Supplemental Appendix for "Education and Selection into Ethnic Identification: Evidence from Roma People in Romania"[†]

Andreea Mitrut ⓡ Gabriel Kreindler ⓡ Margareta Matache
ⓡ Andrei Munteanu ⓡ Cristian Pop-Eleches

September 1, 2026

[†]Mitrut: University of Gothenburg (email: andreea.mitrut@economics.gu.se); Kreindler: Harvard University (email: gkreindler@fas.harvard.edu); Matache : Harvard University (email: mmatache@hsph.harvard.edu); Munteanu: Université du Québec à Montréal (email: munteanu.andrei@uqam.ca); Pop-Eleches: Columbia University (email: cp2124@columbia.edu).

A Additional Exhibits

Table A.1: Census Linkages

	Linkage	All	Roma	Educated
Inconsistent Sex (%)	1992 - 2011	3.0	4.7	3.4
	2002 - 2011	1.3	2.2	1.5
Total Unique Matches	1992 - 2011	4,545,737	95,293	2,096,149
	2002 - 2011	6,286,727	183,922	2,554,153

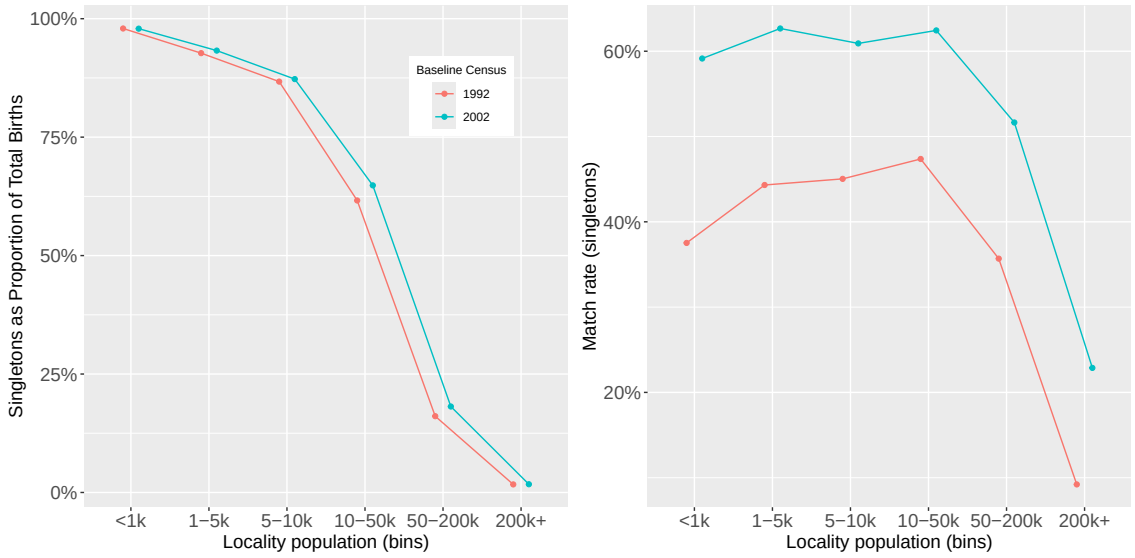
Notes: This table reports the share of linked observations that have inconsistent sex, when using only locality and date of birth to link observations. "Educated" represents individuals who have more than 8 years of schooling at endline (2011).

Table A.2: Census Linkages for Baseline Roma

Education	1992–2011		2002–2011	
	Observations	Inconsistent Sex (%)	Observations	Inconsistent Sex (%)
	(1)	(2)	(3)	(4)
None	15,308	2.2	33,642	1.2
Primary School	28,378	2.2	64,362	1.5
Middle School	36,880	3.6	65,982	1.7
High School or Vocational	13,209	13.2	16,371	6.8
Postsecondary	1,518	29.5	1,440	22.2

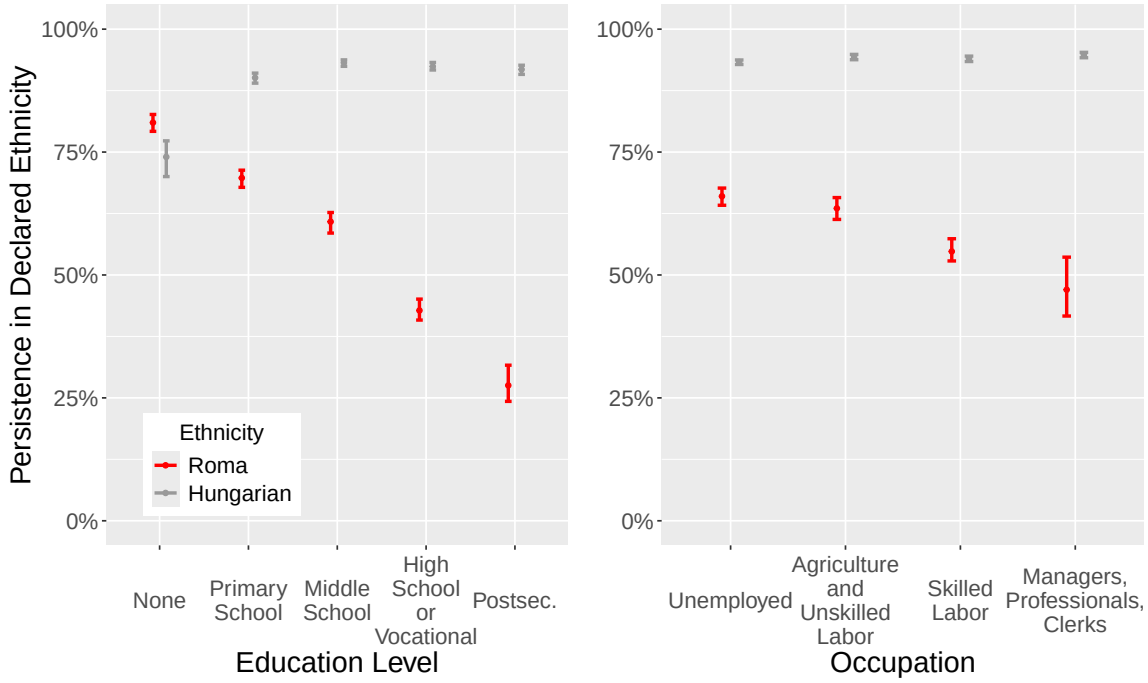
Notes: This table reports the share of linked observations that have inconsistent sex by education level, when using only locality and date of birth to link observations. The sample is individuals with reported Roma ethnicity at baseline.

