Undercounting the Educated Among Marginalized Groups and Why It Matters: Evidence from the Roma People in Romania*

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

September 14, 2024

Abstract

We study how education affects ethnic self-identification and reporting in historically marginalized ethnic groups. We study Europe's largest ethnic minority, the Roma people. We link three rounds of the census and birth vital records in Romania and track reported ethnicity over time to measure "passing" – when a person changes their previously reported ethnicity. We show that passing is widespread and more prevalent for educated Roma: each year of schooling lowers Roma self-identification by 2.6-3.5%. We rule out that mismatch drives our results. We use a structural model to estimate persistent heterogeneity in Roma ethnicity reporting and infer the number of individuals who were ever reported as Roma. We find 72% more people who are ever reported as Roma than self-identified Roma. For postsecondary graduates, there are six times more individuals than in the census. We collect data on perceptions and show that most Romanians are unaware of these patterns, yet are very confident in their ability to recognize ethnicity, even though they rely on malleable markers of identity. Higher rates of passing among educated minority members, like we find here, have implications for how to interpret government statistics and 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 project has IRB approval from Columbia University.

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

1 Introduction

Group identities are central to individuals' economic and market interactions and outcomes (Akerlof and Kranton, 2000). Identity influences how people supply labor (Oh, 2023), how others act toward a person based on their perceived group (Bertrand and Mullainathan, 2004), and how people from different groups work together (Hjort, 2014). This work often conceptualizes group identity as exogenous, yet increasingly, research in economics recognizes that at the individual level, ethnic identity may become a *choice*, including in official race and ethnicity data. Several studies analyze varied forms of ethnic and racial "passing," whereby individuals of a given group no longer identify with it (Cornwell et al., 2017, Dahis et al., 2019, Rademakers and van Hoorn, 2021).[1]

Widespread passing means that the group of people who ever identify with the group is larger than it appears based on official reporting. More important, passing may be *selective*, based on some socio-economic characteristic such as education. This means that those who still identify with the group are no longer representative of the entire group. This type of passing may be particularly important for reinforcing stereotypes about marginalized and vulnerable groups. For instance, when members with positive traits from a marginalized group opt-out of the group at a higher rate, the composition of the self-identified group will skew towards individuals that lack those traits, which may reinforce negative beliefs about the group.

In this paper, we study selective passing by education: how individuals' reported ethnicity varies with educational attainment. We focus on the Roma people, Europe's largest ethnic minority and a highly stigmatized group (European Union Agency for Fundamental Rights, 2023). In Romania, the setting of our study, the Roma people are discriminated against and have lower educational attainments than the rest of the population. This is true across the distribution of educational attainment and particularly striking for high levels of education. In 2011, 0.9% of self-reported Roma finished post-secondary education, compared to 16% for the non-Roma.

We use several large-scale administrative datasets to explore how ethnic identification changes within Roma individuals over time. First, we link records from three waves of the full-count Romanian census based on exact date of birth, locality code, and sex, focusing on unique, two-sided matches (Dahis et al., 2019). This allows us to measure reported ethnicity in this official setting. Second, we use the universe of birth records and the ethnicity that parents record for their children in these vital statistics. We complement this administrative data with a survey of the general population that we conducted to measure beliefs about and awareness of passing.

Our key empirical result is to establish selective passing by education: among individuals reported as Roma in an earlier census round, those who are more educated are less likely to continue to identify as Roma. That is, initially reported Roma who are more educated are more likely to pass. We show this result using two empirical

[1] Research in social psychology highlights the complexity of passing, discussing *contextual passing*, whereby reported group identity depends on the social environment, and that individuals pass towards socially privileged groups (Albuja et al., 2018) .

strategies: a reduced-form approach and a structural model.

First, we use a reduced-form empirical approach to explore the relationship between passing and education. Our sample is restricted to individuals with reported Roma ethnicity at the baseline census (1992 or 2002) and we explore the relationship between reported Roma ethnicity and education at endline (2011). A key concern with this approach is systematic measurement error from mismatches between census records. Mismatch is a major potential threat because a Roma individual who is incorrectly linked will most likely be linked to a non-Roma individual, who on average has higher education. This would lead to spurious selective passing by education. We also worry about attenuation bias from random measurement error in the education variable.

To address these measurement error issues, we use education from the initial census period (baseline) as an instrument for education in the final period (endline). Among mismatched observations, baseline education is not correlated with endline education nor with passing. This means that a higher rate of mismatch proportionally attenuates both the first stage and the reduced form coefficients on baseline education. Hence, the Wald estimator identifies the correct correlation between endline education and passing.

Our main instrumental variable results suggest that the propensity of self-declaring Roma ethnicity decreases by 2.6pp for each additional year of schooling. This is similar to our OLS coefficient of 3.0pp. We show that between 1992 and 2011, a Roma individual with no formal education has a roughly 80% probability of continuing to identify as Roma in 2011. Our coefficient estaimtes imply that for a high school graduate, this likelihood drops to 48.8%. These results are specific to the Roma ethnic group. Passing for ethnic Hungarians and other ethnicities is much lower and does not vary with education.

We also estimate inter-generational passing using data on reported ethnicity in birth records. We use a perfect linkage between mothers' 2011 census records and their children's birth records, provided by the Romanian National Statistics Institute. We find even stronger differential passing by the mother's education: for each additional year of maternal schooling in 1992, mothers are 5.3pp less likely to declare their child's ethnicity as Roma.

A second way to address the potential mismatch issue is by estimating it directly. We link individuals using only date of birth and locality, and use changes in recorded sex to estimate mismatch rates.[2] Mismatch is low, at around 3-10% for key subgroups. We use these estimated mismatch rates to back out true passing rates by education group and find similar results to our instrumental variables approach.

We find a limited amount of heterogeneity in the degree of selective passing rates by education. Sex, parental education, religion and other markers seem to play a small role. Passing seems to occur at fairly similar rates (2-3% per year of schooling between censuses) across our sub-samples and at fairly constant rates throughout individuals' lifecycles.

Our second approach uses a structural model to estimate the number of people ever

[2]Actual sex changes were extremely rare in Romania in our study period.

reported Roma for each level of education attainment, using the passing behavior in our linked data. Our reduced form estimates so far focus on the sample of individuals reported as Roma in an initial census year. When people differ in their time-invariant propensity to identify as Roma, this sample will be skewed towards those who are more likely to identify as Roma every time. This tends to underestimate passing and may affect our selective passing coefficients. We combine the unique feature that in our setting we observe people report ethnicity up to three times, together with parametric assumptions on the distribution of individual propensity to identify as Roma, to estimate this heterogeneity. We estimate this model using the generalized method of moments by matching the population shares of reported ethnicity over the three census years. Our estimated model implies that the number of Roma in Romania is significantly under-reported, with 72% more "ever reported Roma" than officially recorded Roma. Additionally, there are approximately six times more Roma postsecondary graduates and over twice as many high school graduates compared to official statistics. Our findings suggest that when considering ever reported Roma, selective passing reduces the average educational gap between Roma and non-Roma by 13%.

Is the general population aware of these patterns? To shed light on this question, we designed an online survey of the general Romanian population with 2,000 respondents. We find no evidence that on average our respondents are aware of the fact that passing in census data is concentrated among more educated Roma. These results are robust to offering incentives for correct responses, priming respondents about passing among the Roma, focusing on respondents who are confident in their answers, and restricting to samples that pass attention checks. Respondents have highly dispersed beliefs on the correlation between passing and education, and their responses are internally consistent so do not seem to be driven by noise.

Our second result from the survey data is that respondents are typically confident at being able to identify Roma who intend to pass. Respondents describe several markers - such as individuals' way of speaking, physical aspect and clothing, which are malleable characteristics or stereotypes - through which they believe Roma individuals are identifiable even when passing. Together with ethnographic evidence on successful passing (Pantea, 2014), these results suggest that selective passing is a viable and potentially successful strategy.

While social science disciplines have recognized that individuals employ varied strategies regarding identification with their racial and ethnic identity, economists most often have treated race and ethnicity as exogenous and fixed over an individual's lifetime. This study is linked to a small but growing number of empirical studies in economics that document the correlation between identity and social and economic incentives or determinants, such as Antman and Duncan (2015), Cassan (2015), Duncan and Trejo (2011), Cornwell et al. (2016), and Dahis et al. (2020). These studies demonstrate variations in self-reported racial or ethic identity influenced by affirmative action, caste-based land policies, economic status, and wage changes. Moreover, historical analysis by Dahis et al. (2020) highlights "passing" as a response to severe discrimination, where many African American men identified as white between 1880

and 1940. This evolving perspective on identity is complemented by research on peer effects and social learning, as Bursztyn and Yang (2021) document how individuals' perceptions of others influence their behaviors and attitudes, despite widespread misperceptions in various societal contexts.

