

**DOCUMENT DE TRAVAIL / WORKING PAPER**

**No. 2026-06**

**Education and Selection into Ethnic Identification:  
Evidence from Roma People in Romania**

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

**Mai 2026**

---

# Education and Selection into Ethnic Identification: Evidence from Roma People in Romania \*

Andreea Mitrut<sup>†</sup> (r) Gabriel Kreindler<sup>‡</sup> (r) Margareta Matache<sup>§</sup>  
(r) Andrei Munteanu<sup>¶</sup> (r) Cristian Pop-Eleches<sup>||</sup>

October 13, 2025

## Abstract

How does ethnic identification vary with education among disadvantaged minorities? We study this question for Roma people, Europe’s largest ethnic minority, using linked Romanian census data and birth records. We measure how individuals change reported ethnicity over time, or “pass.” Roma identification strongly declines with education, from 80% for those with no education to 40% for postsecondary graduates. We estimate a model with persistent individual heterogeneity and find 3-6 times more Roma postsecondary graduates than in official data. Survey data we collect shows that most Romanians are unaware of these patterns. Such selective passing may reinforce stereotypes about marginalized groups.

---

\*We thank Lukas Althoff, Ben Enke, Rema Hanna, Amanda Pallais, and David Yang for useful conversations, as well as participants at the Junior Applied Micro (JAM) seminar and the ERMAS 2024 conference. The survey in this project is covered by IRB protocol number AAAU5373 at Columbia University. We thank Ioana Veghes for help implementing the survey. The author order was certified randomized.

<sup>†</sup>University of Gothenburg.

<sup>‡</sup>Harvard University.

<sup>§</sup>Harvard University.

<sup>¶</sup>Université du Québec à Montréal.

<sup>||</sup>Columbia University.

## 1 Introduction

Ethnic and racial identity shape economic interactions, influencing labor supply, how individuals are treated by others, and workplace collaboration (Akerlof and Kranton, 2000, Alesina and La Ferrara, 2005, Bertrand and Mullainathan, 2004, Hjort, 2014, Oh, 2023, Pager et al., 2009). While economic analyses typically treat ethnicity and race as fixed, recent work shows that individuals may change their reported identity, or “pass” (Adukia et al., 2025, Cornwell et al., 2017, Dahis et al., 2020, Duncan and Trejo, 2011a,b, 2023, Noghanibehambari and Fletcher, 2025, Rademakers and van Hoorn, 2021).

In this paper, we study whether passing is *selective*. We focus on education, examining whether individuals who self-identify with a disadvantaged ethnic group differ in educational attainment compared to all individuals with ties to that ethnic group.<sup>1</sup> Selective passing is important because it can have aggregate implications. For example, if more educated individuals are more likely to pass, this may reinforce negative beliefs about the group, and may complicate the design and evaluation of policies aimed at members of the group. This argument relies on observers who do not account for selection, so we collect novel data to measure what general observers believe about selective passing.

It is not clear whether members of a disadvantaged group would pass selectively, and if so, in which direction. More educated individuals may be more likely to pass, for example, if ethnic discrimination is stronger in their occupations, or less likely to pass, for example due to affirmative action policies in education, or if they stand to lose more valuable ethnic social networks.

We study how passing varies with education in the context of Roma people in Romania. Roma are Europe’s largest ethnic minority, estimated at 10-12 million people (European Union Agency for Fundamental Rights, 2023). Despite de jure equality and certain affirmative action programs in education, de facto discrimination persists, rooted in a history of institutionalized enslavement, persecution, and forced assimilation. As in other European countries, Roma in Romania have lower educational attainment compared to the non-Roma population, according to official statistics.

For our analysis, we compile large-scale Romanian administrative datasets, and complement these with survey data we collected. To test for the presence and selective nature of passing, we link individual records from the 1992, 2002 and 2011 full-count population censuses using exact date of birth, locality, and sex, focusing on unique, two-sided matches (Dahis et al., 2020). We track ethnicity reported by household respondents over time and define “passing” as a switch from Roma to non-Roma. We also measure parental decisions over intergenerational passing by leveraging a perfect linkage using anonymized national ID numbers between census and birth records, which capture the ethnicity parents report for their children *at birth*.<sup>2</sup> Finally, to

---

<sup>1</sup>We use this definition to recognize individuals’ agency in choosing their identity, while also tracing ties to the group independent of these choices.

<sup>2</sup>Birth records from the Vital Statistics Natality files offer a different context for measuring passing: they are typically completed by the mother, whereas the census is completed by an enumerator (in the presence of the household head and other members). While in both cases data is collected

measure the perceptions and awareness of passing, we collected data from the general population in Romania using an original online survey. Together, these datasets allow us to evaluate both the selectivity of passing and the extent to which observers are aware—or unaware—of this process.

Our key empirical result is that, within the Roma ethnic minority, more educated individuals are significantly less likely to identify as Roma in official registers. We focus on documenting this observational relationship, without making causal claims about the effect of education.

In the first part of our analysis, we use a reduced-form analysis that relates individuals’ educational attainment to changes in their reported ethnicity between census rounds, controlling for baseline locality-by-birth-year effects.

Two data issues could bias our results. First, if a Roma individual is incorrectly linked to a non-Roma (mismatched census records), this may lead to a spurious positive correlation between passing and education, since non-Roma are more educated on average. Second, random measurement error in reported education attenuates the education gradient toward zero.

To address these concerns, we develop a new method to estimate mismatch rates. We link individuals using only date of birth and locality and use changes in recorded sex as a proxy for mismatched records. We rely on the fact that when two records are mismatched, around half of the time the sex will be different in the two census rounds. Mismatch rates are low, around 3-10% for key subgroups. We then use these estimated mismatch rates to back out “true” passing rates by education.<sup>3</sup>

We find that passing increases sharply in education. Among individuals initially reported as Roma, around 80% of those without any formal education are reported as Roma again in 2011. This number falls to 43% for high school or vocational school graduates. Passing is also increasing in occupational skill, although the differences are smaller. Passing is very low, with a negligible education gradient for ethnic Hungarians, the other large ethnic minority in Romania, whose members face less stigma than Roma and have strong political representation.

We use a second strategy to correct for potential mismatch and measurement error, employing baseline education as an instrument for endline education.<sup>4</sup> Our IV results are similar to our previous results. Among individuals reported as Roma at baseline, each additional year of schooling is associated with a 2.6 percentage points (pp) lower likelihood of being reported as Roma at endline, similar to the OLS coefficient of 3.0 pp. Results using a perfectly matched sample of mothers between birth records and the 2011 census yields a  $-2.3$  pp gradient, confirming that the negative education gradient is robust to linkage errors.

We find even stronger *inter-generational* educational gradients. Each additional year of maternal schooling reduces the likelihood of reporting the child as Roma by 5.3 pp in birth records and by 3.5 pp in the census. We also find that the educa-

---

for statistical purposes, birth records may be perceived as a more formal setting.

<sup>3</sup>The method we develop here can be used in other contexts where researchers match individuals across multiple rounds of cross-sectional data.

<sup>4</sup>Mismatch proportionally attenuates both the first-stage and reduced-form, so the Wald estimator identifies the correct correlation between endline education and passing.

tion-passing gradient is about 50% higher for women, larger among individuals with highly educated mothers, and smaller for those with strong language or marriage ties to the Roma ethnicity. Passing occurs at similar rates throughout the life cycle, rather than being concentrated at a single stage such as early adulthood.

In the second part of our analysis, we develop a structural model to estimate the total number of Roma-heritage individuals, accounting for persistent individual differences in the propensity to report Roma ethnicity. Because our regression sample includes only those reported as Roma in the baseline census, it over-represents individuals more likely to identify as Roma every time. This underestimation of the overall passing rate may bias the education gradient. We estimate the model using the three rounds of ethnicity reports.

Our estimated model implies that the number of Roma in Romania is significantly under-reported, with 72% more Roma-heritage individuals than officially recorded Roma in 2011. Under-reporting is highest among the educated, where we find 2.7-6 times more postsecondary graduates and over twice as many high school graduates. Including all Roma-heritage individuals reduces the education gap with non-Roma by 13%.

In the third and final step of our paper, we test whether the general population is aware of passing patterns among Roma, using an online survey of 2,000 Romanians. The average respondent in our sample does not perceive passing to be more common among the educated. This result is robust to incentives, clarifying questions about passing, attention checks, and focusing on respondents from localities with many Roma. Beliefs about the education-passing link are highly dispersed, yet internally consistent and often supported by narrative “theories,” making it unlikely that our findings are driven by random noise. Most respondents believe they can identify Roma even when passing, citing markers such as speech, appearance, skin color, and clothing—traits that are malleable or rooted in homogeneous stereotypes. Combined with ethnographic evidence on successful passing (Pantea, 2014), these results suggest that passing may be a viable strategy for some Roma individuals.

We build on a growing literature on identity choice and passing. In the U.S. between 1880 and 1940, African American men sometimes passed as White in response to discrimination (Dahis et al., 2020), while passing is linked to longer life expectancy among Black Americans (Noghanibehambari and Fletcher, 2025). In the context of immigrant assimilation, among Mexican Americans, identification weakens across generations and through intermarriage with non-Mexicans, and this is correlated with education (Duncan and Trejo, 2011a,b). In work close to ours, Duncan and Trejo (2023) show that enumerators in the 1930 U.S. Census were more likely to classify more educated Mexicans as non-Mexicans, and Adukia et al. (2025) show that U.S. Census enumerators in 1870 were more likely to record respondents with the same skin tone as non-Black if they were wealthier or literate. Our contribution is to show that selective passing reported by households themselves can alter the perceived educational composition of a group in a minority group in a modern, non-immigrant context. Evidence from other contexts show that ethnic switching occurs in Indonesia, India and the U.S., often following inter-ethnic marriage (Rademakers and van Hoorn, 2021), and that wages vary systematically with employer-reported racial classification

of workers in Brazil (Cornwell et al., 2017).

Our results link the recent research on identity choice with work on reinforcing statistical discrimination (Bursztyrn et al., 2017, Glover et al., 2017). We document a new stylized fact: among the Roma minority in Romania, passing is strongly increasing in education. By altering the perceived educational composition of the group, this phenomenon may reinforce low-education stereotypes. The general population appears largely unaware of these patterns, consistent with evidence on systematic misperceptions and selection neglect (Bursztyrn and Yang, 2021, Enke, 2020). Selective passing may also affect Roma educational attainment, as exposure to same-group teachers or role models matters for educational outcomes (Fairlie et al., 2014, Porter and Serra, 2020).