Figure A.1: Singleton Cells and Match Rates by Town Size



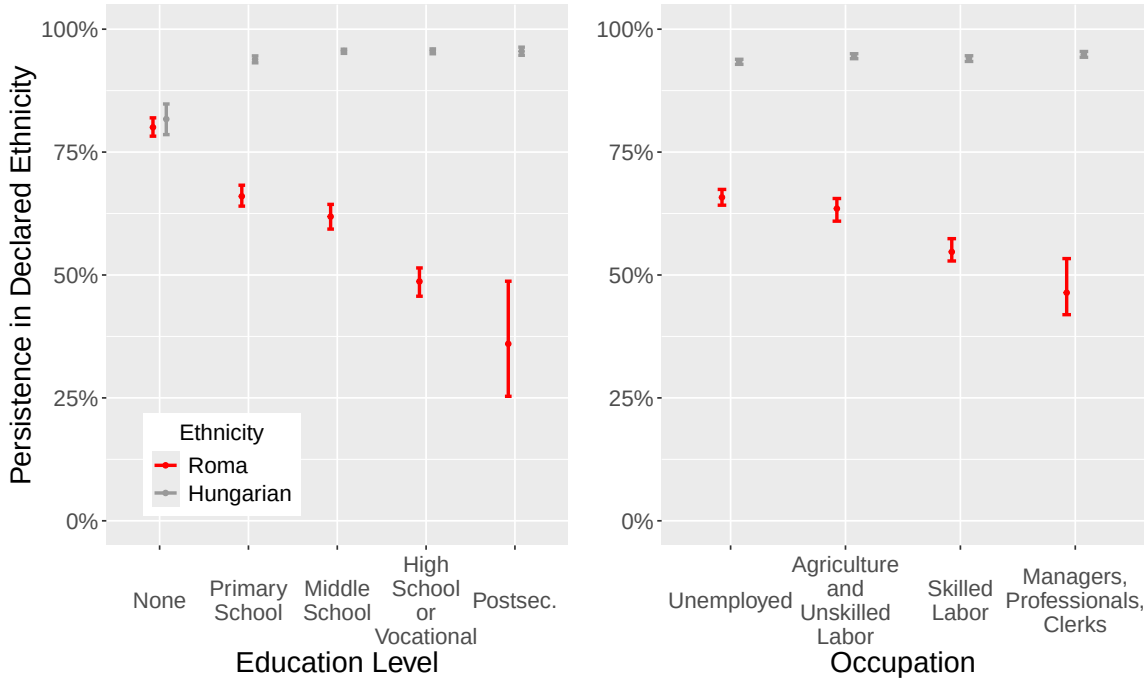
Notes: This figure plots the census record linkage rate by town population bins. More specifically, it plots the percentage of baseline (1992 or 2002) date of birth, locality, and sex singleton cells that are linked to endline (2011) singletons defined by the same variables, yielding unique matches.

Figure A.2: Changes in Ethnic Identification by Education and Occupation, 1992 - 2011



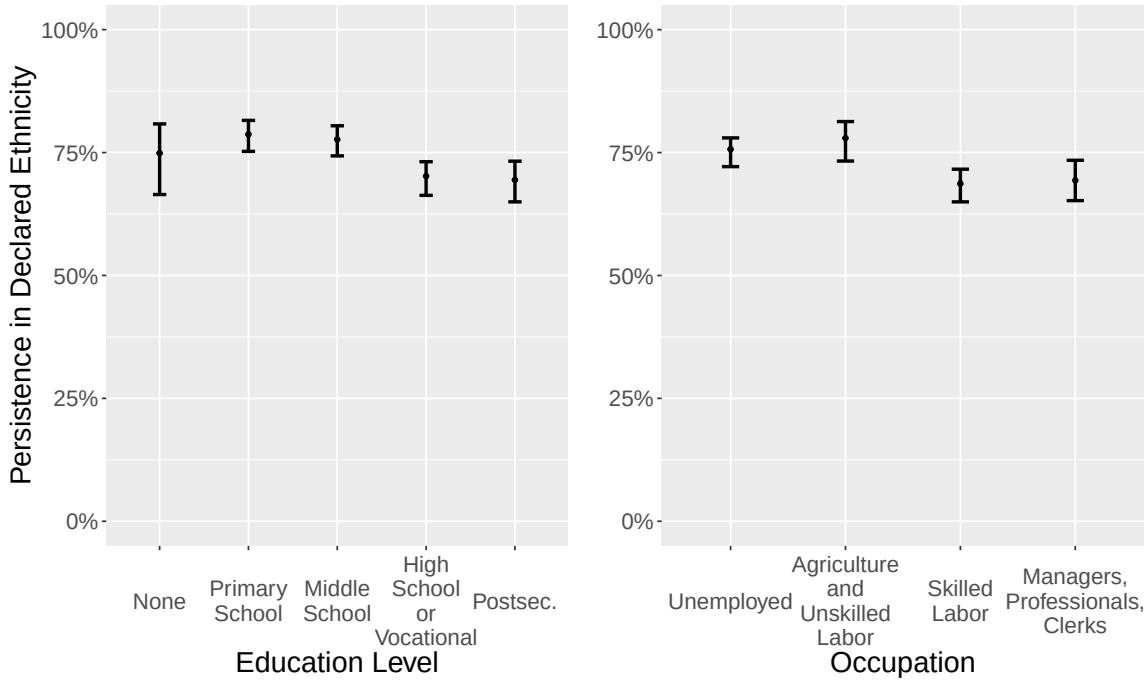
Notes: Replicates Figure 1 using locality-specific gender ratios to more accurately estimate record mismatch.

Figure A.3: Changes in Ethnic Identification by Education and Occupation: Household Heads Only (1992 - 2011)



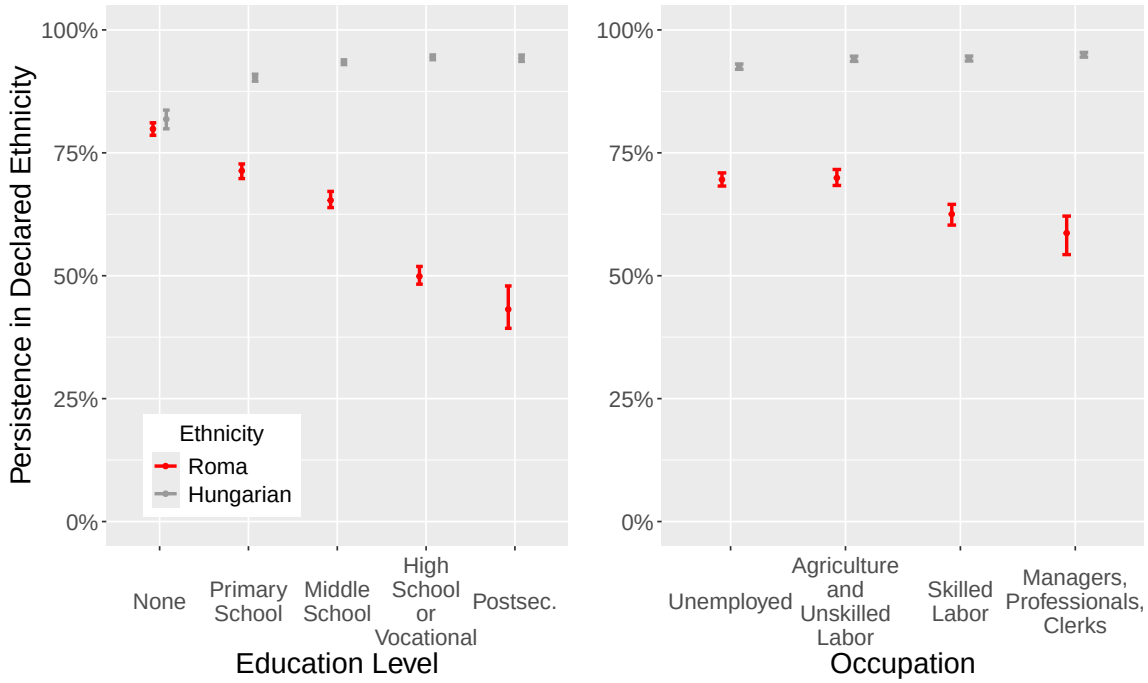
Notes: Replicates Figure 1 with individuals who were household heads (and thus, self-reporting their ethnicity) in 1992 and 2011.