2 Setting and Data

2.1 Setting

With an estimated 10 to 12 million members in Europe, the Roma people are the continent's largest minority group and a heavily stigmatized and discriminated group (European Union Agency for Fundamental Rights, 2023). The Roma people have a history of oppression, including in Romania. In two of the three historical regions of present-day Romania, Roma people were enslaved from the 1300s to the mid-1800s. The Nazis and their allies directly targeted this population during the Holocaust, with 200,000 to 500,000 Roma perishing across Europe. In Romania, approximately 25,000 Roma people were deported and 11,000 killed (Kelso, 2017, Wiesel, 2004). More Roma were forcibly displaced during the communist period and forced to give up their traditional crafts and work in the industrial sector.

In Romania, Roma are an diverse population in terms of dialects, faiths, and traditions. Today, the vast majority of Roma families are settled. Officially, at the 2011 census, more than 600,000 Roma were tallied, representing 3.3% of the total population. On the other hand, some estimates suggest that the real number of Romanian Roma is closer to 1.85 million, or more than 8% of the population, with many Roma individuals not declaring their Roma ethnicity (Zamfir and Preda, 2002).[3]

The Roma people in Romania and other European countries face severe discrimination, residential and educational segregation, and high rates of material deprivation, maternal and infant mortality, poverty, and unemployment (European Union Agency for Fundamental Rights, 2023). They experience significantly lower life expectancy and limited educational access compared to majority populations. In Census data from 2011, Roma adults over 25 have an average of 5.76 years of schooling, compared to 10.59 years for non-Roma. Over 20% of Roma adults have no formal education, while only 0.9% complete postsecondary education, compared to 16% of non-Roma Romanians.

2.2 Identity Choice and Passing

In this section, we discuss how passing has been conceptualized outside economics.

Passing is often defined and studied as an interactive act of presenting oneself and being acknowledged as a member of a different group (Dawkins, 2014, Definition 8). This particular understanding of passing involves an individual passing and an audience and members of an in-group, leading to a "triangular theater of identity"

[3]Zamfir and Preda (2002) obtained this number for the share of Roma in Romania in 1998 by extrapolating *hetero-identified* ethnicity based on data from the catchment areas of 200 randomly selected polling stations.

(Robinson, 1994). In our paper, we study this type of interaction in the specific context of ethnicity reporting between the individual, the enumerator, and the state, for the population census and birth records vital statistics.

Passing can also be understood solely as the "presentation of self" - the action of presenting oneself as a member of a social group one does not belong to (Goffman, 1959). In this context, we could consider census reporting as forms of passing per se, acknowledging that it refers to a particular institutional process and context.

A widely practiced phenomenon around the world, passing is often driven by challenges to gain acceptance into more privileged groups and upward social mobility. While it's commonly associated with racial passing, it is also utilized by individuals from historically marginalized groups, as well as across other social identities such as class, gender, sexuality, religion, or nationality. Passing is understood as proof of the fluidity and diversity of social identities (Robinson, 1994)

Racial and ethnic passing is particularly prevalent and credible when there is a lack of or limited evidence or trace of differences in identity markers or physical, social, and cultural symbols, such as skin complexion, clothing, wealth, or local community history (Dawkins, 2014). In the case of the Roma people in Romania, the variety of ethnic, social, and cultural identity markers and symbols, makes it possible for some Roma individuals to present themselves as members of a more privileged ethnic group Pantea (2014). For example, an individual with lighter skin who lives isolated from their Roma family and community – the evidence of and connection to their ethnic origins – can pass more easily.

The expressions of passing amongst Roma people range on a spectrum between concealing and repression - from individuals amending their Roma identity in particular contexts, situations, and relations to Roma individuals constantly adopting a new, more privileged ethnic identity, particularly Romanian.

2.3 Ethnographic and Other Evidence of Passing among the Roma

Passing among Roma people has been qualitatively documented in ethnographic and other social sciences studies. These studies identify and highlight several reasons why Roma decide to pass, the strategies used, and hint towards passing being widespread. For example, Pantea (2014) examines examples of passing at the workplace, including one individual with lighter skin complexion reporting that she concealed her Roma identity to avoid being dismissed like her darker-skin colleagues. Marin (2023) points to the historical causes of passing, including the Holocaust, as a cause of concealment, particularly because Romanian Roma were deported to Transnistria with the help of census data. Marin's family has spoken the Romani language for generations, self-identify as Roma in their social networks, and continue to practice varied traditions; thus for them, passing is situational, happening mostly in official reporting (Marin, 2023). Other personal reports suggest that Roma are often stereotyped as thieves or beggars, which can lead to a sense of shame or inferiority (Grigore, 2012). The very exonym Romanians use for Roma, namely "t*gan" (G*psy in English) has a pejorative connotation and is often used to offend. In such contexts, even children are coached to withhold their identity and modify their ap-

pearance, clothing, and acting to avoid being labelled and seen as a "G*psy." Yet, in the case of Grigore (2012), we can observe an alteration from passing, a shift from concealing to self-identification and Roma-related writing and advocacy, pointing to the fluidity not only of identity but also of passing.

2.4 Data

Census Data: Our study's main data source is the 1992, 2002, and 2011 Romanian full count censuses. We focus on census questions regarding education, ethnicity, religion, native language, and family characteristics. Information on place of birth, current and previous locality of residence, exact date of birth and sex, and previous town of residence enables us to uniquely link individuals across census years and track changes in ethnicity reporting over time. Family-level variables also allow us to study intergenerational passing, including identifying individuals with links to the Roma ethnicity based on familial declarations in previous or current census waves.

Birth Registry Data: We analyze birth registry data encompassing all births in Romania from 2003 and between 2005 to early 2011. The Romanian National Statistics Institute (INS) links mothers of these children to the 2011 census using their personal identification numbers, ensuring a perfect match. This setup allows us to examine inter-generational passing. Unlike the census, which involves face-to-face interviews with typically male household heads conducted by an enumerator, birth records are individually filled out by mothers shortly after giving birth.

Survey Data: We conducted an online survey with 2,000 respondents to gauge beliefs and perceptions about Roma and passing among the general Romanian population. We describe these data in more detail in section 6 when we analyze the general population's beliefs and perceptions about passing.

2.5 Linking Census Records and Estimating Mismatch

We link individual census records across three waves in Romania. Tracking changes in reported Roma ethnicity over time is a key step for us to measure passing. Matching criteria include date of birth, sex, and place of residence (city, town, or village), using data from the 2011 census on previous residence and moving dates to link individuals who relocated between census periods. Appendix B.1 has details of our procedure.

We focus on unique two-sided matches as in Dahis et al. (2020). We retain an individual record in census t if and only if: (i) no other individual of the same gender in the same town shares their birth date, and (ii) they can be matched to exactly one record in census $t+1$ with the same gender, birth date, and town of current or previous residence.

We match 29% of the 2011 stable population of Romania to the 1992 census, and 39% when we match to the 2002 census.[4] A key reason for excluding observations

[4]If we consider only individuals who were born before the earliest census, our match rates are 32%

from this linked sample is when multiple individuals in a locality are born on the same date and have the same sex, which is more likely to happen in larger towns and cities. (We plan to extend our analysis to matched *groups* in a future version of this paper.) As a consequence, smaller localities are over-represented in our linked sample (Figure A.1). This figure shows that our sample skews towards a more rural, slightly less-educated population.

A major threat to identification is that mismatched records can bias our passing measure towards selective passing by education. We develop a new method to directly estimate the rate of mismatch, which allows us to de-bias our statistics of any potential influence of mismatch. We link individuals based solely on date of birth and locality, then use changes in reported sex as a proxy of mismatch. Assuming stable sex reporting over time and no reporting errors, inconsistent sex across census waves indicates an incorrect link. We estimate mismatch rates by doubling the rate of inconsistent sex, considering that half of the mismatches will have consistent sex by chance.

Our linked records exhibit low mismatch rates (Table A.1). Our correctly matched rates are in excess of 94%, which compares favorably with the 6 to 7% false positive rate observed in recent linkage procedures for historical US Censuses (Price et al., 2021). Using these mismatch rates, we can adjust our measured passing rates for mismatch. Intuitively, the passing rate we measure in the data is a weighted average between "true" passing and spurious passing driven by mismatched records. Assuming conditionally random mismatch, we can back out the true passing rate using the estimated mismatch rates. This is described in more detail in Appendix B.1. In latter sections, we discuss additional methods of addressing mismatch.

