Analysis Plan for Language Learning and Social Integration Project

Cornel Nesseler, Hammad Shaikh, and Jiaqi Zou

Updated: Dec 12, 2025

1 Empirical Framework and Results

Bar plot analysis. We will plot the mean response rate for each text version separately for the football clubs and the rental housing market. This will involve plotting response rate over 6 conditions.

Regression analysis. We will estimate the intent-to-treat effects of acquiring local language proficiency and participating in a local language learning class using the following specification:

$$y_{ir} = \alpha_r + \sum_{j=1}^{5} \beta_j TextVersion_{ir}^{j} + X_{ir}'\Theta + \epsilon_{ir}, \tag{1}$$

where y_{ir} is an indicator for whether the coach or landlord i in region r responds to the message. When emails of realtors or landlords are available, we will also have an indicator for whether the email was opened as another outcome.[1] The variable $TextVersion_{ir}^{j}$ is an indicator for receiving message variation $j \in \{1, 2, 3, 4, 5\}$ (version $j = 0$ is omitted).[2] Since the randomization is carried out within each day of week email is send and the region, we include the day-region strata fixed effects (α_r) in all specifications.[3] The vector X_{ir} includes characteristics of the football club or rental property. When examining social integration using football clubs, we control the league division, whether the team has a website (or a social media page), and whether the team only has males. For the rental housing market, we control the rent, size of property, number rooms, whether it has a balcony, whether it is furnished, whether it has a garage, whether it is furnished, and whether the listing is represented by a professional realtor. Additionally, we control the message sender's gender and foreign group. We will use variables that are commonly variable across all countries for our main analysis.

The main coefficients of interest are the $\beta_j s$, reflecting the differential response rate of receiving text version j from a foreign-sounding name relative to a message written by a native speaker with a native-sounding name. We estimate the model separately for the football clubs and the housing market. For robustness, we also report the results with and without the inclusion of the covariates.

As the randomization is not among clusters, but rather at the unit-level, we will use robust standard errors. Although, as there may be correlation in unobserved determinants of the response rate within a city, we will check the sensitivity our results by clustering the standard errors at the city-level as long as if we have a sufficient number of large cities.

[1] In some countries emails of relators or landlords are not listed in the posting, in this case we need to use the contact form provided and may not be able to track open rates.

[2] Omitted text version is the control group, native sounding name and proficient local language.

[3] Country fixed effects are not included separately as the regions partition the country. We will construct the regions at the city-level, and plan to pool the smaller cities with less observations into a separate strata.

Although our preferred estimation strategy is OLS, we may also report results using probit and logit for sensitivity analysis.

Main Hypotheses: We expect the results to be monotonic in language proficiency and language learning signal for the foreigners. That is we expect the response rate for English to be the worse, and gradually increase as the language learning signal is added and the language proficiency is improved to A2 and Proficient levels. Although we expect that language acquisition and the language learning signal to increase the response rate, we still expect foreigners who communicate in the proficient local language to receive a lower response rate than a native. Overall we hypothesize that $\beta_1 < 0$ (foreign vs. native, proficient local language) and that $\beta_1 > \beta_2 > \beta_3 > \beta_4 > \beta_5$.

1.1 Replicating response rate differentials across foreign vs. native names

We will begin our interpretation of the results by studying response rate differentials across foreign-sounding and native-sounding names as has been done in prior research. That is we will compare the control group with text version 1 (foreign name, proficient local language). We expect to find significant response rate differentials consistent with previous research studies. We will focus on the estimate of $\hat{\beta}_1$ and test $H_0 : \beta_1 = 0$ using a t-test.

1.2 Language proficiency and social integration

The experimental design enables us to isolate the effects of acquiring a higher degree of local language proficiency. We will focus on text versions 0 (native name, proficient local language), 1 (foreign name, proficient local language), 3 (foreign name, broken local language), and 5 (foreign name, proficient English). We can use the previously estimated models and compare coefficient estimates of $\hat{\beta}_1, \hat{\beta}_3$, and $\hat{\beta}_5$. We will test $H_0 : \beta_1 - \beta_5 = 0, H_0 : \beta_1 - \beta_3 = 0$, and $H_0 : \beta_3 - \beta_5 = 0$. separately using a Wald test.[4] Additionally we can test whether there any differences in outcomes across the language proficiency treatments by $H_0 : \beta_1 = \beta_3 = \beta_5 = 0$.