Figure A.4: Changes in Ethnic Identification by Education: Other Ethnicities (1992 - 2011)



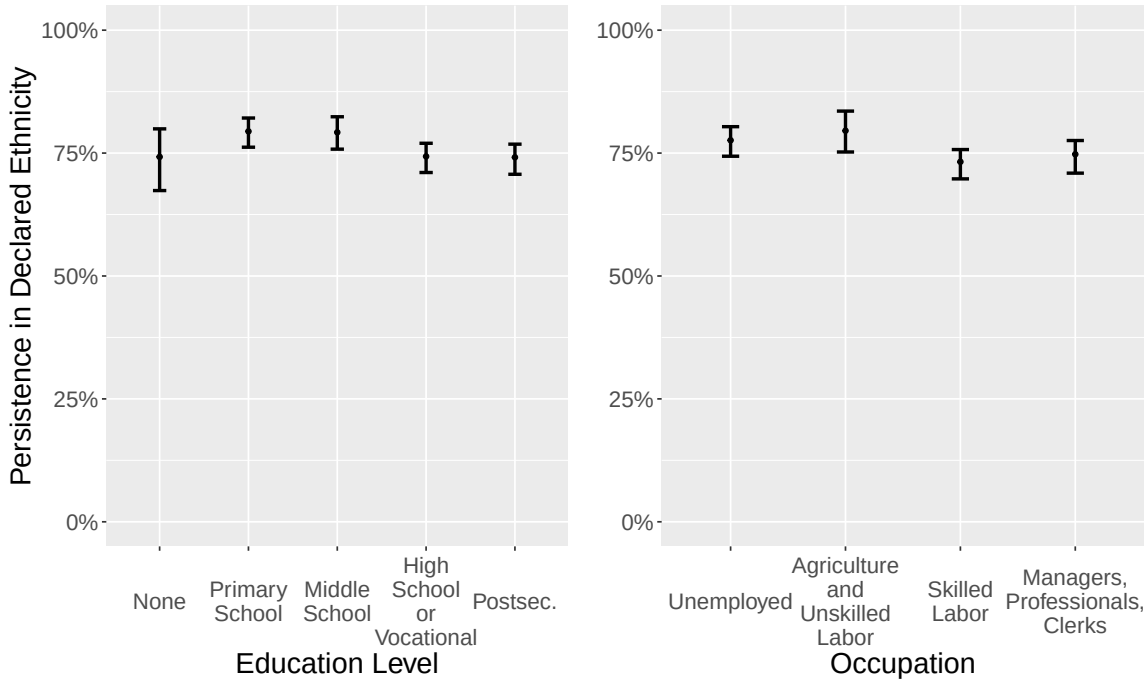
Notes: Replicates Figure 1 for all ethnicities other than Roma and Hungarian, with baseline year 1992. We pool all ethnicities. Changes in ethnic identification are defined as reporting Romanian, Roma, or Hungarian ethnicity.

Figure A.5: Changes in Ethnic Identification by Education (2002 - 2011)



Notes: Replicates Figure 1 with baseline year 2002.

Figure A.6: Changes in Ethnic Identification by Education: Other Ethnicities (2002 - 2011)



Notes: Replicates Figure A.4 with baseline year 2002.

Table A.3: Conditional Independence of Reported Education and Ethnicity for Mismatched Records

	Dependent Variable:			
	Yrs of Schooling (2011)		*Reported Roma (2011)*	
	(All)	(Baseline Roma)	(All)	(Baseline Roma)
	(1)	(2)	(3)	(4)
Baseline Years of Schooling	-0.003	-0.080	-0.001**	-0.003
	(0.006)	(0.060)	(0.000)	(0.004)
N	131,854	4,336	131,854	4,336
R^2	0.70	0.93	0.60	0.92
DV Mean	9.78	9.32	0.04	0.21

Notes: This table shows the relationship between reported education at baseline and i) reported education at endline (1-2) and ii) endline self-reported Roma ethnicity (3-4), conditional on records being mismatched. We use the 1992-2011 sample matched using town of residence and birthdates. We restrict the sample to matched records with different sex between the baseline and endline survey years, virtually guaranteeing that the records are mismatched. Controls include fixed effects for birth year interacted with locality. Standard errors are clustered at the locality level. *p<0.1; **p<0.05; ***p<0.01

Table A.4: Changes in Ethnic Identification by Education (Mothers' Sample)

	OLS '92-'11 (1)	IV '92-'11 (2)	OLS '02-'11 (3)	IV '02-'11 (4)
Schooling Yrs	-0.028***	-0.033***	-0.024***	-0.024***
	(0.002)	(0.012)	(0.002)	(0.004)
Sample	'92-VSN-'11	'92-VSN-'11	'02-VSN-'11	'02-VSN-'11
Outcome	Mothers Mothers Census '11	Mothers Census '11	Mothers Census '11	Census '11
N	19,707	19,707	33,666	33,666
R^2	0.74	0.74	0.70	0.70
DV Mean	0.66	0.66	0.72	0.72
F-stat		424.0		4,981.7

Notes: This table shows the relationship between declaring Roma ethnicity and completed years of schooling, for the subsamples of mothers perfectly linked between the VSN and the 2011 census, that we link back to the 1992 census. Standard errors are clustered at the locality level. *p<0.1; **p<0.05; ***p<0.01.

