

SOEP 2026 - 16th International German Socio-Economic Panel User Conference

Title: Accounting for Migration Selectivity in Wage-Destination Country Language Fluency Estimates

Authors:

Dean R. Lillard, Department of Human Sciences, The Ohio State University, DIW-Berlin, NBER
Xin "Johnathan" Liu*, Department of Economics, The Ohio State University (presenting author)

We study the relationship between a migrant's labor earnings and his fluency speaking, reading, and writing the new language. Countries experience substantial inflows of legal and illegal migrants who seek better economic conditions, physical safety, or personal freedom. Migrants' presence involves complex issues such as competition for jobs requiring fewer skills, increased demand for housing, and difficulties assimilating the destination country's social norms. Careful research must confront complex sample selection issues and measure how societies facilitate or impede a migrant's integration. To address sample selection, we use an IV estimator and an instrument that captures exogenous differences in how difficult it is for a migrant to become fluent. Generically, scholars use the *linguistic "distance"* between a migrant's mother tongue and his destination country's language. Linguists formally measure distance by indices that characterize common features of syntax, roots of words, and shared etymological language development. We extend the previous literature by employing separate distance measures to differentiate between the above three aspects of language. We expect this extension to be fruitful because the labor market return to each plausibly varies across different types of occupations. We also control for pre-migration differences across groups in migrants' access to schooling opportunities and foreign language requirements in their country of origin when they were of secondary school age and the presence of enclaves of country-of-origin migrants in their destination country. We draw panel data from German Socio-Economic Panel (Germany), Panel Study of Income Dynamics (US), and Understanding Society (UK) on migrants' country of origin. Each data set has information on country-of-origin, timing of migration, and the migrant's (self-reported) overall destination language proficiency. Table 1 shows sample sizes and linguistic distance of migrants from just a few of the countries we observe. (To observe country of origin in the PSID we access restricted data.)

Table 1 Sample sizes of respondents, by country of origin

Country	UKHLS		SOEP		Mother tongue
	N	LingDist	N	LingDist	
Turkey	224	96.3	5527	87.8	TURKISH
Poland	809	59	2888	53.1	POLISH
Italy	255	52.5	1616	47.6	ITALIAN
India	2214	68.9	1201	63.8	HINDI
Spain	164	59.3	800	56.8	SPANISH
Nigeria	655	91	771	88	<u>YORUBA/ENGLISH</u>
Pakistan	2025	65.5	477	60.7	URDU
U.S.A	348	0	414	31.3	ENGLISH
China/Hong Kong	412	88.5	315	87.2	MANDARIN
France	234	46.9	257	55.7	FRENCH

Notes: authors' compilation from paneldata.org and UKHLS website