Note on aggregating groups: If we find in the previous section that signaling future language improvement through taking a language class has no effects, then we can just combine them within their language proficiency group, resulting in 4 groups. If we find very small effects between text verson 0 and version 1 (native vs. foreign, proficient language), we may also aggregate the data into the following three language proficiency conditions, combining the language learning signal and foreign-sounding status: (1) Native language, (2) A2 language, and (3) English Language. Then we will carry out the similar sets of tests which are listed above with these aggregated groups instead.

1.3 Taking language classes and social integration

Next, we will investigate whether signaling interest in learning the language, thereby influencing expectations about future language proficiency, can increase response rates. We will focus on text versions 0 (native name, proficient local language), 2 (foreign name, broken local language, local language course), 3 (foreign name, broken local language), 4 (foreign name, proficient English, local language course), and 5 (foreign name, proficient English). We will test the effectiveness of the language learning signal seperately for English and A2, and then check for differential effectiveness of the language learning signal. We can compare the estimates for $\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ and $\hat{\beta}_5$. We will use the Wald test to for test for (a) $H_0 : \beta_2 - \beta_3 = 0$, (b) $H_0 : \beta_4 - \beta_5 = 0$, and (c) $H_0 : \beta_2 + \beta_5 - \beta_2 - \beta_4 = 0$.

[4] We intend to utilize **lincomb** or **test** commands in Stata for implementation after running our main regression model.

1.4 Identifying effects of language proficiency on social integration

In the absence of a randomized experiment, we may expect the language proficiency of foreigners to influence the social opportunities they pursue. For example, foreigners who do not know the local language may only search for social opportunities within the community expat groups, rather than interacting with the native residents. Our experimental design enables us to identify the response of residents when interacting with foreigners of varying language proficiency.

The response rate differential between text version 0 and version $j \in \{1, 2, 3, 4, 5\}$ is identified if $E(TextVersion^{j}_{irc} \cdot \epsilon_{irc} | \alpha_{rc}, X_{irc}) = 0$. Although we expect this to hold by the construction of our experimental design, our results must be interpreted with some caution. It is reasonable to think that the perceptions about the unobserved characteristics of the message sender may be influenced by the contents of the message in unintended ways. For example, being less available due to taking a language class could make some landlords less likely to respond if that is the main time they are available to show their property. Writing proficiently in the local language with a foreign-sounding name may signal to be a citizen of the country, hence better understanding the local culture and being committed to staying in the region for a longer period. We expect the impact of such concerns to be small enough that they do not bias the ordinal rankings of the treatment effects.

Note of robustness check: we could do some additional online Prolofic surveys to address the most common concerns. Here we could just do a survey experiment, and ask questions about the perceptions about the sender after showing the message. We can do this after the study to investigate potential mechanisms if we observe large difference in response rates across treatment groups.

1.5 Heterogeneity analysis

If we have sufficient power to do so, we will estimate the main regression specification separately by each country.

For the football clubs, we may not have sufficient observations for heterogeneity analysis in Hungary and Slovokia, but can compare Austria and Czec Republic. When doing the country-level analysis, we may have to combine versions 2 (foreign name, broken local language, local language course) and 3 (foreign name, broken local language) as 2', and 4 (foreign name, proficient English, local language course) and 5 (foreign name, proficient English) as 3'. This aggregation will result in 4 conditions rather than 6. For the football sample, we will check for treatment effects by the league division to assess whether the returns to acquiring a language depend on how competitive the team is. For the housing market, we should have a large sample of observations per country, providing sufficient power to study treatment effects at the country level.

Additionally, we intend to use AI to estimate demographic characteristics about the football coach and listing realtor/landlord and use this for heterogenity analysis. For example we plan to determine the gender and foreign background using the full name. Then we can investigate heterogeneous treatment effects by gender and country of origin. For example, it may be the case that the response rates in the foreign sounding treatments are higher when the coach or realtor/landlord is themselves foreign. We will also interact the demographic characteristic of the coach or realtor/lanlord with the sender name in our study. For example, a female realtor/landlord is more likely to respond to a female who is requesting to view a property.