3 Main Finding: More Educated Roma Are More Likely to Pass

In Figure 1, we show the persistence in self-declared ethnicity – the opposite of passing – across two census waves, by education groups and by occupation groups.

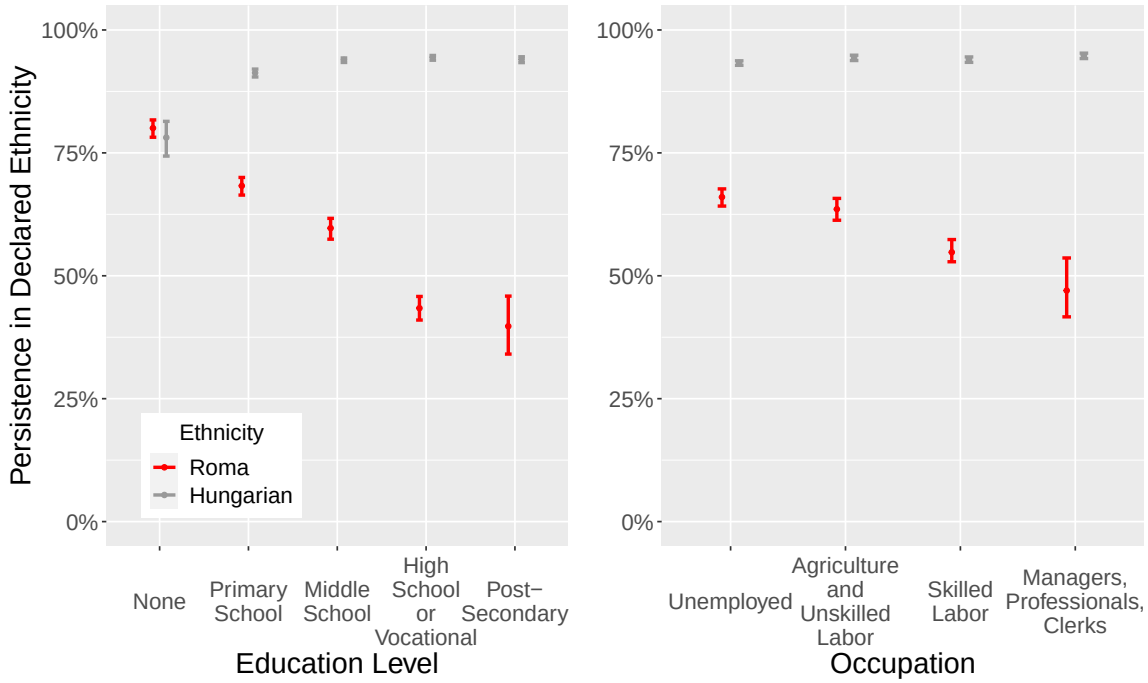
To construct this figure, we start with the sample of respondents who are reported as Roma in 1992 (either by themselves or by others in their household). This is our baseline measure of Roma ethnicity. We then link these responses to the same individuals in 2011, some of whom have moved to form new households, and record their reported ethnicity in 2011, 19 years after our baseline data. We repeat this analysis by education and occupation group. In order to gain precision in these classifications, we measure these attributes in 2011. All our estimates correct for mismatch using the method described in the previous section.

We find that more educated Roma, and those in occupations associated with more education, are more likely to "pass" as Romanian in later censuses. Respondents who self-declare as Roma at baseline (in the 1992 census) are 2.5 to 3.4 pp less likely to continue self-identifying as Roma in the 2011 census with each additional year of schooling completed. Indeed, in the first panel, we show that baseline Roma without

and 40%, respectively. Using this metric, 26% of all individuals that can be matched are uniquely matched across 3 censuses.

formal schooling in 2011 have an 80% probability of still reporting as Roma in 2011. For postsecondary graduates, this probability drops to 40%. In the second panel, we show that baseline Roma who are unemployed or out of the labor force, employed in agriculture or as unskilled labor continue to self-report Roma ethnicity at a 66-69% rate. Managers, professionals in occupations requiring a university degree and clerks, meanwhile, only continue to self-identify as Roma in the 2011 census at rates of 47-58%.

Figure 1: Persistence in Declared Ethnicity by Education and Occupation (1992 - 2011)



Notes: This figure reports persistence in declared ethnicity between 1992 and 2011 using our linked sample. All estimates are corrected for mismatch using mismatch rates estimated using inconsistent sex. Figure A.4 shows results for 2002-2011 instead of 1992-2011. Figures A.2 and A.3 show results for other ethnicities.

The pattern is specific to the Roma minority. To show this, we repeat the analysis for the Hungarian minority, the largest ethnic minority in Romania. We observe essentially no passing for Hungarians, who display a high persistence in their self-identifying as Hungarian at over 90% for most categories other than "no education." There is no education gradient in education of this behavior.[5] Figure A.2 shows results for other ethnic minority groups, who have a moderate degree of passing and no detectable gradient in education.

[5]These patterns are not surprising given the context. Hungarians in Romania have distinctive names, are geographically clustered in a small number of counties, have access to public services and education in Hungarian and have historically strong institutions in the area.

Results are very similar when using 2002 as the baseline year (Figures A.4 and A.3). Passing rates between 1992 and 2011 slightly exceed those between 2002 and 2011 at every education level, suggesting passing is a gradual process. The education gradient of passing is slightly steeper between 1992 and 2011 than between 2002 and 2011, indicating that more educated individuals are increasingly likely to pass over time. We revisit the question of passing over the lifecycle below.

4 Empirical Strategy: Passing and Education Levels

We next turn to a regression based analysis of the link between passing and education. As we did for the results in the previous section, we use a strategy to ensure our results are not driven by mismatch. The regression analysis also allows us to easily conduct heterogeneity analysis in passing behavior.

We quantitatively explore the link between passing and education by focusing on individuals with reported Roma ethnicity at baseline (either in the 1992 or 2002 census) and examine how their reported ethnicity correlates with their education level by 2011. We are interested in the following linear model:

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

Here, $Roma_{i,t=1}$ is an indicator (1 or 0) of declaring Roma ethnicity in the 2011 census, $Educ_{i,t=1}$ represents years of schooling measured in 2011, and $\phi_{L(i),YOB(i)}$ are locality by year of birth fixed effects. A negative β coefficient suggests that more educated individuals are more likely to identify as Romanian rather than Roma.

Several issues may impact this analysis. First, erronously 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 the Roma, this will lead us to incorrectly infer that more educated Roma are more likely to pass for Romanians. This will bias β away from zero. 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. Under a weak assumption that mismatch happens conditionally randomly within locality and year of birth – captures by our locality by year of birth fixed effects – baseline education is not systematically correlated with passing nor with endline education for mismatched records. This means that our first stage – the relationship between endline education and baseline educations – and our reduced form – the relationship between passing and baseline education – will both be attenuated proportionally by the mismatch rate. This implies that the instrumental variable estimates of β will be unbiased even in the present of mismatch.

Table 1 shows our main results. In columns 1 we report the OLS estimates of equation (1), and we find $\beta = -0.03$, meaning that each additional year of schooling is associated with a 3.0 pp decline in the probability of reporting Roma ethnicity. Evaluated at the endpoints of the $Educ_{i,t=1}$ variable, this means that a person reported as Roma in 1992 has a 48.0 percentage point lower probability of reporting

Table 1: Persistence in Declared Ethnicity by Education

	Dependent Variable: *Reported Roma Ethnicity at Endline*			
	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)
Schooling Yrs	-0.030***	-0.026***	-0.053***	-0.035***
	(0.001)	(0.002)	(0.011)	(0.012)
Sample	'92-'11	'92-'11	'92-Birth	'92-Birth-'11
Outcome	Individual	Individual	Child	Mom
N	110,316	107,168	19,707	19,707
R^2	0.56	0.56	0.67	0.71
DV Mean	0.60	0.60	0.30	0.62
F-stat		11,200.1	424.0	424.0

This table shows the relationship between declaring Roma ethnicity and years of schooling, using 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.

Roma in 2011 if they are an university graduate relative to someone without any education.