Table A.5: Changes in Ethnic Identification by Education: Instrumental Variables Approach

	OLS '92-'11	IV '92-'11	OLS 'VSN-'11	IV '92-'11	IV '92-'11
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.818***	0.788***	0.817***	0.632***	0.843***
	(0.010)	(0.016)	(0.016)	(0.068)	(0.046)
Schooling Yrs	-0.034***	-0.029***	-0.022***	-0.058***	-0.039***
	(0.001)	(0.003)	(0.002)	(0.010)	(0.008)
Sample	'92-'11	'92-'11	VSN-'11	'92-VSN-'11	'92-VSN-'11
Outcome	Individual Census '11	Individual Census '11	Mother Census '11	Child VSN	Child Census '11
N	107,168	107,168	30,789	19,707	19,707
R^2	0.08	0.07	0.04	-0.07	0.08
DV Mean	0.60	0.60	0.71	0.30	0.62
F-stat		11,704.1		636.8	636.8

Notes: This table replicates Table 1, using a specification with no fixed effects.

Table A.6: Changes in Ethnic Identification by Education (2002 - 2011)

	OLS '02-'11 (1)	IV '02-'11 (2)	OLS 'VSN-'11 (3)	IV '02-'11 (4)	IV '02-'11 (5)
Schooling Yrs	-0.024***	-0.022***	-0.023***	-0.029***	-0.026***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.004)
Sample	'02-'11	'02-'11	VSN-'11	'02-VSN-'11	'02-VSN-'11
Outcome	Individual Census '11	Individual Census '11	Mother Census '11	Child VSN	Child Census '11
N	210,076	210,076	30,789	33,666	33,666
R^2	0.48	0.48	0.46	0.68	0.67
DV Mean	0.67	0.67	0.71	0.30	0.68
F-stat		38,898.4		4,981.7	4,981.7

Notes: This table shows the relationship between declaring Roma ethnicity and years of schooling. It replicates Table 1 using the 2002 - 2011 linked sample. Controls include fixed effects for birth year interacted with locality. Standard errors are clustered at the locality level. *p<0.1; **p<0.05; ***p<0.01.

Table A.7: Changes in Ethnic Identification by Education – Ethnicity Defined using Romani as Mother Tongue

	OLS '92-'11	IV '92-'11
	(1)	(2)
Schooling Yrs	-0.024***	-0.022***
	(0.002)	(0.004)
Sample	'92-'11	'92-'11
Outcome	Individual Census '11	Individual Census '11
N	40,563	40,563
R^2	0.63	0.63
DV Mean	0.53	0.53
F-stat		3,057.6

Notes: This table shows the relationship between years of schooling and reported Romani native language, using 1992 - 2011 linked records. The sample consists of individuals who report Romani as their native language at baseline (1992). The outcome is reporting Romani as a native language at endline (2011). Controls include fixed effects for birth year interacted with locality. Standard errors are clustered at the locality level. *p<0.1; **p<0.05; ***p<0.01.

Table A.8: Changes in Ethnic Identification by Education and Town Size

	Dependent Variable: Reported Roma Ethnicity (2011)					
	(1)	(2)	(3)	(4)	(5)	(6)
Schooling Yrs (Endline)	-0.028***	-0.025***	-0.016*	-0.025***	-0.020***	-0.017**
	(0.003)	(0.003)	(0.009)	(0.002)	(0.002)	(0.006)
Sample	'92-'11	'92-'11	'92-'11	'02-'11	'02-'11	'02-'11
Town Population	<5k	5k-50k	50k+	<5k	5k-50k	50k+
N	57,858	46,759	2,551	105,626	98,695	5,755
R^2	0.59	0.52	0.54	0.52	0.42	0.41
DV Mean	0.62	0.60	0.32	0.68	0.68	0.39
F-stat	7,060.1	4,260.1	182.0	21,271.9	17,817.7	430.8

Notes: This table shows the relationship between declaring Roma ethnicity and years of schooling across towns with different populations in 2011 an across different linked samples. Controls include fixed effects for birth year interacted with locality. Standard errors are clustered at the locality level. *p<0.1; **p<0.05; ***p<0.01.

Table A.9: Changes in Ethnic Identification by Education – Different Specifications (1992–2011)

	Household FE		Enumerator FE		Non-Declaration	
	OLS	IV	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Schooling Yrs	-0.020***	-0.005*	-0.015***	-0.013***	-0.023***	-0.030***
	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)	(0.002)
N	107,168	107,168	107,168	107,168	107,168	107,168
R^2	0.88	0.88	0.78	0.78	0.57	0.57
DV Mean	0.60	0.60	0.60	0.60	0.66	0.66
F-stat		5,079.2		8,978.4		11,200.1

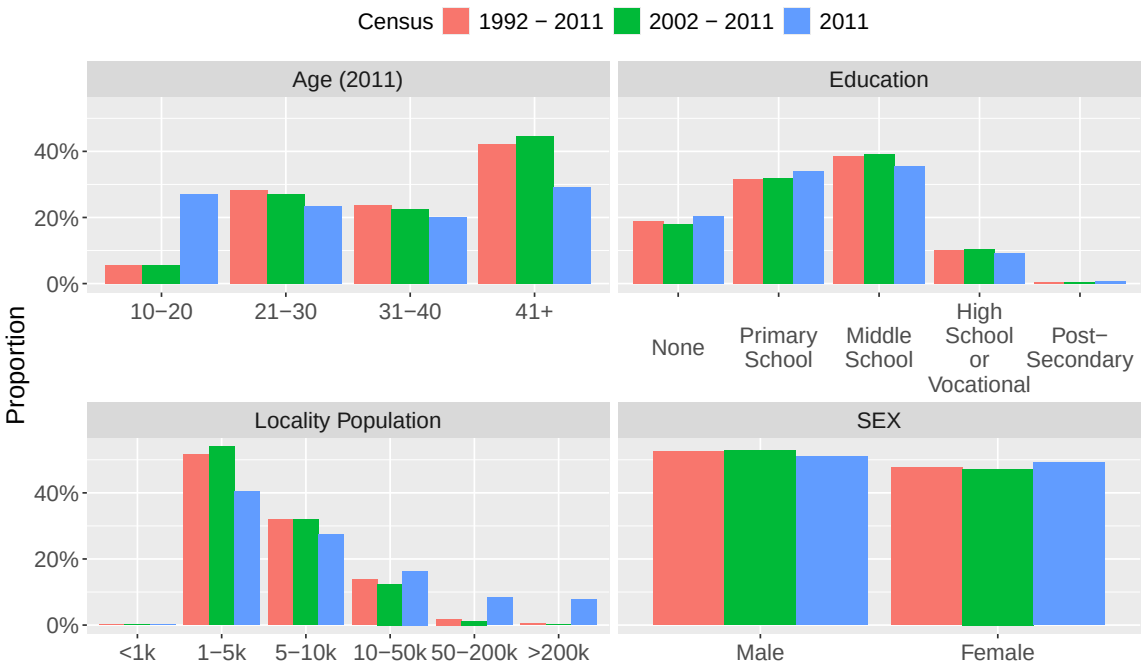
Notes: This table shows the relationship between declaring Roma ethnicity and years of schooling. It provides robustness checks corresponding to the first two columns of Table 1: adding household fixed effects (columns 1 and 2), adding enumerator fixed effects (columns 3 and 4) and omitting non-declaration of ethnicity from changes in ethnic identification by recoding non-responses to the ethnicity question as Roma ethnicity (columns 4 and 5). *p<0.1; **p<0.05; ***p<0.01.