We also suspect the results may vary across big cities and smaller towns. This because the share of foreigners may very across regions. We will do heterogenity analysis by comparing the treatment effects in bigger regions to smaller regions.

2 Analysis of Power for Football Clubs

2.1 Sample size

Assuming that we have around 10 regions per country, then we will have 40 blocks in total. We will include both male and female football clubs in our analysis, as long as female football clubs are present in the country. Across the 4 countries, we expect to have at least 2500 football clubs. To do a power analysis, we will use 2500 realtors or landlords, however, we may collect more observations during the data collection. Observations will be split equally into 6 different conditions so that we will have around 415 observations per condition. We may plausibly be able to collect data on 4000 observations for both the football clubs.

2.2 Minimum detectable effect size (MDES) for main analysis

We will compute the MDES using the PowerUp software provided by Dong, N. and Maynard, R. A. (2013). We will compute the MDES to compare the control group (native, proficient local language) to one of the 5 treatment conditions as these are our primary comparisons of interest in our study. We have 40 blocks (regions) and around 21 observations per block. Assuming constant treatment effects, a significance level of 5%, two-tailed test, partial R^2 for the blocks is 0.1, and power of 80%, we achieve an MDES of 0.18σ. Further including covariates such that the R^2 increases to 0.2 decreases the MDES to 0.17σ. Assuming an average response rate of 40%, the MDES for the response rate differential is around 8 percentage points. Increasing the sample size to a plausible 4000 observations decreases the MDES to 0.14σ or 7 percentage points.

3 Analysis of Power for Housing Market

3.1 Sample size

As rental properties are concentrated in the most prominent cities, assume 5 blocks per country, so 20 blocks in total. Across the 4 countries we expect to have at least 8000 listings.

3.2 Minimum detectable effect size (MDES) for main analysis

Assuming we have around 10000 listings in total, 1666 per group, and 20 blocks. Assume constant treatment effects, 20 blocks, and 166 observations per block. While achieving a power of 80% our MDES is 0.09σ or around 4 percentage points.

3.3 Minimum detectable effect size (MDES) for country-level analysis

At most we will have around 4000 listings per country, and 666 per treatment group. Assume constant treatment effects, 5 blocks, and 266 observations per block. While achieving a power of 80% our MDES is 0.15σ or around 7 percentage points.

4 Survey Sample Size for Native vs. Foreign Sounding Names

Before the implementation of our experiment, we conducted an online survey on Amazon Mechanical Turk to determine which names are the most foreign-sounding to residents of each country in our study. If a respondent guessed whether a name is foreign-sounding or native-sounding, then we could expect a 50% correct classification rate. We would like to include foreign-sounding names that

can be correctly identified by at least 80% of the survey respondents. Using the **power** command in Stata, assuming 5% significant level, 80% power, a sample size of 78 respondents is required.

5 Multiple Hypothesis Testing

In this study, we have 2 outcomes and 5 treatment conditions, resulting in 10 tests of treatment effects. In addition to these 10 tests, we will also perform 6 additional tests (discussed above) for each outcome. [5] In total, our primary analysis will involve carrying out 22 hypothesis tests. As a result, the family-wise error rate for a significance level of 0.05 is at most 68%. Using a smaller significance level of 1% reduces the family-wise error rate to 20%. Only considering the 10 tests corresponding to the main regression estimates further reduces the family-wise error rate to 10%.

In an appendix, we will report p-values adjusted for multiple hypothesis testing. We will use corrections such as Romano-Wolf and Westfall-Young.[6]

6 Analysis Plan for Forecast Data

We are surveying residents in all four countries (both online and in-person) and asking them to forecast the positive initial response rate for those with a foreign-sounding name, with the email written in (1) fluent local language, (2) A2 local language, and (3) English. For simplicity, we omit the language-learning signals. As a benchmark, we provide the positive initial response rate for the control group (native name and fluent local language). We survey both locals and foreigners. The surveys will include three attention-check questions that will ask for details about the provided context. Across all countries, we hope to get a total sample size of 600 participants who pass all the attention checks, with at least 200 foreigners.