## 2 Setting and Data

### 2.1 Roma in Romania

Roma are Europe’s largest minority (European Commission, 2020), and one of the largest ethnic groups in Romania. The 2011 census recorded over 600,000 Roma or 3.1% of the population, up from 1.8% in 1992 and 2.5% in 2002.<sup>5</sup> However, estimates suggest substantial under-reporting of Roma ethnicity (Zamfir and Preda, 2002, Zamfir and Zamfir, 1993).<sup>6</sup> Roma communities experience high poverty and poor health outcomes. In the 2011 census, Roma adults had 5.8 years of schooling on average (20% had none), compared to 10.6 for non-Roma.

These disparities are rooted in centuries of systemic oppression, including slavery until the mid-19th century, persecution during the Holocaust — with 25,000 deported and 11,000 killed in Romania alone (Kelso, 2017, Wiesel, 2004) — and forced assimilation under the communist regime. Today, Roma in Romania and in other European countries continue to face widespread discrimination, including segregation and limited access to health and education (European Union Agency for Fundamental Rights, 2023).

### 2.2 Roma and Passing

Like other ethnic and racial groups facing discrimination worldwide, Roma may engage in passing, employing complex strategies. In Romania, this phenomenon has been documented in several ethnographic studies (Marin, 2023, Pantea, 2014). Key motivations include escaping stigma and negative stereotypes, such as being perceived as thieves or beggars,<sup>7</sup> accessing better opportunities (e.g., employment), or fearing persecution. These ethnographic studies suggest Roma passing often involves adopt-

---

<sup>5</sup>This increase reflects both higher Roma fertility and efforts by civil society to encourage Roma self-identification in official data.

<sup>6</sup>Zamfir and Preda (2002) estimates that in 1998 there were 2-3 times more Roma than in census statistics by extrapolating *hetero-identified* ethnicity based on data from the catchment areas of 200 randomly selected polling stations.

<sup>7</sup>The pejorative exonym “t\*gan” (commonly used for Roma) has a harsh legacy, having once been synonymous with “slave,” and is still often used to offend.

ing non-traditional clothing styles, minimizing the use of Romani language in favor of Romanian, and distancing from traditional markers associated with Roma identity. Passing among Roma is context-dependent and fluid: for some, it is situational, limited to official settings, while cultural identification remains strong (Marin, 2023); others, including children, may be taught to conceal their identity in social interactions, through changes in appearance and behavior. Marin (2023) links some passing behaviors to historical trauma, notably deportations of Romanian Roma during the Holocaust, which relied in part on census records.<sup>8</sup> These patterns coexist with shifts toward self-identification and advocacy (Grigore, 2012). There is little research on how the general population perceives Roma individuals who intend to pass in social and professional interactions. Our survey provides some new evidence on this question.

### 2.3 Administrative Data Sources

Our main data sources are the 1992, 2002, and 2011 Romanian full-count censuses, which cover the entire resident population. In addition to ethnicity and education, the censuses record religion, language, and other socioeconomic characteristics for all household members, as well as exact date of birth, sex, locality of birth, and current and previous localities of residence—allowing us to uniquely link individuals across waves and track changes in reported ethnicity over time. We define an individual as passing if their reported ethnicity changes from Roma in an earlier census to non-Roma in a later census.<sup>9</sup>

We also use Vital Statistics Natality (VSN) files from 2003–2011 (excluding 2004), linked perfectly to the mother’s 2011 census record via personal identification numbers. These birth records offer a different context for the passing of studying: they are typically filled out by mothers without an enumerator present, so they provide an alternative setting to observe ethnic identification free from male reporting and enumerator-related social desirability bias. They may also be perceived as a more formal setting, with potential long-term consequences.

We describe the survey data that we collected in section 5.

### 2.4 Linking Census Records

To track changes in reported Roma ethnicity, we link census records within cells defined using the exact date of birth, locality, and sex. We restrict our sample to unique matches, that is, cells with exactly one individual in each census year, as in Abramitzky et al. (2021), Dahis et al. (2020) (See Appendix B.1.). We uniquely match 33% of the 2011 population to 1992, 39% to 2002, and 26% across all three waves. The sample excludes individuals from the same locality, same sex, and born on the same date. While such coincidences occur by chance, it means our sample puts

---

<sup>8</sup>In our empirical analysis, we control for birth year and town of residence fixed effects, partially to control for mistrust in government authorities.

<sup>9</sup>In 2011, 5.9% of respondents chose “no declared ethnicity” which we code as non-Roma; our results are robust to an alternative definition (Table A.9). In earlier censuses, reporting ethnicity was mandatory.

more weight on smaller localities—which is where most Roma live. We later report heterogeneity by locality population.

### 3 How Selection Into Ethnic Identification Varies by Education

#### 3.1 A Method to Estimate Mismatch Rates

We focus on individuals in our linked sample who were reported as Roma in the 1992 census, either by themselves or by another household member, and track whether they were still reported as Roma in 2011, 19 years later.

Since we examine passing by education, a challenge is that mismatched records can bias passing estimates toward selective passing by education. To address this, we develop a method to estimate and correct for mismatch. We link individuals using only date of birth and locality, then detect mismatches by identifying changes in reported sex across census waves.<sup>10</sup> Assuming sex is stable and accurately reported, an inconsistency indicates an incorrect match. Because half of mismatches will, by chance, have consistent sex, we estimate the mismatch rate as twice the inconsistent-sex rate.

Mismatch rates are low (Table A.1): the correct match rate for Roma is 95%, better than the 6–7% false-positive rate in recent historical U.S. census linkages (Price et al., 2021). Mismatch is somewhat higher for highly educated Roma (Table A.2). We adjust passing rates using these estimates, recognizing that observed passing is a weighted average of actual passing (true changes in reported ethnicity) and spurious passing from mismatches. Mismatch appears conditionally random: among mismatched records, baseline education is uncorrelated with endline education or reported ethnicity, conditional on locality-by-birth-year fixed effects (Table A.3).

Applying this adjustment, Figure 1 plots the share of individuals reported as Roma at baseline who continue to report Roma ethnicity by 2011 education and occupation. Higher education strongly predicts lower Roma identification—that is, higher passing. Only 40% of postsecondary graduates and 43% of high school graduates remain reported as Roma, compared with 80% of those with no formal education. Interestingly, passing also varies by occupation: reported Roma who are unemployed or out of the labor force, or in agriculture/unskilled work have continuation rates around 64%, versus 42–47% among managers, university-trained professionals, and clerks.

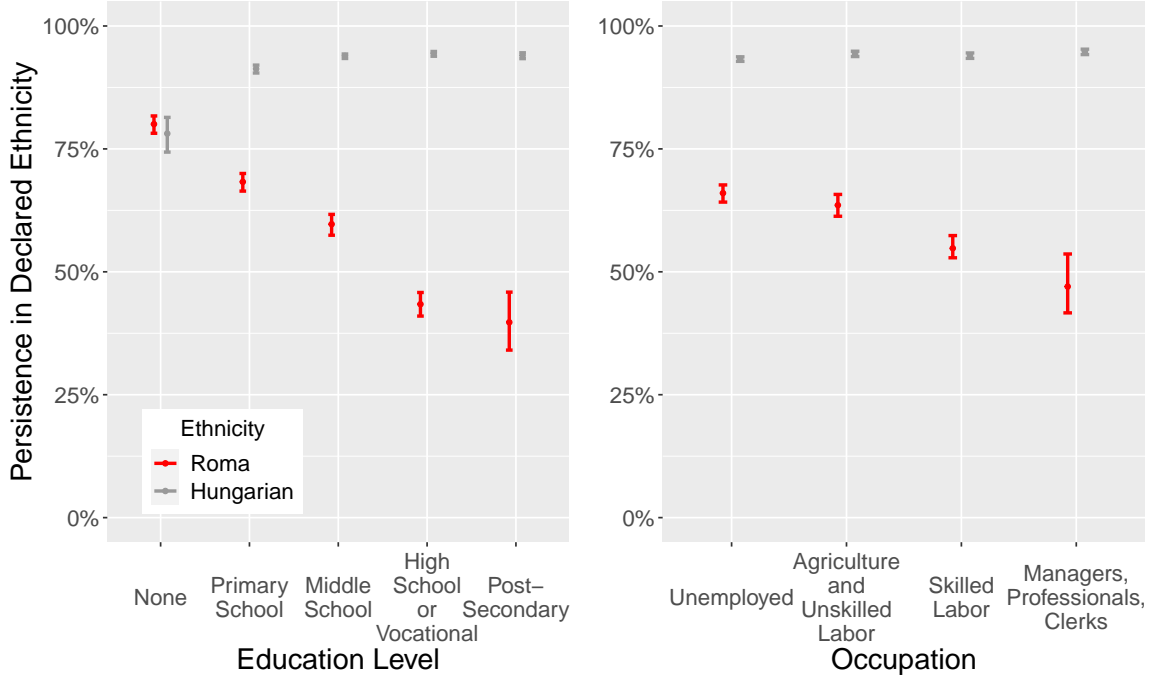
This stark negative education gradient is specific to Roma. There is little evidence of passing for Hungarians—Romania’s other large ethnic minority—who face less stigma and have strong political representation. Over 90% report the same ethnic identity in 2011, with a weakly positive education gradient. This means that our Roma results are not mechanical or a spurious artifact of census linking, and suggests that the relationship between education and ethnic identification varies across groups and contexts.<sup>11</sup>

---

<sup>10</sup>See Appendix B.2 for details. Matching on sex, date of birth, and locality—or only date of birth and locality—yields similar numbers of matches.

<sup>11</sup>Figure A.2 presents results for other minority groups, who exhibit moderate passing and negligible education gradient.

Figure 1: Passing by Education and Occupation, 1992 - 2011



Notes: The linked sample includes individuals reported as Roma at baseline (1992), who are above 25 years old at endline (2011). Education levels and occupations are measured at endline. We apply our mismatch correction method separately for each level of endline education (Appendix B.2). We link records using date of birth and locality and use sex changes between baseline and endline to estimate mismatch rates.

Results are similar for individuals who were household heads at baseline and endline (Figure A.1). In other words, passing is not driven by individuals whose ethnicity was reported by another household member (such as a parent) at baseline. Results are similar using 2002 as the baseline (Figures A.3 and A.4). Passing rates from 1992–2011 slightly exceed those from 2002–2011 across all education levels, with a marginally steeper education gradient, suggesting that more educated individuals increasingly pass over time. We return to life-cycle patterns below.