Table A.10: Between-Sibling Changes in Ethnic Identification

	Dependent Variable: Reported Roma Ethnicity (2011)			
	OLS '92-'11	IV '92-'11	OLS VSN	IV VSN
Schooling Yrs	-0.023***	-0.007	-0.012***	-0.083**
	(0.002)	(0.010)	(0.001)	(0.042)
Sample	'92-'11	'92-'11	VSN	'92-VSN-'11
N	47,043	47,043	15,018	2,601
DV Mean	0.62	0.62	0.64	0.72
F-stat		506.7		16.5

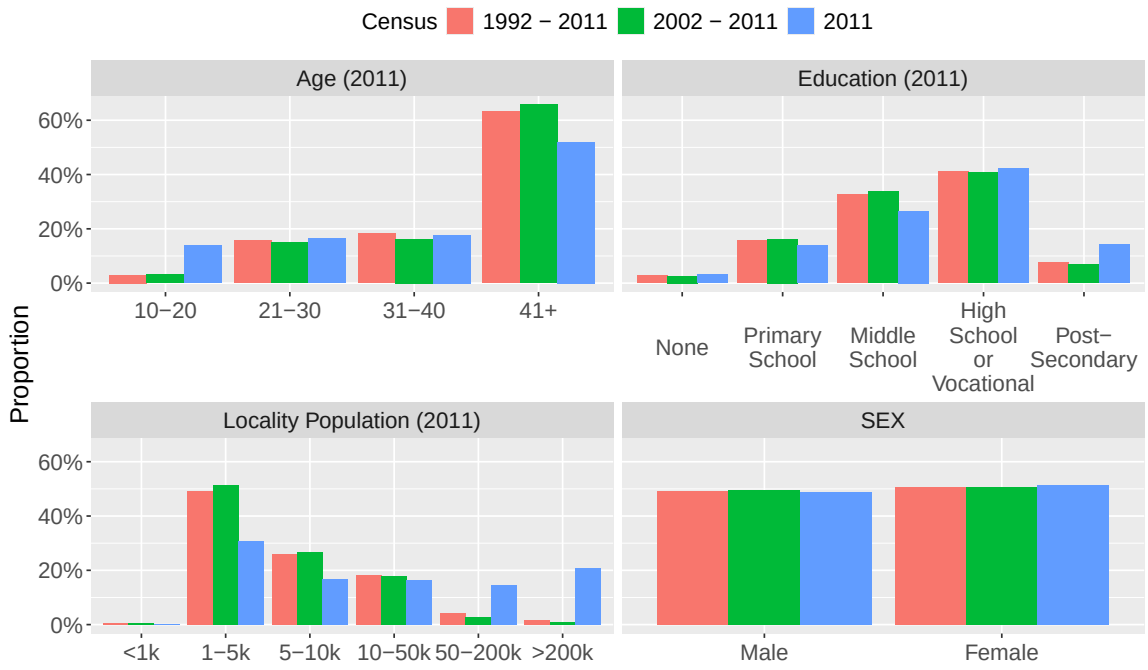
Notes: This table shows the relationship between declaring a child's Roma ethnicity and maternal completed years of schooling within sibling pairs. Columns 1–2 use the 1992–2011 linked census panel, restricted to individuals living with a parent in 1992 in households containing at least two such panel members; the outcome is self-reported Roma ethnicity in 2011. Column 2 instruments 2011 schooling with 1992 schooling. Both include commune × baseline birth-year and 1992 household fixed effects, clustered at the commune level. Columns 3–4 use the VSN (2003–2011) second- and later-born children of mothers whose first child was declared Roma on the birth certificate; the outcome is the non-first child's birth-certificate Roma indicator. Column 3 uses maternal schooling imputed from the VSN school code; column 4 restricts to mothers linked to the 1992–2011 panel and Roma in 1992, instrumenting their 2011 schooling with their 1992 schooling. Both include child birth-year fixed effects and cluster at the mother level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Figure A.7: Sample Representativeness: Self-Reported Roma Individuals



Notes: This figure compares the linked samples of baseline Roma individuals to the entire self-declared Roma population of the 2011 census born in 1991 and before on four dimensions: age (in 2011), educational attainment (in 2011), sex and population of town of residence (in 2011). The sample is restricted to self-declared Roma in 2011.

Figure A.8: Sample Representativeness: All Individuals



Notes: This figure compares the linked samples to the entire population of the 2011 census born in 1991 and before on four dimensions: age (in 2011), educational attainment (in 2011), sex and population of town of residence (in 2011).

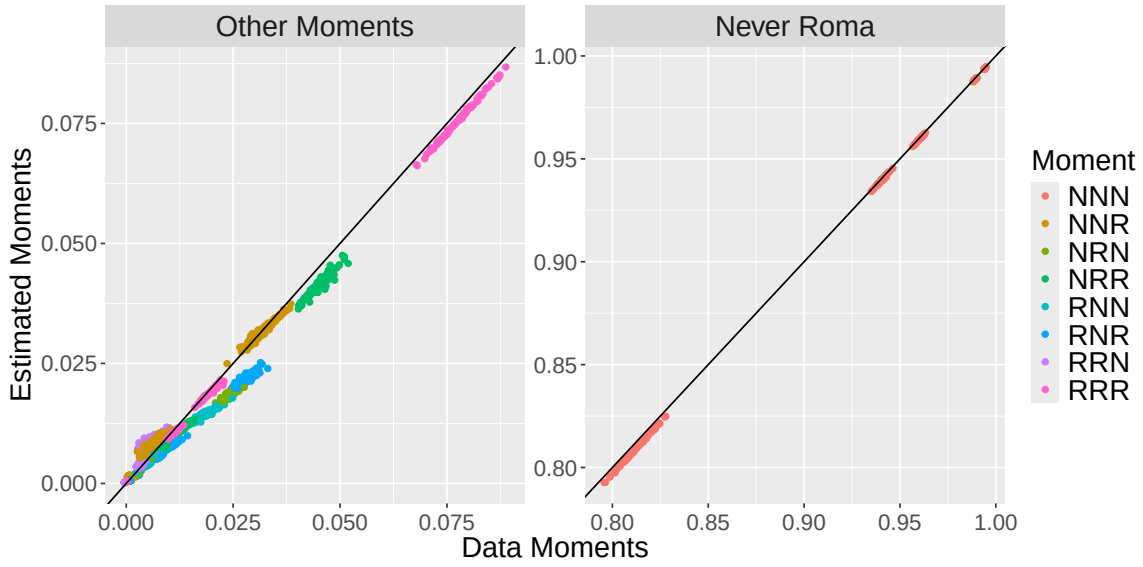
Table A.11: Intermarriage and Ethnic Identification (2002–2011)