Our IV estimates using education in 1992 as an instrument, holding the estimation sample unchanged, are in column 2. The instrument is strong, and the coefficient drops slightly to -0.026, which slightly lowers the above number to a 41.6 percentage points difference. This result confirms that mismatch in our census linking procedure is not responsible for these results.[6]

We next analyze inter-generational passing by looking at a matched sample of mothers reported as Roma in 1992, their children's birth records between 2005 and 2011, and the mother's record in the 2011 census. We define "passing" as the child's ethnicity not reported as Roma on the birth record in column 3, and as the mother's ethnicity not reported as Roma in the 2011 census in column 4. Both columns report instrumental variables estimates. We find that each additional year of the mother's schooling is associated with a 5.3 percentage point decline in the probability that the child's ethnicity is reported as Roma (column 3). This magnitude is especially striking given the low average rate of reporting children in this sample as Roma, at only 30%. These results show that there is a steep gradient of inter-generational passing with respect to maternal education. For the mother's reported ethnicity in

[6]We show additional evidence for this result in Table A.3, where we focus on results from a matched sample of mothers whose reported ethnicity is recorded in both the 2011 census and their children's birth records (from 2003 to 2011). The linking was done by the Romanian National Statistics Institute based on national ID number, which was recording in the 2011 census, and hence is a perfect matching. The results show that, as in our main results, passing between the birth records and census records is correlated with education. Our estimate of β is smaller in this case, because our sample is defined based on Roma reported ethnicity in the birth records, which is rare and hence a stronger marker of identification with the ethnicity.

2011, we find $\beta = -0.035$, which is a third larger than our benchmark coefficient in column 2.

Our results have several implications. The minimal implication is that, in official statistics, both the number and educational attainment of people with Roma background are underestimated. This would holds even if passing in these official documents were not to be correlated with passing behavior in day-to-day life. This is relevant given the policy importance of these figures. For example, in Romania the fact that there are fewer than 1% of Roma with university education in Census statistics has attracted a lot of attention. Our results show that the numbers are much larger, as this group, even if they self-identified as Roma at baseline, only continue doing so less than 50% of the time.

Under the assumption that the passing patterns observed in the data also apply to day-to-day behavior and that passing as non-Roma can be successful in real life, our results have deeper implications. To the extent that stereotype creation and statistical discrimination is partly based on observational learning, high-education Roma passing as non-Roma may lead to overly-negative assumptions about Roma education and occupational status. In the following sections, via a survey, we show that there are potential benefits to passing, that successful passing is possible, given that malleable characteristics are used by non-Roma to identify Roma and that Romanian respondents are unaware of the link between passing and education. We also investigate heterogeneity in the relationship between passing and education.

4.1 Determinants of Passing and Selective Passing

In Table 2, we investigate the heterogeneity in passing behavior to better understand potential mechanisms.

In Panel 1, we find that individuals with more educated mothers are more likely to pass as non-Roma, with this likelihood increasing significantly with their own education, while paternal education shows no such pattern. Marriage plays a crucial role in assimilation, as those with non-Roma spouses are significantly more likely to pass than those with Roma spouses, with a steeper education gradient for those married to non-Roma.

Panel 2 examines the relationship between passing and personal characteristics such as sex, migrant status, native language, and religion. Females, non-Romani-speakers, and followers of the majority Christian Orthodox faith exhibit steeper education gradients for passing compared to males, Romani-speakers and people of other religions. Romani speakers show a stronger attachment to Roma identity, with only 27% passing, in contrast to 45% among non-Romani speakers. Meanwhile, migrants are more likely to pass (57%) compared to non-migrants (37%).

Table 2: Heterogeneity in Passing

	Panel 1: Passing and Family Characteristics *Dependent Variable: Reported Roma Ethnicity (2011)*							
	Mom Education		Dad Education		Roma Parents		Spouse	
	≥ 8 Yrs	< 8 Yrs	≥ 8 Yrs	< 8 Yrs	Both	Not Both	Roma	Non-Roma
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Schooling Yrs	-0.030***	-0.018***	-0.023***	-0.020***	-0.025***	-0.026***	-0.021***	-0.033***
	(0.005)	(0.003)	(0.004)	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)
N	29,334	70,003	38,855	49,580	79,973	131,931	133,897	78,007
R^2	0.58	0.55	0.56	0.59	0.50	0.56	0.57	0.57
F-stat	1,465.7	4,736.7	2,246.1	2,588.5	5,069.5	28,702.4	25,753.1	5,885.3
Roma	0.53	0.64	0.57	0.64	0.62	0.62	0.74	0.41

	Panel 2: Passing and Individual Characteristics *Dependent Variable: Reported Roma Ethnicity (2011)*							
	Sex		Migrant		Native Language		Orthodox Religion	
	Male	Female	Yes	No	Romani	Non-Romani	Yes	No
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Schooling Yrs	-0.025***	-0.033***	-0.013	-0.026***	-0.017***	-0.026***	-0.027***	-0.015***
	(0.002)	(0.002)	(0.009)	(0.001)	(0.003)	(0.002)	(0.002)	(0.004)
N	107,804	104,100	11,263	200,641	82,395	129,509	175,842	36,062
R^2	0.57	0.59	0.83	0.49	0.52	0.53	0.50	0.59
F-stat	15,126.4	12,136.2	431.0	31,760.3	7,501.7	23,150.6	25,850.8	4,813.7
Roma	0.63	0.61	0.43	0.63	0.73	0.55	0.62	0.62

	Panel 3: Baseline Age and Household Head Status *Dependent Variable: Reported Roma Ethnicity (2011)*						
	Head Baseline	Not Head Baseline	Head Always	10-20 y.o. Baseline	20s Baseline	30s Baseline	40+ y.o. Baseline
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Schooling Yrs	-0.021***	-0.029***	-0.022***	-0.025***	-0.026***	-0.025***	-0.024***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	50,018	161,011	43,508	57,294	46,876	33,565	36,741
R^2	0.69	0.49	0.71	0.52	0.55	0.60	0.66
Roma	0.63	0.62	0.65	0.61	0.64	0.64	0.60
F-stat	13,541.8	18,133.5	12,821.3	5,874.1	10,951.6	10,184.6	10,129.2

Each column represents a regression estimated on a different sample, where the dependent variable is an indicator for self-reporting Roma ethnicity in the 2011 census. The main sample is 1992 - 2011 linked census records who with reported Roma ethnicity at baseline. *p<0.1; **p<0.05; ***p<0.01

In Panel 3, we explore heterogeneity by household head status and age. If household heads respond to the census for the entire family, the reported ethnicity may differ compared to what individuals would report themselves. This may be especially true for young children. We find little difference in passing rates by household head status or age, though non-household heads pass slightly more with additional education. We split the sample by age in 1992, and find that passing occurs consistently across different age groups as young as 10 years old to 40 years or older. This suggests that passing is a continuous and probabilistic process throughout individuals' lives,

rather than a change concentrated in a particular stage of life.

5 Model: Estimating Roma Population and Education Levels

Now, we shift focus to estimating the size of the population with a Roma background and its educational distribution. Our estimates so far do not directly enable us to do this because in all the analysis so far, we condition on a sample of individuals who declare Roma at baseline. If people differ in a time-invariant likelihood of declaring themselves as Roma, this initial sample will be skewed towards those with high Roma likelihood. This will under-estimate passing rates overall and may bias the gradient of passing with education.

In this section, we use a model where we estimate the degree of heterogeneity in time-invariant propensity to identify as Roma using the fact that we observe (up to) three rounds of self-identification.

5.1 Model Setup and Estimation

Our model has individuals i with education group e. Some individuals are connected to the Roma ethnicity in the sense that at some point they have a probability to report (or have reported by someone in their household) their ethnicity as Roma. For brevity we henceforth call these individuals "ever reported Roma" and write $i \in \mathcal{R}$. We use this definition to recognize individual's ethnic self-identification, while tracing connections to the Roma ethnicity independently from self-reported ethnicity. The share of "ever reported Roma" in group e is $p_e^{\mathcal{R}}$. We call all other individuals non-Roma.

We observe self-reported ethnicity in three census periods $t = 0, 1, 2$. An ever reported Roma individual i in period t declares themselves as Roma with i.i.d. probability

$$\pi_{iet}^{\mathcal{R}} = \Pr(Roma_{iet} = 1 \mid i \in \mathcal{R}) = c(\pi_{et} + \pi_i)$$

where the function $c(x) = \min(1, \max(0, x))$ returns x censored to the interval $[0, 1]$. Self-reported ethnicity depends on education (e) and time period (t) through π_{et}, and on individual i's time-invariant propensity to self-identify as Roma π_i, which we assume is drawn i.i.d. from a normal distribution with mean 0 and standard deviation σ_e. (We vary the shape of this distribution in sensitivity analysis.) 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 pair of time periods and reported ethnicity in both periods. See Appendix C for details.

We estimate the share of ever Roma, $p_e^{\mathcal{R}}$, and the passing parameters π_{et} and σ_e, as well as mismatch probabilities, using panel data $(Roma_{it}, Sex_{it})_{i,t=0,1,2}$ on reported ethnicity and sex. We use the generalized method of moments (GMM) targeting the eight shares of ethnicity triples $(Roma_0, Roma_1, Roma_2) \in \{0, 1\}^3$ and the shares of inconsistent sex by a pair of census periods that includes 2011 and reported ethnicity in all three time periods. For inference, we bootstrap the entire procedure at the town level.

