sociological science

Supplement to:

Abramitzky et al. 2023. "The Refugee Advantage: English-Language Attainment in the Early Twentieth Century." Sociological Science 10: 769-805.

Appendix Table 1. Coding scheme for variables collected from oral histories.

Variable Name	Label	Description/Coding	Data Type	Coding Notes
id	Numbered list of entries	Number	Integer	
source	Oral history archive (e.g. Ellis Island Foundation)	Organization name	Open String	See "sources" tab
firstname	Respondent's first name at time of interview	Name	Open String	
lastname	Respondent's middle name at time of interview	Name	Open String	
birthplace_general	Country of birth noted at top of oral history	Country	Open String	
birthplace_specific	Specific place of birth given by respondent when asked where they were born	Town/Area, County/Country	Open String	
birth_year	Birth year	yyyy	Integer	N/A if missing
arrival_year	Year that respondent arrived	yyyy	Integer	N/A if missing
age_at_arrival	Age the respondent was when they came to the US	Age	Integer	N/A if missing
interview_year	Year of interview	yyyy	Integer	N/A if missing
age_at_interview	Age of respondent at time of interview	Age	Integer	N/A if missing
religion	Respondent's religion	Name (i.e. catholic, orthodox, jewish, etc)	Open String	
refugee	Respondent left country for reasons of violence, religious/ethnic/racial persecution, or an event such as famine	Yes/No	Constrained String	
pm_urban_rural	Grew up in rural or urban area before migration	Urban/Rural	Constrained String	Rural = agricultural area or small town or village. Urban = city or large town, Urban-rural = grew up in some mix of urban and rural places
pm_occupation	Last job of respondent in home country before they migrated	Name (i.e. farm laborer, helper, etc)	Open String	This is before they came to the US
pm_father_occ	Respondent's father's job in home country	Name (i.e. farmer, blacksmith, etc)	Open String	
pm_mother_occ	Respondent's mother's job in home country	Name (i.e. farmer, blacksmith, etc)	Open String	

pm_educ_level	Highest level completed before the interviewee migrated	Options: none, elementary school, some high school, high school, some college, trade school, associate's degree, bachelor's degree, master's degree, doctoral or professional degree, N/A	Constrained String	If respondent migrated during school, indicate level at migration (e.g. someone who started high school abroad and finished high school in US would be "some high school")
pm_married	Already married when they migrated?	Yes/No	Constrained String	
language_native	Respondent's native language	Language (e.g. English, Russian etc)	Open String	
pm_english	Did the interviewee know English before migrating?	Yes/No	Constrained String	
migrate_alone	Did the interviewee migrate alone?	Yes/No	Constrained String	"No" if they travelled alone but were meeting someone, or if someone else in their party was following immediately after (within 6 months)
documented	"Yes" if respondent talked about a visa, greencard, etc; "No" if respondent was undocumented when they migrated	Yes/No	Constrained String	"NA" if not discussed
refugee	Was the respondent a refugee?	Yes/No	Constrained String	
family_sponsored	Did respondent's relatives financial support their migration or apply for visas for them?	Yes/No	Constrained String	
org_sponsored	Did respondent receive aid and/or support from a NGO to migrate? (e.g. church sponsoring or hosting refugees; educational scholarship from a university or nonprofit...)	Yes/No	Constrained String	
learn_english	If the respondent did not speak English before migration, how did they learn English? (e.g. school, friends, college course)	Open String	Not a Yes/No	
total_children	number of children the respondent had (including adopted)	0, 1, 2...	Integer	
am_married	Respondent got married after migration	Yes/No	Constrained String	"No" if the respondent got married before migrating and did not remarry

spouse_same _nationality	Spouse born in same country as respondent	Yes/No	Constrained String	Answer can be for a spouse married before or after migration
am_occ_first	Respondent's first job in US	Name of job	Open String	
am_latest_occ	What is the respondent's current job? If retired or unemployed, what was their most recent job?	name of job	Open String	
am_educ_level	Highest educational level completed at time of interview	Options: none, elementary school, some high school, high school, some college, trade school, associate's degree, bachelor's degree, master's degree, doctoral or professional degree, N/A	Constrained String	If the interviewee completed no further school after migration, response should match pm_educ_level answer
am_loc_first	Location of first residence in US	location in US (e.g. Portland, Maine)	Open String	
imm_destination	Was respondent's first location somewhere where they had a community of immigrants from the same country?	Yes/No	Constrained String	
am_loc_current	Location of current residence in US (at time of interview)	Location in US (e.g. Portland, Maine)	Open String	
ever_visit	Has the respondent been back to their home country since migrating?	Yes/No	Constrained String	
citizen	Did respondent ever become an American citizen?	Yes/No	Constrained String	

Note: This table provides the coding scheme used by research assistants to convert transcripts obtained from oral histories to data. "Open String" refers to RAs being able to enter the information that they come across in the oral histories, whereas "Constrained String" refers to RAs picking between the provided options.