### 3.2 Education Gradient in Passing: Regression Analysis

We next conduct a regression-based analysis of the link between education and passing, employing an alternative strategy to address potential record mismatch. As before, we restrict the sample to individuals reported as Roma in the 1992 baseline census and estimate a linear model relating their reported ethnicity in 2011 to their education level:

$$Roma_{i,t=1} = \beta Educ_{i,t=1} + \phi_{L(i),YOB(i)} + \epsilon_{i,t=1}. \quad (1)$$

Here,  $Roma_{i,t=1}$  is an indicator for Roma reported ethnicity in the 2011 census. We measure years of schooling  $Educ_{i,t=1}$  in 2011 to capture education levels accurately for individuals still in school at baseline. The terms  $\phi_{L(i),YOB(i)}$  are baseline locality by year-of-birth fixed effects, which control for unobserved cohort-specific and local factors such as school quality, local stigma or identity norms, as well as variation in mismatch rates. We thus measure the relationship between education and ethnic identification net of any such factors.

We call the coefficient  $\beta$  the “education gradient.” A negative value means that more educated individuals are more likely reported as non-Roma rather than Roma.

Several issues may impact this analysis. First, erroneously linked census records may bias our results. If we randomly mismatch a Roma individual at baseline to a Romanian individual at endline, this will appear as passing in our data and, to the extent that Romanians are more educated on average than Roma, this will lead us to incorrectly infer that more educated Roma are more likely to pass for non-Roma. This will bias  $\beta$  towards more negative values. Secondly, we also worry that random measurement error in the education variable will attenuate our results.

To address these concerns, we use baseline education as an instrument for endline education. The intuition is that, if mismatch happens conditionally randomly within locality and year of birth, then among mismatched records, baseline education is uncorrelated with passing and with endline education. That is, both the first stage and reduced form are zero for mismatched records. Consequently, in data that includes a share  $P$  of correctly matched records and  $1 - P$  of mismatched records, the first stage and reduced form will be attenuated by the same factor  $P$ . Table A.3 shows that among mismatched records, baseline education is uncorrelated with endline education, and with passing.<sup>12</sup>

Table 1 reports results, starting with the OLS estimate of equation (1) of  $\beta = -0.030$  in column 1. Each additional year of schooling is associated with a 3.0 percentage points (pp) lower probability of reporting Roma ethnicity. At the endpoints of the education variable, a university graduate in 2011 is 48 pp less likely to report Roma ethnicity than someone with no education.

Using baseline education as an instrument (column 2) yields a strong first stage and a similar coefficient of  $-0.026$ , implying a 41.6 pp reduction when comparing university graduates to those with no schooling. The close agreement between OLS and IV confirms that census mismatches are not driving the strong negative relationship between education and passing.

Column (3) replicates the OLS specification using a perfectly matched sample between the VSN and the 2011 census. Because the linkage is based on unique national ID numbers, there is no risk of mismatching. The estimated coefficient of  $-0.023$  remains highly significant and close in magnitude to the main results, confirming that the strong negative education gradient is not an artifact of imperfect linkage.<sup>13</sup>

---

<sup>12</sup>Rather exceptionally, we can thus directly test the exclusion restriction. However, we cannot control for mismatch in our main specification because we cannot observe mismatched records where reported sex is the same by chance.

<sup>13</sup>We find comparable results using linked census records for a similar sample of mothers. Table

Table 1: Passing by Education: Instrumental Variables Approach

<i>Dependent Variable:</i>					
<i>Reported Roma Ethnicity (2011)</i>					
	OLS '92-'11	IV '92-'11	OLS VSN-'11	IV '92-'11	IV '92-'11
	(1)	(2)	(3)	(4)	(5)
Schooling Yrs	-0.030*** (0.001)	-0.026*** (0.002)	-0.023*** (0.001)	-0.053*** (0.011)	-0.035*** (0.012)
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 <sup>2</sup>	0.56	0.56	0.46	0.67	0.71
DV Mean	0.60	0.60	0.71	0.30	0.62
F-stat		11,200.1		424.0	424.0

Notes: This table shows the relationship between reported Roma ethnicity and years of schooling (measured in 2011). The sample in columns 1 and 2 is all individuals reported as Roma in the 1992 census that can be uniquely matched to the 2011 census. In column 3 we use a perfectly matched sample between the VSN and 2011 census and the sample is mothers reported as Roma in VSN. In column 4 and 5, the sample is a mother-child pair, provided the mother is reported as Roma in the 1992 census and uniquely matched to the 2011 census, and the child is born between 2003 and 2011 (excluding 2004), based on records from the VSN. The outcome is census reported ethnicity in 2011 in columns 1, 2 and 3. In column 4, the outcome is the child's reported ethnicity in the VSN. In column 5, the outcome is the child's reported ethnicity in the 2011 census. The specification is equation (1) and it includes fixed effects for birth year interacted with locality at baseline. Standard errors are clustered at the locality level. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01.

So far we focused on how individuals' own education relates to their reported ethnicity, using increasingly precise linkages between records. We now turn to inter-generational passing: education may also be related to passing across generations, either through how parents report their children's ethnicity or how children later self-report. Here, we study the first channel - parental reporting- using VSN birth records, while we analyze the second channel in the following subsection.

Our sample includes women identified as Roma in the 1992 census, whose census records in 1992 and 2011 we match to their children's birth records (2003–2011) in the VSN files. We define passing as a child's ethnicity being reported as non-Roma or undeclared. Birth record ethnicity is typically reported by the mother, a key contrast with the census, where an enumerator records information from the household head, who is often male.

Results show that each additional year of maternal schooling is associated with a 5.3 pp lower probability that the child is reported as Roma in birth records (column 4). When measured in the 2011 census, the gradient is  $-3.5$  pp per year of maternal education (column 5). The effect is concentrated among women whose spouse was

A.4 replicates columns 1 and 2 for 1992–VSN–2011 and 2002–VSN–2011 linkages.

never reported as Roma (Table A.5). These findings highlight the prevalence of intergenerational passing and its steep education gradient.<sup>14</sup>

**Robustness Results.** Results using 2002 as the baseline year are similar (Table A.6). The slightly smaller magnitudes over this shorter time horizon are consistent with the graphical evidence from Figure A.3. We find a similar pattern for the education gradient when we use reported Romani native language instead of reported ethnicity, highlighting language as another dimension passing (Table A.7).

Our benchmark sample is skewed towards smaller localities, where matches based on the date of birth are more likely to be unique. Table A.8 runs the analysis from column 2 in Table 1 by town size. Most of our sample of matched individuals reported as Roma at baseline lives in towns below 50,000 inhabitants, similar to the entire Roma population (Figure A.5).<sup>15</sup> Comparing column 2 to 1, and 6 to 5, reveals that the education gradient in larger localities is negative and precisely estimated but 10–20% smaller in magnitude. (Our gradients in even larger localities are imprecisely estimated.)

Table A.9 reports additional checks. The education–passing link holds within households, though with smaller magnitude: more educated siblings pass at higher rates than less educated ones (columns 1–2). The gradient remains about half as large with census tract fixed effects, which proxy for enumerator-specific behaviors, suggesting that roughly half of the relationship within localities reflects variation across neighborhoods and enumerators, with the remainder operating at smaller spatial scales. Finally, our main analysis codes undeclared ethnicity (an option only in 2011) as passing. Results are similar when undeclared is instead coded as Roma (columns 5–6).

**Heterogeneity Results.** We next examine heterogeneity in the education–passing relationship (Table A.10). The education gradient is 50% higher for women than for men. Parental education also matters, consistent with our intergenerational results in Table 1, with mother’s education particularly important. Both the average passing rate and the strength of the gradient are lower among individuals with stronger ties to the Roma minority—proxied by Romani language use and having a Roma spouse—suggesting that such individuals are less marginal in their ethnic identification decisions.<sup>16</sup>

We then examine passing over the life cycle. Passing could occur early—if young people are reported as Roma by others in their household and later change their self-report—or it could be a continuous, probabilistic process shaped by contextual factors such as the enumerator. Panel 3 of Table A.10 shows that a large education gradient exists even among household heads in both rounds, who report their own ethnicity. Splitting the sample by age at baseline reveals no clear pattern in either

---

<sup>14</sup>Reporting context matters: in the VSN data, 30% of children are reported as Roma, compared with 62% of the same children in the 2011 census.

<sup>15</sup>See Figure A.6 for similar matching statistics by town size for the entire population.

<sup>16</sup>Passing is higher for individuals of Orthodox religion and for migrants, though the education gradient for migrants is imprecise due to the small sample.

passing rates or the education gradient.<sup>17</sup> While we cannot separate age from cohort effects, the evidence supports passing as a probabilistic, continuous process, a view that informs our model in the next section.

## 4 Model: Estimating Roma Population and Education Levels

We next estimate the size and educational distribution of the population with Roma heritage. Previous analyses condition on individuals reported as Roma at baseline, which overrepresents those persistently identified as Roma in any period and thus underestimates overall passing and may bias the education gradient.<sup>18</sup> To address this, we estimate a model with heterogeneity in the persistent likelihood of being identified as Roma, exploiting the fact that we observe three rounds of ethnic identification per individual.

### 4.1 Model Setup and Estimation

Our model considers individuals  $i$  in education group  $e$ . Some have a connection to the Roma ethnicity, which we call “Roma-heritage,” denoted  $i \in \mathcal{R}$ . The share of Roma-heritage in group  $e$  is  $h_e^{\mathcal{R}}$ . All others are classified as non-Roma.

We observe reported ethnicity  $Roma_{iet} \in \{0, 1\}$  in three census periods  $t = 0, 1, 2$ . A Roma-heritage individual  $i$  in period  $t$  is reported Roma with i.i.d. (possibly zero) probability

$$\pi_{iet} \equiv \Pr(Roma_{iet} = 1 \mid i \in \mathcal{R}) = c(\pi_{te} + \pi_{ie}),$$

where the function  $c(x) = \min(1, \max(0, x))$  censors  $x$  to the interval  $[0, 1]$ . Reported ethnicity depends on education ( $e$ ) and time period ( $t$ ) through  $\pi_{te}$ , and on individual  $i$ ’s time-invariant propensity  $\pi_{ie}$ , which we assume is drawn i.i.d. from a distribution with mean 0 and standard deviation  $\sigma_e$ . We later estimate the model varying the shape of this distribution: normal, uniform, or log-normal. Non-Roma individuals always report non-Roma ethnicity. We allow individuals in our panel data to be mismatched to a random individual in group  $e$ . Mismatch probabilities are allowed to vary by the pair of time periods and reported ethnicity in both periods (Appendix C).