	Panel 1: Own Ethnic Identification					
	Dependent Variable: Self Reported Roma					
	Women			Men		
	Roma Spouse	Non-Roma Spouse	No Spouse	Roma Spouse	Non-Roma Spouse	No Spouse
	(1)	(2)	(3)	(4)	(5)	(6)
Schooling Yrs	-0.028***	-0.029	-0.031***	-0.016***	0.006	-0.026***
	(0.003)	(0.022)	(0.004)	(0.002)	(0.016)	(0.005)
N	31,281	4,358	50,280	31,424	4,198	51,021
R^2	0.75	0.92	0.66	0.75	0.93	0.61
DV Mean	0.76	0.46	0.60	0.76	0.57	0.63
F-stat	6,892.7	453.7	3,512.9	9,263.6	900.6	3,312.6

	Panel 2: Inter-generational Ethnic Identification					
	Dependent Variable: Child Reported Roma					
	Birth Certificate (VSN)			Census 2011		
	Roma Spouse	Non-Roma Spouse	No Spouse	Roma Spouse	Non-Roma Spouse	No Spouse
	(1)	(2)	(3)	(4)	(5)	(6)
Schooling Yrs	-0.019***	-0.040**	-0.022*	-0.018***	-0.016	-0.029*
	(0.006)	(0.017)	(0.013)	(0.007)	(0.028)	(0.017)
N	14,901	3,243	4,861	14,901	3,243	4,861
R^2	0.74	0.77	0.84	0.76	0.90	0.89
DV Mean	0.33	0.19	0.33	0.79	0.42	0.50
F-stat	2,140.2	513.0	325.6	2,140.2	513.0	325.6

Notes: All columns are IV regressions instrumenting 2011 schooling with 2002 schooling. Panel A uses the sample of Roma individuals linked across the 2002 and 2011 censuses; the dependent variable is own Roma self-declaration in 2011. Panel B uses the sample of VSN children linked to their Roma mothers; columns (1)–(3) use children's ethnicity reported in the VSN as outcomes, while columns (4)–(6) use children's reported ethnicity in the 2011 census as outcomes. Roma spouse is defined as a spouse reported as Roma at baseline (2002). The sample includes only unmarried individuals at endline (2011) or individuals who are married and whose spouses are linked to baseline records, allowing us to observe their ethnicity at baseline. Controls include fixed effects for birth year interacted with locality. Standard errors are clustered at the locality level. *p<0.1; **p<0.05; ***p<0.01.

Figure A.9: Structural Model Fit



Notes: This figure plots the moments used to estimate the model from section 4, in the data (X axis) and the estimated model (Y axis), across education groups and bootstrap samples. For example, moment NRN represents the share of observations in the triple linked sample with reported ethnicity $(Roma_{i,1992}, Roma_{i,2002}, Roma_{i,2011}) = (0, 1, 0)$.

Table A.12: Model: Fraction of Roma-heritage Individuals with a Null Probability of Roma Identification

		Panel 1: Normal		
		$p = 0$ Roma self-report		
Education	σ	1992	2002	2011
None	0.38	0.11	0.07	0.04
Primary	0.24	0.02	0.01	0.00
Middle School	0.20	0.01	0.00	0.00
High School	0.19	0.00	0.01	0.01
Postsecondary	0.21	0.17	0.17	0.17
		Panel 2: Uniform		
		$p = 0$ Roma self-report		
Education	σ	1992	2002	2011
None	0.97	0.17	0.07	0.00
Primary	0.38	0.40	0.16	0.00
Middle School	0.24	0.61	0.24	0.00
High School	0.21	0.70	0.28	0.00
Postsecondary	0.21	0.68	0.27	0.00
		Panel 3: Lognormal		
		$p = 0$ Roma self-report		
Education	σ	1992	2002	2011
None	0.74	0.00	0.00	0.00
Primary	0.45	0.00	0.00	0.00
Middle School	0.37	0.00	0.00	0.00
High School	0.50	0.00	0.00	0.00
Postsecondary	0.52	0.00	0.00	0.00

Note: This table shows how our structural model captures Roma-heritage individuals who have a null probability of ever identifying as Roma. The table contains three panels, one for each version of the model. σ corresponds to the variance of the individual-specific shock that impacts self-reporting for different levels of endline schooling. The last three columns report the estimated fractions of individuals who have a null probability of ever self-identifying as Roma.

Table A.13: Respondent Views Internally Consistent but Divergent

Are Roma Who Change Their Ethnic Identification More Educated?	Do Educated Roma Change Their Ethnic Identification?		
	No	Yes	Total
No	170	63	233
Yes	54	204	258
Total	224	267	491

Note: This table cross-tabulates responses to two questions about the relationship between changes in ethnic identification and education. Only respondents in the "explicit" group ($N = 1,000$) received these questions. We drop observations where the two groups are rated the same (equally likely to change their ethnic identification, same education), and when the respondent is not sure.

Table A.14: Perceived Markers of Roma Ethnicity

Characteristic	Proportion
Way of Speaking	0.677
Physical Aspect	0.548
Skin Color	0.509
Name	0.288
Clothing	0.267
Family History	0.255

Notes: Proportions represent the share of respondents citing each trait as indicative of Roma ethnicity, among those who reported it is very easy or relatively easy to identify Roma who change their ethnic identification. $N = 1,255$.

B Model details

B.1 Signing the Bias from Conditioning on Baseline Reported Ethnicity

In this section, we show formally that our benchmark regression analysis based on a sample of individuals with Roma reported ethnicity at baseline is conservative, in the sense that the estimated coefficient on education is biased towards zero.

Consider a population of individuals with Roma heritage. Assume a linear probability model for the probability that individual i reports their ethnicity as Roma at time t :

$$\Pr(Roma_{it} = 1) = \alpha + \beta e_i + \mu_i + \epsilon_{it}. \tag{2}$$

Here, α is a constant, β measures how Roma identification varies with education $e_i \in \{0, 1\}$. (Our hypothesis is $\beta < 0$.) The fixed effect μ_i allows some individuals to be more likely to declare themselves as Roma in all time periods. We assume that μ_i is mean-zero and distributed according to probability density function f. Unlike in our main model from section 4, here we assume that f is the same for all levels of education. ϵ_{it} is an iid, mean-zero, random shock.