Appendix Table 2. Summary statistics of measures of linguistic ability.

	N	Mean	Std Dev	Min	Max
Age of Acquisition	1100	4.86	0.33	4	6
Mean Sentence Length	1100	12.20	4.54	2	42
Accent	915	0	1	-4	3
Syllables Per Minute	884	163.76	39.47	53	274

Note: This table shows the summary statistics of the four measures of linguistic ability: Age of Acquisition (AoA), Mean Sentence Length (MSL), Accent and Syllables Per Minute (SPM).

Appendix Table 3. Correlation between measures of linguistic ability.

	AoA	MSL	Accent	Years of Schooling	Arrival Age	Income	Syllables Per Minute
Age of Acquisition	1.0000						
Mean Sentence Length	0.3812	1.0000					
Accent	0.3199	0.2577	1.0000				
Years of Schooling	0.5394	0.2006	0.3202	1.0000			
Arrival age	-0.1551	-0.1381	-0.5214	-0.2068	1.0000		
Income	0.2431	0.1371	0.0355	0.2389	0.0548	1.0000	
Syllables Per Minute	0.1222	0.1434	0.2652	0.2064	-0.1957	0.0275	1.0000

Note: This table depicts correlation between the four measures of linguistic ability: Age of Acquisition (AoA), Mean Sentence Length (MSL), Accent and Syllables Per Minute (SPM) and other characteristics of migrants from our data such as years of schooling, arrival age and income. N= 438.

Appendix Table 4. Summary of non-refugee and refugees in the National Immigrant Survey

	Non-refugee		Refugee	
	Observations	Mean	Observations	Mean
Speaks English	2,994	.74	169	.85
		(.44)		(.36)
Speaks English well	2,994	.39	169	.33
		(.49)		(.47)
English class, current or within the last year	2,871	.28	169	.33
		(.45)		(.47)
English class before arrival in US	2,830	.34	168	.26
		(.47)		(.44)
Age	3,148	40.36	171	39.93
		(14.75)		(14.62)
Female	3,157	.54	171	.51
		(.50)		(.50)
Years of schooling	3,148	12.10	171	11.37
		(4.72)		(4.11)
Rural	3,151	.42	171	.41
		(.49)		(.49)
Catholic	3,157	.32	171	.12
		(.47)		(.33)
Christian Orthodox	3,157	.11	171	.18
		(.31)		(.38)
Protestant	3,157	.15	171	.26
		(.35)		(.44)
Muslim	3,157	.11	171	.12
		(.32)		(.33)
Other Religion	3,157	.14	171	.17
		(.35)		(.38)
No Religion or declined to answer	3,157	.12	171	.15
		(.33)		(.35)
Year of departure	3,125	2000	168	1997
		(6.73)		(5.34)
Europe and Central Asia	3,157	.10	171	.26
		(.30)		(.44)
Russia, Ukraine and Poland	3,157	.03	171	.36
		(.18)		(.48)
Middle East and North/Sub-Saharan Africa	3,157	.16	171	.16
		(.37)		(.37)
Rest of World	3,157	.71	171	.22
		(.46)		(.41)
N	3,157		171	

Note: This table presents the characteristics of the 2003 cohort of surveyed immigrants in NIS, split by non-refugee and refugee status. This cohort is a random sample of adults receiving legal permanent residence between May and November of 2003. Refugee = 1 for immigrants with Refugee or Asylee visa status. The standard deviation of each variable is in parentheses.

Appendix Table 5. Association between refugee status and linguistic outcomes, immigrants arriving after age 12.

