WORKING PAPER

Re-Evaluating the Returns to Language Skills Using Latent Trait Estimates

James V. Marrone

RAND Labor & Population

WR-1212
October 2017
This paper series made possible by the NIA funded RAND Center for the Study of Aging (P30AG012815) and the RAND Labor and Population Unit.

RAND
CORPORATION

For more information on this publication, visit www.rand.org/pubs/working_papers/WR1212.html

Published by the RAND Corporation, Santa Monica, Calif.

Re-Evaluating the Returns to Language Skills Using Latent Trait Estimates

James Marrone*

RAND Corporation

October 9, 2017

Abstract

Past studies have established a sizable wage premium for immigrants who learn the language of their destination country. However, these measurements are usually based on self-reported ordinal measures of spoken language fluency, which are problematic and inconsistent in several ways. Using detailed survey data from United States and France, this paper constructs more robust latent trait measures of language fluency using well-established psychometric methods, and re-evaluates the evidence for the wage premium. The results show that measures of spoken fluency alone conflate those with and without non-verbal skills (reading, writing, and comprehending) and therefore overestimate the wage premium for speaking. I find that the additional wage premium attributable to full verbal/nonverbal fluency is as large as that for verbal fluency alone. In addition, I provide evidence that the skills separating verbal-only from full fluency are generally related to education and training before immigration rather than to skills acquired after. Finally, I show that systematic differences in survey response between demographic groups are not a major source of measurement bias.

*1200 S. Hayes St., Arlington, VA 22202, James_Marrone@rand.org. Thank you to Christopher Doss, Steven Durlauf, Casey Mulligan, and Kevin Murphy for helpful comments.

1 Introduction

Language skills are an important component of immigrants' human capital, correlating with immigrants' wages (McManus *et al.* , 1983; Chiswick, 1991; Borjas, 2014), propensity for return migration (Dustmann & Görlach, 2016), neighborhood sorting (Lazear, 1999), employment status (Dustmann & Fabbri, 2003), occupational choice (Schoellman, 2010), and skill complementarity (Berman *et al.* , 2003). Measuring all of these outcomes is, in turn, vital for understanding the full economic impact of immigration on both immigrants and their host countries. Nevertheless, it is well-established that the raw data used to measure language skills – typically self-reported fluency on an ordinal Likert scale – can be unreliable for several reasons, including misclassification bias, recall bias in retrospective questions, or reference bias due to systematic differences in reference points between groups.

This paper uses standard statistical methods to re-evaluate the results of some seminal papers on the wage premium for linguistic human capital, using latent indices of language skill that account for several of these problems. In particular, the latent indices account for potential multidimensionality in language skills and systematic differences in response patterns across groups and across survey questions. Past research focused on correcting for misclassification bias in fluency data, for instance when respondents over- or under-estimate their fluency. This is especially problematic with self-reported survey questions (Hausman *et al.* , 1998), and Dustmann & van Soest (2001, 2004) use latent factor methods to correct the bias in both panel and cross-sectional data. But correcting for misclassification error may be insufficient for two reasons, both illustrated by Table 1 using data from the New Immigrant Survey (one of the two surveys used in this paper). First, Panel 1a shows that using different measures of fluency may lead to different estimates of underlying latent skill, since respondents' own evaluation of their English skills may contradict the interviewers' evaluations. Most studies use the former measure, while Dustmann & van Soest (2004) use the latter. Which is more appropriate? Ideally, we would use information from both questions while controlling for misclassification bias both *within* and *across* the two questions. Second, Panel 1b shows that even without misclassification bias, a single measure of fluency can be mis-informative. The panel shows that respondents' own evaluation of their skills in speaking and understanding do not perfectly align. This does not imply misclassification: people can indeed have different skills in one versus the other. But how should researchers incorporate the information from both variables?