We are interested in estimating β. First, assume we had access to a representative sample of Roma individuals i. We could estimate β as (the expectation is over ϵ_{it}):

$$\widehat{\beta}_{FULL} = \mathbb{E}\Pr(Roma_{it} = 1 \mid e_i = 1) - \mathbb{E}\Pr(Roma_{it} = 1 \mid e_i = 0).$$

In reality, we only observe the reported ethnicity $Roma_{it}$ in different time periods. Our regression approach is to condition on a sample of individuals reported as Roma in a previous period. We compute:

$$\begin{aligned}\widehat{\beta}_{COND} = &\Pr(Roma_{it} = 1 \mid Roma_{i(t-1)} = 1 \;\&\; e_i = 1) - \\ &\Pr(Roma_{it} = 1 \mid Roma_{i(t-1)} = 1 \;\&\; e_i = 0).\end{aligned}$$

We are interested in whether β_{COND} is biased relative to β. First, when there is no persistent heterogeneity, that is, when $\mu_i = 0$ for all i, there is no bias because conditioning does not affect probabilities. Within each education group, we condition on ϵ_{it} and the realization of the probability, both of which are iid across time, so we recover the unconditional Roma probability.

However, conditioning on $Roma_{i(t-1)} = 1$ will typically introduce a bias when the distribution of μ_i is non-degenerate. This is because individuals with higher μ_i are more likely to satisfy $Roma_{i(t-1)} = 1$ and hence enter the conditional sample. The

key question is whether this effect "cancels out" between the high- and low-education groups. We now show that, in general, it does not. In the linear probability model, we can write the conditional probability as

$$\mathbb{E}\Pr(Roma_{it} = 1 \mid Roma_{i(t-1)} = 1, e_i = 0) = \alpha + \int \mu \; f(\mu \mid Roma_{i(t-1)} = 1, e_i = 0) d\mu$$

By Bayes' rule, the conditional PDF f is given by

$$\begin{aligned} f(\mu \mid Roma_{i(t-1)} = 1, e_i = 0) &= \frac{\Pr(Roma_{i(t-1)} = 1 \mid \mu, e_i = 0) f(\mu)}{\int \Pr(Roma_{i(t-1)} = 1 \mid \eta, e_i = 0) f(\eta) d\eta} \\ &= \frac{\alpha + \mu}{\alpha + \overline{\mu}} f(\mu) \\ &= \Big(1 + \frac{\mu}{\alpha}\Big) f(\mu). \end{aligned}$$

This expression is intuitive: the sample with $Roma_{i(t-1)} = 1$ is skewed towards individuals with large μ. Plugging back into the conditional probability, we get:

$$\mathbb{E}\Pr(Roma_{it} = 1 \mid Roma_{i(t-1)} = 1 \;\&\; e_i = 0) = \alpha + \int \mu \; \Big(1 + \frac{\mu}{\alpha}\Big) f(\mu) d\mu$$

Similarly, for the high education group we get:

$$\mathbb{E}\Pr(Roma_{it} = 1 \mid Roma_{i(t-1)} = 1 \;\&\; e_i = 1) = \alpha + \beta + \int \mu \; \Big(1 + \frac{\mu}{\alpha + \beta}\Big) f(\mu) d\mu$$

Putting these together we get the bias:

$$\begin{aligned} \widehat{\beta}_{COND} &= \beta + \int \mu \; \Big(1 + \frac{\mu}{\alpha + \beta}\Big) f(\mu) d\mu - \int \mu \; \Big(1 + \frac{\mu}{\alpha}\Big) f(\mu) d\mu \\ &= \beta \left(1 - \int \mu^2 \frac{1}{\alpha(\alpha + \beta)} f(\mu) d\mu\right) \\ &= \beta \left(1 - Var(\mu) \frac{1}{\alpha(\alpha + \beta)}\right) \end{aligned}$$

This is a bias towards zero, assuming that $Var(\mu) < \alpha(\alpha + \beta)$, i.e. the variance of μ is not too large. (This is a weak condition given that the term in equation (2) must be between 0 and 1.)

In our case we estimate $\widehat{\beta}_{COND} < 0$, hence the true β is even larger in absolute value (i.e. more negative). This means that through the lens of this model with heterogeneity, the regression analysis is conservative in terms of the strength of the association between education and changes in ethnic identification.

B.2 Model Setup

The model consists of two sub-populations: Roma-heritage (comprising a proportion p_R) and non-Roma heritage (with a proportion $1-p_R$). We focus on three census period indexed by $t = 0, 1, 2$. Individuals i who are Roma, with endline educational level e declare their ethnicity $Roma_{iet} = 1$ with probability π_{iet}. With remaining probability $1 - \pi_{iet}$ they declare non-Roma (0) ethnicity.

The self-declaration probability π_{iet} can be decomposed into two parts, such that $\pi_{iet} = \pi_{te} + \pi_{ie}$. The first component ($\pi_{te}$) is a time-varying (or census-varying) component capturing the mean propensity to self-declare as Roma for the group with education level e. The second component (π_{ie}) is an individual-specific idiosyncratic but time-constant "taste" for identifying as Roma for individual i. This parameter π_{ie} is drawn from a distribution. We estimate the model for a normal distribution (mean zero with estimated standard error), a uniform distribution (mean zero with estimated range), and log-normal distribution with normalized mean and estimated standard deviation.

Additionally, we allow mismatch in our model. In other words, any self-reported ethnicity history H we observe $H_i = \{Roma_{i0}, Roma_{i1}, Roma_{i2}\}$ could be submitted by one individual whose census records we correctly link across the three census periods, or could be the result of mismatched links between the censuses. We focus our analysis on the declared ethnicity on the last wave of the census. Therefore, within each educational category e, we define a $m_{He}(0)$ probability of mismatch ($M_0 = 1$) between census waves at $t = 0$ and $t = 2$ and a m_{He1} probability of mismatch ($M_1 = 1$) between census waves at $t = 1$ and $t = 2$ for individuals with education e for each different type of ethnic reporting history H. For example, for records with postsecondary schooling as endline attainment, and self-reporting histories $H = \{1, 0, 1\}$, we estimate m_{He0} and m_{He1}.

These mismatch rates are estimated by perfectly matching the rate of inconsistencies in sex reported across linked census records with declared ethnicities H across respondents with endline education e. For example, a 2% rate of inconsistent sexes implies a mismatch rate of 4%. In case of mismatch, we assume that a census record is linked to a random census record in another census wave within the same education category. Because the Roma represent only a small minority of individuals, this means that there is a very high likelihood for census records to be matched to a non-Roma record (with probability $1 - p_R$).

We now have all the elements to express the probability of observing any ethnic self-reporting history in the census. There are four possible mismatch scenarios:

1. the most straightforward scenario where there is no mismatch, with probability $p_{M_0=0,M_1=0} = (1 - m_{He0}) \times (1 - m_{He1})$;
2. mismatch between waves 0 and 2 only, with probability $p_{M_0=1,M_1=0} = m_{He0} \times (1 - m_{He1})$;
3. mismatch between waves 1 and 2 only, with probability $p_{M_0=0,M_1=1} = (1 - m_{He0}) \times m_{He1}$;
4. mismatch between both waves 0 and 2 and 1 and 2, with probability $p_{M_0=1,M_1=1} = m_{He0}m_{He1}$.