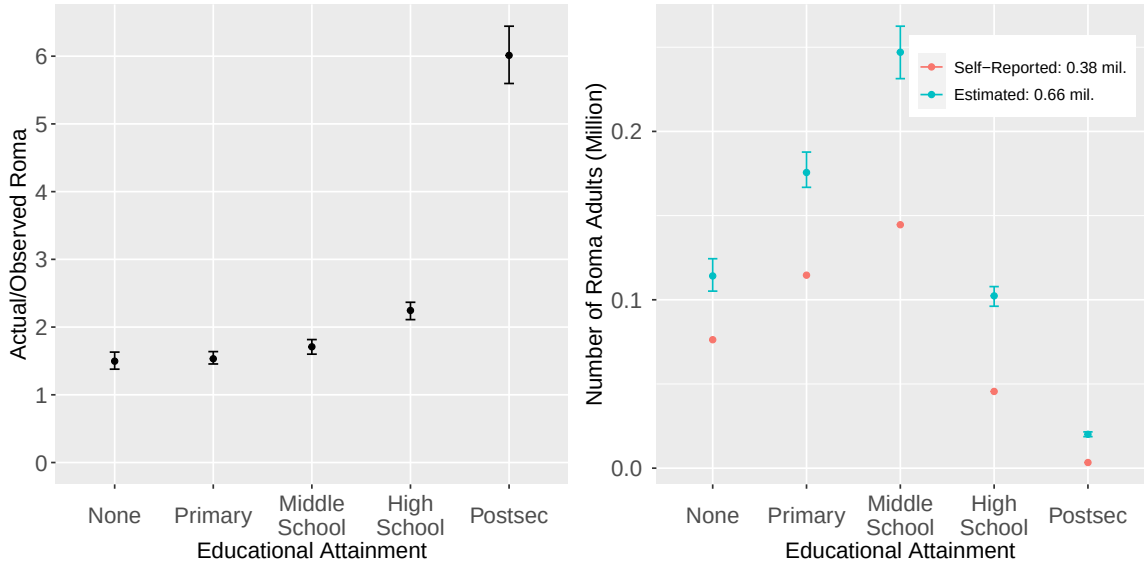
To understand how our model is identified, consider first the case with no heterogeneity ($\sigma_e = 0$) and no mismatch, and two census rounds $t = 0, 1$. We can recover the mean passing probability using a conditional probability that we observe in the data:

$$\widehat{\overline{\pi}}^{\mathcal{R}}_{ie1} = Pr(Roma_{i1} \mid Roma_i = 1).$$

The right hand side is the share of self-reported Roma conditional on self-reported Roma at $t = 0$. This corresponds to the reduced form analysis in previous sections. We can recover the share of ever reported Roma from the share of reported Roma: $\widehat{p}^{\mathcal{R}}_e = \overline{Roma_{ie1}}/\widehat{\overline{\pi}}^{\mathcal{R}}_{ie1}$.

In the presence of heterogeneity ($\sigma_e > 0$), $\Pr(Roma_{i1} \mid Roma_i = 1)$ is an underestimate of $\overline{\pi}^{\mathcal{R}}_{ie1}$, because the sample of reported Roma at $t = 0$ will over-represent individuals with large π_i. Because we have *three* observations per individual, we can detect statistically persistence in an individual's likelihood to declare Roma ethnicity and estimate σ_e. Appendix C has more formal details and a more extensive discussion of identification.

Figure 2: Estimated Roma Population



Notes: The left figure plots the ratio between all ever reported Roma individuals – as inferred in the estimated model – and the number of reported Roma in the 2011 census. This is $1/\widehat{\overline{\pi}}^{\mathcal{R}}_{e,2011}$, the inverse mean probability to report Roma ethnicity of ever reported Roma individuals. The right figure plots the number of reported Roma in the 2011 census (red) and the model estimate of the number of all ever reported Roma (blue).

5.2 Estimation Results

Our estimation results follow. First, the model fit is good, despite being over-identified (Figure A.5). Targeted and model moments are very close, with a mean absolute deviation of 0.0014.

The estimated model suggests that the number of individuals with Roma backgrounds in Romania is significantly under-reported. In the left panel of Figure 2, we find that there are roughly six times more Roma postsecondary graduates than in official data, more than twice as many high school graduates, and close to 1.5-2 times more middle school graduates, primary school graduates and Roma without formal schooling.

In absolute terms (right panel), there are roughly 275,000 more adult Roma than reported at the 2011 census in our matched sample . This means that within our matched sample, the proportion of Roma 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 ever reported Roma adults completed 6.2 years of schooling, versus a reported 5.8 years of schooling. This represents an under-reporting of Roma average education levels of 11%. Accounting for this explains roughly 13% of the Roma - non-Roma education gap.

6 General Population Beliefs about Passing: Survey Evidence

So far, we showed that self-reported Roma have lower education than the full set of individuals ever reported Roma.

What does the general population believe about passing and its relationship with education? If passing is successful and the general population is unaware of passing and forms beliefs uniquely based on self-reported Roma, this could negatively affect beliefs about the Roma in general.

We collected data on beliefs about Roma passing and its relationship with education from 2,000 Romanian respondents of an online survey. Respondents were drawn from a database of a major Romanian survey firm and stratified by nine regions in Romania. Similar to other online surveys, the sample over-represents urban areas and educated respondents.

In our survey, we elicit participant beliefs about selective passing. To do this, we inform all participants that 50.8% of self-reported Roma individuals in the 2011 census graduated middle school. We then elicit beliefs about the share of middle school graduates among Roma individuals more broadly, regardless of their reported ethnicity at the census, and offer respondents financial incentives for correct answers.[7]

Our headline result suggests that Romanians, on average, are not aware that more educated Roma are more likely to pass. The first row in Table 3 reports results from our incentivized measure and shows that only 28.3% of our sample believes that Roma who pass are more educated than those who are self-declared as Roma. If anything,

[7]In our instructions to respondents, we specifically define what we mean by "Roma individuals" as someone whose ethnicity was reported in the earlier 1992 census as Roma. This follows the approach in our reduced form analysis. The specific questions we use are included in Appendix **XXX**.

a larger share incorrectly believes that Roma who pass are less educated (42%). A further 29.6% believes there is no selection either way.

One possible concern with this result is that respondents do not understand exactly what we ask about. We rule out this explanation using three strategies (Table 3). First, we randomly select respondents in half of our sample and ask a series of questions that explicitly refer to Roma who pass in the Census, before the incentivized question discussed above. We call these respondents the "priming" group. Priming group respondents are more likely to answer that Roma people who pass are more likely to be educated (35.4% vs 21.3%), yet overall this group is still evenly split between the different response options. Second, right after the incentivized question, we ask respondents to recall which group of people our question referred to. Around half of our sample answers correctly, yet they are about as likely to correctly answer "more" (37.8%) as to incorrectly answer "less" (39.1%). Third, we also ask all respondents how confident they are in their response to our incentivized question. The answers are essentially similar among confident and not confident respondents.

Respondents seem to have internally consistent theories of how passing and education are related. For the passing sub-sample, we elicit this correlation in two different ways. First, we ask whether more educated Roma are more or less likely to pass. Second, we ask whether Roma who pass are more or less educated. Respondent answers to these two questions are internally consistent 76% of the time, yet respondents disagree on the direction (Table A.4).[8]

The second main result from our survey is that respondents are highly confident in their ability to identify Roma who pass. First, 68% of respondents say it is very easy or relatively easy to tell if a person is Roma if that person is not declaring their ethnicity. We then ask how it is possible to tell whether someone is Roma. The most often cited markers of Roma ethnicity are way of speaking (68%), physical look (55%), skin color (51%), names (29%), and clothing (27%). Only 25% of respondents say they use a person's family or community background. Arguably, the top mentioned traits are malleable (manner of speaking, of dressing, and physical look). Moreover, in Romania there is a considerable degree of overlap in skin tones between Romanians and Roma, and the Roma people do not have distinctive names. This suggests that passing by manipulating these features may be successful, and educated Roma may pass more easily.

[8]Respondents also had to fill out open text responses explaining their answers to both of these questions. In these responses, respondents typically proposed "theories" that supported their response. One of the most commonly mentioned words was "shame," yet some respondents think that more educated Roma are less ashamed of their ethnicity, while others claim the contrary.