The current approaches to multidimensionality in studies of immigrant language skills are relatively ad-hoc. Some studies take means over different measures of fluency, some use multiple separate measures (Dustmann, 1994), and some sort respondents into discrete groups based on combinations of survey answers (McManus, 1985). It is not obvious that any given approach is superior to any other. Regardless, discrete variables impede inference anyway. When the distribution of responses is of interest (as opposed to the mean or a binary summary of the data), changes in the distribution can be mis-identified or misinterpreted, as illustrated by Figure 1. Suppose a survey asks "How well do you speak English?" with two possible answers: "Well" or "Not well." The left panel shows the initial continuous distribution of fluency, with the threshold above which people answer "Well." The conclusion from the data would be that 80% of people speak well. The right panels show three possible changes at the time of a second survey, including a change in the perceived threshold as to what constitutes "speaking well." Each of these new distributions leads to the conclusion that 95% of people speak well. However, the changes in fluency for infra-marginal respondents, as well as the reasons behind the changing mean, are vastly different and cannot be distinguished, even though such phenomena may be of even interest in their own right.

The methods used in this paper, standard in the psychometric literature, yield latent measures of language skill that are continuous rather than discrete, incorporate information about multiple aspects of language skill, and control for certain differences across groups. The workhorse method is the graded response model (GRM), specifically designed for extracting information from responses to test questions with ordinal answer scales (Samejima, 1969; Muraki & Carlson, 1995). The latent measures are also robust to misclassification caused by systematically different response patterns between groups, conditional on the latent trait. This form of misclassification, called differential item functioning (DIF), occurs when two respondents in different demographic groups but with the same latent language skill give different answers to the same question (see Zumbo, 2007, for a historical overview of DIF detection methods). DIF may be caused, for instance, because respondents with the same grammatical skill and vocabulary have different reference points for what it means to speak English "well." Someone with many native-speaking friends may rate their English as "worse" than with the same vocabulary and grammar skill who mainly has foreign friends. This will result in systematic differences in responses to the question "How well do you speak English?" based on one's social network.

The latent traits are calculated using detailed survey data from the New Immigrant Survey (NIS) and the Trajectories and Origins (TeO) survey. The NIS is a representative survey of new legal permanent residents in the United States, while TeO is represents all immigrants in France. The virtue of using such detailed surveys is that they incorporate multiple questions about language fluency, which is not true of the larger, more commonly-used surveys such as the Census. Using two surveys offers a valuable comparison for the formation of and returns to linguistic human capital in different developed countries and for different languages. After calculating latent-trait measures of language skills, I replicate analyses from seminal papers measuring the return to language skills on the labor market. These papers study returns to language skills in both correlational and causal frameworks, although they use datasets that contain only one (or at most two) questions on language skill, making them inadequate for leveraging the methods used here. Thus, I replicate the previous papers' regression analyses as closely as possible in my survey data, verify that they yield the same inference as those original papers, and then compare to the results using the continuous latent traits.

The latent-trait results from both the NIS and TeO datasets revise the standard inferences from previous literature in similar and important ways. First, language fluency is not unimodal, but rather has a multi-humped distribution with long tails. This is because the relevant dimensions of language fluency include not just verbal skills but also nonverbal, such as reading and writing. In both surveys, respondents fall into roughly one of three categories: (1) those with poor skills in all areas, (2) those with high verbal but low nonverbal skills, and (3) those with good skills in all areas. But the most commonly-used measure of fluency in previous research is generally a binary summary of verbal skill only, which would be insufficient to capture the differences between the groups identified by the latent-trait estimates. In turn, using such a binary measure leads to mis-estimation of the returns to language skills. Most importantly, a binary measure of speaking skill will, by construction, conflate those in aforementioned groups (2) and (3), and thus misattribute to speaking what might be returns to reading or writing. In fact, I estimate that the additional income premium associated with having full fluency in all skills is at least as large as the premium just for speaking. This raises questions for future research, such as the degree to which lack of reading or writing skill may be a barrier to employment, and the relative importance of policy interventions to improve verbal versus nonverbal human capital. As suggestive evidence for the latter questions, I show that respondents in groups (2) versus (3) are distinguished by their exposure to formal language training prior to migration, and their access to social networks that allow them procure jobs prior to moving, implying that such immigrants may be selected before arrival and would have developed the relevant human capital abroad.

The paper proceeds as follows: Section 2 describes the statistical methods used for calculating latent traits, Section 3 describes the data, Section 4 outlines the results of the latent variable estimation, and Section 5 describes the results of the replication exercises and provides evidence for the underlying sources of differences in linguistic skill. The appendix provides evidence for the robustness of the latent trait measurements to systematic differences in response patterns across groups.

