Data Dictionary: Recoding Three Surveys for "A Human Capital Model of Linguistic and Cultural Assimilation"

James V. Marrone*

September 9, 2016

Contents

1 Using This Document **3**
1.1 Navigating the Survey Data . . . 3

2 Common Variables Among All Surveys **4**
2.1 Demographics . . . 4
2.2 Language Skills . . . 7
2.3 Linguistic Proximity . . . 8
2.4 Migration History & Visa Status . . . 9
2.5 Family and Spousal Variables . . . 10
2.6 Education . . . 12
2.7 Employment History and Income . . . 13
2.8 Social Variables . . . 15

3 NIS-Only Variables **16**
3.1 Employment History . . . 16
3.2 Language Use . . . 17
3.3 Religion . . . 19

4 ENI-Only Variables **19**
4.1 Demographics . . . 19
4.2 Language Use and Regional Language Skills . . . 19
4.3 Social Variables & Relationship with Birthplace . . . 20
4.4 Merging with Linguistic Data . . . 21

5 TeO-Only Variables **21**
5.1 Demographics . . . 21
5.2 Language Skills & Daily Use . . . 22
5.3 Education . . . 23

*University of Chicago, 1126 E. 59^{th} St., Chicago, IL 60637, jmarrone@uchicago.edu

5.4 Social Variables & Self-Identity 23

6 Tables **24**

1 Using This Document

This document provides instructions and rationale for re-coding three surveys so that they are commensurable and may be used to replicate results in "A Human Capital Model of Linguistic and Cultural Assimilation" (Marrone, 2016). The surveys are:

1. New Immigrant Survey (NIS) from the USA, a nationally-representative survey of recent legal permanent residents.
2. National Immigrant Survey (*Encuesta Nacional de Inmigrantes*, or ENI) from Spain, a nationally-representative survey of immigrants to Spain.
3. Trajectoires et Origins (TeO) from France, a nationally-representative survey of immigrants from abroad, migrants from French territories, children of these two groups, and children of French natives.

Each section describes how to construct variables used in the final paper, using variables from the three surveys. In the first section, variables common to all surveys are described with separate instructions for each survey. Variables unique to each survey are described in the subsequent sections. Variables are arranged by topic, not necessarily corresponding to the order in the original surveys; the Table of Contents provides additional guidance. Variable names used for the paper are in **bold** font, with labels in *italics*. Original variable names from the survey datasets are in `teletype` font. When necessary, a brief rationale or explanation is provided.

Unless otherwise noted, variables should be coded as missing if they are missing in the original surveys, or if the respondents answered "Do Not Know" or "Did Not Answer."

1.1 Navigating the Survey Data

NIS survey data comes in several publicly-available files, which may be merged using the identifier `PU_ID`. See `http://nis.princeton.edu/` and Jasso *et al.* (2005). Note the variable coding convention: datasets are identified by the first letter of the variable name. For example, `J1` is in the dataset `j_adult.dta`, whereas `H1` is in `h_adult.dta`, etc.

The ENI survey comes in one publicly-available file, and therefore requires little preparation before use. See `http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t20/p319&file=inebase&L=1` and Reher & Requena (2009). The individual observations are labeled by the identifier `IDQ`. However, there are some unique naming conventions that are explained in the relevant sections below. To begin, note that for demographic variables, each suffix `xx` corresponds to a certain person in the household, while the variable `NPELEG` indicates the suffix for the respondent. Hence, as noted below, one must often look for variables named with the proper suffix.

Most of the variables in the TeO dataset are publicly available; see `http://teo_english.site.ined.fr/`, Beauchemin & Simon (n.d.) and Beauchemin *et al.* (2010). Merge the files `indiv_brut.dta` and `indiv2.dta` using the identifying variable `ident`. For religious and political variables, the dataset `polrel2.dta` may be requested through an application process.

2 Common Variables Among All Surveys

2.1 Demographics

respID *respondent ID*

1. NIS:
   - = `PU_ID`
2. ENI:
   - = `IDQ`
3. TeO:
   - = `ident`

perwt *person weight*

1. NIS:
   - = `NISWGTSAMP1`
2. ENI:
   - = `FACTORP`
3. TeO:
   - = `poidsi`

bp *birthplace (country code)*

1. NIS:
   - "missing" if `A9Amo`=-1 or -2 ("refused" or "don't know")
   - drop respondent from data if `A9Amo`=218 ("USA")
   - otherwise = `A9Amo`
2. ENI:
   - = `PNACxx` where `xx`=`NPELEG`
3. TeO:
   - "missing" if `regionnaise2`=6101 or 9999
   - otherwise = `regionnaise2`

bp_group *birthplace region* The three surveys have different sets of options for birthplace; NIS is the least detailed, so the surveys are standardized to match NIS using the groupings below:

1. **bp_group** = 1: South/East Asia and Pacific
   - from NIS: **bp** = 44 (China), 98 (India), 111 (Korea), 164 (Philippines), 224 (Vietnam), or 302 (South/East Asia and Pacific)

- from ENI: **bp** = 404 (Bangladesh), 407 (China), 409 (Philippines), 410 (India), 411 (Indonesia), 415 (Japan), 424 (Nepal), 430 (South Korea), 431 (North Korea), 432 (Singapore), 434 (Sri Lanka), 435 (Thailand), 437 (Vietnam), 438 (Taiwan), 498 (Other), or 499 (Other)
- from TeO: **bp** = 3101 (Vietnam), 3201 (Cambodai), 3301 (Laos), or 3501 (Other Asia)