Additionally, we define $N_{possible}(H, M_0, M_1)$, a function which indicates which reporting histories H can be associated to a non-Roma-endline individual i, given mismatch scenarios M_0 and M_1. For example, when there is no mismatch between the three records, self-declaring as Roma in *any* census guarantees that the record belongs to a Roma individual. Only when an individual self-declares as non Roma in *all* of the censuses can the person possibly be a non-Roma. However, with mismatch, non-Roma at time $t = 2$ could be associated with any self-declaring history where $H_2 = 1$, as long as the record is matched to an appropriate record j at time $t = 0, 1$. The variable $N_{possible}$ allows us to identify, for each mismatch scenario, which self-reporting histories can be associated to non-Roma individuals. We can express $N_{possible}(H, M_0, M_1)$ as:

$$N_{possible}(H, M_0, M_1) = (1 - Roma_2) \times (1 - Roma_1)^{1-M_1} \times (1 - Roma_0)^{1-M_0}$$

We can now write out the proportion of each observed ethnic declaration history H for education group e as:

$$\begin{aligned} P_H^e = \sum_{M_{0,1} \in \{0,1\}} \Big(& p_{M_0,M_1} p_R^e \prod_{t \in \{0,1,2\}} \left[(\pi_{iet})^{Roma_t} (1 - \pi_{iet})^{1-Roma_t} \right]^{1-M_t} \\ & \prod_{t \in \{0,1\}} \left[(p_R^e \pi_{jet})^{Roma_t} (p_R^e (1 - \pi_{jet}) + (1 - p_R^e))^{1-Roma_t} \right]^{M_t} + \\ & N_{possible}(H)(1 - p_R^e) \prod_{t \in \{0,1\}} \left[(\pi_{jet})^{Roma_t} (1 - \pi_{jet})^{1-Roma_t} \right]^{M_t} \Big) \end{aligned}$$

where M_2 is always equal to 0.[‡] The first term gives the probability of an observed history of declared ethnicities for records of Roma individuals i, who are in proportion p_R^e in the population and who self-declare ethnicity H_t following

[‡]Since a record at time $t = 2$ cannot be mismatched to itself.

$(\pi_{iet})^{Roma_t}(1-\pi_{iet})^{1-Roma_t}$.

The second term indicates the probability of observing census responses in case of Roma individuals at $t=1$ who are incorrectly matched to another census record. In this case, they are mismatched with probability p_R^e to a Roma individual j, who may declare their ethnicity at rate π_{jet} or change their ethnic identification at rate $1-\pi_{jet}$. Or, there could be incorrect mismatch to a non-Roma individual, with probability $1-p_R^e$. Note that both these scenarios are possible when $Roma_t = 0$, whereas $Roma_t = 1$ can only occur if the record at time t belongs to a Roma individual.

Lastly, the third term shows us the probability of a declaration history stemming from a non-Roma record at time $t=2$. In this case, which occurs with probability $1-p_R^e$, is non-zero only when $N_{possible}(H)=1$.

As an illustration, when $H=\{1,0,0\}$, $M_0=1$ and $M_1=0$, the expression simplifies to:

$$P_H^e = p_{1,0}\ [p_R^e(p_R^e\pi_{je0})(1-\pi_{ie1})(1-\pi_{ie2}) + (1-p_R^e)(p_R^e\pi_{je0})]$$

where $p_{1,0}$ is the joint probability of $M_0=1, M_1=0$, the first term in the summation is associated to an endline-Roma individual declaring non-Roma status at rate $1-\pi_{ie2}$, being mismatched to a Roma individual j at $t=0$ at rate p_R^e, who declares their Roma ethnicity at rate π_{je0}. They are also correctly matched to their own record at time $t=1$, where they changed their ethnic identification with rate $1-\pi_{ie1}$. The second term in the summation corresponds to the individual at time $t=2$ being non-Roma (with probability $1-p_R^e$), and being mismatched to a self-declared Roma at time $t=0$, who self-declares at rate π_{je0}.

C Survey Questions

This section reports the survey questions used to ascertain beliefs about the levels of schooling of Roma who changed their ethnic identification and those who did not.

The main, incentivized, question on beliefs about the educational attainments of Roma who change their ethnic identification is included below.

Incentivized question, English.

> Attention! If your answer to the following question is correct, you will win a bonus of 100 points, in addition to the base compensation for this questionnaire.
>
> **Information:** The 2011 census shows that 50.3% of adult individ-

uals who *identified as Roma* had completed at least the 8th grade.

Bonus Question: What percentage of all adult Roma individuals (regardless of how they reported their ethnicity) had completed at least the 8th grade in 2011?

Note: The correct answer is calculated based on the education of individuals who had declared themselves as Roma in the 1992 census, regardless of how they declared their ethnicity in 2011.

Options:

a. Less than 50.3%

b. Exactly 50.3%

c. More than 50.3%"

Incentivized question, Romanian.

Atentie! Daca raspunsul dumneavoastra la intrebarea urmatoare este corect, veti castiga un bonus de 100 puncte, in plus fata de compensatia de baza pentru acest chestionar.

Informatie: Recensamantul din 2011 arata ca 50.3% dintre persoanele adulte care s-au declarat de etnie roma au absolvit cel putin clasa a 8-a.

Intrebare (cu bonus): Ce procent dintre toate persoanele adulte de etnie roma (indiferent cum și-au declarat etnia) au absolvit cel putin clasa a 8-a in 2011?

(**Precizare:** raspunsul corect este calculat pe baza educatiei persoanelor care s-au declarat de etnie roma anterior, la recensamantul din 1992, indiferent de cum si-au declarat etnia in 2011.)

Optiuni:

a. Mai putin de 50.3%

b. Exact 50.3%

c. Mai mult de 50.3%

The question used to ascertain the attentiveness of respondents is included below.[§]

Attention question, English.

Verification Question: What were the questions on the previous page referring to?

a. Roma individuals who *identified as Roma* in the 2011 census

b. Roma individuals who *did not identify as Roma* in the 2011 census

[§]The interface did not allow users to scroll back to the previous question.

c. Roma individuals, *regardless of how they declared their ethnicity* in the 2011 census

d. I am not sure."

Attention question, Romanian.

Intrebare de verificare. La ce s-au referit intrebarile de pe pagina anterioara?

a. Persoane rome care s-au declarat de etnie roma la recensamantul din 2011

b. Persoane rome care NU s-au declarat de etnie roma la recensamantul din 2011

c. Persoane rome, indiferent de etnia declarata la recensamantul din 2011

d. Nu sunt sigur