Table 3: Beliefs Regarding Roma Passing and Education

	"Passing Roma Are ____ Educated Compared to Self-Declared Roma (%)" (Incentivized)			
	Less	Equally	More	Obs.
Full Sample	42.0	29.6	28.3	2,000
Priming	39.3	25.3	35.4	1,000
No Priming	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
Confident	42.2	29.4	28.4	1,200
Not Confident	41.8	30.0	28.2	800

Note: Survey responses to an incentivized question about education of passing and non-passing Roma. "All" indicates all respondents. "Priming" are respondents who were primed with questions and information about specific groups of passing and non-passing Roma. "Attentive" indicates respondets who passed an attentiveness check. "Confident" indicates respondents who were confident about their answers.

7 Discussion

In this study, we highlight selective passing by education among the Roma community in Romania. We have demonstrated that Roma individuals with higher levels of education are more likely to underreport their ethnicity in official records, a trend that leads to significant underreporting of Roma numbers and their educational attainment. We estimate that in Romania, there are roughly 72% more Roma-background adults than in official statistics. In particular, there are six times more Roma post-secondary graduates and more than two times more high school graduates.

We find that the Romanian population is not aware of the link between education and passing. Additionally, we show suggestive evidence that markers used to identify Roma, such as speech and clothing, are malleable and may result in lower costs of passing for educated Roma.

We document these results for the largest minority group in Europe, the Roma people. Our results might extend to other marginalized groups, although the precise pattern of passing by education or other positive traits is an empirical question that may differ by setting.

Our study contributes to the literature on endogenous identity choice, and our main contribution is to highlight how large-scale and selective passing may lead the subset of an ethnic group who identifies with that group to be significantly different than the wider ethnic group. When the general population is not aware of these pat-

terns, this may contribute to negative beliefs about and attitudes towards the entire ethnic group, which may ultimately lead to a a vicious cycle feedback by encouraging passing.

References

George A Akerlof and Rachel E Kranton. Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753, 2000.

Analia F Albuja, Diana T Sanchez, and Sarah E Gaither. Fluid racial presentation: Perceptions of contextual "passing" among biracial people. *Journal of Experimental Social Psychology*, 77:132–142, 2018.

Francisca M Antman and Brian Duncan. Incentives to identify: Racial identity in the age of affirmative action. *The Review of Economics and Statistics*, 97(3):710–713, 2015.

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.

Guilhem Cassan. Identity-based policies and identity manipulation: Evidence from colonial Punjab. *American Economic Journal: Economic Policy*, 7(4):103–131, 2015.

Christopher Cornwell, Luis Rivera, and Ian Schmutte. Racial identity and labor market outcomes. *Journal of Human Resources*, 51(3):656–683, 2016.

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.

Renan Dahis, Claudia Olivetti, and M Daniele Paserman. Passing as white: Economic behavior and the social construction of racial identity. *American Economic Journal: Applied Economics*, 12(3):87–129, 2020.

Ricardo Dahis, Emily Nix, and Nancy Qian. Choosing racial identity in the united states, 1880-1940. Technical report, National Bureau of Economic Research, 2019.

Marcia Alesan Dawkins. *Clearly invisible: Racial passing and the color of cultural identity*. Baylor University Press, 2014.

Brian Duncan and Stephen J Trejo. Who remains Mexican? Selective ethnic attrition and the intergenerational progress of Mexican Americans. *Demography*, 48(2):152–178, 2011.

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.

Erving Goffman. *The presentation of self in everyday life*. Anchor Books, 1959.

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.

R. 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.

Suanna Oh. Does identity affect labor supply? *American Economic Review*, 113(8): 2055–2083, 2023.

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.

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.

Amy Robinson. It takes one to know one: Passing and communities of common interest. *Critical inquiry*, 20(4):715–736, 1994.

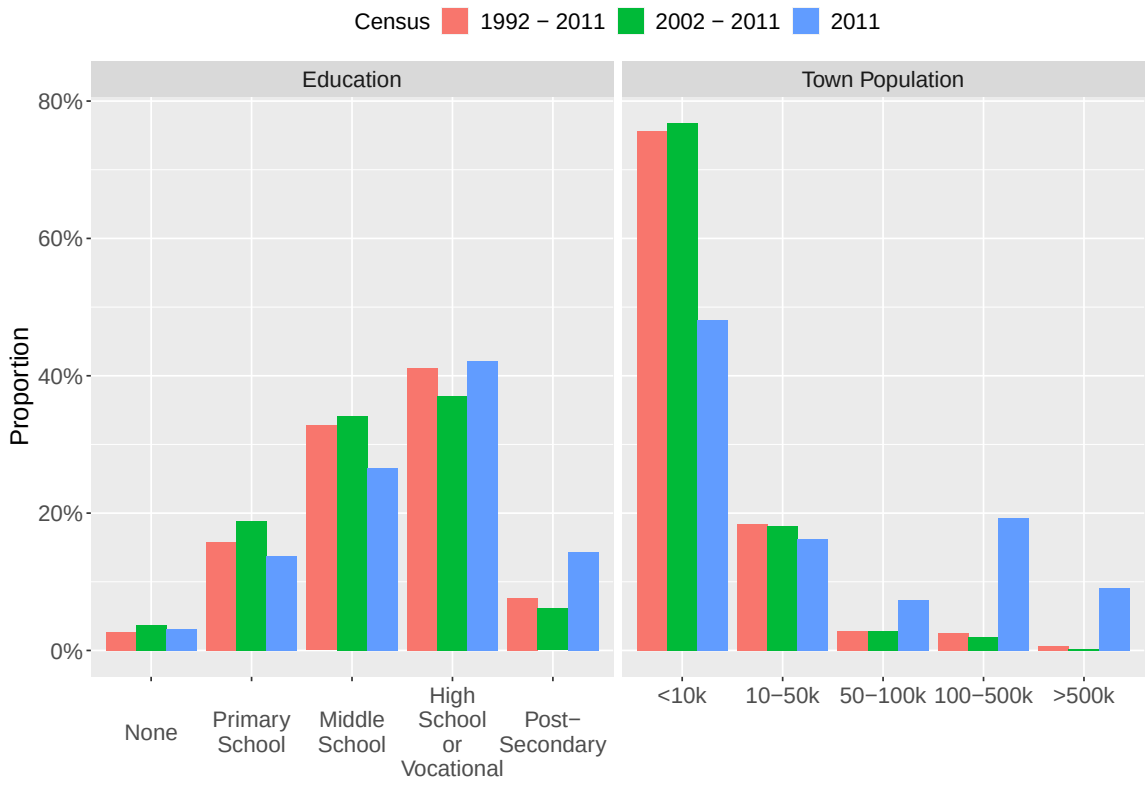
Elie Wiesel. Final report. Technical report, International Commission on the Holocaust in Romania, 2004.

Cătălin Zamfir and Marian Preda. *Romii în România*. Editura Expert Bucharest, 2002.

A Online Appendix

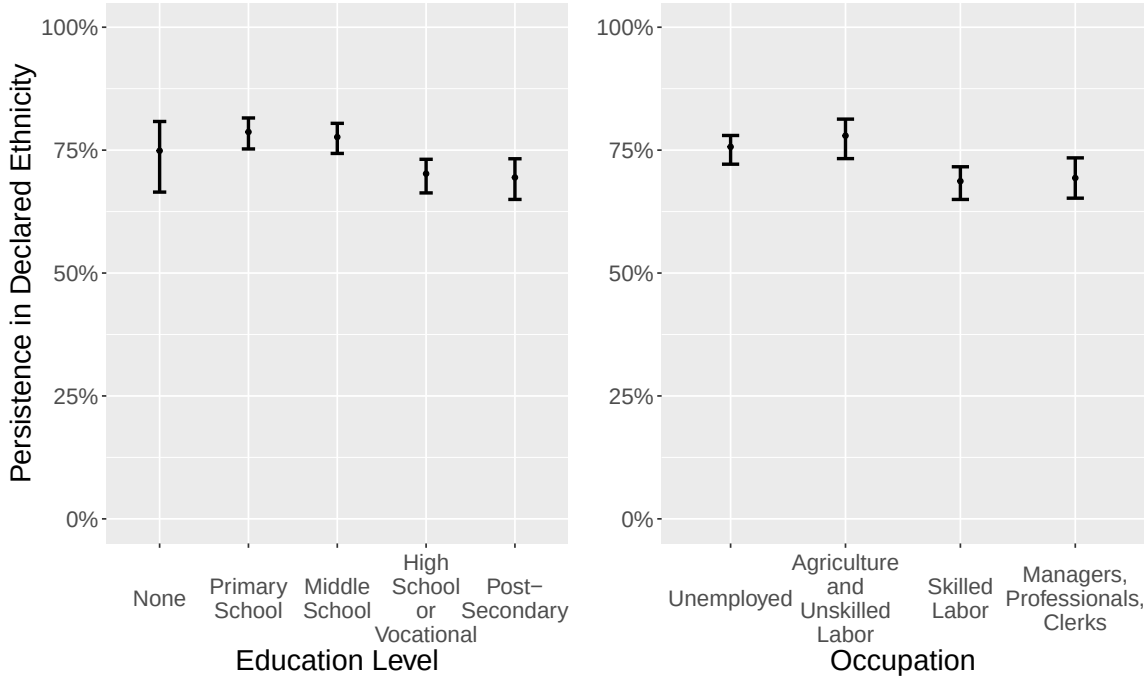
A.1 Additional Figures and Tables

Figure A.1: Educational and Town Size Distribution Comparison: 2011 Census and Linked Samples