2. **bp_group** = 2: Sub-Saharan Africa
   - from NIS: **bp** = 69 (Ethiopia), 152 (Nigeria), or 306 (Sub-Saharan Africa)
   - from ENI: **bp** = 201 (Burkina Faso), 202 (Angola), 204 (Benin), 207 (Cape Verde), 208 (Cameroon), 210 (Rep. of Congo), 211 (Ivory Coast), 214-221 (Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Equatorial Guinea, and Kenya), 223 (Liberia), 225 (Madagascar), 227 (Mali), 229 (Mauritius), 231 (Mozambique), 233 (Niger), 234 (Nigeria), 235 (Central African Rep.), 236 (South Africa), 239 (Senegal), 241 (Sierra Leone), 243 (Sudan), 245 (Tanzania), 246 (Chad), 247 (Togo), 250 (DRC), 298 (Other), or 299 (Other)
   - from TeO: **bp** = 2401 (Senegal) or 2403≤**bp**≤2601 (Gambia, Guinea-Bissau, Guinea, Mali, Burkina Faso, Niger, Chad, Ivory Coast, Ghana, Togo, Benin, Nigeria, Cameroon, Central Afr. Rep., Gabon, Rep. of Congo, DRC, Equatorial Guinea, and Other)
3. **bp_group** = 3: Middle East and North Africa
   - from NIS: **bp** = 307 (Middle East and North Africa)
   - from ENI: **bp** = 203 (Algeria), 213 (Egypt), 224 (Libya), 228 (Morocco), 230 (Mauritania), 248 (Tunisia), 414 (Israel), 420 (Lebanon), 426 (Pakistan), 433 (Syria), 436 (Turkey), or 449 (Palestine)
   - from TeO: 2101≤**bp**≤2301 (Algeria, Morocco, and Tunisia), or **bp** = 2402 (Mauritania), 3401 (Turkey), or 5106 (Middle East)
4. **bp_group** = 4: Latin American and Caribbean
   - from NIS: **bp** = 47 (Colombia), 55 (Cuba), 62 (Dominican Republic), 65 (El Salvador), 88 (Guatemala), 92 (Haiti), 135 (Mexico), 163 (Peru), or 305 (Latin America and Caribbean)
   - from ENI: 303≤**bp**≤398 (Mexico, Costa Rica, Cuba, Dominica, El Salvador, Guatemala, Haiti, Honduras, Nicaragua, Panama, Dominican Republic, Trinidad & Tobago, Argentina, Bolivia, Brazil, Colombia, Chile, Ecuador, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela, Puerto Rico, Aruba, and Other)
   - from TeO: **bp** = 5104 (Central America) or 5105 (South America)
5. **bp_group** = 5: Europe and Central Asia
   - from NIS: **bp** = 166 (Poland), 172 (Russia), 215 (Ukraine), or 301 (Europe and Central Asia)
   - from ENI: 101≤**bp**≤199 but **bp**≠125 (Albania, Austria, Belgium, Bulgaria, Cyprus, Denmark, Finland, France, Greece, Hungary, Ireland, Italy, Luxembourg, Malta, Monaco, Norway, Netherlands, Poland, Portugal, Andorra, Germany, Romania, Sweden, Switzerland, Ukraine, Latvia, Moldova, Belarus, Georgia, Estonia, Lithuania, Czech Republic, Slovakia, Bosnia, Croatia,

Slovenia, Armenia, Russia, Macedonia, Montenegro, Serbia, and Other) or **bp**=442 (Azerbaijan), 443 (Kazakhstan), or 447 (Uzbekistan)

- from TeO: 4101≤**bp**≤4801, except 4408 (UK)

6. **bp_group** = 6: English-Speaking Countries and Oceania
   - from NIS: **bp** = 38 (Canada), 105 (Jamaica), 217 (UK), 304 (North America), or 308 (Oceania)
   - from ENI: **bp** = 125 (UK), 301 (Canada), 302 (USA), 501 (Australia), or 504 (New Zealand)
   - from TeO: **bp** = 4408 (UK), 5103 (North America), or 5107 (Oceania)

yrBorn *year of respondent's birth*

1. NIS:
   - = dropped if `A7` = -1 or -2
   - otherwise = `A7`
2. ENI:
   - = `ANACxx` where `xx`=`NPELEG`
3. TeO:
   - = `anaise`

AaM *age at migration to destination country*

1. NIS: The respondent's migration trajectory needs to be examined in order to identify the age at which they first moved to the USA
   - = `K4_ii`-**yrBorn**, where `ii` is the first suffix from 01 to 17 for which `K6_ii`=218 (in other words, the age corresponding to the first move to the United States)
2. ENI:
   - "missing" if `EDLLEG`=999
   - otherwise, = `EDLLEG`
3. TeO:
   - =`arrivag`

