observations are:

- The biggest disparities are is women 55-64 and 65+ without education. For instance, the February survey has 0 women in the latter category.
- We have oversampled young women, especially 25-34 years old with university education, and undersampled men without university education across all age categories except 55- and 65+.

Weights with Survey package

To compute post-stratification weights we rely on the `postStratify` function from the `survey` package. The function adjusts the sampling and replicate weights so that the joint distribution of a set of post-stratifying variables matches the known population joint distribution. **However, the package documentation does not describe how exactly the adjustment is implemented.**

March

Table 3: March Survey PostStratify Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	13863.25	24875.41	54076.05	73292.34	107672	347290.3

The table below explores shows strata that were assigned the highest weight.

Table 4: March Survey, Top Five Rows by Weight

age_group	gender	university_education	weight_poststratify
65+	Woman	BA-	347290.3
18-24	Man	BA-	271803.2
25-34	Man	BA-	191770.9
45-54	Woman	BA-	137495.7
35-44	Woman	BA-	135884.6

February

Table 5: February Survey PostStratify Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	8670.35	19144.08	37348.11	73292.34	87350.22	1649629

Note: some strata had no observations in the survey (NA on education for some age gender groups). This means we had to ignore them in producing weights.

Table 6: February Survey, Top Five Rows by Unique Weight

age_group	gender	university_education	weight_poststratify
65+	Woman	BA-	1649629.1
55-64	Woman	BA-	695189.1
18-24	Man	BA-	430355.1
25-34	Man	BA-	404849.8
65+	Woman	BA+	256902.0

August

Table 7: August Survey PostStratify Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	18133.48	27510.75	33083.25	74666.57	112266.5	574250.5

For August, we also see that some weights are much larger than others. As you can see in the graphs below, the distribution of weights is similarly skewed and the disparities between the bulk of the distribution and its tales are in the same orders of magnitude. However, the largest weight in Feb survey is three times bigger than the largest weight in Aug survey.

The largest weights in both surveys relate to different population groups.

Table 8: August Survey, Top Five Rows by Weight

age_group	gender	university_education	weight_poststratify
65+	Man	BA-	574250.5
55-64	Woman	BA-	273862.4
65+	Woman	BA-	199955.0
35-44	Woman	BA-	150708.4
45-54	Woman	BA-	148495.4

July

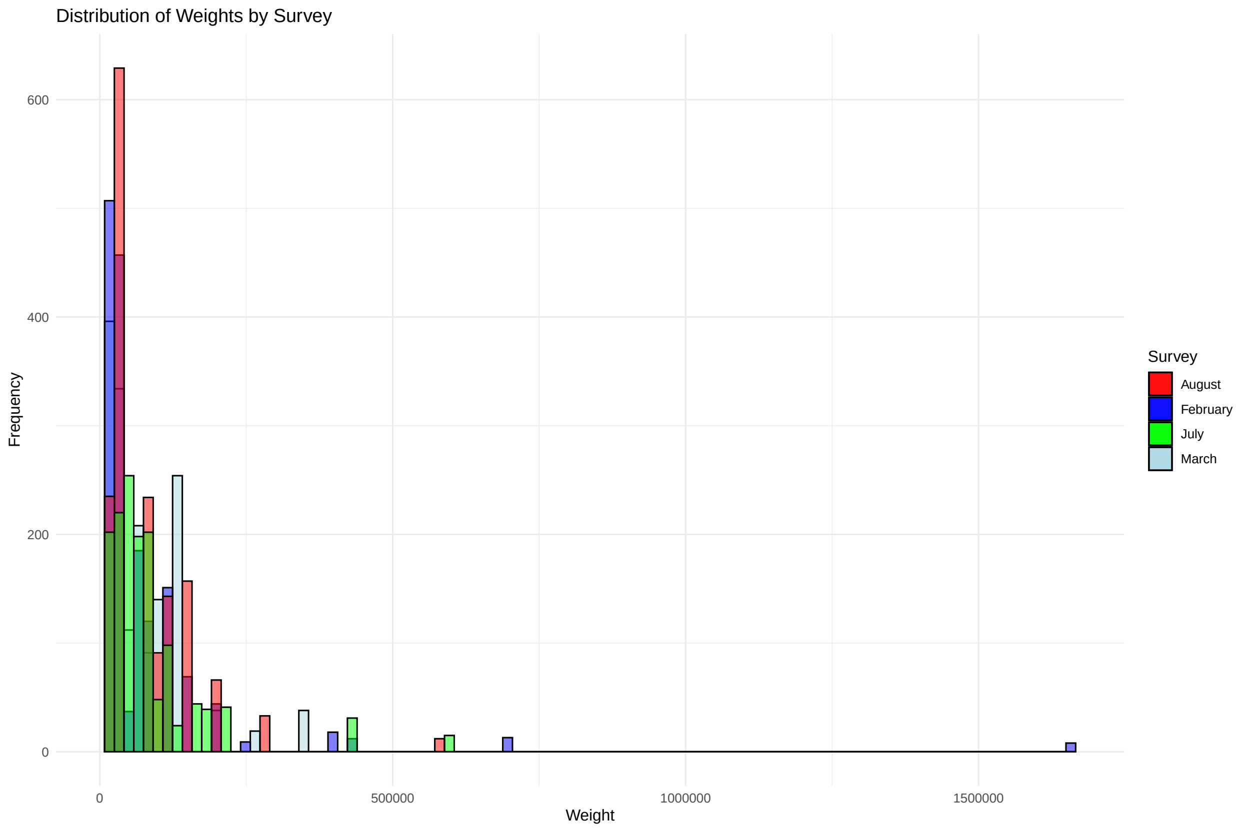
Table 9: August Survey PostStratify Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	15910.55	37565.66	66620.95	84369	79836.23	591827.9