2 Method

2.1 Constructing Latent Traits: Graded Response Model

As with any latent trait analysis of language skills, the basic assumption for this exercise is that a person's linguistic human capital stock L is unobserved, but manifests certain observable skills and behaviors. The measures of those skills and behaviors must be used to extrapolate the trait.

Survey data on language skills and language use are ideal inputs into the graded response model (GRM), developed originally for binary data by Samejima (1969) and expanded to polytomous ordinal data by Muraki & Carlson (1995). Because GRM comes from the psychometrics literature and may be unfamiliar to labor economists, this section briefly reviews it. Further details can be found in Reckase (2009, §2.1.3.3 and §4.1.2.3).

The components of the graded response model are survey questions $q \in \{1, \ldots, Q\}$, each of which has possible responses $r \in \{1, \ldots, R_q\}$. Responses are hierarchical. In other words, response r also implies or subsumes response $r-1$. Likert scales fit this framework perfectly, since a choice of one rating category necessarily subsumes or outranks all lower categories.

The data for each individual i consists of responses $\{x_{qi}\}_{q=1}^{Q}$. The probability of rating at least r on question q is $\mathrm{P}_q\left[r \middle| L\right]$. The probability of scoring exactly r is $\mathrm{p}_q\left[r \middle| L\right] = \mathrm{P}_q\left[r \middle| L\right] - \mathrm{P}_q\left[r+1 \middle| L\right]$, which can be plotted as a function of L to give category response curves. I adopt the standard framework and use the logistic form of the GRM, so that the probabilities are given by equation 1. The question-specific discrimination parameter α_q determines how well question q distinguishes different latent skills L, and alters the width of the category response curves. The response-specific difficulty parameter β_{qr} determines the latent skill L that is most likely to answer the question correctly, thereby shifting the peak of the category response curve right or left.

$$\mathrm{P}_q\left[x_{qi} \geq r \middle| L_i\right] = \begin{cases} 1 & r = 0 \\ \frac{1}{1+\exp(-\alpha_q(L_i - \beta_{qr}))} & 0 < r < R_q \\ 0 & r = R_q \end{cases} \tag{1}$$

Assuming the distribution of latent traits is $f(\cdot)$, the maximum likelihood contribution of respondent i is equal to the the probability of i's vector of responses (equation 2). Parameters can be estimated via maximum likelihood by aggregating the individual likelihood contributions for all respondents.

$$\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \mathrm{p}\left[\boldsymbol{x}_i \middle| \boldsymbol{\alpha}, \boldsymbol{\beta}, L_i\right] f(L_i)\, dL_i \tag{2}$$

2.2 Accounting for Group Differences: Differential Item Functioning

Differential item functioning (DIF) refers to a difference between groups in their response pattern on a question or set of questions, controlling for latent skill. The standard method for DIF detection is a χ^2

likelihood ratio test (Swaminathan & Rogers, 1990). Given an estimate of respondents' latent traits L_i and their groups g_i, $g_i \in \{1, \ldots, G\}$, cumulative logit models determine whether the probability of different responses depends on the latent trait only (model 1 in equation 3), the trait and the group membership (model 2), or an interaction between the two (model 3). Using each model's log-likelihood values, a likelihood ratio test between models 1 and 2 identifies uniform DIF (where the differences between groups are constant at all levels of the trait) and a test between models 2 and 3 identifies non-uniform diff (where the difference depends trait level).

$$\text{Model 1:} = \alpha + \beta L_i + \epsilon_i \tag{3a}$$

$$\text{Model 2:} = \alpha + \beta L_i + \delta g_i + \epsilon_i \tag{3b}$$

$$\text{Model 3:} = \alpha + \beta L_i + \delta g_i + \gamma L_i g_i + \epsilon_i \tag{3c}$$

