Data construction synthesis

Timing of citizenship acquisition and immigrants' children educational outcomes: a family fixed-effects approach

Marie Labussière

Introduction

Our original dataset was composed of yearly observations of all second-generation individuals legally residing in the Netherlands between 1995-2016 (dataset A, cf. Table 12 p.1). Statistics Netherlands (CBS) defines second-generation individuals as children born in the Netherlands to at least one foreign-born parent. In our analysis, we focus instead on those born to *two* foreign-born parents (hereafter the second generation "narrowly defined"). Selecting individuals whose two parents are registered as first-generation immigrants requires identifying parents' generation.

We first describe how we matched respondents with their parents' characteristics to obtain parents' generation. To do so, we had to recode a number of missing values for parental identifiers and parental generation, mostly due to administrative gaps. In most cases, these gaps can be linked to the outward or circular migration moves of the parents.

We then describe how the naturalisation variable was constructed based on the recoding of the nationality variables, which are also subject to administrative gaps. We identified a number of atypical/non-standard naturalisation trajectories involving loss or re-acquisition of Dutch citizenship over time.

Finally, we give an overview of the main data manipulations leading to the dataset used for the descriptive statistics and the analysis (*full_sample.dta*).

Contents

Matching of children and parents' characteristics 2
1.1. Deletion and recoding of missing parental identifiers 2
1.2. Selection of the second-generation narrowly defined 4
2. Nationality variables 6
2.1. System-missing values 6
2.2. Atypical trajectories 9
3. Overview 11

Matching of children and parents' characteristics

1.1. Deletion and recoding of missing parental identifiers

To match children with their parents' characteristics, we first looked at parental identifiers. Not all children had valid parental identifiers during the whole period, and some of them did not even have one. Overall, we had over **12.2 million observations** for which both parental identifiers were missing, and more than **1.3 million (61.5%) individuals** had at least once (i.e. one year) both parental identifiers missing.

We firstly noted that in a very substantial number of cases, parental identifiers were missing because the child was **present in the dataset before birth** with an unknown age, probably due to a data artefact. We deleted these artificial years so that children were only observed after birth; this led to the deletion of more than 5.5 million observations[1] (dataset B).

We had still about 6.7 million observations with both parental identifiers missing and more than 585 thousands (27.5%) individuals with both parental identifiers missing at least once. We noticed several interesting patterns for the latters: half of them had at least one recorded move outside the Netherlands (versus 15% in the total population), and 22.4% of them had died before the end of the observation period (versus 4.3% in the total population). Checks confirmed that parental identifiers were no longer provided when a child was **registered as living outside the Netherlands or as deceased**[2]. Besides, when looking at missing observations for individuals that had no recorded move or death, we could see that 89.2% of the observations were registered at age zero. This temporal pattern suggests **administrative lags** in the recording of parental identifiers, which is plausible considering the information provided by the variable *XKOPPELNUMMER*. This variable indicates that in certain cases, CBS has difficulties associating a child with their parents, especially when the latters are not mentioned in the marriage registers[3]. Furthermore, we noticed that 10% (versus 4.5% on average) of the observations were registered in 1995, year corresponding to the introduction of municipal population registers, which supports further the hypothesis of an administrative lag.

Following these considerations, we decided to recode missing parental identifiers for those who had both parental identifiers missing at least once. To decrease computing time, we only recoded for **individuals born in 1990**; our population was therefore reduced to the relevant birth cohorts, leading to the **dataset C**.

However, there were **two cases** to take into account before recoding:

- **Individuals who changed parental identifiers over the period**: this likely refers to children changing *legal guardians* over the period. In case of gaps in the recording, it is quite tricky to complete the missing values of the child's parental identifiers. Besides, these cases were substantially complicating the recoding. As a result, they have been deleted, entailing the loss of 470 individuals, corresponding to 6,556 observations (dataset D).

- **Individuals whose parental identifiers were missing during the entire observation period**: as it was not possible to match them with their parents, they have thus been deleted. This entailed the loss of 9,850 individuals, corresponding to 172,362 observations (dataset E).

[1] As a result, 12 individuals whose age was always unknown have been deleted as well.
[2] Note that individuals who die during the observation period are not removed from the dataset.
[3] *XKOPPELNUMMER* is a very detailed variable (more than a hundred categories) recording whether and how a given child was connected to his or her parents by the administrative authorities. The variable was not present in the current dataset, but may be used later to bring further support to the aforementioned hypotheses.