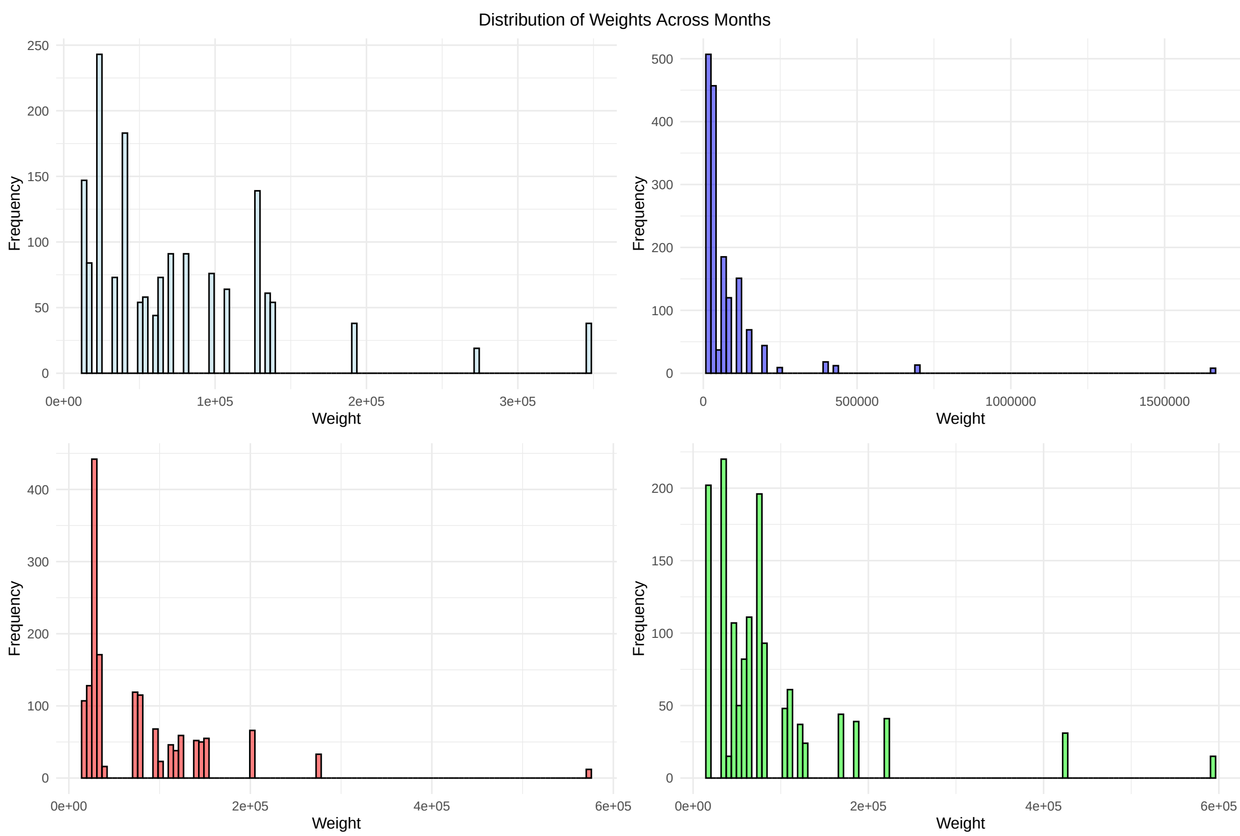
As with previous surveys, the July survey's distribution of weights is also skewed and the disparities between the bulk of the distribution and its tales are in the same orders of magnitude. The largest weight (and biggest disparity to population census 2020 is men without university education aged 35-44).