Since DIF detection conditions on latent traits, it is important to match respondents on the correct traits, so that real differences between groups are not misidentified as DIF. However, matching is not a trivial exercise. When DIF is present then any calculation of the latent trait will be "impure" and cannot necessarily provide a valid matching criterion – yet some sort of matching criteria must be used in order to identify DIF in the first place. To solve this conundrum, several "purification" options have been proposed, each with their own limitations. I use the iterative method of Crane *et al.* (2006), in which latent traits are estimated on the full sample, then questions with DIF are identified as described above. Next, GRM parameters are estimated separately for questions found to have DIF (group-specific estimates) versus those that do not (full sample estimates); using these newly-estimated parameters the DIF questions are re-flagged. The procedure is repeated until the same set of questions is flagged. This approach preserves the information from all questions, rather than discarding those questions with DIF, and also avoids misidentification of DIF in early stages of the loop. The method of Choi *et al.* (2011), implemented in the R package `lordif`, implements this method and yields DIF-free estimates of the latent trait L_i.

Dustmann & van Soest (2001) have performed a related analysis, accounting for misclassification in a panel data setting. Building on Lee & Porter (1984) and Hausman *et al.* (1998), they assume that the latent trait L_i is discrete and that the probability of a respondent misclassifying themselves is a function only of their true latent trait and their reported trait. In other words, $\Pr\left[x_{qi} = r \middle| L_i = \tilde{r}\right]$ is a function only of r and $\tilde{r}$, and not of individual characteristics. The DIF detection used in this paper, on the other hand, allows for the survey instrument to interact with the respondent's characteristics, in the sense that people may differentially misclassify themselves depending on personal characteristics.

The type of misclassification addressed by Dustmann & van Soest (2001) is also different. They identify misclassification by leveraging questionable time variation in response patterns.[1] They use the panel data structure of the German Socioeconomic Panel (GSOEP) to perform fixed-effects regression analyses, and find that misclassification error results in downward bias of regression coefficients. However, the GSOEP only contains information only on speaking ability. Thus, while this study must resort to cross-sectional analyses, it adds to the analysis of Dustmann and Van Soest by illuminating an additional source of measurement error – namely, that caused by using only one measure of language skills, which necessarily misses the differences between verbal-only and full fluency that are illustrated below, and which form the substantive part of this

[1] In particular, they note that some respondents downgrade their own spoken fluency from one year to the next, which is taken as misclassification in one or both years, since it is unlikely that people would "forget" German after having moved to Germany. See their Table 1 for evidence analogous to Table 1 of this paper.

analysis.

3 Data

Robust measures of language fluency require data with more variables than are found in many large surveys. The patterns documented by Table 1 can only be characterized when data is sufficiently rich as to allow for multiple measures of language skill and use. Ideally, survey data will allow for corroboration of responses using multiple forms of evidence (to account for misclassification of the sort suggested by Panel 1a and for measurement of skill along multiple margins (to account for the multi-faceted nature of linguistic skills suggested by Panel 1b. Unfortunately, many commonly-used surveys, such as the U.S. Census and the American Community Survey, fail to satisfy both conditions.

Data must be also be rich in variation in order to run the GRM and DIF algorithms. For DIF detection, data must be split into demographic cells, and questions' response vectors must not be too correlated across respondents within any given cell. In addition, in order to be informative, each question must have sufficient variation across responses within each cell. The need for variation puts limitations on the questions used to identify the latent trait and the groupings used to flag for DIF.

To counter the data problems of many surveys, I use data from the New Immigrant Survey (NIS) from the United States and the Trajectories and Origins (*Trajectoires et Origins*, or TeO) survey from France. Both surveys offer detailed information on each respondent's language skills as well as their entire migration history, employment history, education, marriage, visa status. The NIS is representative of new legal permanent residents (see Jasso *et al.* , 2005), while TeO is nationally-representative of all immigrants in France (see Beauchemin & Simon, n.d., for the project description). Complete details for constructing variables may be found in the author's data dictionary, available on request.