We estimate the share of Roma-heritage,  $h_e^{\mathcal{R}}$ , and the passing parameters  $\pi_{et}$  and the standard deviations  $\sigma_e$ , as well as mismatch probabilities, using panel data  $(Roma_{iet}, Sex_{iet})_{t=0,1,2}$  on reported ethnicity and sex. We use the generalized method of moments (GMM) targeting the following moments for each education group  $e$ . We match eight shares of ethnicity triples  $Roma_{ie} \equiv (Roma_{ie0}, Roma_{ie1}, Roma_{ie2}) \in \{0, 1\}^3$ . For mismatch, we match the share of observations with inconsistent sex between  $t \in \{0, 1\}$  and  $t = 2$ ,  $Sex_{iet} \neq Sex_{ie2}$ , separately by each triple  $Roma_{ie}$ , for a total of sixteen moments. For inference, we bootstrap the entire procedure at the

<sup>17</sup>These results help address concerns that differential mortality drives our findings. For instance, if high-education Roma had higher mortality, they might appear less often in the linked sample. The similar gradients among younger groups with low mortality reduce this concern.

<sup>18</sup>In a simple model, heterogeneity in time-invariant Roma reporting biases the education coefficient toward zero, making earlier estimates conservative (Section C.1).

town level.<sup>19</sup>

To understand how our model is identified, consider first the case with no heterogeneity ( $\sigma_e = 0$ ) and no mismatch, and two census rounds  $t = 0, 1$ . We can recover the mean passing probability using

$$\overline{\pi_{ie1}} = \Pr(Roma_{ie1} \mid Roma_{ie0} = 1).$$

This conditional probability corresponds to the reduced form analysis in previous sections. The share of Roma-heritage is  $h_e^R = \overline{Roma_{ie1}} / \overline{\pi_{ie1}}$ .

In the presence of heterogeneity ( $\sigma_e > 0$ ),  $\Pr(Roma_{ie1} \mid Roma_{ie0} = 1)$  is an underestimate of  $\overline{\pi_{ie1}}$ , because the sample of reported Roma at  $t = 0$  will over-represent individuals with large  $\pi_i$ . Because we have *three* observations per individual, we can statistically detect persistence in an individual’s likelihood to declare Roma ethnicity and estimate  $\sigma_e$  (see Appendix C).

The estimated model suggests that the number of individuals with Roma heritage in Romania is significantly under-reported. In the left panel of Figure 2, we find that there are more than twice as many Roma-heritage high school graduates, and 1.5-2 times more middle school graduates, primary school graduates and Roma without formal schooling. For postsecondary graduates, there are roughly six times more Roma-heritage than in official data, when the shape of individual heterogeneity is normal or uniform. In these cases, the model infers a sizable share who never report Roma ethnicity,  $\pi_{iet} = 0$  (Table A.11). Because it is impossible to learn about this group of people using our data, we also report estimation results using the lognormal distribution, which leads all Roma-heritage individuals to have strictly positive  $\pi_{iet}$ . This is our most conservative specification, and we find there are 2.7 times more postsecondary graduate Roma-heritage individuals.

In absolute terms (right panel), there are roughly 275,000 more adult Roma-heritage individuals than reported Roma at the 2011 census in our matched sample. This means that within our matched sample, the proportion of Roma-heritage in the Romanian adult population is roughly 3.7%-4.1% of the total Romanian population, rather than a reported 2.3%. In terms of educational achievement, we estimate that Roma-heritage adults completed 6.2 years of schooling, versus 5.8 years for reported Roma. Accounting for this discrepancy explains roughly 13% of the Roma - non-Roma education gap.

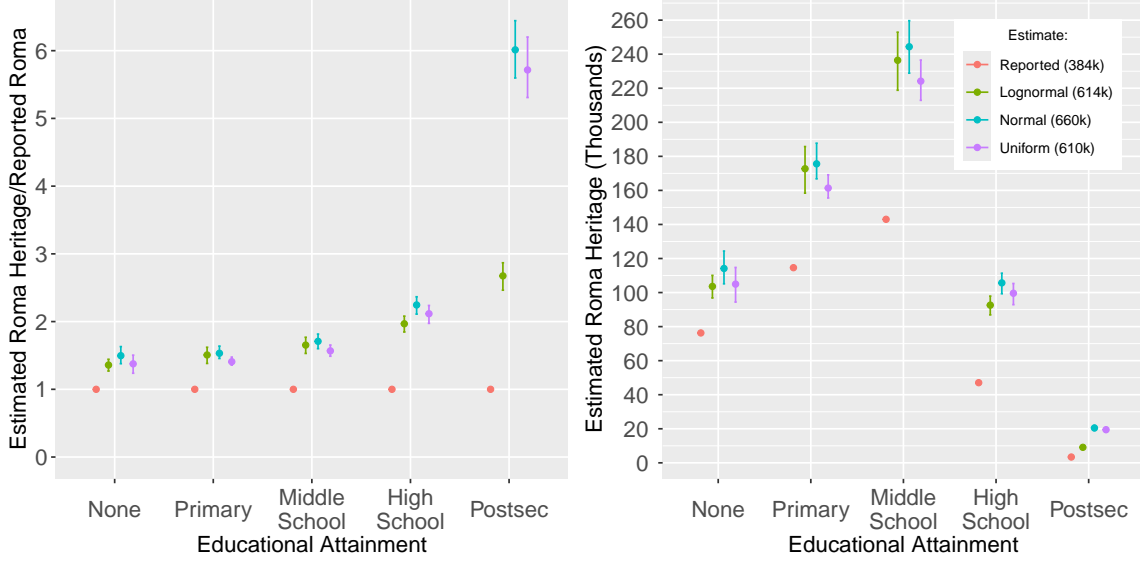
## 5 General Population Beliefs about Passing: Survey Evidence

What does the general population believe about passing and its relationship with education? We collected data on beliefs about Roma passing from 2,000 respondents from Romania, using an online survey. Respondents were drawn from a database of a major Romanian survey firm and stratified by nine regions in Romania.<sup>20</sup>

<sup>19</sup>The model fit is good, despite being over-identified (Figure A.7).

<sup>20</sup>Similar to other online surveys, the sample over-represents educated respondents. 30% live in rural areas, and among urban residents, 78% live in towns or cities where the share of Roma is above the national median.

Figure 2: Estimated Roma Population Accounting for Heterogeneity



Notes: This figure plots the model-inferred Roma-heritage population, and the reported Roma population in the 2011 census, in ratios and levels in the left and right panels, respectively. The model is estimated for individual heterogeneity following a normal, uniform, and log-normal distribution.

In our survey, we elicit participant beliefs about selective passing. Participants are first informed that 50.3% of reported Roma individuals in the 2011 census graduated middle school or above. We then elicit beliefs about the share of middle school or above graduates among Roma individuals broadly, regardless of their census-reported ethnicity, and offer financial incentives for correct answers.<sup>21</sup>

Our headline result is that Romanians, on average, are unaware that more educated Roma are more likely to pass. The first row in Table 2 reports results from our incentivized measure and shows that only 28.4% of our sample believes that Roma who pass are more educated than those who are reported as Roma. If anything, a larger share incorrectly believes that Roma who pass are less educated (42%). A further 29.6% believe there is no selection either way.

To ensure respondents understand the question, we implement two checks (Table 2). First, we randomly assign half the sample to a “explicit” group, who answer a series of questions explicitly referencing Roma who pass in the census before the incentivized question. Explicit group respondents are roughly evenly split across response options. Second, immediately after the incentivized question, we ask all respondents an attention check question (see Appendix D). About half answer correctly; among them, 37.8% answer “more” (correct) and 39.1% answer “less” (incorrect). Results are similar for rural and urban respondents, and among urban respondents, as a function of the share of Roma in their locality.

<sup>21</sup>A “Roma individual” is defined as someone whose ethnicity was reported as Roma in the 1992 census, consistent with our reduced-form analysis. See Appendix D for the exact wording.

Respondents appear to hold internally consistent views on the relationship between passing and education. In the priming group, we elicit this correlation in two ways: by asking whether more educated Roma are more or less likely to pass, and whether Roma who pass are more or less educated. Answers to these two questions align 76% of the time, though respondents disagree on the direction (Table A.12). Open-text explanations typically offer theories supporting the respondent’s choice. A common theme is “shame,” with some arguing that more educated Roma feel less ashamed of their ethnicity, and others asserting the opposite.

Table 2: Beliefs Regarding Roma Passing and Education

Sample	“Passing Roma Are ___ Educated Compared to Self-Reported Roma (%)” (Incentivized)			Obs.
	Less	Equally	More	
Full Sample	42.0	29.6	28.4	2,000
Explicit questions	39.3	25.3	35.4	1,000
No explicit questions	44.7	34.0	21.3	1,000
Attentive	39.1	23.2	37.8	855
Not Attentive	44.2	34.5	21.3	1,145
Rural	40.5	31.2	28.3	600
Urban	42.6	29.0	28.4	1400
Above Median Roma	43.2	28.3	28.5	590
Below Median Roma	42.2	29.5	28.3	810

Note: Survey responses to an incentivized question about education of passing and non-passing Roma. “Explicit questions” refers to respondents who were randomly assigned to answer multiple questions explicitly referencing passing and non-passing Roma, prior to the main question. “Attentive” indicates respondents who passed an attentiveness check. In last two rows, we split respondents in urban areas by the median value of the Roma population in their locality.

The second main result from our survey is that respondents express high confidence in their ability to identify Roma who pass (Table A.13). Overall, 68% report it is very or relatively easy to tell if someone is Roma when that person does not declare their ethnicity. When asked how they could tell, the most frequently cited markers are way of speaking (68%), physical appearance (55%), skin color (51%), names (29%), and clothing (27%). Only 25% mention family or community background. Most of these traits are either malleable (e.g., clothing, physical appearance) or stereotypical (e.g., a distinctive Roma way of speaking). In Romania, distinctly Roma names are rare, and skin tone varies, overlapping substantially between Roma and non-Roma.

These findings suggest that Roma individuals may be able to pass successfully.

## 6 Discussion

A conservative implication of our results is that official statistics underestimate both the number and educational attainment of people with Roma heritage, with significant policy implications. For example, less than 1% of the reported Roma are documented as university graduates - a figure that has drawn significant policy attention ([The World Bank, 2014](#), [United Nations Development Programme, 2012](#)) - yet our analysis suggest that passing may increase this share 3-6 times.

If the passing patterns observed in official records extend to everyday social and professional interactions, our results have deeper implications. The strong positive educational gradient in passing, coupled with the apparent lack of awareness within the general population, suggests that passing may reinforce existing negative stereotypes about Roma. Indeed, to the extent that stereotype creation and statistical discrimination is partly based on observational learning, the relative lack of visibility of highly educated Roma may skew perceptions about the entire group. Over time, these misperceptions can perpetuate systemic inequalities and limit opportunities for upward mobility, even for individuals who have attained higher education.