Table 10: July Survey, Top Five Rows by Weight

age_group	gender	university_education	weight_poststratify
35-44	Man	BA-	591827.9
65+	Woman	BA-	425710.7
55-64	Woman	BA-	220425.8
25-34	Man	BA-	186853.7
45-54	Man	BA-	167718.7



Distribution of Weights by Survey
Frequency
Weight
Survey
August
February
July
March
0
200
400
600
0
500000
1000000
1500000



Distribution of Weights Across Months
Frequency
Weight

Weights created manually

To check the plausibility of resulting weights, we create alternative weights based on the population frequencies of the combination of the same strata (Yana's approach). The weights are calculated for each category:

$$\text{weight}_i = \frac{\text{population frequency}_i}{\text{sample frequency}_i}$$

March

Table 11: March Survey Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	0.19	0.51	0.85	1.24	1.76	4.74

February

Table 12: February Survey Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	0.12	0.48	1.39	2.81	2.27	22.51

August

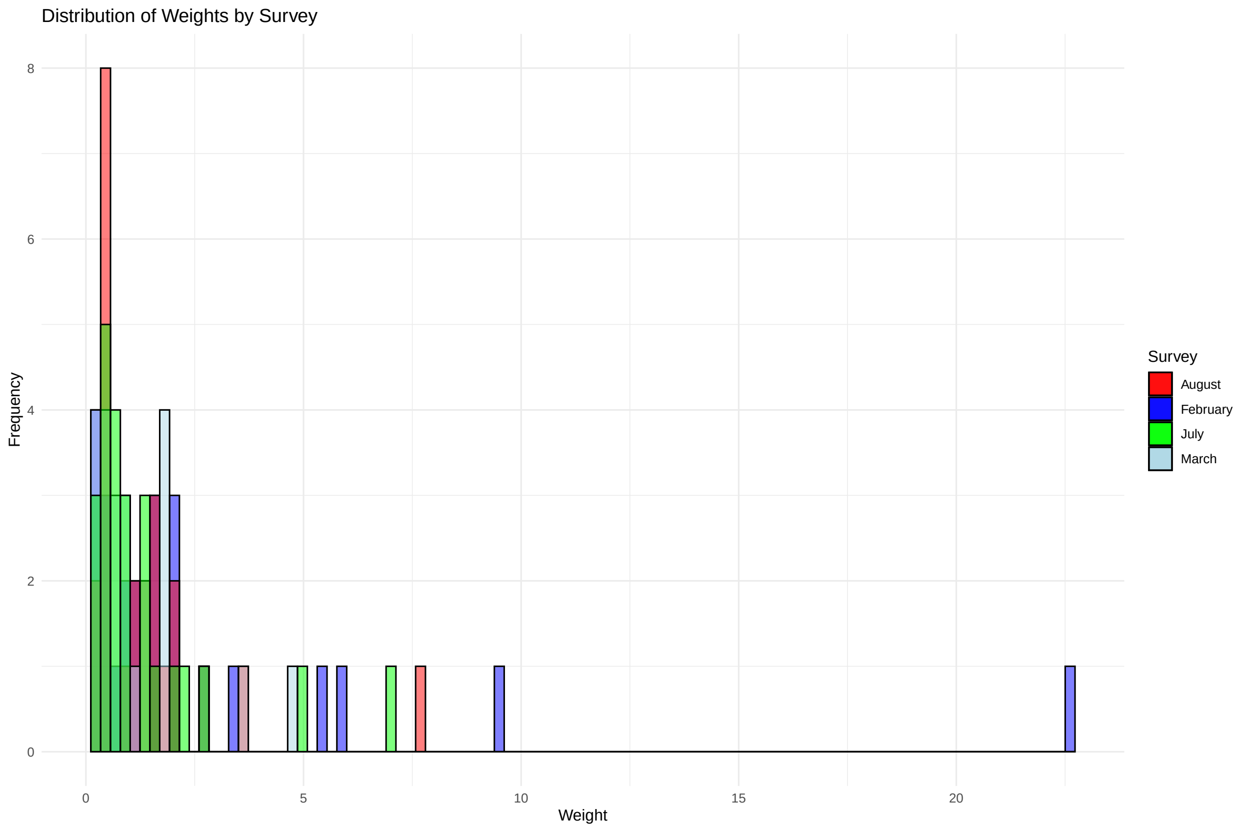
Table 13: August Survey Weights Summary

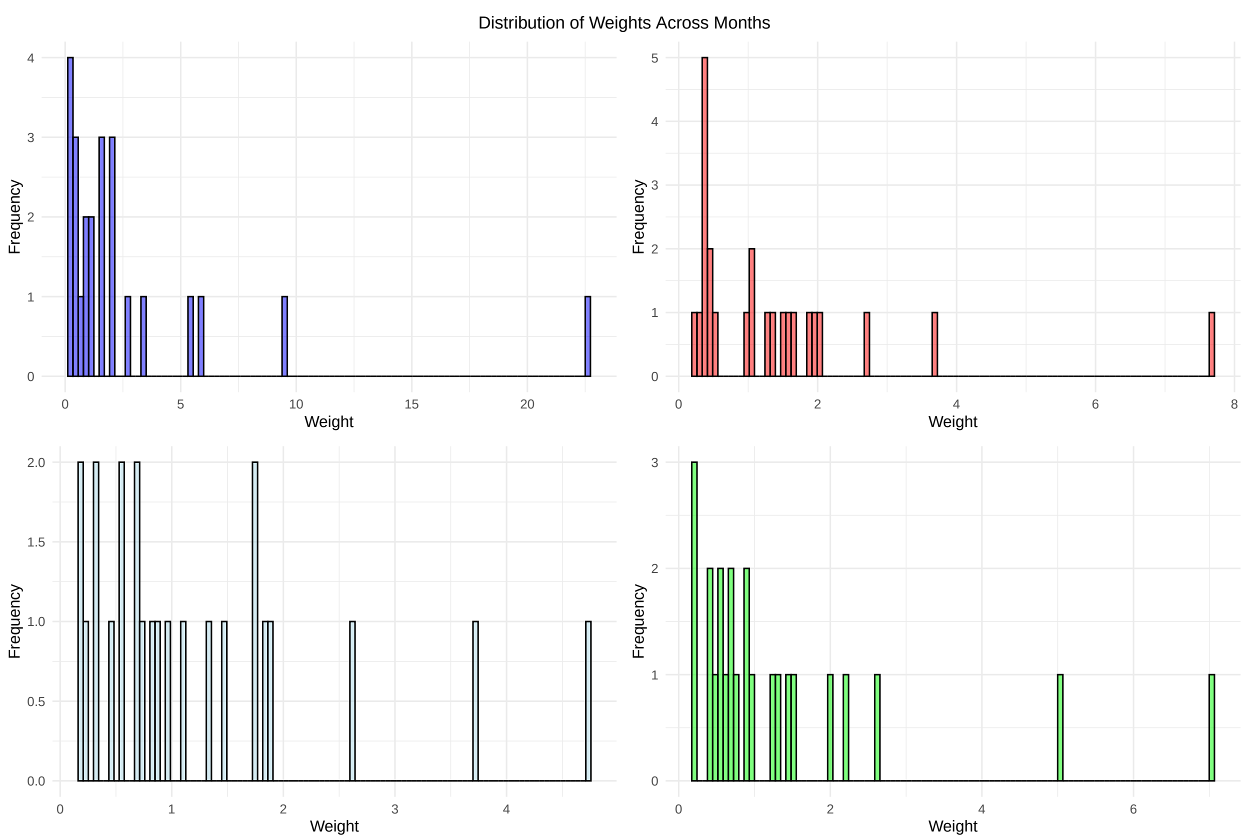
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	0.24	0.4	1.05	1.43	1.72	7.69

July

Table 14: July Survey Weights Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	0.19	0.52	0.83	1.38	1.49	7.01





Calculating weights this way gives us the same overall picture. The main difference is in the orders of magnitude between the median weight and the tail (highest weight).