female *indicator variable for female respondent*

1. NIS:
   - = `A6`-1

   ENI:
   - = `SEXOxx-1`

   TeO:
   - = `sexee`-1

2.2 Language Skills

natLang *respondent's native language/mother tongue*

1. NIS: The mother tongue is not explicitly recorded. Languages used at age 10 are listed, but the most-commonly used language is not; hence, it is assumed to be the first language provided (`J3_1mo`). Since most languages are not listed explicitly, some are imputed based on the currency in which the person was paid in their first job. Note that for some countries, the imputation is based on which languages are explicitly coded versus which are not. For example, in Algeria, the most common non-European language (Arabic) is explicitly coded as a language; hence someone who speaks another non-European language is imputed as speaking Berber, the next-most-common such language in Algeria.
   - = `J3_1mo` if available
   - = 1 (English) if `J3_1mo` not available and `J1`=2 (person never spoke anything other than English)
   - = recode according to Table 1 below, based on value of `J3_1mo`, if applicable:
2. ENI:
   - = `LMAT`
3. TeO: the mother tongue is not explicitly recorded, although a "reference language" (`lref`) is calculated as the most-used childhood language other than French. Mother tongue will be imputed using the following steps, which are similar to the calculation of `lref` while accounting for the fact that French may in fact be the most-used childhood language.
   - `l_lnfrea` records the language which the respondent most prefers; it is one of the 4 possible languages used with mother or father as a child
   - **natLang**=`lnm1` (1st language used with mother) if `l_lnfrea`=1
   - **natLang**=`lnm2` (2nd language used with mother) if `l_lnfrea`=2
   - **natLang**=`lnp1` (1st language used with father) if `l_lnfrea`=3
   - **natLang**=`lnp2` (1st language used with father) if `l_lnfrea`=4
   - **natLang**=`lnm1` if `l_lnfrea` is missing

natLang_fluency *respondent is fluent in native language (binary scale)*

1. NIS: Not available
2. ENI:
   - = 1 if `LMAT`=101 (native speaker of Spanish)
   - otherwise = (`ALFAB`==1)
3. TeO:
   - = (`l_nivlr`$\geq$3)

native_speaker *respondent's mother tongue is English/Spanish/French*

1. NIS:
   - = (**natLang**==1) (native language is English)
2. ENI:
   - = (`LMAT`==101) (native language is Spanish)
3. TeO:
   - = (**natLang**==10) (native language is French)

loc_fluency *fluency in local language (1-4 scale)*

1. NIS:
   - = missing if `J14`=-1 or -2
   - = otherwise, =5-`J14`, to create a 1-to-4 scale, with 1 being "Not at all" and 4 being "Very well"
2. ENI: For language variables, the ENI records information on up to 6 languages, in addition to the mother tongue. In cases when the mother tongue is *not* Spanish, the suffix `xx` below refers to the number for which `IDIOxx`=101; in other words, which of the 6 languages was entered as Spanish.
   - fluency = 4 if `LMAT`=101 (all respondents with Spanish as mother tongue report being fluent)
   - otherwise, fluency =`(COMPxx+LEExx+HABLAxx+ESCRIxx)/4` (average of understanding, speaking, reading, and writing)
   - otherwise = 1 if Spanish not reported as a language the respondent knows
3. TeO:
   - = 4 if **TeO_french_arriv**=4 (if person was fluent on arrival, they did not get asked about language skills again; see TeO section for instructions on **TeO_french_arriv**)
   - = 4 if **AaM**$\leq$3 (immigrants under 3 were not asked the question, and are assumed to be fluent now)
   - = 4 if **natLang**=10 (native speakers)
   - otherwise = `frauj`+1 if available
   - otherwise = 4 if `frauj` missing but `q_franc`=1 (interviewer felt they were fluent)
   - otherwise = 2 if `frauj` missing but `q_franc`=2 (interviewer felt there was some difficulty)
   - otherwise = 1 if `frauj` missing but `q_franc`=3 (interviewer felt there was much difficulty)

fluentYN *fluency on binary scale* = (**loc_fluency**$\geq$3)

2.3 Linguistic Proximity

lang_proxim *linguistic proximity of native language to local language* This variable requires languages to be matched to languages in the Ethnologue database (Lewis *et al.* , 2016). Most languages are unambiguous and can be matched easily. Sometimes the surveys code languages with different names from Ethnologue,

or code multiple languages in one category. Those cases are disambiguated in Table 2 below. After all languages are matched to Ethnologue, linguistic proximity can be calculated using the method in Fearon (2003).

2.4 Migration History & Visa Status

For each respondent in each survey, the migration history can be constructed as a matrix vector with the number of columns equal to the number of reported moves between countries, plus a column capturing birth information.

Migration Trajectories

1. NIS: This survey is detailed enough for each trajectory to contain three rows: year of birth, followed by the year of each move; next, country of birth, followed by the country to which the respondent moved each time; finally, visa status for each time respondent moved to the USA. To construct the trajectory, use variables `K4_1` through `K4_17` (year of each move), `K6_1` through `K6_17` (countries to which respondent moved), and `K11_1` through `K11_17` (visa status, if moved to the USA).
   - Moves may be recorded out of order, so data should be sorted according to years reported in `K4_1` through `K4_17`
   - Birth country is given by variable **bp** above
   - Year born equals `A7`
   - Visa status equals 0 for moves to countries other than the USA
2. ENI: Each trajectory consists of two rows: year of birth, followed by the year of each move; country of birth, followed by the country code for each subsequent move. To construct the trajectory, use variables `NOP01` through `NOP12` (number of each move), `EDPA01` through `EDPA12` (age at which respondent moved), and `PTRS01` through `PTRS12` (country codes for each move).
   - Birth country is given by variable **bp** above
   - Year born equals `ANACyy` year born, where `yy=NPELEG` records the respondent's number in the household.
   - The year of move number `xx` can be (approximately) calculated by adding `EDPAxx` to `ANACyy`.
3. TeO: The two rows for each trajectory contain year of birth, followed by the year of each move; and the code for birthplace, followed by the country code for each subsequent move. To construct each trajectory, use the file `trajmig2` and variables `m_nlig` (move number), `m_anfin` (year of departure), `m_agfin` (age at departure), and `regionmig2` (country code for destination).
   - Birth country is given by variable **bp** above
   - Year of birth equals `anaise`
   - Using the sequence of moves defined by `m_nlig`, the rows of the trajectory will equal the relevant values of `m_anfin` and `regionmig2`

tenure *total years in destination country* This variable equals the total number of years spent in the destination country, as determined by the migration trajectory for each person. To check the robustness of empirical patterns, this may alternatively be defined as the number of years in the most recent stint.