One promising path for future research is understanding why more educated Roma pass more. We outline three hypotheses. First, the returns from a non-Roma identity may be higher for more educated Roma, perhaps due to higher discrimination or because of higher dispersion in returns for skilled individuals. Second, the cost of passing may be lower for the more educated; our finding that the general population uses stereotypes and malleable markers to recognize Roma lends some support to this hypothesis. Third, more educated Roma may increasingly feel less attached to the Roma ethnic identity, among other factors, reflecting processes of implicit cultural assimilation. Our data does not allow us to sharply distinguish between these possibilities. Another important avenue is estimating the impact of government policy and non-government interventions that encourage ethnic identification on the level of passing and on its educational gradient.

## References

- Ran Abramitzky, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. Automated Linking of Historical Data. *Journal of Economic Literature*, 59(3):865–918, September 2021. ISSN 0022-0515. doi: 10.1257/jel.20201599.
- Anjali Adukia, Richard Hornbeck, Daniel Keniston, and Benjamin Lualdi. The Social Construction of Race during Reconstruction. Technical Report w33502, National Bureau of Economic Research, February 2025.
- George A Akerlof and Rachel E Kranton. Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753, 2000.
- Alberto Alesina and Eliana La Ferrara. Ethnic diversity and economic performance. *Journal of Economic Literature*, 43(3):762–800, 2005.
- Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.
- Leonardo Bursztyn and David Y Yang. Misperceptions about others. *Annual Review of Economics*, 13:383–407, 2021.
- Leonardo Bursztyn, Thomas Fujiwara, and Amanda Pallais. ‘Acting wife’: Marriage market incentives and labor market investments. *American Economic Review*, 107(11):3288–3319, 2017.
- Christopher Cornwell, Jason Rivera, and Ian M Schmutte. Wage discrimination when identity is subjective: Evidence from changes in employer-reported race. *Journal of Human Resources*, 52(3):719–755, 2017.
- Ricardo Dahis, Emily Nix, and Nancy Qian. Choosing racial identity in the United States, 1880-1940. Technical Report 26465, National Bureau of Economic Research, 2020. URL <https://www.nber.org/papers/w26465>.
- Brian Duncan and Stephen J. Trejo. Intermarriage and the Intergenerational Transmission of ethnic identity and human capital for Mexican Americans. *Journal of Labor Economics*, 29(2):195–227, April 2011a. ISSN 0734-306X. doi: 10.1086/658088.
- Brian Duncan and Stephen J. Trejo. Who remains Mexican? selective ethnic attrition and the intergenerational progress of Mexican Americans. In David L. Leal and Stephen J. Trejo, editors, *Latinos and the Economy: Integration and Impact in Schools, Labor Markets, and Beyond*, pages 285–320. Springer, New York, NY, 2011b. doi: 10.1007/978-1-4419-6682-7\_14.
- Brian Duncan and Stephen J Trejo. Which Mexicans are white? Enumerator-assigned race in the 1930 census and the socioeconomic integration of Mexican Americans. Technical report, National Bureau of Economic Research, 2023.

- Benjamin Enke. What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398, 2020.
- European Commission. A new EU Roma strategic framework. Fact-sheet. [https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combating-discrimination/roma-eu/roma-equality-inclusion-and-participation-eu\\_en](https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combating-discrimination/roma-eu/roma-equality-inclusion-and-participation-eu_en), 2020. Accessed: July 2025.
- European Union Agency for Fundamental Rights. *Roma in 10 European countries: Main results - Roma survey 2021*. Publications Office of the European Union, 2023. ISBN 978-92-9489-125-9. doi: 10.2811/221064. URL <https://doi.org/10.2811/221064>.
- Robert W Fairlie, Florian Hoffmann, and Philip Oreopoulos. A community college instructor like me: Race and ethnicity interactions in the classroom. *American Economic Review*, 104(8):2567–2591, 2014.
- Dylan Glover, Amanda Pallais, and William Pariente. Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, 132(3):1219–1260, 2017.
- Cristina Grigore. The Gypsy in me. *The New Yorker*, June 2012. URL <https://www.newyorker.com/books/page-turner/the-gypsy-in-me>. Published on June 21, 2012.
- Jonas Hjort. Ethnic divisions and production in firms. *The Quarterly Journal of Economics*, 129(4):1899–1946, 2014.
- Michelle Kelso. “And Roma were victims, too.” The Romani genocide and Holocaust education in Romania. In *Holocaust Education*, pages 75–92. Routledge, 2017.
- Rowena Marin. *Who Am I in the World?: A Story of Becoming*. New Degree Press, 2023. ISBN 9798889266358. URL <https://books.google.com/books?id=0QSUzwEACAAJ>.
- Hamid Nohanibehambari and Jason Fletcher. Passing as white: Racial identity and old-age longevity. Working Paper 33394, National Bureau of Economic Research, January 2025. URL <http://www.nber.org/papers/w33394>.
- Suanna Oh. Does identity affect labor supply? *American Economic Review*, 113(8): 2055–2083, 2023.
- Devah Pager, Bart Bonikowski, and Bruce Western. Discrimination in a low-wage labor market: A field experiment. *American Sociological Review*, 74(5):777–799, 2009.
- Maria-Carmen Pantea. On pride, shame, passing and avoidance: An inquiry into Roma young people’s relationship with their ethnicity. *Identities*, 21(5):604–622, 2014.

- Catherine Porter and Danila Serra. Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, 12(3):226–254, 2020.
- Joseph Price, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. Combining family history and machine learning to link historical records: The Census Tree data set. *Explorations in Economic History*, 80:101391, 2021.
- Robbert Rademakers and André van Hoorn. Ethnic switching: Longitudinal evidence on prevalence, correlates, and implications for measuring ethnic segregation. *Journal of Development Economics*, 152:102694, 2021.
- The World Bank. Diagnostics and policy advice for supporting Roma inclusion in Romania, 2014. URL <https://www.worldbank.org/content/dam/Worldbank/document/eca/romania/OutputEN.pdf>.
- United Nations Development Programme. Roma education in comparative perspective, 2012.
- Elie Wiesel. Final report. Technical report, International Commission on the Holocaust in Romania, 2004.
- Cătălin Zamfir and Marian Preda. *Romii în România*. Editura Expert Bucharest, 2002.
- Cătălin Zamfir and Elena Zamfir. *Țigani în ignorare și îngrijorare*. Alternative, 1993.

## A Online Appendix

### A.1 Additional Figures and Tables

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

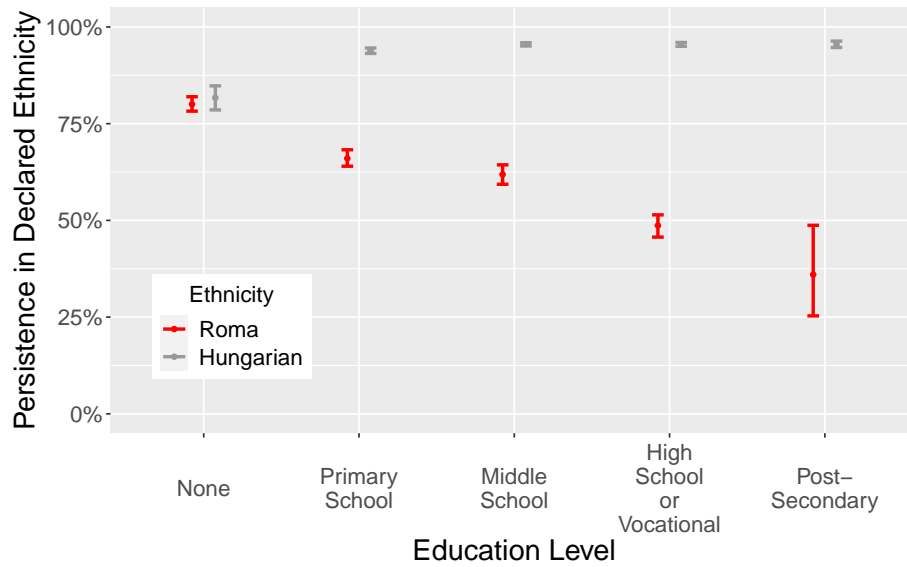
Notes: This table reports the share of linked observations that have inconsistent sex, when using only locality and date of birth to link observations. “Educated” represents individuals who have more than 8 years of schooling at endline (2011).

Table A.2: Census Linkages for Baseline Roma

Education Level	1992-2011		2002-2011	
	N	Inconsistent Sex (%)	N	Inconsistent Sex (%)
None	15,308	2.2	33,642	1.2
Primary School	28,378	2.2	66,487	1.6
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

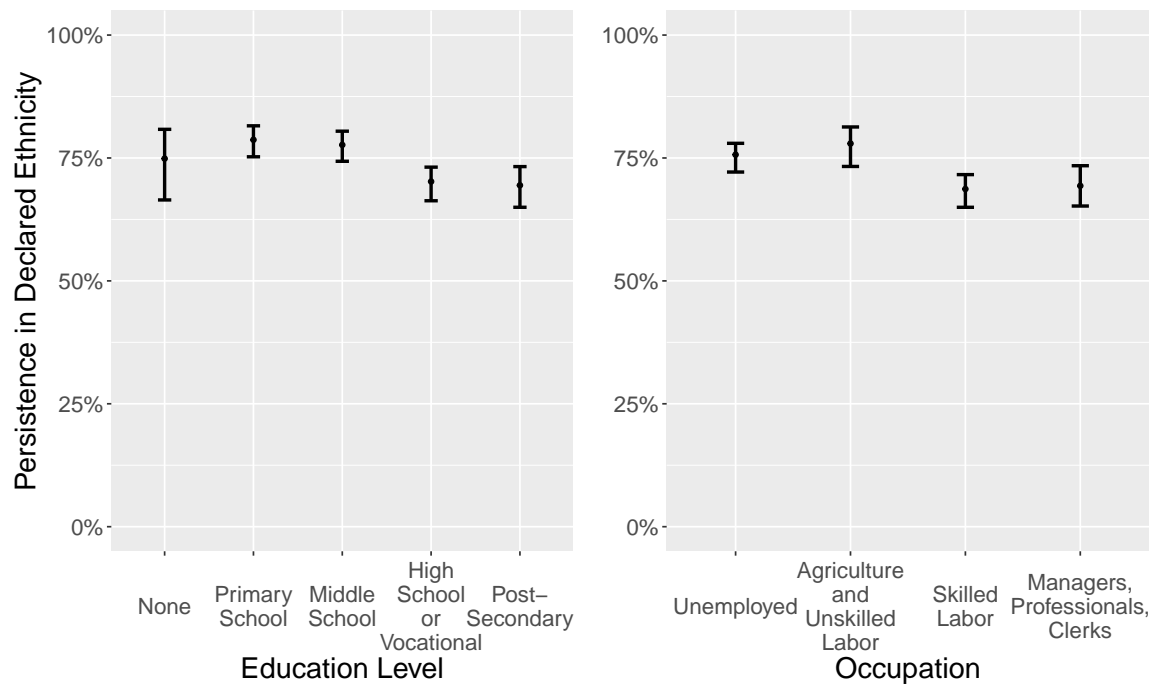
Notes: This table reports the share of linked observations that have inconsistent sex by education level, when using only locality and date of birth to link observations. The sample is individuals with reported Roma ethnicity at baseline.