	AoA	AoA (drop words)	MSL	Accent	SPM	AoA	AoA(drop words)	MSL	Accent	SPM
Refugee or mixed reasons	0.386***	0.352***	0.127	-0.00634	0.160	0.241**	0.215*	0.0227	-0.134	0.0921
	(0.098)	(0.098)	(0.100)	(0.101)	(0.110)	(0.110)	(0.112)	(0.110)	(0.121)	(0.126)
Laborer						-0.138	-0.149	-0.124	0.490***	0.00876
						(0.176)	(0.178)	(0.175)	(0.189)	(0.199)
Skilled						0.0840	0.0763	0.0973	0.250	0.0398
						(0.141)	(0.142)	(0.140)	(0.153)	(0.161)
White Collar						0.321**	0.310**	0.149	0.156	-0.137
						(0.150)	(0.152)	(0.150)	(0.164)	(0.173)
Urban						0.175*	0.173*	-0.0446	-0.0444	0.0222
						(0.104)	(0.105)	(0.103)	(0.113)	(0.120)
Catholic						-0.265	-0.272	-0.441***	-0.0983	-0.433**
						(0.166)	(0.167)	(0.165)	(0.175)	(0.182)
Jewish						0.0941	0.0525	-0.0847	0.220	-0.108
						(0.163)	(0.165)	(0.162)	(0.174)	(0.182)
Protestant						-0.247	-0.252	-0.380**	-0.108	-0.694***
						(0.162)	(0.164)	(0.161)	(0.172)	(0.182)
Orthodox						-0.514**	-0.533**	-0.615***	-0.337	-0.0177
						(0.223)	(0.225)	(0.221)	(0.242)	(0.252)
Outcome mean	-0.118	-0.127	-0.119	-0.513	-0.145	-0.109	-0.120	-0.157	-0.525	-0.111
R^2	0.346	0.334	0.151	0.123	0.0941	0.384	0.369	0.197	0.162	0.170
N	454	454	454	393	383	360	360	360	317	309

Note: This table reports the underlying coefficients for Figure 5. Linguistic measures have been standardized. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are not included in this sample. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Results are unweighted.
$p < 0.1$; **$p < 0.05$; ***$p < 0.01$; standard errors are shown in parentheses.

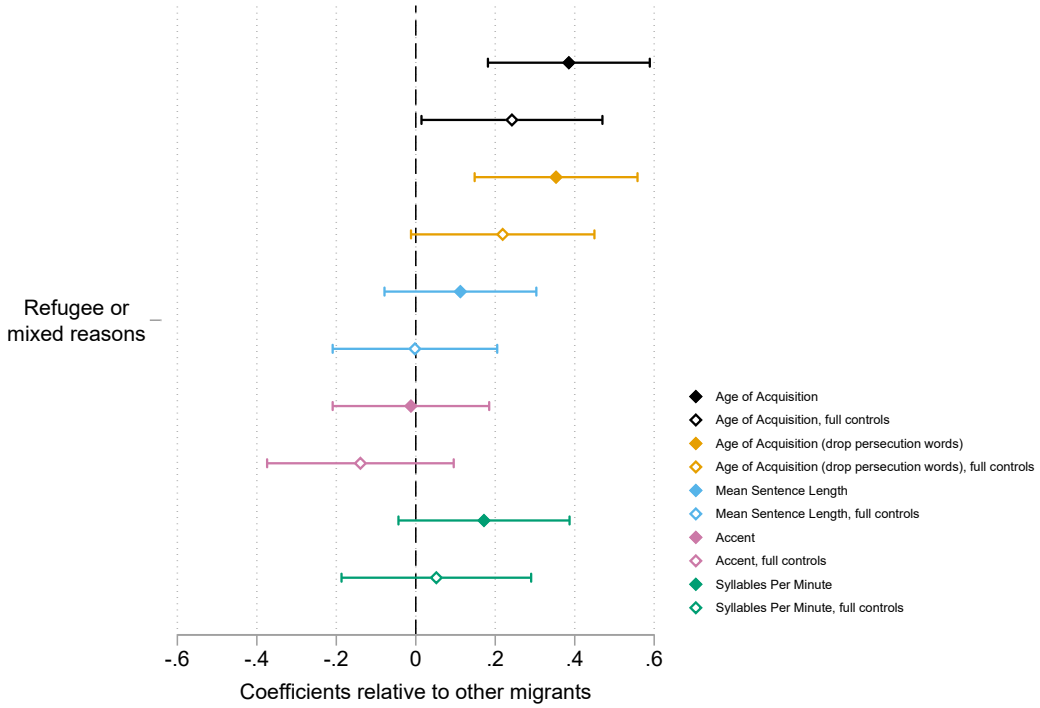
Appendix Table 6. Persecution related words in Age of Acquisition measure.