citizen *respondent is a citizen of destination country*

1. NIS: By construction, all respondents in this survey have LPR status and are not US citizens.
2. ENI:
   - = `(NESPxx==1)` where `xx`=`NPELEG` records the respondent's number in the household
3. TeO:
   - = `(inat_lab<3)`

visa *respondent's visa status*

1. NIS: **visa** is coded as the last visa they had upon entry into the US (i.e. before they had LPR status). This is the last value of `K11_xx`, which was used to define the migration matrix above.
2. ENI:
   - "missing" if `DOCUM`=0
   - = 0 if **citizen**=1
   - otherwise = `DOCUM`
3. TeO:
   - "missing" if `m_carte`=9 or 10
   - otherwise = `m_carte`

refugee *respondent has refugee visa or status*

1. NIS: Since everyone in this survey has LPR status, this variable indicates they entered the country with refugee status
   - = (**visa**==1)
2. ENI: Include asylees as well
   - = (`DOCUM`==6—`DOCUM`==7)
3. TeO:
   - = (`m_carte`==1)

2.5 Family and Spousal Variables

cohabit *respondent lives with spouse or partner*

1. NIS:
   - = `(A52==1|A52==2)`

2. ENI:
    - = `(CONVES==1)`
3. TeO:
    - = `(couplee==1)`

spouse_bp *birthplace of respondent's partner/spouse (country code)*

1. NIS: Note: The variable should be coded for the current spouse. The NIS survey requires that answers be recorded for up to 4 marriages with codes `X=1,2,3,4`. The variables below should be filled in with the answer for the most recent marriage: hence `X=4` if applicable; if not, `X=3`; if not, `X=2`, etc.
    - = `A145_X`
2. ENI:
    - = `PNACES`
3. TeO: In this survey "spouse" refers to a married or unmarried partner. Some questions are only asked if the partner lives with the respondent (**couplee**=1). Thus, the following variables should be coded as "missing" if the partner does not live with the respondent (**couplee**=2).
    - same coding strategy as **bp**, using `lnaisc` instead of `lnaise`; `depnc` instead of `depne`; and `regionnaisca2` instead of `regionnaise2`

spouse_bpGroup *spouse's birthplace region* All surveys: recode **spouse_bp** using the same rules as for **bp_group**

endogamy_bp *respondent's partner/spouse is from same birthplace* All surveys: = (**bp**==**spouse_bp**)

endogamy_eth *respondent's partner/spouse has same ethnicity* All surveys: = (**bp_group**==**spouse_bpGroup**)

children *no. of biological children*

1. NIS:
    - = `A232`
2. ENI:
    - "missing" if `TOTHI` missing or `TOTHI`=99
    - otherwise = `TOTHI`
3. TeO:
    - "missing" if e_nbenf=88 or 99
    - otherwise = e_nbenf

ownHouse *respondent owns place of residence*

1. NIS:
    - = (`H1A`==1)

2. ENI:
   - = (TENV≤3)
3. TeO:
   - = (l_antsta==1)

2.6 Education

educ_HS *respondent's highest level of schooling is a high school degree*

1. NIS: Code as 1 if respondent's highest degree is high school, with at most 12 years of school
   - = (A24==3&A20≤12)
2. ENI:
   - = (4≤MNIV≤5&TITULO==1)
3. TeO: For education categories, the TeO survey records highest level of education for those who completed education outside of France (f_finniv) and in France (f_fincla). Only one variable should be filled in for each person. These categories are used to categorize education here and below.
   - = 1 if f_finniv=6 and f_finniv not missing
   - otherwise = 1 if 10≤f_fincla≤16 and f_fincla not missing
   - otherwise = 0

educ_someColl *respondent's highest level of schooling is some college, but no degree*

1. NIS: Code as 1 if respondent's highest degree is high school but has more than 12 years of school
   - = 1 if (A24==3&A20>12)
   - = 1 if (A24==4—A24==999)
   - otherwise = 0
2. ENI:
   - = (MNIV>5&TITULO==0)
3. TeO: **educ_someColl** *respondent attended college but did not graduate*
   - = 1 if f_finniv=7 and and **educ**=7 f_finniv not missing
   - otherwise = 1 if 24≤f_fincla≤28 and f_fincla not missing
   - otherwise = 0

educ_coll *respondent has a college degree or more*

1. NIS:
   - = 1 if (A24≥5&A24 =999)

- otherwise = 0

2. ENI:
   - = (`MNIV`>5&`TITULO`==1)
3. TeO:
   - = 1 if `f_finniv`=7 and and **educ**<7 `f_finniv` not missing
   - otherwise = 1 if 17≤`f_fincla`≤23 and `f_fincla` not missing
   - otherwise = 0

educ_finishAfter *respondent finished education in destination country*

1. NIS:
   - = (`A25mo`==218) (most recent degree is from the USA)
   - otherwise = 1 if `C1`=7 (currently in school)
   - otherwise = 0
2. ENI:
   - = 1 if `TERESP`=1
   - otherwise, = 0 if `TERESP`=6
3. TeO:
   - = 1 if `f_fincla` not missing (highest grade achieved was in France)
   - otherwise = 1 if `f_claact`≤28 (currently in degree program)
   - otherwise = 0