For **February** survey, the difference (max weight / median weight) with `postStratify` is **44.17**, but with simple manually created weights it is **16.19**. For **August** survey, the difference with `postStratify` is **17.36**, but with simple manually created weights it is **7.32**.

Survey July Randomisation check

To check for randomisation, I first create a categorical variable that captures treatment assignment:

```
survey_july <- survey_july %>%
  mutate(
    treatment1 = case_when(!is.na(DV1A) ~ "A",
                           !is.na(DV1B) ~ "B",
                           !is.na(DV1C) ~ "C",
                           !is.na(DV1D) ~ "D",
                           !is.na(DV1E) ~ "E",
                           !is.na(DV1F) ~ "F",
                           !is.na(DV1H) ~ "I",
                           !is.na(DV1I) ~ "I",
                           TRUE ~ NA_character_)
```

Then, I run a multinomial logit to see if the odds of a treatment category assignment relative to control (or first treatment category) can be predicted as a function of the base demographic controls, namely age, gender, university education, and size of a city of residence.

As I see from the tables below, we do not have any violation of random assignment, as indicated by the absence of any relationships between treatment assignment and demographic variables that can be statistically different from zero.

Table 15: Randomisation Check Protest

	Dependent variable:					
	B	C	D	E	F	I
	(1)	(2)	(3)	(4)	(5)	(6)
age	-0.02^{**}	-0.01	-0.02^{**}	0.003	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
genderMan	-0.31	0.11	-0.05	-0.59^{***}	-0.08	-0.24
	(0.22)	(0.22)	(0.22)	(0.23)	(0.22)	(0.19)
university_educationBA+	-0.39^{*}	-0.41^{*}	-0.09	-0.29	-0.30	-0.18
	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)	(0.19)
urbancity_250_500_thousand	0.01	-0.10	-0.48	-0.93^{**}	-0.57	-0.13
	(0.46)	(0.45)	(0.46)	(0.43)	(0.46)	(0.40)
urbancity_500_thousand_to_1_million	0.12	0.18	-0.08	-0.57	0.15	0.28
	(0.46)	(0.45)	(0.45)	(0.42)	(0.44)	(0.40)
urbancity_more_than_1_million	-0.07	-0.44	-0.30	-0.91^{**}	-0.55	-0.12
	(0.43)	(0.43)	(0.42)	(0.39)	(0.42)	(0.37)
urbancity_or_village_less_than_100_thousand	-0.18	-0.47	-0.21	-0.78^{**}	-0.50	0.05
	(0.43)	(0.43)	(0.42)	(0.39)	(0.42)	(0.37)
urbandifficult_to_answer	-1.13	0.33	-0.44	-1.64	0.41	-0.20
	(1.28)	(0.94)	(1.07)	(1.27)	(0.91)	(0.93)
urbanMoscow	0.11	0.03	-0.25	-0.96^{**}	-0.18	0.01
	(0.47)	(0.46)	(0.47)	(0.45)	(0.46)	(0.41)
urbanno_answer	-0.55	-12.56^{***}	-12.87^{***}	-13.48^{***}	-0.57	-0.94
	(1.47)	(0.0000)	(0.0000)	(0.0000)	(1.47)	(1.46)
urbanSaint_Petersburg	-0.91	-0.17	-0.38	-1.43^{**}	-0.35	-0.12
	(0.65)	(0.55)	(0.56)	(0.59)	(0.54)	(0.48)
Constant	1.21^{**}	0.83	1.14^{**}	1.01^{**}	0.93^{*}	1.25^{***}
	(0.51)	(0.51)	(0.50)	(0.48)	(0.50)	(0.45)
n	1416	1416	1416	1416	1416	1416
Akaike Inf. Crit.	5,461.07	5,461.07	5,461.07	5,461.07	5,461.07	5,461.07

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16: Randomisation Check Education Experiment 1