Figure A.1: Passing by Education: Household Heads Only (1992 - 2011)



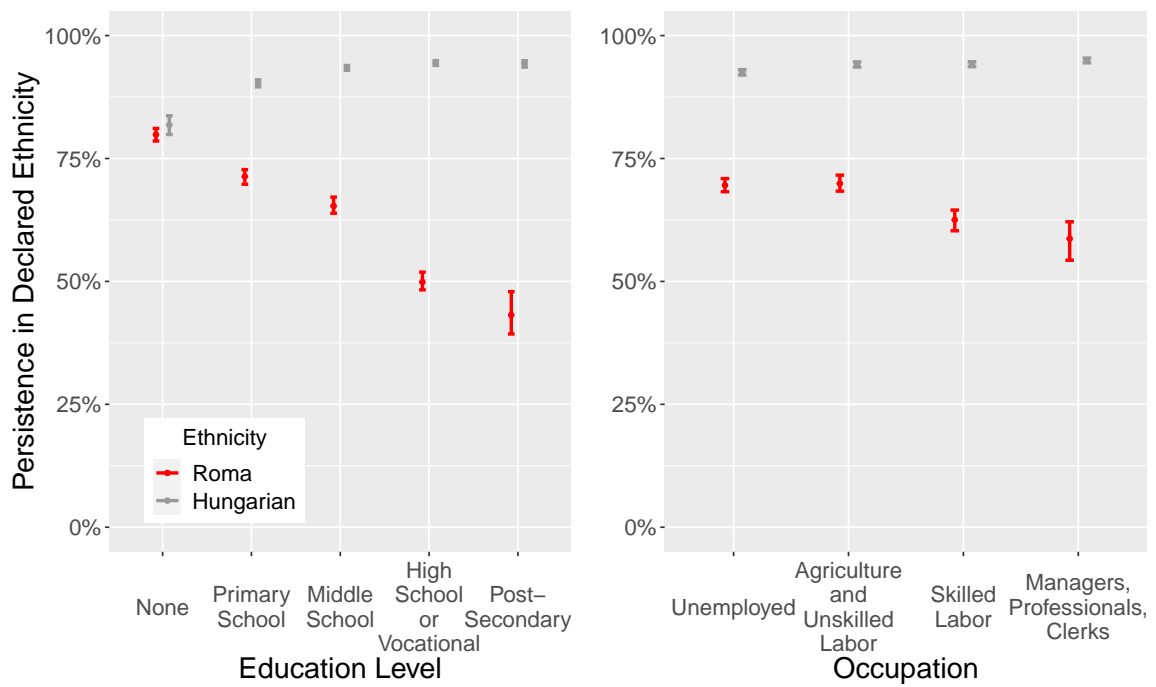
Notes: Replicates Figure 1 with individuals who were household heads (and thus, self-reporting their ethnicity) in 1992 and 2011.

Figure A.2: Passing by Education: Other Ethnicities (1992 - 2011)



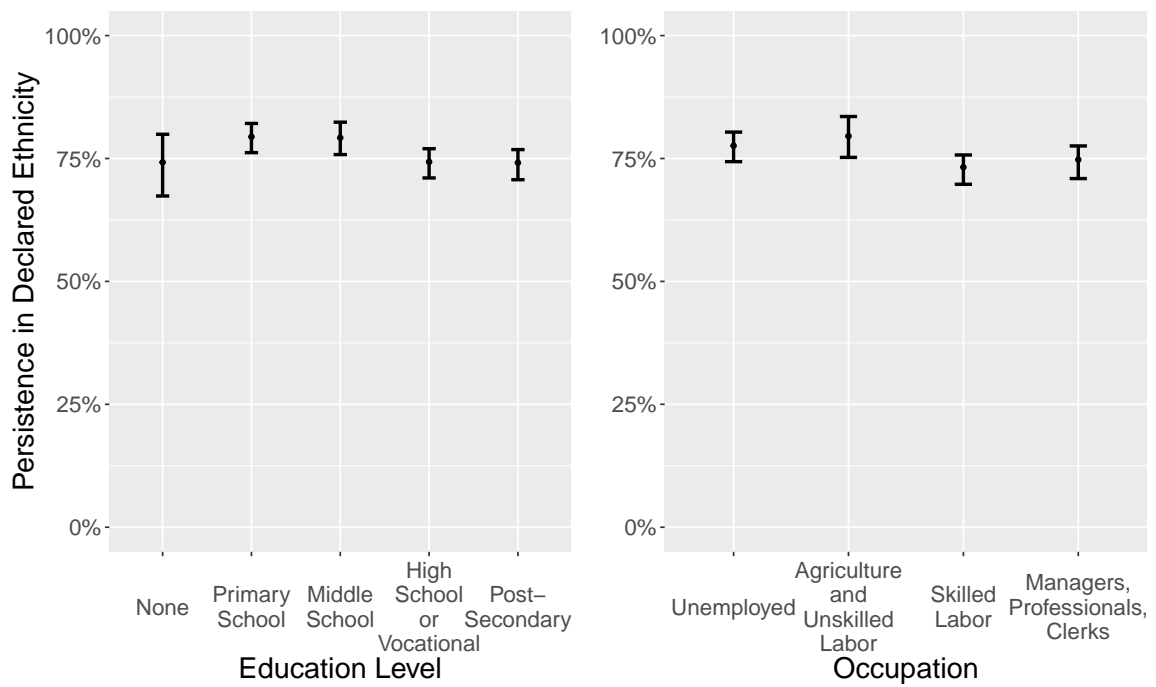
Notes: Replicates Figure 1 for all ethnicities other than Roma and Hungarian, with baseline year 1992. We pool all ethnicities. Passing is defined as reporting Romanian, Roma, or Hungarian ethnicity.

Figure A.3: Passing by Education (2002 - 2011)



Notes: Replicates Figure 1 with baseline year 2002.

Figure A.4: Passing by Education: Other Ethnicities (2002 - 2011)



Notes: Replicates Figure A.2 with baseline year 2002.

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

	<i>Dependent Variable:</i>			
	<i>Yrs of Schooling (2011)</i> (All) (1)	<i>(Baseline Roma)</i> (2)	<i>Reported Roma (2011)</i> (All) (3)	<i>(Baseline Roma)</i> (4)
Baseline Years of Schooling	-0.003 (0.006)	-0.080 (0.060)	-0.001** (0.000)	-0.003 (0.004)
N	131,855	4,336	131,855	4,336
R <sup>2</sup>	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: Passing by Education (Mothers' Sample)

<i>Dependent Variable:</i>				
<i>Self-Reported Roma Ethnicity (2011)</i>				
	OLS '92-'11	IV '92-'11	OLS '02-'11	IV '02-'11
	(1)	(2)	(3)	(4)
Schooling Yrs	-0.028*** (0.002)	-0.033*** (0.012)	-0.024*** (0.002)	-0.024*** (0.004)
Sample	'92-VSN-'11	'92-VSN-'11	'02-VSN-'11	'02-VSN-'11
Outcome	Mothers Census '11	Mothers Census '11	Mothers Census '11	Mothers Census '11
N	19,707	19,707	33,666	33,666
R <sup>2</sup>	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: Intergenerational Passing by Education and Father’s Ethnicity

	<i>Dependent Variable:</i>			
	<i>Reported Roma Ethnicity (2011)</i>			
	Roma Father (1)	non-Roma Father (2)	Roma Father (3)	non-Roma Father (4)
Schooling Yrs (Endline)	0.012 (0.013)	-0.044* (0.026)	-0.016*** (0.004)	-0.027*** (0.009)
Sample	'92-'11	'92-'11	'02-'11	'02-'11
N	11,315	8,485	24,522	10,056
R <sup>2</sup>	0.72	0.81	0.71	0.82
DV Mean	0.93	0.21	0.83	0.28
F-stat	221.1	108.7	3,686.1	1,013.0

Notes: This table shows the relationship between declaring a child’s Roma ethnicity and maternal completed years of schooling years of schooling. We show results for i) different linked samples and ii) Romani and non-Romani ethnicity of the father. “Roma father” is the sample where the spouse is reported as Roma in at least one census or has at least one Roma reported parent. 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.6: Passing by Education (2002 - 2011)

	<i>Dependent Variable:</i>			
	<i>Reported Roma Ethnicity (2011)</i>			
	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)
Schooling Yrs	-0.024*** (0.001)	-0.022*** (0.001)	-0.029*** (0.003)	-0.026*** (0.004)
N	210,076	210,076	33,666	33,666
R <sup>2</sup>	0.48	0.48	0.68	0.67
DV Mean	0.67	0.67	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: Passing by Education – Ethnicity Defined using Romani as Mother Tongue

<i>Dependent Variable:</i> <i>Romani Native Language</i> <i>at Endline</i>		
	OLS (1)	IV (2)
Schooling Yrs	-0.024*** (0.002)	-0.022*** (0.004)
N	40,563	40,563
R <sup>2</sup>	0.63	0.63
DV Mean	0.53	0.53
F-stat		3,057.6