educ_finishBefore *respondent finished education before immigration*

- All surveys: = 1-**educ_finishAfter**

2.7 Employment History and Income

employed *respondent currently employed*

1. NIS:
   - = `(C1==1)`
2. ENI:
   - = `(TRAB2==1)`
3. TeO:
   - = `(situae==1)`

multiJob *respondent works more than one job*

1. NIS:
   - = (C17>1)
2. ENI:
   - = (PLUR1==1)
3. TeO:
   - = (p_sanbw==1)

jobThruRel *respondent procured their job through a friend or relative*

1. NIS:
   - = (C29_1==1—C29_2==1)
2. ENI:
   - = (FORM04==1)
3. TeO:
   - = (p_sacmtw==2—p_sacmtw==3)

jobOfferBefore *respondent had a job offer prior to immigration*

1. NIS:
   - = (C25_1==1—C25_2==1)
2. ENI:
   - = (PROPTR==1)
3. TeO:
   - = (m_atrav==1)

currJob_incperyr *annual income in current job* After following the instructions below, ENI data (in Euros) may be converted to US dollars by multiplying by 1.371 (the average dollar/Euro exchange rate in 2007, the year of the survey). TeO data (also in Euros) may be converted to US dollars by multiplying by 1.4726 (the average exchange rate in 2008). All incomes may then be converted into current US dollars using the CPI, noting the relevant date each survey was conducted.

1. NIS: There are multiple ways to calculate current income in the NIS survey. To most closely mirror calculations of other income variables, use the employment section (Section C). Information is reported separately for two jobs, corresponding to the suffix _1 (main job) and _2 (secondary job), as well as for all other jobs combined. Total income is found via the following steps:
   - Calculate annual income for each job X=1,2 using reported income (C48_X) and time period (C48A_2 or C48a3_X). If applicable, incorporate hours per week (33_X) and weeks per year (C37_X)
   - Convert to US dollars if necessary, using currency (C48A_X or C48A1_X) and the average exchange rate in 2003

- Calculate annual income for all other jobs using reported income (`C58`), time period (`C58A2` or `C58B`), hours per week (`C59`), and currency (`C58a` or `C58A1`)
- Add the incomes from all three sources, if applicable, to get total annual income

2. ENI:
   - = 12·`INGRE` (income per month) if available
   - otherwise, equal to the mean of **currJob_incperyr** for people in the same bracket (defined by `INGRTR`) with the same occupational code (`OCUP2`)
3. TeO:
   - = 12·`p_salnet` (monthly wage income) if available and `p_salnet`≠888888 or 999999
   - otherwise = 12·`p_nsremu` (self-employment income) if available and `p_nsremu`≠888888 or 999999
   - otherwise, if `p_salest` is available (estimated wage income), impute income using the average of either **currJob_incperyr** for people in the same estimated income range, with the same job skill level `p_class`, who reported `p_salnet`
   - otherwise, if `p_nsest` is available (estimated self-employment income), impute impute income using the average of **currJob_incperyr** for people in the same estimated income range, with the same job skill level `p_class`, who reported `p_nsremu`
   - Adjust currency to Euros, if it's in Swiss Francs. Currency is denoted by `p_saluni` if `p_salnet` was used for income, and `p_nsuni` if `p_nsremu` was used for income. To convert from Swiss Francs to Euros, use the currency variable defined below and divide Francs by 1.485 (the average exchange rate in 2008).

2.8 Social Variables

relig_YN *respondent is religious*

1. NIS: NIS allows people to list up to three religions. As long as the first answer is not "none," the person is considered religious.
   - = (`J30_1mo`!=8)
2. ENI:
   - = (`PNXREL`==1)
3. TeO:
   - = 1 if `r_relsoi`=1
   - = 0 if `r_relsoi`=2
   - otherwise "missing"

relig_which *respondent's religion*

1. NIS: The person's first answer is assumed to be their main religion

- = J30_1mo

2. ENI: Not available
3. TeO:
   - = relego1 (in restricted dataset)

healthy *respondent reports good or very good overall health*

1. NIS:
   - = (d1==1—d1==2)
2. ENI: Not available
3. TeO:
   - = (s_etat==1—s_etat==2)

3 NIS-Only Variables

3.1 Employment History

Below, the prefix **firstJob** refers the first job ever worked, usually before immigration; **leaveJob** refers to the last job held before coming to the USA (possibly the same as **firstJob**); **arrivJob** refers to the first job held after arriving in the USA (possibly the same as **leaveJob** or **currJob**).

NIS_firstJob_ctry *country worked first job*

- = B27mo

NIS_firstJob_occ *OCC code for first job*

- = B29oc

NIS_firstJob_cen *Census code for first job*

- = B29in

NIS_firstJob_selfemp *Self-employed in first job*

- = 1 if B32=1 or 3
- = 0 if B32=2
- otherwise, "missing"

NIS_firstJob_incperyr *Annual income in first job*

- Use reported income (B42) and time period (B43 or B45). If needed, use hours per week (B36) and weeks per year (B37)
- Convert to US dollars if necessary, using currency (B40 or B41) and the average exchange rate in the year the person first had the job (B26)

Note: if B50=2 then the respondent did not have another job before arriving in the USA. Fill in the variables below with the same values as the corresponding variables above. Otherwise, if B50=1, follow the instructions below.

NIS_leaveJob_ctry *country worked last pre-US job*

- = B52mo

NIS_leaveJob_occ *OCC code for last pre-US job*

- = B54oc

NIS_leaveJob_cen *Census code for pre-US job*

- = B54in

NIS_leaveJob_selfemp *Self-employed in last pre-US job?*

- = 1 if B57=1 or 3
- = 0 if B57=2
- otherwise, "missing"

NIS_leaveJob_incperyr *amount earned in last pre-US job*

- Use reported income (B66) and time unit (B67 or B69). If needed, use hours per week (B61) and weeks per year (B62)
- Convert to US dollars if necessary, using currency (B65) and the average exchange rate in the year the person first had the job (B51)

Note: if B74=2 then the respondent did not have another job in the US prior to their current job. If so, fill in the variables below using the same values for the corresponding variables in Section 2.7. Otherwise, follow the instructions below.

