**Using a Susceptible-Infected-Removed Discrete Time (SIR) model to predict the Future Proportion of English Speakers Under High-Migration Assumptions**

**Model**

Drawing from a paper which uses an SIR model to predict indigenous language extinction[[1]](#endnote-1), I use an epidemiological model that treats English as a spreadable pathogen to predict the long-term proportion of English-speakers in the United States under the assumption of high migration from non-English-speaking immigrants.

The basic model is an open population discrete time Susceptible-Infected model which categorizes people unable to speak English as ‘susceptible’ and those who speak English as ‘infected.’ There is no ‘removed’ or ‘recovered’ group, since individuals are assumed to remain English speakers (ESs) for life after acquiring the language, and those who leave the population through death are assumed not to have any influence on the spread of English. Non-English speakers (NESs) can acquire English through contact with ESs; once they are ‘infected’, they enter the ‘infected’ group of the population in the next year. This assumption does not imply that NESs usually become proficient speakers within a year; some research shows that a reasonable estimate is 3-5 years.[[2]](#endnote-2) However, acquiring a language is a gradual process which may involve roping other NESs into learning, so one person’s time spent learning English should, in my view, be considered part of the ‘infectious period’. Furthermore, NESs in this model are defined as those who speak English less than ‘very well’, so individuals in the ‘susceptible’ group may be at differing points in the process of learning English.

There are two communities in this model, the natives and the immigrants, who interact with one another at differing rates (in SIR models, this is called non-homogenous mixing). All natives are assumed to speak English at all times and give birth to English-speaking children, and some proportion of immigrants are assumed to speak English at t = 0. Every year, some positive number of migrants arrive, none of whom are assumed to speak English, and all of whom enter the immigrant community.

Since the typical application of SIR models is to infectious diseases, the infected are usually a minority of the population to begin with. The spread of the disease is a function of the transmissibility, the number of the infected, and number of susceptible individuals, and the contact rate between members of the population. Transmissibility and contact are typically represented by the parameter *β*, and the growth of the infected population, assuming no recovery, is usually formulated as:

,

where *I* represents the number of infected individuals and *S* represents the number of susceptible individuals. In this study, the infected vastly outnumber the susceptible to begin with. Under the ordinary definitions of the transmission parameter, *β* would have to be very small in order for any susceptible individuals to avoid infection for more than a few periods, even under the assumption of non-homogenous mixing. This does not lend itself well to interpretation, nor does it intuitively represent the situation we are attempting to model.

In this two-community model, immigrants are assumed to make contact with the native community at a rate that is independent of the number of natives. Therefore, the transmissibility parameter should not be a function of the number of natives. The number of ‘infected’ immigrants is important however… [CHANGE MODEL EQUATIONS TO INCORPORATE X]

The model and its parameters are defined below:

: Native population at time *t,* assumed to be 100% English-speaking.

: Immigrant population at time *t*.

: Total population at time *t*.

: English-speaking immigrant population at time *t*.

: Total Non-English-speaking population at time *t*.

: Total English-speaking population at time *t*.

: Number of arrived migrants at time *t,* assumed to be entirely non-English speaking.

: Annual birth rate

: Annual death rate

: Contact parameter between immigrant and native communities.

: Transmission rate of English from the native community to immigrant individuals.

+: Weighted transmission rate under non-homogenous mixing.

: Discrete time evolution of *N.*

Model Equations:



**Data and Calibration**

The birth and death rates are assumed to be fixed across years and identical between immigrant and native communities. For both, I use the 2021 values from Macrotrends.net[[3]](#endnote-3), which are 1.2% and 0.9% respectively. For immigrant and native populations, I use the American Community Survey (ACS) one-year estimates available through the Integrated Public Use Microdata Series (IPUMS), an individual-level population database. By employing a residual statistical method created by economist Christian Gunadi, I estimate the immigrant population, legal and illegal, at the state level for the 2005-2019 period[[4]](#endnote-4). Estimates for yearly migration come from the Department of Homeland Security’s Yearbook of Immigration Statistics[[5]](#endnote-5), which record the number of immigrants obtaining lawful permanent residence in the United States by state and year. While this data does only reflect legal immigration … [still thinking about how to address this given that there isn’t reliable data on new arrivals from illegals]

Estimates for the share of English-speaking immigrants are more difficult to obtain for the full 2005-2019 period. The U.S. Census Bureau’s ACS Table B16005 records the number of foreign-born residents who claim to speak English ‘exclusively’, ‘very well’, ‘well’, ‘not well’, or ‘not at all’ from 2010-2019. For the purpose of this model, I assume those who speak English less than ‘very well’ to be the in the ‘susceptible’ group. In order to fill in the missing data for 2005-2009, I impute the missing shares of the foreign-born population who speak English using a linear trend calculated from the 2010-2019 values.