Additional remarks on selection bias due to deletion of individuals

In the first case the deletion is not really problematic to the extent that we can assume these children have very specific experiences related to the change of legal guardians (e.g. death of both parents), that may also impact school outcomes and even legal status. As we seek to limit unexplained heterogeneity within our population, it seems preferable to restrict our analysis to children who had the same legal guardians over the period.

In the second case the deletion is more a matter of concern: while it leads to the removal of a small – though non-negligible – number of children (0.9%), we cannot assess with great certainty its impact in terms of selection bias due to the absence of external information. However the available variables suggest that the absence of valid parental identifiers is strongly correlated with outward migration and foreign-citizenship: a move is recorded for 94.8% of these children (versus 15.6% in the total population), and only 34.4% of them are Dutch citizens at the first year of observation (versus 74.2% in the total population). The absence of missing parental identifiers may thus indicate complex familial configurations where children were born in the Netherlands but have lived in the country for a very short period before going abroad, making it difficult for administrative authorities to connect them with their parents. As we study families that stay long enough in the Netherlands to be eligible for Dutch citizenship, these children are probably not part of our population of interest.

Our aim was to **recode cases where both parental identifiers were missing**, so that to be able to match children with at least one legal guardian[4]. When recoding we had **two categories of children**: those who had two parental identifiers registered at least once, and those who had only one parental identifier registered at least once.

Recoding strategy:

- **For those who had both parental identifiers registered at least once**, we simply replaced the missing values by the valid parental identifiers recorded either beforehand or afterwards.

- **For those who had only one parental identifier registered at least once**, we only recoded the years in which *both* parental identifiers were missing, in order to be able to match the child with at least one parent. Where necessary, we therefore replaced one of the two missing parental identifiers, by either the one of the mother or the one of the father[5]. As the majority of children had at least once a valid mother identifier – in 85.1% of cases the father identifier was never recorded for those with an atypical pattern, versus 6.1% for mothers' identifiers – we gave priority to the identifier of the mother. The identifier of the father was only used when no mother identifier was available over the whole period.

Table 1 summarizes the distribution of patterns of missing parental identifiers in the dataset D, and shows which categories were recoded.

[4] Note that we did not recode for those who had only one parental identifier missing (6.3%), because children do not necessarily have two legal guardians; they may have not been acknowledged from birth by one of the parent, and thus have only one registered parental identifier. Even if a second legal guardian is recorded over time, this person has not necessarily acknowledged the child from birth. It would then be quite arbitrary to recode these parental identifiers, and difficult as well to distinguish these atypical family configurations from administrative lags. Furthermore, one parental identifier is sufficient to match these children with their household characteristics; they are therefore kept in the dataset without any recoding.

[5] Knowing that in 61 cases, the same parental identifier is alternatively coded as being a mother or a father identifier. In any case the notions of mothers and fathers need to be taken carefully: all parents recorded as mothers (resp. fathers) are not necessarily women (resp. men), which may reflect administrative variations in the way parents are associated with their children, especially in the case of homoparental or single-parent families.

Table 1 Patterns of missing parental identifiers (PID) (dataset D)

Pattern	Frequency	Percent	Manipulations
1) No missing PID	833,055	73.4	-
2) Both PID missing at least once (only)	187,601	16.5	Recoded in (1)
3) One PID missing at least once (only)	72,061	6.3	-
4) Atypical pattern combining (2) and (3)	32,842	2.9	Recoded in (3)
5) Both PID always missing	9,850	0.9	Deleted
Total	1,135,409	100.0	

1.2. Selection of the second-generation narrowly defined

After the recoding of missing parental identifiers described above, every child had at least one valid parental identifier for each year, and could therefore be matched with (one of) their parents' characteristics. While the matching did not entail complications, we faced some troubles selecting second-generation children born to two foreign-born parents.

Our aim was to restrict our population to children whose two parents were first-generation according to CBS (see Table 2). This represents 527,796 children (51.7%) if the selection is based on parents' generation at the first observation year of the child. However, we figured out that 9,370 individuals had their parents recorded as being first-generation at least once during the observation period, but not in the first year of observation. This was mainly related to fathers' generation, containing a significant number of missing values[6]. These missing values did not seem to be due to outward migration but rather to administrative lag, to the extent that in roughly half of the cases, generation was only missing for the first four years of observations or below. Although we had little information on these missing values, we decided to recode them because generation is based on one's – and one's parents' – place of birth (see Table 2, p.5); this is not supposed to change over time. Yet, we did not recode generation for a small number of individuals whose parents change generation over time, as it suggested administrative error and that it would have been arbitrary to decide which one to retain.