The New Immigrant Survey consists of 8,573 adult respondents who received legal permanent resident (LPR) status in the year preceding the survey, of whom the 8467 born abroad and currently residing in the USA are used in this paper. Data collection for the NIS took place in 2003 and 2004, sampling from immigrants who were 18 years and older at admission to LPR, with four sampling strata based on visa category: spouses of U.S. citizens, employment principals, diversity principals (i.e. lottery winners), and all other immigrants. By construction, the NIS is not representative of all immigrants, but the TeO survey offers affirmation of NIS results based on a more representative sample. Language skills are captured in 4 questions that were asked of every respondent:

1. "How well do you speak English?" (1-4 scale)

2. "How well do you understand English?" (1-4 scale)

3. Interviewer's assessment of respondent's spoken English (1-5 scale)

4. Interviewer's assessment of how well respondent understood the questions (1-3 scale)

Following previous studies that used binary measures of speaking skill, "speaking well" is defined as a score of 4 out of 4 for Question 1. Question 4 is an imperfect measure of English skill, because it could conceivably apply to a respondent's understanding of questions that were translated into another language. However, more questions allows for more reliable identification of DIF. To check the robustness of the results to the inclusion of this question, the latent trait was calculated using only the first three questions (without purifying for DIF), after substituting the question "Did the respondent require any translation?" (yes-or-no question) as a proxy for Question 4, and after adding the question on translation to make 5 questions total. The latent traits from all scenarios were highly correlated (more than 99%) with the original estimates.

The Trajectories and Origins Survey encompasses 21,761 respondents. The survey took place in 2008 and 2009 with a stratified sampling strategy to capture representative samples from five different segments of the French population, of which I use the first, consisting of immigrants (N=8734). Besides representing the full population of immigrants, TeO improves on NIS by including more questions regarding language skill. Those questions are as follows:

1. "How well do you speak French?" (1-4 scale)
2. "How well do you understand French?" (1-4 scale)
3. "How well do you read French?" (1-4 scale)
4. "How well do you write French?" (1-4 scale)
5. Interviewer's assessment of respondent's French (1-3 scale)
6. To what extent was the interview translated? (1-3 scale)
7. "Do you have difficulty giving your name and phone number in French?" (binary)
8. "Do you have difficulty answering simple questions in French?" (binary)
9. "Do you have difficulty asking for information or services in French?" (binary)
10. "Do you have difficulty taking part in a conversation in French?" (binary)
11. "Do you have difficulty describing something in detail in French?" (binary)

As with the NIS, I define the binary measure of "speaking well" in the TeO survey as a response of 4 out of 4 to Question 1. The paths through the TeO survey are somewhat complicated, and result in many questions being asked only of certain sub-populations. For example, those who report speaking "very well" were not subsequently asked if they had difficulty asking for information or services. Answers are filled in accordingly when the questions were not asked (details available in data dictionary), but calculations were performed two ways: only the first 6 questions, and using all 11. Since the resulting latent trait was virtually indistinguishable, the more parsimonious approach with 6 questions was used throughout the remainder of this paper.

Importantly, I do not drop survey respondents from either survey based on their mother tongue or year of immigration. Immigrants who arrived very early or who learned English/French as a native language will have response profiles indicating the highest-possible level of skill for every question. Thus, these respondents essentially reflect native-born native-speakers, at least linguistically. This group provides a baseline against which to calculate latent traits for the rest of the sample.

Income is calculated from combinations of survey questions, depending on employment status. In the NIS, income is calculated as the sum of the individual's earnings from wages and tips, or self-employment earnings, whichever applies. In the TeO survey, income is calculated as twelve times the reported monthly earnings from all jobs, or from self-employment, whichever applies. In both surveys, when respondents opt to report their income as a bracket, income is imputed as the survey-weighted average income among all other respondents within income bracket/occupational skill/(self-)employment cells.

Table 2 provides summary statistics for both surveys. The table additionally compares groups based on the most-commonly-used measure of language skill: those who speak well versus those who do not. Unsurprisingly, speaking well correlates with having used some English or French as a child; having better education; finishing education in the destination country; marrying someone from outside one's birthplace; being male, non-refugee, or employed; holding a higher-skill job and earning more; migrating at a younger age; and having a longer tenure in the destination country. There are salient differences between surveys as well, particularly in the demographic makeup of immigrants. France has far more immigrants from the Middle East and North Africa; they are more likely than other non-Europeans to speak well, reflecting the linguistic heritage of certain countries as well as likely selection in immigration flows. The USA has more Latin American immigrants, who are relatively less likely to speak English well. In the US, employment visas correlate with speaking well, while in France family reunification visas predict poor speaking skill. There is some indication that those with poor skills are more likely to have found jobs through social networks, while those with strong skills are more likely to have procured a job before migrating. Overall, these statistics affirm the general findings that language assimilation is associated with labor force participation, income, and other indicators of social assimilation.