NIS_arrivJob_occ *OCC code for first post-immigration job*

- = B278oc

NIS_arrivJob_cen *Census code for first post-immigration job*

- = B78in

NIS_arrivJob_incperyr *annual income in first job after migration*

- Use reported income (B91) and time period (B92). If needed, use hours per week (B85) and weeks per year (B86)
- Convert to US dollars if necessary, using currency (B89) and the average exchange rate in the year the person first had the job (B75)

3.2 Language Use

NIS_eng_only *only speak English*

- = (J1==2)

NIS_eng_everAge10 *ever used English at age 10*

- = 1 if any of J3_1 through J3_10 = 1
- otherwise = 0

NIS_eng_everSpouse *sometimes speaks English with spouse* The variable should be coded for the current spouse. The NIS survey requires that answers be recorded for up to 4 marriages with codes X=1,2,3,4. The variables below should be filled in with the answer for the most recent marriage: hence X=4 if applicable; if not, X=3; if not, X=2, etc.

- = 1 if any of A148_X, A150_X, A152_X, A154_X, A156_X, or A158_X = 1 for the appropriate value of X

NIS_eng_onlySpouse *respondent only speaks English with spouse*

- = 1 if only one of A148_X, A148_X, A148_X, A148_X, A148_X, A148_X = 1 and the others are missing or coded 0 ("none")
- otherwise, = 0

NIS_eng_everHome *ever use English at home*

- =1 if any of J5_1 through J5_10 = 1
- =0 otherwise

NIS_eng_everWork *ever use English at work*

- =1 if any of J7_1 through J7_10 = 1
- =0 otherwise

NIS_eng_everFriends *ever use English with friends*

- =1 if any of J10_1 through J10_10 = 1
- =0 otherwise

NIS_eng_everChurch *ever have English spoken in church*

- "missing" if J39≠1 (not a member of a church)
- otherwise = 1 if any of J45_1mo through J45_10mo = 1

NIS_eng_classBefore *took English class before immigration*

- = (J27==1)

NIS_eng_classLastYr *took English class in last year*

- = (J28==1)

NIS_eng_classNow *enrolled in English class now*

- = 1 if A31==1
- otherwise = 1 if A39 indicates English or ESL class

- otherwise = 0

3.3 Religion

NIS_relig_preimm *how often attended religious services before immigration*

- = missing if `J38n`=-1 or -2
- = `J38n` otherwise

NIS_relig_usa *how often attend religious services in USA*

- = missing if `J38O`=-1 or -2
- = otherwise, convert `J38O` to the scale of `J38n` based on how long the respondent has been in the USA

NIS_relig_member *member of a religious organization (church) in USA?*

- = `(J39==1)`

NIS_church_homogCtry *percent of adults at church from same country of origin*

- = `J42` if `J39`=1
- otherwise "missing"

NIS_church_homogLang *percent of adults at church who speak respondent's native language*

- =`J43` if `J39`=1
- otherwise "missing"

4 ENI-Only Variables

4.1 Demographics

Note: The respondent's region of residence determines if there is a second official language for which their skills should be recorded.

ENI_reg *respondent's state/autonomous region of residence*

- = `CCAA`

4.2 Language Use and Regional Language Skills

Note: for regional languages (Valencian/Catalan, Galician, Basque), the survey coding is the same as for Spanish (described above for calculating **loc_fluency**): when the regional language is not the mother tongue, the suffix `xx` corresponds to the number for which `IDIOxx` equals the appropriate code. The relevant ENI language codes are: Catalan = 102, Valencian = 103, Basque = 104, Galician = 105. The variables below should be filled in with the appropriate language: Valencian if **reg**=10; Catalan if **reg**=10 and Valencian

is not recorded; Catalan if **reg**=3, 4, 9; Galician if **reg**=12; Basque if **reg**=15 or 16. For respondents from other regions, the variables should be empty.

ENI_reg_speak *respondent's ability to speak regional autonomous language (1-4 scale)* =

- = 5-`HALECA`

ENI_reg_compYN *respondent comprehends regional language (binary)* = `(COMPxx==1)`

ENI_reg_readYN *respondent reads regional language (binary)* = `(LEExx==1)`

ENI_reg_writeYN *respondent writes regional language (binary)* = `(ESCRIxx==1)`

ENI_reg_speakYN *respondent speaks regional language (binary)* = `(HABLAxx==1)`

ENI_reg_fluency *respondent's overall fluency in regional language (1-4 scale)*

- = 4 if **natLang** is the regional language
- otherwise =`(reg_compYN+span_readYN+span_writeYN+span_speakYN)`
- otherwise = 0 if relevant regional language was not recorded

4.3 Social Variables & Relationship with Birthplace

ENI_bp_contact *does respondent ever talk to people in birthplace?*

- = `(CONFAM==1)`

ENI_bp_lastVisit *year of last visit to birthplace*

- = `ANOULT` if `ANOULT`> 0
- otherwise = `ALLExx` where `xx`=`NPELEG` (year of immigration)

ENI_remitYN *does respondent remit money to birthplace?*

- = `(REMESA==1)`

ENI_remitFreq *frequency of remittance*

- = 0 if **ENI_remitYN**=0
- otherwise = `REFRQ`

ENI_remitAmt *amounted remitted last year*

- = 0 if **ENI_remitYN**=0
- = "missing" if `RECANT`=99999
- otherwise = `RECANT`