As a result, we recoded generation for 10,305 fathers and 1,522 mothers, which resulted in the addition of 9,238 individuals in our sample of interest, resulting in the dataset F of 537,034 individuals (versus 527,796 before the recoding).

[6] We had 10,422 (1,623) individuals that had their father (mother) recorded as first-generation at least once during the observation period but not in the first year of observation, and for 98.9% (93.8%) of them is was due to missing values. For the others, it was due to change in parental generation (from Dutch to first-generation in 9 cases out of 10).

Table 2 Distribution of parental generation among second-generation children at the first year of observation (dataset E)

			Father's generation			
			Native	1st	2nd	Total
Mother's generation	Native	Count	**64,398**	**135,922**	**3,306**	**203,626**
		%	6.3%	13.3%	0.3%	20.0%
	1st	Count	**161,546**	**527,796**	**48,667**	**738,009**
		%	15.8%	51.7%	4.8%	72.4%
	2nd	Count	**3,000**	**75,355**	**6**	**78,361**
		%	0.3%	7.4%	0.0%	7.7%
Total		Count	**228,944**	**739,073**	**51,979**	**1,019,996**
		% of Total	22.4%	72.5%	5.1%	100.0%

Remark: individuals with missing values (n = 105,563) are not included in the table.

CBS definitions of generation

Native: individual whose two parents were born in the Netherlands

First-generation: individual who is born abroad from at least one foreign-born parent

Second-generation: individual who is born in the Netherlands from at least one foreign-born parent

Remark

Table 2 brings light on how Statistics Netherlands defines the second generation, both conceptually and empirically. It shows that, by defining the second generation as those born in the Netherlands from at least one foreign-born parent, some children are categorized as second-generation although both of their parents are natives (6.3%). This likely refers to situations where **parents are born abroad from native-born parents**, which puts them in the category of natives while their children are considered as being second-generation. We can also notice that, due to the **strict criteria applied for the definition of natives**, individuals born from one second and one first-generation parents are still considered as being second-generation, rather as third-generation or natives (12.2%). Only children born from two second-generation parents should in theory be counted as natives (the 6 recorded cases suggest administrative error in that regard).

Table 3 gives a summary of the data recoding for parental identifiers and generation.

Table 3 Overview of data recoding – parental identifiers and generation

Operation	**Recoded observations**	**Corresponding individuals**
Missing parental identifiers (*1_TGmissingPID.sps*)		
Both parental identifiers are missing	2,662,003	187,601
Atypical trajectories	421,656	32,842
Missing parental generation (*3_merge.sps)*		
Mother's generation	6,888	1,522
Father's generation	58,367	10,305

2. Nationality variables

We did not have any variable recording whether and when a naturalisation had occurred in someone's trajectory, thus we constructed our own variable. We relied on two main variables for that purpose: *gbanationaliteit1* recording one's nationality and *gbanationaliteit2* recording one's other nationality when applicable, priority being given to Dutch nationality[7]. Our naturalisation variable was constructed in the following way: we created a dummy when a change was observed from a non-Dutch to a Dutch citizenship in *gbanationaliteit1*. We then summed this dummy at the individual level to see whether an individual had naturalised or not during the observation period (and how many times, see *infra*).

However, we had to deal with two types of issues: 1) a high number of system-missing values for both nationality variables 2) atypical trajectories that led to misleading counts of naturalisations.

2.1. System-missing values

Nationality variables had two types of missing values (see Table 4): system-missing (".") and user-missing values ("0" and "9999"). Whereas the latter are recorded in CBS codebook and correspond to a situation where one's nationality is declared as "unknown" by Dutch authorities, the former are due to administrative gaps in the recording.

In the case of user-missing values, there are good reasons to think that a nationality declared as "unknown" by Dutch authorities is necessarily non-Dutch, since administrations contributing data to CBS are supposed to have reliable information on their nationals[8]. For this reason, we considered that a change from an unknown to a Dutch nationality was to be counted as a naturalisation.

By contrast, it seemed quite arbitrary to make a similar choice in the case of system-missing values, since the causes of the gaps are unknown. As a result, we first tried to determine the most plausible reasons for the gaps.