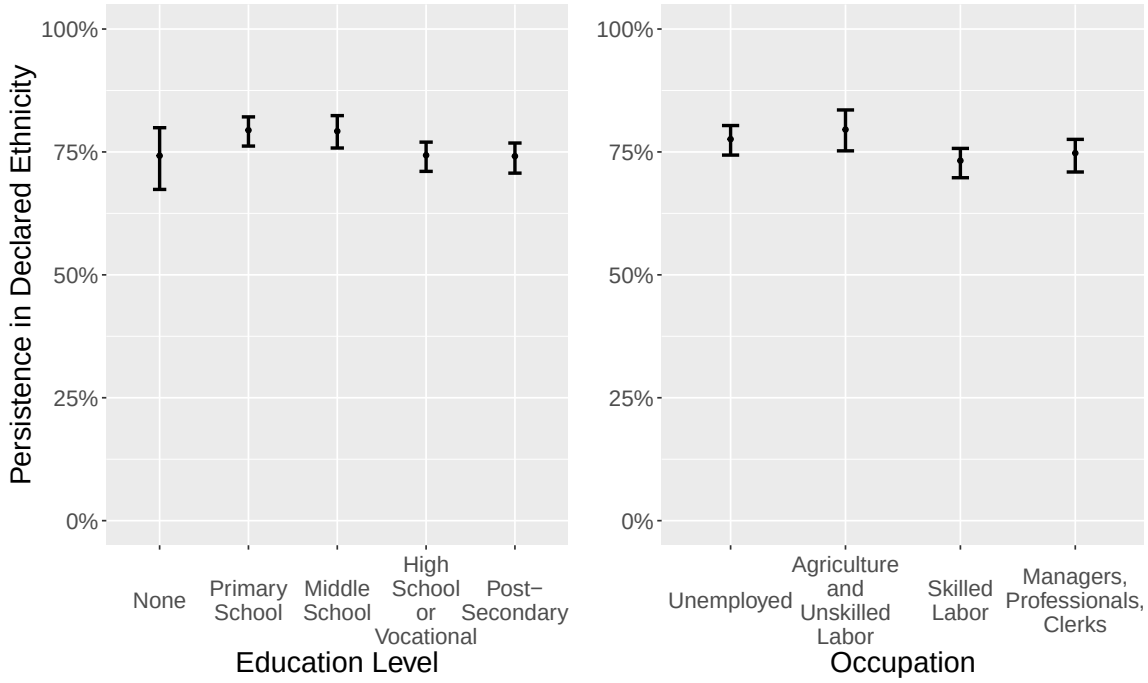
Notes: This figure compares the linked samples to the entire 2011 census sample based on education and town population.

Figure A.2: Persistence in Declared Ethnicity by Education: Other Ethnicities (1992 - 2011)



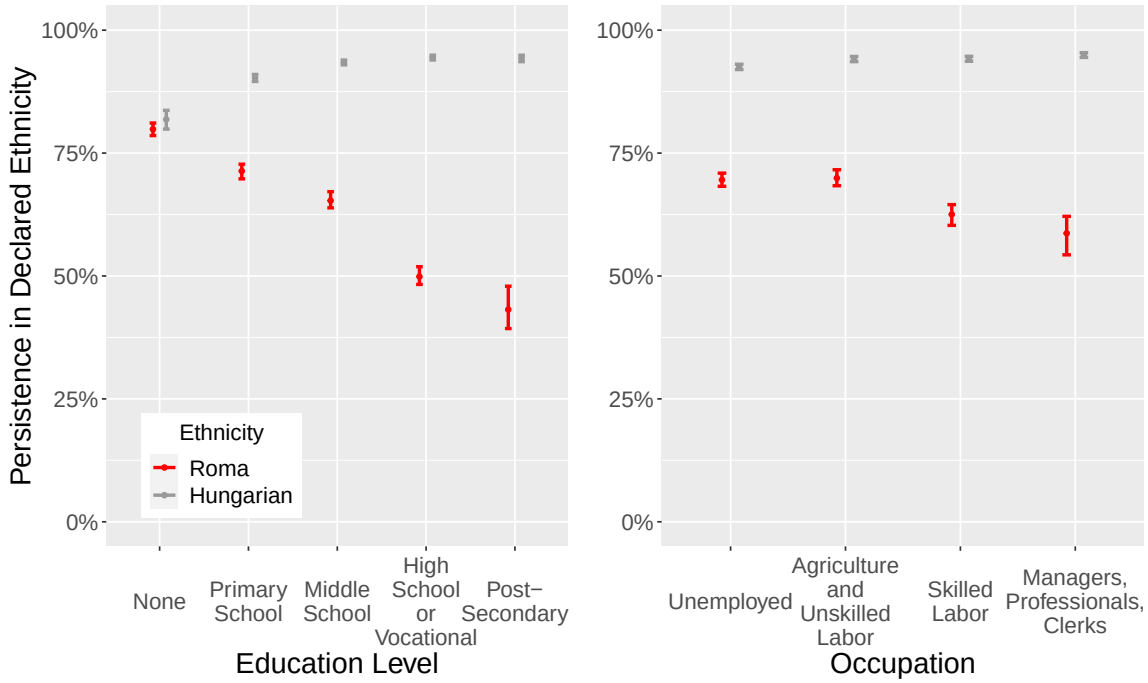
Notes: Replicates Figure 1 for all ethnicities other than Roma and Hungarian, with baseline year 1992. We pool all ethnicities to passing is measured as reporting Romanian, Roma, or Hungarian ethnicity.

Figure A.3: Persistence in Declared Ethnicity by Education: Other Ethnicities (2002 - 2011)



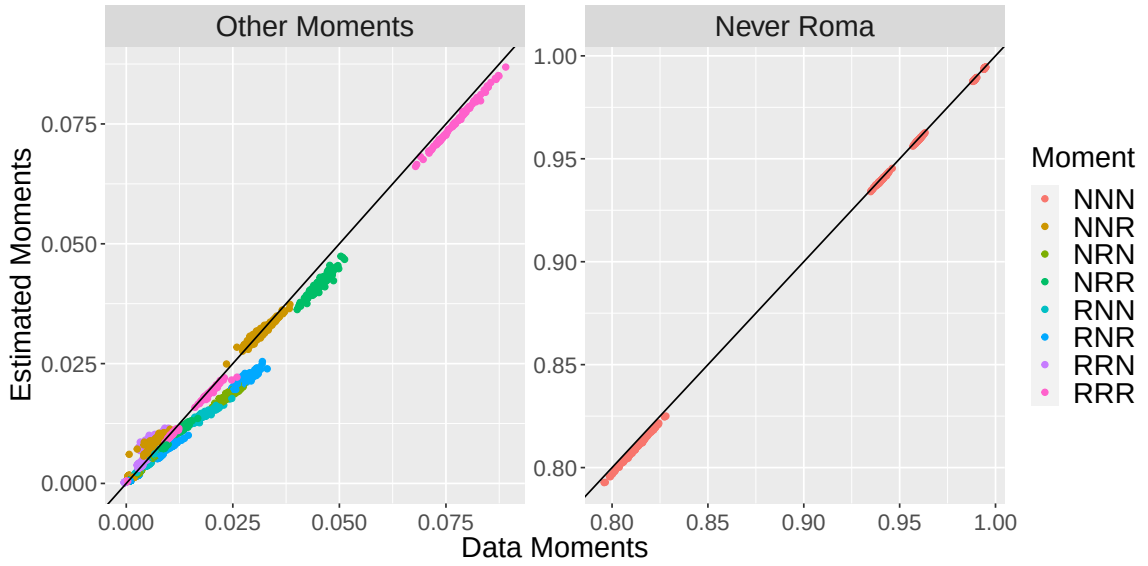
Notes: Replicates Figure A.2 with baseline year 2002.