ENI_mig_contactYN *did respondent have someone to contact on arrival?*

- = `(LLCONT==1)`

ENI_mig_contactFam *did respondent contact family on arrival?*

- = (LLFAM==1)

ENI_mig_contactFriend *did respondent contact friend/acquaintance on arrival?*

- = (LLAMIG==1|LLCONO==1)

ENI_mig_influYN *did someone from home influence respondent to immigrate?*

- = (INFLU==1)

4.4 Merging with Linguistic Data

ENI_reg_proxim *linguistic proximity between mother tongue and regional language* Linguistic proximity to regional languages can be calculated using the same method as for **lang_proxim** (see instructions above). This variable only applies for respondents living in certain regions. When **ENI_reg**=3, 4, 9, or 10, the proximity to Catalan should be calculated; when **ENI_reg**=12, proximity to Galician. Proximity to Basque is irrelevant, as it is always 0: Basque is a language isolate that shares no branches of the phylogenetic tree with any other language.

5 TeO-Only Variables

5.1 Demographics

TeO_strat *respondent's stratum in the survey (immigrant, French territory, 1st gen., 2nd gen.)*

- = lienmig
- replace = 1 if **bp**>2100 (respondent born abroad, assuming tier is coded incorrectly)

TeO_mom_bp *mother's birthplace*

- Use the same coding strategy as for **bp**, but with regionnaism2 instead of regionnaise2

TeO_mom_bpGroup *mother's birthplace category*

- Use same strategy as for **bp_group** but with **mom_bp** rather than **bp**

TeO_dad_bp *father's birthplace*

- Use the same coding strategy as for **bp**, but with regionnaisp2 instead of regionnaise2

TeO_dad_bpGroup *father's birthplace category*

- Use same strategy as for **bp_group** but with **dad_bp** rather than **bp**

Note: The ethnicity variables below are meant to capture each respondent's heritage, based on where they or their parents were born. This allows for more detailed characterization of native French respondents based on their parents' birthplaces. However, it also means that categories are not mutually exclusive.

TeO_eth *respondent's heritage, coded by group*

- = 1 (South/East Asia and Pacific) if (**bp_group**==1|**TeO_mom_bpGroup**==1|**TeO_dad_bpGroup**==1)
- = 2 (Sub-Saharan Africa) if (**bp_group**==2|**TeO_mom_bpGroup**==2|**TeO_dad_bpGroup**==2)

- = 3 (Middle East and North Africa) if (**bp_group**==3|**TeO_mom_bpGroup**==3|**TeO_dad_bpGroup**==3)
- = 4 (Latin America and Caribbean) if (**bp_group**==4|**TeO_mom_bpGroup**==4|**TeO_dad_bpGroup**==4)
- = 5 (Europe and Central Asia) if (**bp_group**==5|**TeO_mom_bpGroup**==5|**TeO_dad_bpGroup**==5)
- = 6 (English-Speaking Countries/Oceania) if (**bp_group**==6|**TeO_mom_bpGroup**==6|**TeO_dad_bpGroup**==6)
- = 7 (DOM/TOM - only for TeO respondents) if (**TeO_strat**==2|`cpidom_m`==2|`cpidom_p`==2)

5.2 Language Skills & Daily Use

TeO_french_arriv *ability to speak French on arrival (1-4 scale)*

- = `frarri`+1 if `frarri` available
- otherwise = 1 if **AaM**≤3 (question not asked if person was less than 3 years old at time of arrival, so consider them as not speaking any)
- otherwise = 3 if `ilang`=0 & `frarri` missing & **TeO_strat**<3 (only French spoken as child, question not asked, and respondent is an immigrant)

TeO_french_classYN *respondent has taken a French course since immigrating*

- = `(l_courfr==1)`

refLang_fam *respondent speaks reference language with family? (1-4 scale)*

- "missing" if l_lrfami=8
- otherwise = `5-l_lrfami`

refLang_friend *respondent speaks reference language with friends? (1-4 scale)*

- "missing" if l_lrvois=8
- otherwise = `5-l_lrvois`

french_work *respondent speaks French at work? (binary)*

- = `(p_salang==1)` if applicable (respondent works as a regular employee)
- otherwise = `(p_ailang==1)` if applicable (respondent works as an aide)
- otherwise = `(p_nslang==1)` if applicable (respondent is self-employed)

french_spouse *respondent speaks French with spouse? (binary)*

- = `(lncj1_gr==10|lncj2_gr==10)`

french_fam *respondent speaks French with family? (binary)*

- = `(lnfrea_gr==10|lnfreb_gr==10)`

french_child *respondent speaks French with children? (binary)*

- = `(lnenf1_gr==10|lnenf2_gr==10)`

5.3 Education

TeO_educ_onArr *respondent's education at time of arrival*

- = `nivarri`, if available (only asked of people who went to school abroad, then in France, i.e. `f_pisco`=3)
- otherwise = 1 if `etudi`=1 and `f_finniv`≤2 (education completed on arrival, and never went past primary)
- otherwise = 2 if `etudi`=1 and `f_finniv`=3 or 4 (education completed on arrival, and never went past lower secondary)
- otherwise = 3 if `etudi`=1 and `f_finniv`=5 or 6 (education completed on arrival, and never went past higher secondary)
- otherwise = 2 if `etudi`=1 and `f_finniv`=7 (education completed on arrival, and has higher education)
- otherwise = 0 if `f_deban`=0 (never went to school at all)
- otherwise = 0 if `f_pisco`=1 (never went to school outside France)