Table 4 Counts of user- and system-missing for nationality variables

	System-missing "."	**User-missing "0"**	**User-missing "9999"**
Gbanationaliteit1	313,616	108,985	31
Gbanationaliteit2	313,616	150,325	2,160,422
Gbanationaliteit1_pa (fathers)	234,317	78,464	1
Gbanationaliteit2_pa (fathers)	234,317	109,161	1,388,833
Gbanationaliteit1_ma (mothers)	187,831	95,122	0
Gbanationaliteit2_ma (mothers)	187,831	94,768	1,584,501

We first counted the number of system-missing values per individuals and figured out that 20.4% of our population had *at least one* system-missing value during the observation period. Among them, 40.8% had *only one* missing value while at the other end, 11.6% had 10 or more and up to 21.

After some inquiries, we identified three main causes of gaps.

[7] Meaning that when an individual is Dutch, Dutch citizenship is necessarily recorded as his/her first citizenship.
[8] This was confirmed through our exchanges with CBS (see email from Clemens Siermann, October 24, 2018).

- **Outward migration:** when an individual is abroad, there is a gap in the recording of his/her nationality: it is missing during the period spent abroad and recorded again upon return. Outward migration is particularly associated with individuals having 2 system-missing values or more (up to 17, see Table 5).
- **Administrative lag:** some individuals have system-missing values from the first year of observation onwards, which corresponds in most cases to the year of birth. Nationality is only missing for a short period (one year in 93.1% of the cases) before being properly recorded in the dataset.
- **Death:** nationality is missing from the year of death onwards. It is the main reason for individual cases associated with a large number of missing values (e.g. it explains all cases above 17 missing values, see Table 5)

Table 5 gives for each cause identified the number of individual cases concerned; note that percentages are calculated on the remaining cases[9]. We can see that once we take into account outward migration, death and administrative lag at birth, **we are left with 402 unexplained cases**. This amount, representing 0.6% of those with at least one missing, seemed quite reasonable considering our limited access to external information. Based on this understanding, we completed the nationality variables for individuals with system-missing values.

Table 5 Plausible causes for system-missing values (children)

Number missing	Associated with outward migration		Associated with death		Administrative lag at birth		Remaining (Unexplained)	Total
	counts	percent	counts	percent	counts	percent	counts	counts
0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	297,347
1	9,412	32%	0	0%	19,783	99.8%	48	29,243
2	5,665	99%	47	64%	n.a.	n.a.	26	5,738
3	4,388	98%	51	59%	n.a.	n.a.	36	4,475
4	4,217	98%	48	74%	n.a.	n.a.	17	4,282
5	4,326	98%	58	64%	n.a.	n.a.	33	4,417
6	4,367	98%	64	74%	n.a.	n.a.	23	4,454
7	3,975	97%	57	52%	n.a.	n.a.	52	4,084
8	3,675	97%	67	69%	n.a.	n.a.	30	3,772
9	2,832	97%	68	81%	n.a.	n.a.	16	2,916
10	2,144	97%	57	**79%**	n.a.	n.a.	15	2,216
11	962	92%	81	92%	n.a.	n.a.	7	1,050
12	775	93%	54	93%	n.a.	n.a.	4	833
13	651	90%	65	93%	n.a.	n.a.	5	721
14	600	86%	69	73%	n.a.	n.a.	25	694
15	686	88%	75	82%	n.a.	n.a.	17	778
16	772	90%	64	77%	n.a.	n.a.	19	855
17	791	88%	78	73%	n.a.	n.a.	29	898
18	1	2%	52	100%	n.a.	n.a.	0	53
19	0	0,0%	85	100%	n.a.	n.a.	0	85
20	0	0,0%	57	100%	n.a.	n.a.	0	57
21	0	0,0%	67	100%	n.a.	n.a.	0	67
Total	50,239	-	1,264	-	19,783	-	**402**	**369035**

[9] For example, death explains 79% of the cases with 10 system-missing values once we have removed cases associated with outward migration (=57/(2216-2144)*100).

Completing the gaps

To complete the nationality variables, we identified three types of gaps: bottom- and top-edge gaps, when missing values appear respectively on the lower or upper edge; and middle gaps. These gaps are exemplified in Table 6, knowing that some individual trajectories combine different types of gaps.

We completed the nationality variables based on the following reasoning:

- **Bottom-edge gaps** are mostly due to administrative lag in the recoding, meaning that we can replace missing values by the first non-missing values encountered.
- **Middle gaps** are in the great majority of cases due to outward migration; if the individual has the same nationality before and after the recorded migration move, the nationality variable is completed to ensure continuity. If the individual has changed nationality while being abroad, we code the missing values as ".a" to indicate a mismatch. If the individual has acquired Dutch citizenship upon return, we code the missing values as ".n" in order to be able to take into account these specific cases of naturalization.
- **Top-edge gaps** are mostly due to deaths: rows corresponding to post-mortem years will be deleted from our dataset, while missing values that are not attributable to death are recoded as ".z".