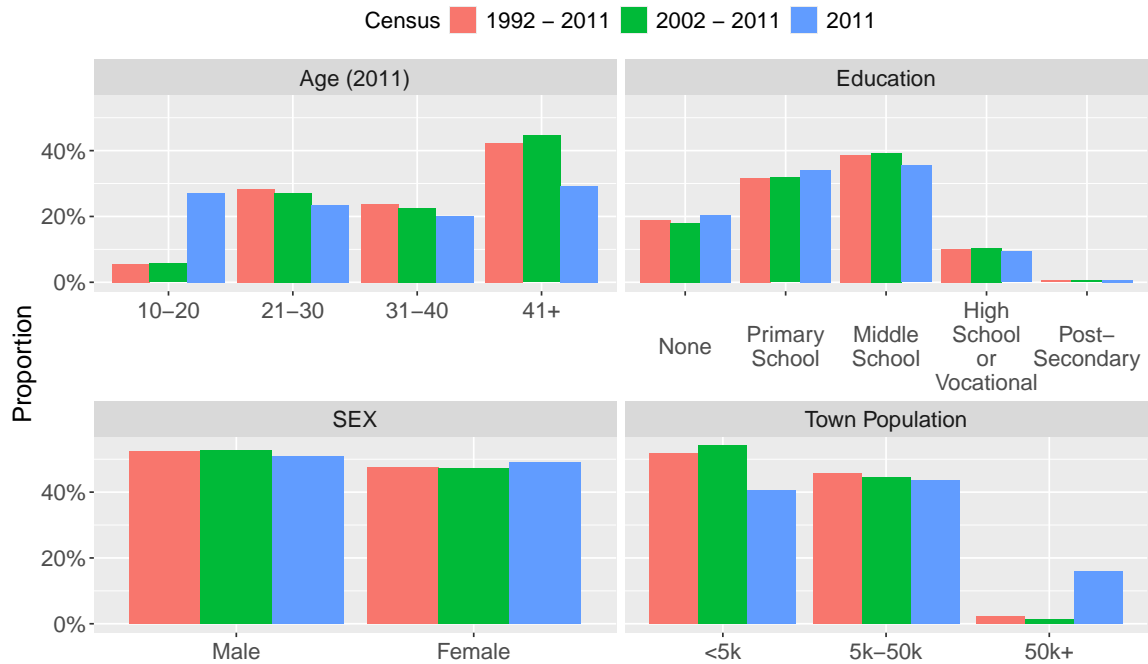
Notes: This table shows the relationship between years of schooling and reported Romani native language, using 1992 - 2011 linked records. The sample consists of individuals who report Romani as their native language at baseline (1992). The outcome is reporting Romani as a native language at endline (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: Passing by Education and Town Size

	<i>Dependent Variable:</i> <i>Reported Roma Ethnicity (2011)</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
Schooling Yrs (Endline)	-0.028*** (0.003)	-0.025*** (0.003)	-0.016* (0.009)	-0.025*** (0.002)	-0.020*** (0.002)	-0.017** (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 <sup>2</sup>	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

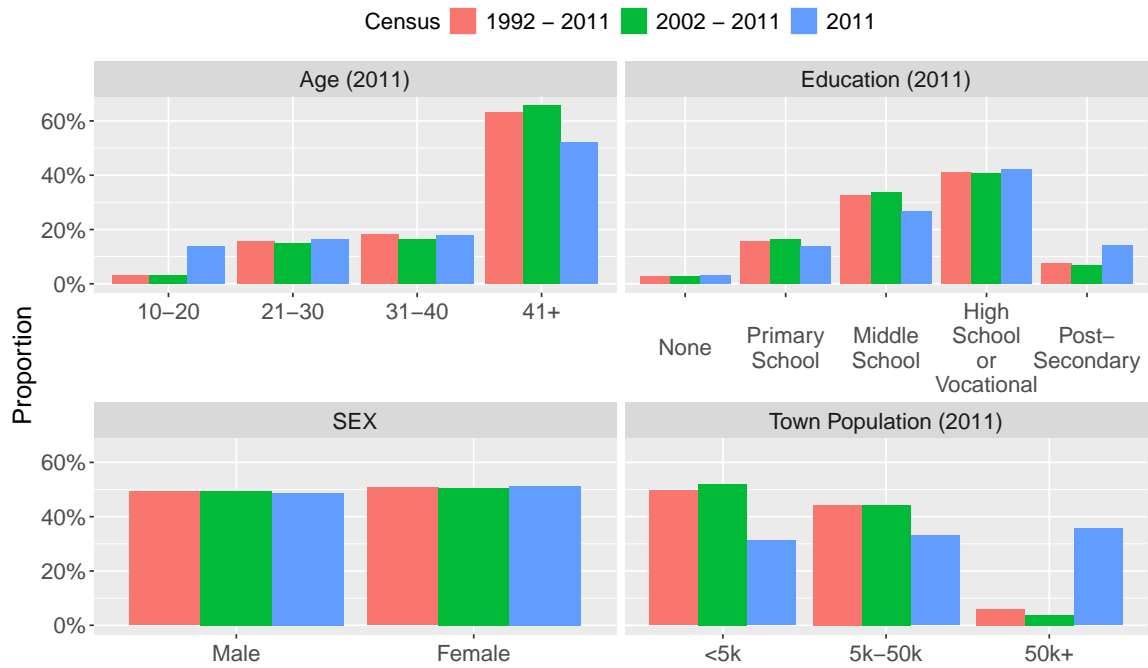
Notes: This table shows the relationship between declaring Roma ethnicity and years of schooling across towns with different populations in 2011 an across different linked samples. Controls include fixed effects for birth year interacted with locality. Standard errors are clustered at the locality level.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

Figure A.5: 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.6: Sample Representativeness



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.9: Passing 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*** (0.001)	-0.005* (0.003)	-0.015*** (0.001)	-0.013*** (0.002)	-0.023*** (0.001)	-0.030*** (0.002)
N	107,168	107,168	107,168	107,168	107,168	107,168
R <sup>2</sup>	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

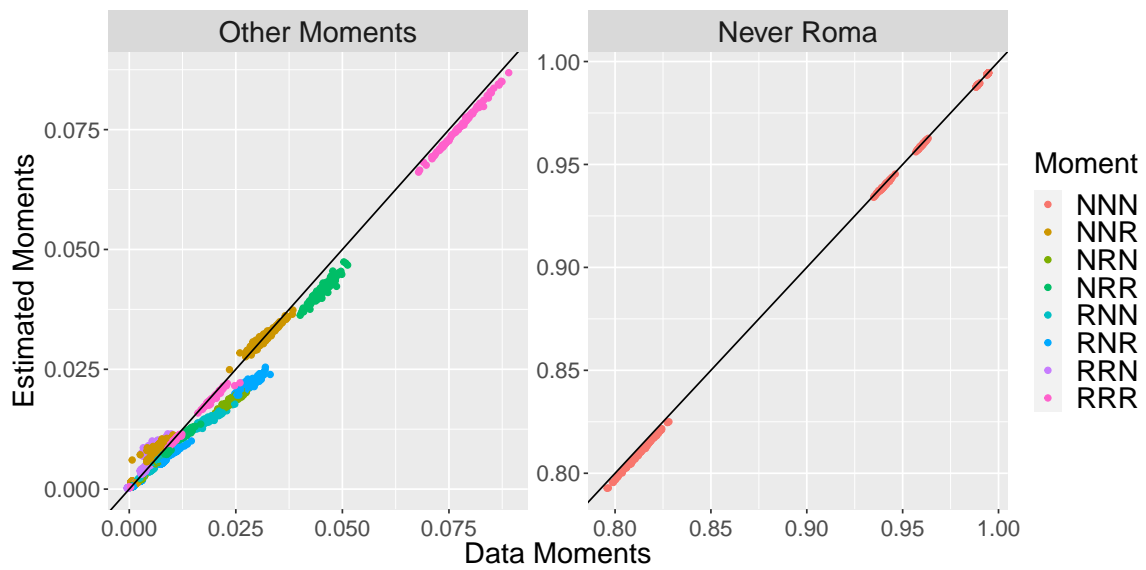
Notes: This table shows the relationship between declaring Roma ethnicity and years of schooling. It provides robustness checks corresponding to the first two columns of Table 1: adding household fixed effects (columns 1 and 2), adding enumerator fixed effects (columns 3 and 4) and omitting non-declaration of ethnicity from passing 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: Passing by Education – Heterogeneity

Panel 1: Passing, Gender and Parental Education								
<i>Dependent Variable: Reported Roma Ethnicity (2011)</i>								
	Sex		Sex & Parental Educ.		Mom Education		Dad Education	
	Male	Female	Male	Female	≥ 8 Yrs	< 8 Yrs	≥ 8 Yrs	< 8 Yrs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Schooling Yrs (Endline)	-0.024*** (0.003)	-0.036*** (0.004)	-0.019** (0.007)	-0.026*** (0.009)	-0.043*** (0.009)	-0.018*** (0.004)	-0.029*** (0.007)	-0.024*** (0.005)
Schooling Years (Mother)			-0.006*** (0.002)	-0.003 (0.002)				
Schooling Years (Father)			-0.004** (0.002)	-0.002 (0.002)				
N	55,401	51,767	29,597	22,222	17,631	41,777	24,084	30,534
R <sup>2</sup>	0.66	0.68	0.65	0.66	0.65	0.62	0.62	0.66
DV Mean	0.62	0.59	0.62	0.57	0.52	0.63	0.56	0.63
F-stat	5,470.4	3,998.0	757.1	443.8	385.6	2,063.8	608.6	1,269.2
Panel 2: Passing and Other Individual and Family Characteristics								
<i>Dependent Variable: Reported Roma Ethnicity (2011)</i>								
	Migrant		Native Language		Orthodox Religion		Roma Spouse	
	Yes	No	Romani	Non-Romani	Yes	No	Yes	No/N.A.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Schooling Yrs (Endline)	-0.006 (0.017)	-0.027*** (0.002)	-0.016*** (0.004)	-0.027*** (0.003)	-0.026*** (0.002)	-0.019*** (0.006)	-0.020*** (0.002)	-0.033*** (0.004)
N	7,363	99,805	42,183	64,985	90,467	16,701	64,414	42,754
R <sup>2</sup>	0.86	0.58	0.59	0.61	0.58	0.66	0.69	0.63
DV Mean	0.42	0.62	0.73	0.52	0.61	0.60	0.76	0.36
F-stat	216.5	11,318.3	3,072.2	7,999.6	9,163.0	1,630.6	10,269.3	1,598.7
Panel 3: Baseline Age and Household Head Status								
<i>Dependent Variable: Reported Roma Ethnicity (2011)</i>								
	Head	Not Head	Head	10-20 y.o.	20s	30s	40+ y.o.	
	Baseline	Baseline	Always	Baseline	Baseline	Baseline	Baseline	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Schooling Yrs		-0.019*** (0.004)	-0.031*** (0.003)	-0.020*** (0.004)	-0.031*** (0.004)	-0.021*** (0.003)	-0.030*** (0.003)	-0.022*** (0.005)
N		20,342	86,483	18,006	26,875	19,045	15,628	12,783
R <sup>2</sup>		0.78	0.57	0.80	0.53	0.59	0.61	0.74
DV Mean		0.62	0.60	0.63	0.59	0.63	0.62	0.57
F-stat		4,696.0	6,295.7	4,464.6	2,522.5	4,786.1	4,164.5	2,563.0