5.4 Social Variables & Self-Identity

remit_YN *did respondent send remittance last year? (binary)* = `(a_piver==1)`

agree_feelFrench *respondent agrees that "I feel French" (1-4 scale)*

- "missing" if `x_apparf`=5 or 6
- otherwise = `5-x_apparf`

agree_feelBP *respondent agrees that "I feel that I am (demonym of birthplace)" (1-4 scale)*

- "missing" if `x_appare`=5 or 6
- otherwise = `5-x_appare`

agree_feelMomBP *respondent agrees that "I feel that I am (demonym of mother's birthplace)" (1-4 scale)*

- "missing" if `x_apparm`=5 or 6
- otherwise = `5-x_apparm`

agree_feelDadBP *respondent agrees that "I feel that I am (demonym of father's birthplace)" (1-4 scale)*

- "missing" if `x_apparp`=5 or 6
- otherwise = `5-x_apparp`

relig_child *religion was important growing up? (1-4 scale)*

- "missing" if `r_impedu`=8 or 9
- otherwise = `r_impedu`

relig_import *importance of religion now (1-4 scale)*

- "missing" if `r_impvie`=8 or 9

- otherwise = `r_impvie`

relig_clothing *does respondent wear religious clothing? (1-3 scale)*

- "missing" if `r_ostent`=8 or 9
- otherwise = `4-r_ostent`

relig_diet *does respondent observe religious diet? (1-3 scale)*

- "missing" if `r_miam`=8 or 9
- otherwise = `4-r_miam`

6 Tables

Table 1: Rules for imputing variable **natLang** in the NIS survey

Language at Age 10 (`J3_1mo`)	Currency of first job (`B40`)	Implied Country	Implied Language
16 (Other European)	4 (Austrian Schilling)	Austria	German
16	15 (Cyprus Pound)	Cyprus	Greek
16	16 (Czech Koruna)	Czech Rep.	Czech
16	17 (Danish Krone)	Denmark	Danish
16	18 (Dutch Guilder)	Netherlands	Dutch
16	21 (Finnish Marka)	Finland	Finnish
16	23 (German Mark)	Germany	German
16	25 (Greek Drachma)	Greece	Greek
16	28 (Hungarian Forint)	Hungary	Hungarian
16	29 (Israeli Shekel)	Israel	Hebrew
16	33 (Italian Lira)	Italy	Italian
16	44 (Norwegian Krone)	Norway	Norwegian
16	55 (Swedish Krona)	Sweden	Swedish
17 (Other Non-European)	1 (Algerian Dinar)	Algeria	Berber
17	24 (Ghana Cedi)	Ghana	Akan
17	29 (Israeli Shekel)	Israel	Hebrew
17	34 (Japanese Yen)	Japan	Japanese
17	36 (Kenyan Schilling)	Kenya	Swahili
17	41 (Malaysian Ringgit)	Malaysia	Malay
17	52 (Singapore Dollar)	Singapore	Malay
17	53 (South African Rand)	South Africa	Zulu
17	59 (Thai Baht)	Thailand	Thai
18 (Other Philippines)	ANY	Philippines	Filipino
19 (Other India)	ANY	India	Tamil

Table 2: Matching Survey Language Names to Ethnologue Languages

Survey	Survey Language Name	Corresponding Ethnologue Name
ENI	Yaounde	French
ENI	Flemish	Dutch
ENI	Rifeno	Berber
ENI	Tamazight	Berber
ENI	Fante	Akan
ENI	Valencian	Catalan
ENI	Bosnian/Croatian/Serbian	Serbo-Croatian
ENI	Moldovan	Romanian
TeO	Other Slavic	Russian
TeO	Indo-European	Bulgarian
TeO	Other Niger-Congo	Wolof
TeO	French Patois	French Creole
TeO	Other	Nilo-Saharan
TeO	Austronesian	Pangasinan
TeO	Indo-Iranian	Farsi
TeO	Other Indo-European	Romanian

References

Beauchemin, Cris, & Simon, Patrick. *TeO Project Description.* `http://teo_english.site.ined.fr/fichier/s_rubrique/20309/teo.note.eng.en.pdf`.

Beauchemin, Cris, Hamelle, Christelle, & Simon, Patrick. 2010 (October). *Trajectories and Origins Survey on Population Diversity in France: Initial Findings.* Report. Institut national d'études démographiques.

Fearon, James D. 2003. Ethnic and Cultural Diversity by Country. *Journal of Economic Growth*, **8**(2), 195–222.

Jasso, Guillermina, Massey, Douglas S., Rosenzweig, Mark R., & Smith, James P. 2005. The U.S. New Immigrant Survey: Overview and Preliminary Results Based on the New-Immigrant Cohorts of 1996 and 2003. *In:* Morgan, Beverley, & Nicholson, Ben (eds), *Immigration Research and Statistics Service Workshop on Longitudinal Surveys and Cross-Cultural Survey Design: Workshop Proceedings.* Crown Publishing.

Lewis, M. Paul, Simons, Gary F., & Fennig, Charles D. 2016. *Ethnologue: Languages of the World, Nineteenth Edition.* Online Version http://www.ethnologue.com.

Marrone, James V. 2016. *A Human Capital Model of Linguistic and Cultural Assimilation.* Working Paper.

Reher, David, & Requena, Miguel. 2009. The National Immigrant Survey of Spain: A new data source for migration studies The National Immigrant Survey of Spain: A New Data Source for Migration Studies in Europe. *Demographic Research*, **20**(March), 253–278.