Table 6 Examples of trajectories with bottom, middle or top gaps before and after completion
("." stands for missing, 1 for Dutch and any other letter stands for non-Dutch)

	Bottom edge gap		**Middle gap with continuity**		**Missing .a**		**Missing .n**		**Top edge gap**	
Year	before	after	before	after	before	after	before	after	before	after
1	.	X	X	X	X	X	X	X	X	X
2	.	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X
4	X	X	.	X	.	.a	.	.n	X	X
5	X	X	.	X	.	.a	.	.n	X	X
6	X	X	.	X	.	.a	.	.n	.	.d/.z
7	X	X	X	X	Y	Y	1	1	.	d/.z
8	X	X	X	X	Y	Y	1	1	.	d/.z

Table 7 gives an overview of the number of observations recoded for each type of gap. As regards the missing values attributable to death, the deletion of all post-mortem years led to the deletion of 15,477 observations (dataset I, see Table 12 p.11).

Table 7 Number of system-missing observations recoded (children)

	Nationality 1	**Nationality 2**
bottom edge gaps	33,407	33,407
middle gaps	264,128	264,128
continuity	258,401	258,745
mismatch (.a)	2,646	5,383
mismatch (.n)	3,081	0
deaths (.d)[10]	15,428	15,428
top edge gaps (.z)	653	653
Total	312,963	312,963

[10] Note that some individuals had still their nationality recoded after deaths, corresponding to 49 observations. These too have been recoded as ".d" missing values, leading to a total of 15,477 observations labelled as such.

Parental gaps

The same procedure was followed to recode parental naturalisation variables. However, we did not complete bottom edge-gaps because the hypothesis of an administrative lag seemed less credible for parents than for their children. Parents' first year of observation does indeed not necessarily correspond to their year of arrival in the Netherlands (especially for those who arrived prior to 1995, who are only observed from 1995 onwards) and outward migration appeared as a predominant factor regardless of the number of missing values[11]. As a result, we recoded this system-missing values as ".z", which is the same label that was used for top-edge gaps.

After taking into account outward migration and deaths, we account for most of the missing observations, leaving 99 (0.3%) unexplained cases for mothers and 106 (0.2%) for fathers.

Table 8 gives an overview of the number of observations recoded for each type of gap. Note that years observed after deaths have not been deleted because our unit of observation is the child: parents should stay in the observation window as long as their children are alive, even though they are dead.

Table 8 Number of system-missing observations recoded (parents)

	Fathers		**Mothers**	
	nationality 1	**nationality 2**	**nationality 1**	**nationality 2**
bottom/top edge gaps (.z)	16,483	16,483	16,193	16,193
middle gaps	177,751	177,751	156,840	156,840
continuity	172,830	172,742	152,707	152,924
mismatch (.a)	2,856	5,009	981	3,916
mismatch nat (.n)	2,065	0	3,152	0
deaths (.d)[12]	40,083	40,083	14,798	14,798
sum	234,317	234,317	187,831	187,831

2.2. Atypical trajectories

A last issue was related to atypical trajectories including sudden changes of nationality. We indeed noticed that some individuals had a different nationality recorded for only one year, before having their previous nationality being recorded again. This was particularly problematic when Dutch nationality was involved: individuals loosing their Dutch citizenship or acquiring it for a one-year period were counted as naturalized citizens, which they were probably not. Consequently, these trajectories were smoothed to avoid taking into account dubious naturalisations. Table 9 gives for each variable the number of recoded observations and shows that this decision is expected to have a very negligible impact owing to the small number of casess involved.

We have noticed other atypical trajectories that have not been recoded due to the lack of external information. As some changes in nationality are difficult to explain but may correspond to "rare but plausible" individual cases where individuals lose their Dutch citizenship, we have just made sure to make them identifiable in the dataset. Again, these atypical trajectories only concern a small number of individuals, as shows Table 10. However, it was important to tag them to facilitate the construction of variables such as the date and age at naturalisation, which cannot be time-constant in case of double naturalisations.

[11] For children outward migration was only associated with 32.2% of the cases with only one system-missing value (see Table 5, p.6). By contrast, it was associated with more than 90% of these cases for parents – a percentage similar to those of the other categories.