Justification for dropping attention-check failures. To check whether those who fail the attention checks provide lower-quality forecast data, we will check the variance of the forecasts across both groups.[7] We hypothesize that those who failed at least one attention check will have higher variability in forecasts than those who passed all attention checks. Additionally, we also hypothesize that those who failed at least one attention check will be more likely to forecast fluent $\times$ foreign name to be higher than fluent $\times$ native name (possible in reality but unlikely). If our hypothesis holds (at least to a reasonable extent), we will drop those who failed at least one attention check for the main analysis presented below.

The perceived returns to local language learning. Our main outcome of interest is the perceived returns to local language learning. For expositional simplicity, we will primarily focus on the returns to becoming fluent (relative to English). All other combinations of language-learning returns (e.g., from English to A2) will be considered secondary. The perceived returns to fluency will be defined as:

$$\text{Perceived returns to fluency} = \text{Forecast for Fluent - Forecast for English.}$$

We choose this measure because it reflects the perceived benefits of learning a language for an English-speaking foreigner. If a foreigner overestimates the response rate to using only English

[5] We may carry out more tests and will adjust the multiple hypothesis testing accordingly

[6] We will use Stata commands **rwolf2** and **wyoung** for implementation that allow for multiple treatments and strata fixed effects.

[7] We can use **sdtest** in Stata and perform a ratio-of-variances test.

and underestimates the response rate to becoming fluent, that implies a low perceived return to learning the language.

Perceived Returns vs. Actual Returns. We will compare the perceived returns to becoming fluent to the actual returns. The actual returns to fluency estimated from our RCT:

$$\text{Actual returns to fluency} = \text{RCT Estimate for Fluent - RCT Estimate for English.}$$

We will then compare how the forecasts of perceived returns to fluency differ from our estimates. Our hypothesis is that, on average across all countries, the perceived returns will be less than the actual returns to becoming fluent. We can use a bar chart to illustrate the results (i.e., plot the mean perceived returns to become fluent vs. the actual returns from the RCT). For the bar plot, the confidence interval for the actual returns can be constructed using the standard Wald standard error formula for an unpooled difference in proportions. If it is not visually obvious whether the perceived returns are different from the actual returns, then we may consider a bootstrap procedure for hypothesis testing.

Similarly, when looking within a country, we can also plot the kernel distribution of the perceived returns to becoming fluent with a vertical line representing the actual returns. The kernel distribution will give us a sense of the percentage of people who underestimate (or overestimate) the perceived returns to becoming fluent. Alternatively, when pooling the countries together, we can also calculate the ratio of perceived returns to becoming fluent to the actual returns, and then plot the kernel distribution of this ratio (e.g., < 1 implies underestimation).

Heterogeneity in perceived returns. We will check for heterogeneity in the perceived returns to fluency by (1) foreigners vs. natives (based on survey questions), and (2) by each country. Although the results will be noisier, we expect to have at least 100 observations in each group. We hypothesize that foreigners will have lower perceived returns to learning the language than natives (since natives may be biased toward not expecting any ethnic discrimination). However, both foreigners and natives, we believe, will underestimate the true returns to becoming fluent. We also hypothesize that the magnitude of the perceived returns to learning the language will be proportional to the actual returns across countries (the rankings will align). The country with the largest actual returns will also be the country with the largest perceived returns. We can also plot the percentage of those who underestimate the fluency return by country.

Exploring mechanisms behind perceived returns. To better understand the mechanism underlying the perceived returns to learning the language, we will compare the forecasted response rate with the actual response rate across the three language groups. We can make a bar chart or a coefficient plot to illustrate this. We hypothesize that forecasters will overestimate the response rate from English and underestimate the response rate for fluency (relative to the truth as estimated in our RCT). As long as there is sufficient data available, we will break down this result by foreigners vs. natives, and by each country.

Dealing with variation in sample size. The sample sizes of the survey will naturally vary across countries. Additionally, the sample sizes in the RCT will vary across countries. We need to account for these differences in sample sizes when examining the forecast survey data on its own and when comparing the forecasts to the estimates in the RCT. We will use weights to make the comparisons comparable, with a preference for assigning equal weights across countries. We may also consider inverse-variance weights for robustness.