Notes: This table reports heterogeneity in passing and its relationship with education. Each column reports an instrumental variables regression (Table 1, column 2) estimated on a different sample, where the dependent variable is an indicator for Roma reported ethnicity in the 2011 census. The main sample corresponds to Table 1, column 2. The sample is further restricted to records where individual and family variables (parental education, parental and spousal reported ethnicity, native language, religion, household age status) are available in the baseline census. Roma spouse is defined as a spouse reported as Roma in at least one census, or has a Roma reported parent. 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.7: 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.11: Fraction of Roma-background Individuals with a Null Probability of Roma Self-Reporting

Panel 1: Normal				
$p = 0$ Roma self-report				
Education	$\sigma$	1992	2002	2011
None	0.39	0.12	0.07	0.04
Primary	0.23	0.01	0.01	0.00
Middle School	0.21	0.01	0.00	0.00
High School	0.19	0.00	0.01	0.01
Postsecondary	0.21	0.17	0.17	0.16
Panel 2: Uniform				
$p = 0$ Roma self-report				
Education	$\sigma$	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	$\sigma$	1992	2002	2011
None	0.64	0.00	0.00	0.00
Primary	0.45	0.00	0.00	0.00
Middle School	0.38	0.00	0.00	0.00
High School	0.51	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-background individuals who have a null probability of ever self-identifying as Roma. the table contains three panels, one for each version of the model.  $\sigma$  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.12: Respondent Views Internally Consistent but Divergent

		Are Roma Who Pass More Educated?		
Are Educated Roma More Likely to Pass?	No	Yes		Total
No	<b>170</b>	63		233
Yes	54	<b>204</b>		258
Total	224	267		491

Notes: This table cross-tabulates responses to two questions about the relationship between passing and education. Only respondents in the “priming” group ( $N = 1,000$ ) received these questions. We drop observations where the two groups are rated the same (equally likely to pass, same education), and when the respondent is not sure.

Table A.13: Perceived Markers of Roma Ethnicity

Marker	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 pass.  $N = 2,000$ .

## B Additional Results

### B.1 Census Linking Procedure

This section describes our census linking procedure. We attempt to match each record in our endline (2011) census to a record in a baseline census (either 1992 or 2002). We do not directly link the 1992 and 2002 censuses because only the 2011 census has a rich set of questions regarding individuals geographic mobility, including an individual’s previous locality of residence and moving year. These questions allow us to link movers by identifying the locality of residence at baseline. The 2002 census does not have comparable information so we do not use it as an endline period.

We link records based on birth date and locality. In some specifications we also use the individual’s sex. For non-movers (i.e. individuals who live in the locality of their birth), or individuals who last changed locality before the baseline census, we use their locality of residence at baseline and endline to match on locality. For movers who moved between the baseline and endline censuses, we match the locality at baseline to the locality reported as the previous place of residence in the endline census. For this group of movers, we rely on the assumption that the previous locality of residence mentioned in the 2011 census matches the locality of residence in the baseline census. (This assumption may be wrong if the individual moved more than once between censuses.)

We only retain unique two-sided matches. In other words, for each birth date-locality(-sex) cell, we retain it only if there is exactly one record in the baseline census in the cell, and exactly one record in the endline census in the cell. This procedure excludes cells that have the same number  $K > 1$  of records in both time periods, which corresponds to multiple individuals.<sup>22</sup> It also excludes cases where the cell has different numbers of records in the baseline and endline censuses, which may occur because of mortality or errors.

When using birth date, locality and sex, we uniquely match 5,479,009 records between the 1992 and the 2011 censuses and 7,516,703 records between the 2002 and 2011 censuses. 4,454,321 records are uniquely matched across all three census waves.

---

<sup>22</sup>It is possible to use non-unique matches. However, accounting for differences in the number of records across censuses and adjusting for mismatch becomes more involved. For simplicity, we focused on the unique matches.

## B.2 Adjusting for Mismatch

To adjust our results for mismatch, we use the matching procedure described above using only birth date and localities, and not sex. We match 4,545,737 records for 1992-2011, 6,286,727 records for 2002-2011, and 3,630,390 records across all three censuses.

Because we do not use an individual’s sex for linking records, we can use the sex variable reported in the baseline and endline censuses to identify mismatched records. When sex differs across census waves within linked records, we consider the linked records to be mismatched.<sup>23</sup> Note that this only captures approximately half of all mismatches records, because for around half of all the mismatched records, by chance reported sex will be the same in the two waves. Consequently, we estimate the mismatch rate by doubling the fraction of records with inconsistent sex.

We now explain how we use the estimated mismatch rates to correct our estimates of passing for mismatch. We do this procedure separately for a group  $g$ , for example the set of individuals with a certain education level in the 2011 census.

The passing rate in the linked data is a mix of spurious passing among mismatched records, and actual passing in correctly matches records. Consider a sample of individuals  $i$  in group  $g$  who were reported as Roma at baseline, as in our main analyses. We can write the persistence in reported Roma ethnicity among this sample as:

$$\mathbb{E}[Roma_{ig}^{\text{data}}] = \underbrace{\mu_g \times p_g^{\text{Roma}}}_{\text{random mismatch}} + (1 - \mu_g) \times \mathbb{E}[Roma_{ig}]$$

where matched census record  $i$  in group  $g$  has probability  $\mu_g$  of being mismatched,  $p_g^{\text{Roma}}$  is the probability of being linked to an endline Roma record if mismatched, and  $Roma$  and  $Roma^{\text{data}}$  are indicators for the observed and real (adjusted for mismatch) declared Roma ethnicity at endline. If we know the mismatch rate  $\mu_g$  we can recover the true persistence in self-declared ethnicity:

$$\mathbb{E}Roma_{ig} = \frac{\mathbb{E}Roma_{ig}^{\text{data}} - \mu_g p_g^{\text{Roma}}}{1 - \mu_g}$$

The key assumption in this procedure is that among mismatches records are conditionally random, so the chance of matching to another Roma individual is  $\mu_g^{\text{Roma}}$ . Table A.3 shows evidence in support of this assumption.

We use this procedure to adjust passing rates for mismatch in Figure 1. We

---

<sup>23</sup>Actual sex changes were extremely uncommon in Romania during this period.

repeat the following procedure for each group  $g$  defined by education level in 2011. We estimate matching rates  $\hat{\mu}_g = 2p_g$  where  $p_g$  is the proportion of records with inconsistent sex in group  $g$ . We estimate  $\hat{p}_g^{\text{Roma}}$  as the fraction of endline records with reported Roma ethnicity among the records with inconsistent sex across census waves.

## C Model details

### C.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 connected to the Roma ethnic group.<sup>24</sup> Assume a linear probability model for the probability that individual  $i$  reports their ethnicity as Roma at time  $t$  :

$$\Pr(\text{Roma}_{it} = 1) = \alpha + \beta e_i + \mu_i + \epsilon_{it}. \quad (2)$$

Here,  $\alpha$  is a constant,  $\beta$  measures how Roma identification varies with education  $e_i \in \{0, 1\}$ . (Our hypothesis is  $\beta < 0$ .) The fixed effect  $\mu_i$  allows some individuals to be more likely to declare themselves as Roma in all time periods. We assume that  $\mu_i$  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.  $\epsilon_{it}$  is an iid, mean-zero, random shock.

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

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

In reality, we only observe the reported ethnicity  $\text{Roma}_{it}$  in different time periods. Our regression approach is to condition on a sample of individuals reported as Roma

---

<sup>24</sup>Defining a person's true ethnicity is likely impossible and conceptually problematic. In this section, we abstract from these difficulties to explain clearly the econometric implications of our approach.

in a previous period. We compute:

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

We are interested in whether  $\beta_{COND}$  is biased relative to  $\beta$ . 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  $\epsilon_{it}$  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  $\mu_i$  is non-degenerate. This is because individuals with higher  $\mu_i$  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 the given by

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

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

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

Similarly, for the high education group we get:

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

Putting these together we get the bias:

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

This is a bias towards zero, assuming that  $\text{Var}(\mu) < \alpha(\alpha + \beta)$ , i.e. the variance of  $\mu$  is not too large. (This is a weak conditions 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  $\beta$  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 passing.

## C.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  $\pi_{iet}$ . With remaining probability  $1 - \pi_{iet}$  they declare non-Roma (0) ethnicity.

The self-declaration probability  $\pi_{iet}$  can be decomposed into two parts, such that  $\pi_{iet} = \pi_{it} + \epsilon_i$ . The first component ( $\pi_{it}$ ) 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 ( $\epsilon_i$ ) is an individual-specific idiosyncratic but time-constant “taste” for self-declaring Roma for individual  $i$ . This parameter  $\epsilon_i$  is drawn from a normal distribution with mean 0 and standard deviation  $\sigma^e$ .<sup>25</sup>

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 ( $M_0 = 1$ )

---

<sup>25</sup>We also estimated the model using a uniform distribution for  $\sigma$ . Our results are very robust to the distributional assumption for  $\epsilon_i$ .

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:

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

where  $M_2$  is always equal to 0.<sup>26</sup> 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  $(\pi_{iet})^{Roma_t} (1 - \pi_{iet})^{1-Roma_t}$ .

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  $\pi_{jet}$  or pass 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

---

<sup>26</sup>Since a record at time  $t = 2$  cannot be mismatched to itself.

Roma ethnicity at rate  $\pi_{je0}$ . They are also correctly matched to their own record at time  $t = 1$ , where they passed with rate  $1 - \pi_{ie1}$ . The second term in the summation corresponds to the individuals 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  $\pi_{je0}$ .

## D Survey Questions

This section reports the survey questions used to ascertain beliefs about passing and non-passing Roma levels of schooling.

The main, incentivized, question on beliefs about the educational attainments of Roma who pass 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 individuals 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 puțin clasa a 8-a în 2011?

(**Precizare:** răspunsul corect este calculat pe baza educației persoanelor care s-au declarat de etnie roma anterior, la recensământul din 1992, indiferent de cum și-au declarat etnia în 2011.)

**Opțiuni:**

- a. Mai puțin 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.<sup>27</sup>

**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
- 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 întrebările de pe pagina anterioară?

- a. Persoane roma care s-au declarat de etnie roma la recensământul din 2011
- b. Persoane roma care NU s-au declarat de etnie roma la recensământul din 2011
- c. Persoane roma, indiferent de etnia declarată la recensământul din 2011
- d. Nu sunt sigur

---

<sup>27</sup>The interface did not allow users to scroll back to the previous question.