Figure A.4: Persistence in Declared Ethnicity by Education (2002 - 2011)



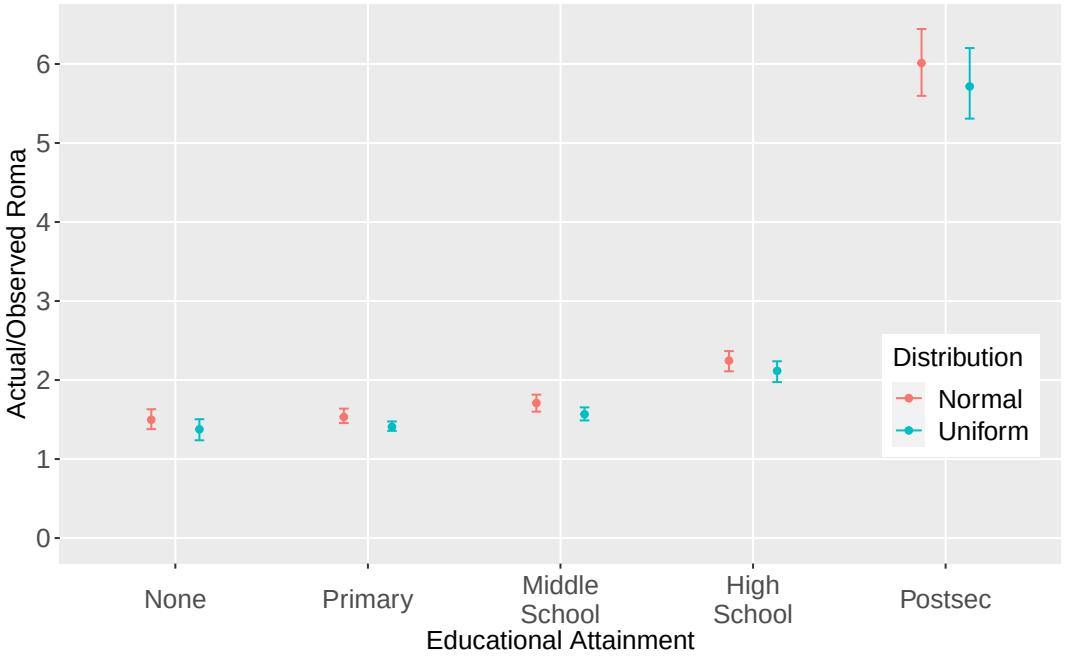
Notes: Replicates Figure 1 with baseline year 2002.

Figure A.5: Structural Model Fit



Notes: This figure plots the moments used to estimate the model from section 5, 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 $(Roma_{i,1992}, Roma_{i,2002}, Roma_{i,2011}) = (0, 1, 0)$.

Figure A.6: Structural Model Comparison: Different Distribution in Individual Heterogeneity in Self-Declaration Rates



Notes: This figure replicates Figure 2 for a model where the individual propensity to declare Roma ethnicity π_i is distributed according to a uniform distribution centered at zero rather than normal distribution, and we estimate the parameters σ_e, the width of the distribution, for each education group e.

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

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

	1992-2011		2002-2011	
Education Level	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
Post- Secondary	1,518	29.5	1,440	22.2

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.

Table A.3: Passing With Perfect Linkages

	Dependent Variable: Reported Roma Ethnicity at Endline			
	(1)	(2)	(3)	(4)
Schooling Yrs	-0.017***	-0.016***	-0.017***	-0.014***
	(0.002)	(0.003)	(0.002)	(0.004)
Town FE		x		x
Birthyear FE			x	x
Town x Birthyear FE				x
N	6,232	6,232	6,232	6,232
R^2	0.02	0.33	0.03	0.77
DV Mean	0.80	0.80	0.80	0.80

Persistence in ethnic self-declaration rates among Roma women who declared their child as Roma on birth records (2003-2011) at the 2011 census. Records are perfectly linked between birth records and census data using unique national ID numbers. Standard errors are clustered at the town level.

Table A.4: Respondent Views Internally Consistent but Divergent

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

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.

B Additional Results

B.1 Adjusting for Mismatch

The passing rate in the data is a mix of mismatch and actual passing. We can write the persistence in self-declared Roma ethnicity as:

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

If we know the mismatch rate μ_g we can recover the true persistence in self-declared ethnicity:

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

C Model details

C.1 Model Setup

The model consists of two sub-populations: Roma (comprising a proportion p_R) and non-Roma (with a proportion $1-p_R$). We focus on three census period indexed by $t = 0, 1, 2$. Roma individuals indexed by i and with endline educational level e declare their ethnicity (E) as Roma (1) with probability d_{it}^e. With remaining probability $1-d_{it}^e$ they declare non-Roma (0) ethnicity.

The self-declaration probability d_{it}^e can be decomposed into two parts, such that $d_{it}^e = d_t^e + \epsilon_i$. The first component ($d_t^e$) 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 (ϵ_i) is an individual-specific idiosyncratic but time-constant "taste" for self-declaring Roma for individual i. This parameter ϵ_i is drawn from a normal distribution with mean 0 and standard deviation σ^e.[9]

Additionally, we allow mismatch in our model. In other words, any self-reported ethnicity history we observe $E_i = \{E_{i1}, E_{i2}, E_{i3}\}$ 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 $m_{E,1}^e$ probability of mismatch between census waves 1 and

[9] We also estimated the model using a uniform distribution for σ. Our results are very robust to the distributional assumption for ϵ_i.

3 and a $m^e_{E,2}$ probability of mismatch between census waves 2 and 3 for individuals with education e for each different type of ethnic reporting history E. For example, for records with postsecondary schooling as endline attainment, and self-reporting histories $E = \{1, 0, 1\}$, we estimate $m^e_{E,1}$ and $m^e_{E,2}$.

These mismatch rates are estimated by perfectly matching the rate of inconsistencies in sex reported across linked census records with declared ethnicities E 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 or 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 $p_{M_1=0,M_2=0} = (1 - m^e_{E,1}) \times (1 - m^e_{E,2})$;
2. mismatch between waves 1 and 3 only, with probability $p_{M_1=1,M_2=0} = m^e_{E,1} \times (1 - m^e_{E,1})$;
3. mismatch between waves 2 and 3 only, with probability $p_{M_1=0,M_2=1} = (1 - m^e_{E,1}) \times m^e_{E,2}$;
4. mismatch between both waves 1 and 3 and 2 and 3, with probability $p_{M_1=1,M_2=1} = m^e_{E,1} m^e_{E,2}$.

Additionally, we define $N_{possible}(E, M_1, M_2)$, a function which indicates which reporting histories E can be associated to a non-Roma-endline individual i, given mismatch scenarios M_1 and M_2. 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 $E_2 = 1$, 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 $N_{possible}(E, M_1, M_2)$ as:

$$N_{possible}(E, M_1, M_2) = (1 - E_2) \times (1 - E_2)^{1-M_1} \times (1 - E_0)^{1-M_1}$$

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

$$P_E^e = \sum_{M_{1,2}\in\{0,1\}} \Big(p_{M_1,M_2} p_r^e \prod_{t\in\{0,1,2\}} \left[(d_{it}^e)^{E_t}(1-d_{it}^e)^{1-E_t}\right]^{1-M_t} \prod_{t\in\{0,1\}} \left[(p_r^e d_{jt}^e)^{E_t}(p_r^e(1-d_{jt}^e)+(1-p_r^e))^{1-E_t}\right]^{M_t} + N_{possible}(E)(1-p_r^e) \prod_{t\in\{0,1\}} \left[(d_{jt}^e)^{E_t}(1-d_{jt}^e)^{1-E_t}\right]^{M_t} \Big)$$

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 E_t following $(d_{it}^e)^{E_t}(1-d_{it}^e)^{1-E_t}$.

The second term indicates the probability of observing census responses in case of Roma individuals at $t = 2$ 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 d_{jt}^e or pass at rate $1 - d_{jt}^e$. 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 $E_t = 0$, whereas $E_t = 1$ 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 $N_{possible}(E) = 1$.

As an illustration, when $E = \{1, 0, 0\}$, $M_1 = 1$ and $M_2 = 0$, the expression simplifies to:

$$P_E^e = p_{1,0} \ \left[p_r^e(p_r^e d_{j0}^e)(1-d_{i1}^e)(1-d_{i2}^e) + (1-p_r^e)(p_r^e d_{j0}^e)\right]$$

where $p_{1,0}$ is the joint probability of $M_1 = 1, M_2 = 0$, the first term in the summation is associated to an endline-Roma individual declaring non-Roma status at rate $1-d_{i2}^e$, being mismatched to a Roma individual j at $t = 0$ at rate p_r^e, who declares their Roma ethnciity at rate d_{j0}. They are also correctly matched to their own record at time $t = 1$, where they passed with rate $1 - d_{i1}^e$. The second term in the summation corresponds to the individua 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 d_{j0}.