With these data, we can see the real observed trend in the English-speaking population over time by state. Using the 2005 observed values as the t = 0 values for my model, I calibrate the model by running it for 2005-2019 by state for all valid combinations (to three decimal points) of the unknown parameters: the transmissibility parameter *β* and the native-immigrant contact parameter *c*. I calculate the sum of squared errors (SSE) for the proportion of ESs produced by each of these iterations of the model and take the parameter combination which minimizes the SSE across the entire sample as the chosen values. The values obtained are 0.055 for the transmissibility parameter and 0.995 for the contact parameter. Many other value combinations work well, but the product of the two parameters always stays within a narrow range. This suggests the intuitive fact that a high transmissibility rate may compensate for a low inter-community contact rate and vice versa. The fact that the contact parameter is close to 1 or that the transmissibility rate is only 0.05 is somewhat arbitrary; a transmissibility of 0.08 and a contact rate of 0.57 function nearly as well in terms of matching the observed data, for example.

**Projection**

The only variable which is not predicted within the context of the model for years beyond 2019 is annual migration. For this variable, I impute a linear trend using the 2005-2019 observed values. This uniformly results in a stark upward trend in yearly migration for each state, which may not be realistic as it almost certainly overstates the inflow of migrants in future years. However, the purpose of this research is to show that even under unrealistically strong assumptions regarding NESs (the other two of which are that all migrants are NSEs and all children born to migrants are NESs), English remains proficiently spoken by the vast majority of the population for many decades under this model.

Using the parameter values obtained above and the observed data for 2005 as values for t = 0, I then project the proportion of the total population able to speak English at least ‘very well’ out to 2100 for each state and for the nation (Figure X). The long-run trend for most states and the nation is a nearly a flat line. Even under these strong assumptions, the transmission of English counteracts increasing migrant flows of NSEs as well as the birth of new NSEs within the population, and English remains proficiently spoken by the vast majority of the population until the end of the model. Additionally, the slight decline observed in most states eventually flattens out, despite increasing migration which in some states leads the immigrant communities to eventually dominate the native communities in terms of size (Figure Y).

Figure Xa: Estimated Proportion of the Population able to Speak English (2005-2100)



Figure Xb: Estimated Proportion of the Population able to Speak English (2005-2100)



Figure Ya: Estimated Population (2005-2100)



Figure Yb: Estimated Population (2005-2100)



1. <https://www.econstor.eu/bitstream/10419/210691/1/10.21307_stattrans-2019-035.pdf> [↑](#endnote-ref-1)
2. https://web.stanford.edu/~hakuta/Publications/(2000)%20-%20HOW%20LONG%20DOES%20IT%20TAKE%20ENGLISH%20LEARNERS%20TO%20ATTAIN%20PR.pdf [↑](#endnote-ref-2)
3. [https://www.macrotrends.net/countries/USA/united-states](https://www.macrotrends.net/countries/USA/united-stateste). While one might expect the death rate to be higher in 2021 due to the COVID-19 pandemic, the trend in the US death rate appears not to have been impacted by the pandemic. [↑](#endnote-ref-3)
4. . IPUMS USA, <https://usa.ipums.org/usa/>; Christian Gunadi, “On the Association between Undocumented Immigration and Crime in the United States,” Oxford Economic Papers 73, no. 1 (January 2021): 200–24; Michelangelo Landgrave and Alex Nowrasteh, “Illegal Immigrant Incarceration Rates, 2010–2018: Demographics and Policy Implications,” Cato Institute Policy Analysis no. 890, April 21, 2020; Alex Nowrasteh, Andrew C. Forrester, and Michelangelo Landgrave, “Illegal Immigration and Crime in Texas,” Cato Institute Working Paper no. 60, October 13, 2020 (accepted for a chapter in the forthcoming book *On Inequality and Freedom*, edited by Lawrence Eppard, to be published by Oxford University Press in 2021). [↑](#endnote-ref-4)
5. https://www.dhs.gov/immigration-statistics/yearbook [↑](#endnote-ref-5)