[12] Note that some individuals had still their nationality recoded after deaths, corresponding to 98 (45) observations for fathers (mothers). These too have been recoded as ".d" missing values, leading to a total of 40,181 (14,843) observations labelled as such.

Table 9 Number of observations recoded for atypical trajectories

"x x 1 x x" configurations		"1 1 x 1 1" configurations	
nationality 1	39	nationality 1	32
nationality 2	29	nationality 2	26
nationality 1 (*fathers*)	30	nationality 1 (*fathers*)	25
nationality 2 (*fathers*)	30	nationality 2 (*fathers*)	24
nationality 1(*mothers*)	25	nationality 1(*mothers*)	9
nationality 2(*mothers*)	25	nationality 2(*mothers*)	8

Table 10 Counts of atypical trajectories that are not recoded (individual level)

	Children	Mothers	Fathers
Double naturalisations *Individuals who are not Dutch at the first year of observation and natuturalise twice* Ex. x x x 1 1 1 x x x 1 1 1 1 1 1	70	15	52
Loss and reacquisition of Dutch citizenship[13] *Individuals who are Dutch at the first year of observation, loose and subsequently reacquire Dutch citizenship* Ex. 1 1 1 1 1 1 x x x x 1 1 1 1 1	28	11	53
Total	**98**	**26**	**105**

Finally, in Table 11 we give a brief overview of how missing values were coded for naturalisation variables.

Table 11 Labels of missing values for naturalisation variables

Missing value	Naturalisation	Date at naturalisation/Age at naturalisation
.a	*Atypical naturalisations (see Table 10)*	*Those whose naturalisation date/age is unclear: those who have atypical naturalisations or who naturalise while being abroad i.e. from .n, see Table 6 p.8)*
.b	*Individuals being Dutch from the first year of observation onwards (i.e. from birth in the case of children)*	*Individuals being Dutch from the first year of observation onwards (i.e. from birth in the case of children)*
.c	*n.a.*	*Those who do not naturalise over the period*

[13] Note that this second category has a slightly different meaning for parents and children. For the latter it refers to children who are Dutch from birth and subsequently loose and reacquire Dutch citizenship. By contrast, the former may have naturalized prior to the observation period and may be more comparable to those who naturalized twice.

3. Overview

Table 12 Overview of data selection steps

N*	Details	Deleted individuals	Deleted observations	Total individuals	Total observations
A	Selection of the second generation			2,134,568	42,252,918
B	Exclusion of the observations with unknown age	12	5,518,462	2,134,556	36,734,456
C	Selection of the cohorts born in 1990 and after	998,677	21,939,765	1,135,879	14,794,691
D	Exclusion of respondents with changing parental identifiers	470	6,556	1,135,409	14,788,135
E	Exclusion of respondents with both parental identifiers always missing	9,850	172,362	1,125,559	14,615,773
F	Selection of respondents with two first-generation parents	588,525	7,270,751	537,034	7,345,022
G	Exclusion of respondents with parents originating from the former Dutch colonies	69,366	1,082,061	467,668	6,262,961
H	Exclusion of the post-2010 cohorts	98,633	350,226	369,035	5,912,735
I	Exclusion of the years observed *after* death	7	15,477	369,028	5,897,258
J	Exclusion of individuals who are not present in the educational registers	21,512	439,945	347,516	5,457,313
K	Selection of respondents who take the Cito test between 2008 and 2015 (for the first time)	251,255	3,828,176	96,261	1,629,137
L	Selection one row per respondent	0	1,532,876	96,261	96,261
M	Exclusion of respondents registered in institutional households	72	72	96,189	96,189
N	Exclusion of respondents enrolled in practical education	1,174	1,174	95,015	95,015
O	Exclusion of respondents with missing sociodemographic characteristics	288	288	94,727	94,727

Remark: Dataset O corresponds to the dataset used for the replication do-files, *full_sample.dta*.

Table 13 Overview of data recoding

Operation	Recoded observations	Corresponding individuals
Missing parental identifiers (*1_TGmissingPID.sps*)		
Both parental identifiers are missing	2,662,003	187,601
Atypical trajectories	421,656	32,842
Missing parental generation (*3_merge.sps)*		
Mother's generation	6,888	1,522
Father's generation	58,367	10,305
Missing nationality *(1_nat_ready.do)*		
Children's nationality	312,963	*Time-varying*
Missing parental nationality *(3_nat_parents.do)*		
Mother's nationality	187,831	*Time-varying*
Father's nationality	234,317	*Time-varying*