Word	Age of Acquisition	Total Usage
annexation	13.85	1
anti	9.37	170
armistice	14	9
armistices	14	1
army	7.15	861
attack	6.58	65
battalion	11.95	29
bomb	8	54
camps	5.78	70
communism	11.74	14
communist	13.22	54
communists	13.22	48
duty	7.15	76
fascism	14.33	3
fascist	14.68	8
freedom	7.05	175
genocide	13.2	23
ghetto	10.15	65
kill	6.35	217
killed	6.35	458
navy	7.15	127
oven	5.67	177
pogrom	14.33	33
pogroms	14.33	53
recession	13.74	11
refugees	10.56	107
revolution	10	68
socialist	13.61	25
socialists	13.61	16
survive	7.11	73
survived	7.11	121

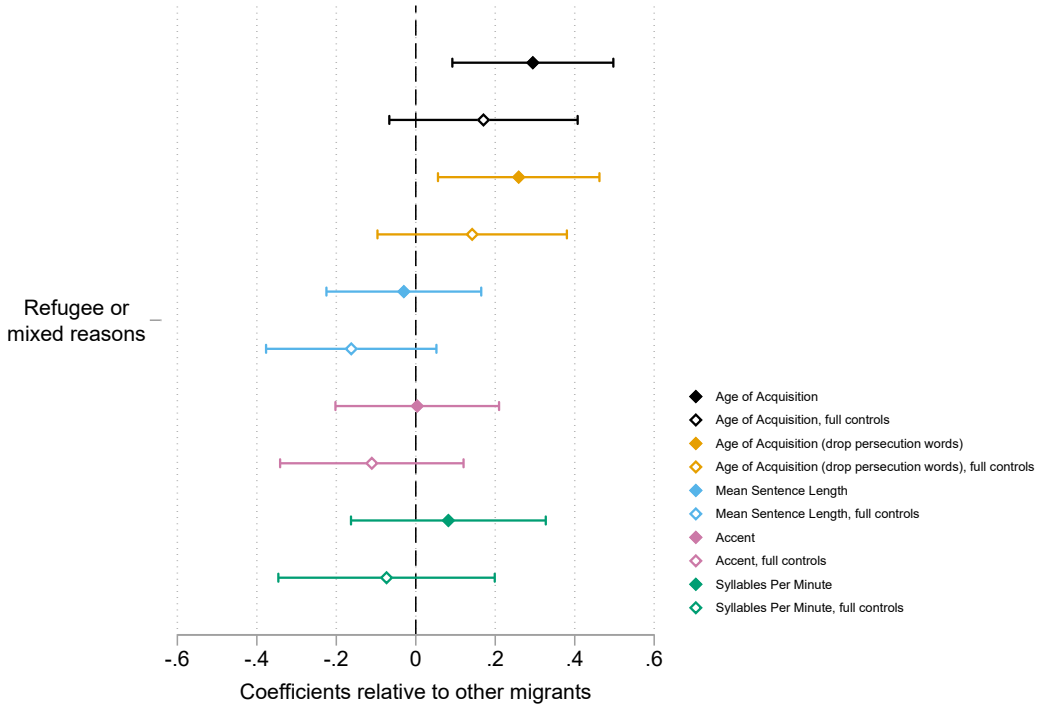
Note: This table lists persecution related words used in oral histories. These words were dropped to create a robust Age of Acquisition measure, referred to as "Age of Acquisition (drop persecution words)" in main figures.

ID_base	row	first_name	middle_name	last_name	birthplace_general	birthplace_specific	birth_date	interviewer	interview_location	interview_year
1167	1	Jenny	NA	Bohsung	Germany	Westhausen, Germany	NA	Margo Nash	NA	1973
1184	2	John (Jack)	Brenden	Brady	Ireland	Ballyhaise, Ireland	7/29/23	Paul E. Sigrest	Ellis Island Recording Studio	1995
1190	3	Rose	NA	Breci	Italy	Carletini, Italy	NA	Dr. Willa Appel	Ellis Island Recording Studio	1985
1371	4	Tilda	NA	De Mello	Brazil	Manaus, Brazil	4/12/17	Janet Levine	Albertson, New York	1992
1494	5	Rita	Costa	Finco	Italy	Asiago, Italy	7/16/16	Janet Levine	Cudahy, Wisconsin	1996
1642	6	Julia	Barlas	Groulx	Greece	NA		Nancy Dallet	NA	1989