	Dependent variable:	
	B	C
	(1)	(2)
age	0.01*	0.002
	(0.005)	(0.004)
genderMan	−0.16	0.05
	(0.14)	(0.14)
university_educationBA+	0.07	0.20
	(0.13)	(0.13)
urbancity_250_500_thousand	0.05	0.13
	(0.27)	(0.27)
urbancity_500_thousand_to_1_million	0.11	0.01
	(0.26)	(0.26)
urbancity_more_than_1_million	0.22	−0.02
	(0.25)	(0.25)
urbancity_or_village_less_than_100_thousand	0.30	0.32
	(0.25)	(0.25)
urbandifficult_to_answer	−0.47	0.08
	(0.65)	(0.56)
urbanMoscow	0.01	−0.10
	(0.27)	(0.27)
urbanno_answer	0.97	0.21
	(1.25)	(1.43)
urbanSaint_Petersburg	0.57	0.76**
	(0.37)	(0.36)
Constant	−0.46	−0.34
	(0.29)	(0.29)
n	1416	1416
Akaike Inf. Crit.	3,138.87	3,138.87
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 17: Randomisation Check Education Experiment 2

	Dependent variable:	
	B	C
	(1)	(2)
age	-0.01^{**}	-0.01
	(0.004)	(0.005)
genderMan	0.25^{*}	-0.09
	(0.14)	(0.14)
university_educationBA+	0.06	0.13
	(0.13)	(0.13)
urbancity_250_500_thousand	-0.32	-0.59^{**}
	(0.28)	(0.27)
urbancity_500_thousand_to_1_million	-0.08	-0.45^{*}
	(0.27)	(0.26)
urbancity_more_than_1_million	-0.26	-0.45^{*}
	(0.26)	(0.25)
urbancity_or_village_less_than_100_thousand	-0.36	-0.68^{***}
	(0.26)	(0.25)
urbandifficult_to_answer	-0.22	0.10
	(0.67)	(0.60)
urbanMoscow	-0.07	-0.35
	(0.29)	(0.28)
urbanno_answer	0.79	-12.19^{***}
	(1.18)	(0.0000)
urbanSaint_Petersburg	0.03	-0.55
	(0.35)	(0.36)
Constant	0.52^{*}	0.68^{**}
	(0.30)	(0.30)
n	1416	1416
Akaike Inf. Crit.	3,129.12	3,129.12
Note:	*p<0.1; **p<0.05; ***p<0.01	

Survey July Missingness

To check for missing responces, I pull together all treatment variables of the same experiment and treat answers "I do not know" and "refuse to answer" as missing. Then, I estimate if base control demographic variables predict missingness in the experiment. The table below shows the results.

As I see from the table, younger people, women, and people who could not answer about the population size of their place of residence are less likely to answer. People with university education are more likely to provide an answer to the experiments questions.

Table 18: Missings Check Experiments

	Dependent variable:		
	missing_treatment1	missing_treatment2	missing_treatment3
	(1)	(2)	(3)
age	-0.002^{**}	-0.002^{***}	-0.002^{**}
	(0.001)	(0.001)	(0.001)
genderMan	-0.09^{***}	-0.03^{*}	-0.11^{***}
	(0.02)	(0.02)	(0.02)
university_educationBA+	-0.11^{***}	-0.02	-0.09^{***}
	(0.02)	(0.02)	(0.02)
urbancity_250_500_thousand	0.11^{**}	0.002	-0.04
	(0.05)	(0.04)	(0.04)
urbancity_500_thousand_to_1_million	0.05	0.02	-0.01
	(0.04)	(0.03)	(0.04)
urbancity_more_than_1_million	0.03	0.003	-0.02
	(0.04)	(0.03)	(0.04)
urbancity_or_village_less_than_100_thousand	0.08^{*}	0.03	0.01
	(0.04)	(0.03)	(0.04)
urbandifficult_to_answer	0.31^{***}	0.22^{***}	0.24^{**}
	(0.10)	(0.08)	(0.10)
urbanMoscow	0.01	0.03	-0.01
	(0.05)	(0.04)	(0.04)
urbanno_answer	-0.03	0.08	0.45^{**}
	(0.22)	(0.17)	(0.20)
urbanSaint_Petersburg	0.10^{*}	0.02	0.001
	(0.06)	(0.05)	(0.06)
Constant	0.39^{***}	0.25^{***}	0.40^{***}
	(0.05)	(0.04)	(0.05)
Observations	1,416	1,416	1,416
R^2	0.05	0.03	0.06
Adjusted R^2	0.04	0.02	0.05
Residual Std. Error (df = 1404)	0.43	0.33	0.40
F Statistic (df = 11; 1404)	7.05^{***}	3.56^{***}	7.54^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01