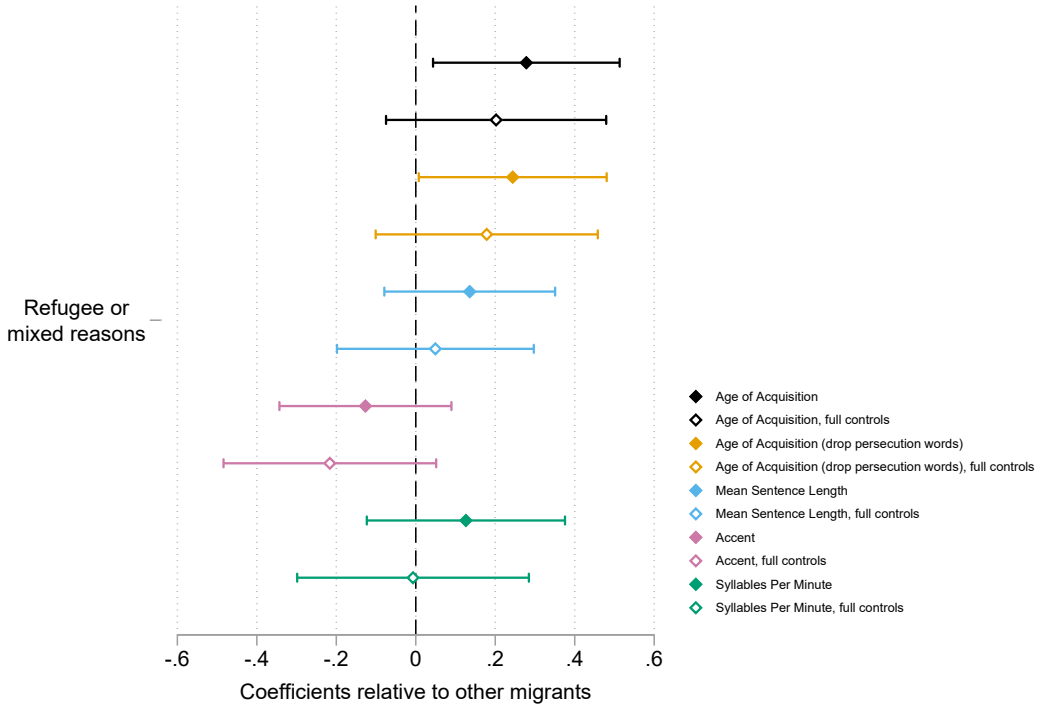
Appendix Figure 1. Image of standardized template used to manually extract data from oral histories. Image of part of template used by coders to fill in information for individuals in oral histories.



Appendix Figure 2. Association between refugee status and linguistic outcomes for immigrants arriving after age 12, English speakers included. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are included in this sample. The five linguistic measures are: Age of Acquisition (N= 512, N= 413 with full controls), Age of Acquisition calculated after dropping persecution related words (N= 512, N= 413), Mean Sentence Length (N= 512, N= 413), Accent (N= 449, N= 368) and Syllables Per Minute (N =438, N=359). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. A more positive accent score indicates an accent closer to that of the US born. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Significance is at the 5% level.



Appendix Figure 3. Association between refugee status and linguistic outcomes for immigrants arriving after age 12, dropping migrants arriving after 1933, English speakers included. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are included in this sample. The four linguistic measures are: Age of Acquisition (N= 424, N= 340 with full controls), Age of Acquisition calculated after dropping persecution related words (N= 424, N= 340), Mean Sentence Length (N= 424, N= 340), Accent (N= 372, N= 302) and Syllables Per Minute (N =362, N=294). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Results are unweighted. Significance is at the 5% level.



Appendix Figure 4. Association between refugee status and linguistic outcomes for immigrants arriving after age 14, English speakers included. Refugee = 1 indicates that an immigrant was coded as a refugee in both the first and second round of coding. Mixed reasons = 1 indicates that an immigrant was coded as a refugee in only the first round or only the second round. Here, we combine the two into an indicator = 1 if an immigrant was either coded as a refugee or as moving for mixed reasons. English speaking immigrants, those from Britain and Ireland, are included in this sample. The five linguistic measures are: Age of Acquisition (N= 415, N= 331 with full controls), Age of Acquisition calculated after dropping persecution related words (N= 415, N= 331), Mean Sentence Length (N= 415, N= 331), Accent (N= 363, N= 296) and Syllables Per Minute (N =355, N=288). Linguistic measures have been standardized to have a mean of zero and standard deviation of one. Controls for all regressions include age, age squared, arrival period, birthplace and gender. Added controls in regressions with "full controls" include father's pre-migration occupation, pre-migration urban status and religion. Significance is at the 5% level.