Figure 2 plots the implied density of skills based on responses to certain survey questions above. The distribution varies for different skills, with varying degrees of skewness. In the TeO survey, the reported skill in writing is nearly bimodal. Collapsing these variables into binary measures may imperfectly characterize the heavy left tail as being more homogeneous than it really is; and using any one of the four measures omits information about the complementarity and substitutability between different skills. Latent trait estimates can help determine the degree to which different skills matter.

4 Latent Trait Estimation Results

The baseline latent trait estimate is calculated from equations 1 and 2 using the questions listed above. To test the feasibility of the GRM framework, it is important to verify the unidimensionality of the latent trait. Figure 3 plots the principal components calculated from the response vectors to the relevant questions from each survey. Variables were scaled to have unit variance. The figure shows that the majority of variation is explained by the first component, supporting the use of a unidimensional framework.

Figures 4 and 5 show the category response curves for each question, after GRM parameter estimation was performed. As is expected for the graded response model, the probability of giving a response associated with higher fluency is increasing with the latent trait. It is also clear that the usage of different question scales differs. For example, interviewers in the NIS generally treat speaking as a binary skill, using the two most extreme ratings more often than any of the three in between (Figure 4c). This is not a technical problem for estimation, but it does reduce the heterogeneity in the resulting latent trait estimates.

It is useful to compare the latent trait measures against the commonly-used binary measure of "speaking well." Figure 6 plots the density of the estimated latent trait for individuals who report speaking well versus those who do not. It is immediately apparent that for both NIS and TeO the binary and continuous measures of fluency imply very different distributions of skills. Among those who report speaking well, the distribution of the latent trait is actually bimodal. This is further evidence of what was already shown in Table 1, in terms of speaking ability capturing only a partial picture of complete language skills.

Based on the distribution of latent traits, much of the remaining analysis focuses on three sub-groups, labeled accordingly on Figure 6: those who report speaking poorly (Group 1); those who report speaking well but whose latent trait scores fall in the lower tail of the corresponding distribution; and those who speak well and whose latent trait scores are in the right tail of the distribution. Effectively, Group 2 consists of respondents who report speaking well but who may have difficulty – to a greater or lesser degree – in speaking, reading, writing, etc. Group 3 consists of respondents who claim to speak well and also report maximum fluency across reading, writing, comprehending, and interviewer assessments. The three groups can roughly be classified as "non-proficient," "verbally proficient," and "fully proficient." For future reference, Table 3 lists the mean and median latent trait scores of each group, along with the threshold values used to define Group 2 versus Group 3. dValues are listed before and after the latent traits were scaled to be between 0 and 1.

Most importantly, Figure 6 shows that these various skill groups embody large differences in income. In both surveys, conditional on being employed, respondents in Group 2 earn about 10% more than those in Group 1. But in TeO, respondents from Group 3 earn 40% more than those in Group 2; in the NIS Group 3 earns nearly double Group 2. The premium for full proficiency, then, appears to be much higher than the premium for verbal proficiency only. These differences between groups will be explored in more depth below.

The baseline estimates shown in Figure 6 do not account for DIF. Due to their sample sizes, the surveys cannot be split into many cells. Binary or ternary groupings would be optimal, in order to preserve cell sizes that are sufficiently large to have enough variation for calculations. While DIF could potentially be related to many different variables: age at immigration, tenure, birthplace, education, gender, employment status, marital status, propensity to remit money, refugee status, and neighborhood immigrant density could all feasibly affect a person's view of what it means to "speak well," because they are evidence of different references or frames for the question. Flagging variables for DIF at the 1% significance level, I find that grouping on any of the aforementioned variables yields purified latent trait measures that are highly correlated with both the raw estimate and with each other (upwards of 99% correlation in each case). When considering survey design, one may still want to account for this DIF; but in this setting, where the overall latent trait estimate is of primary interest, DIF has little bearing on the outcomes being studied.

Because I find that DIF has virtually no effect on the latent trait estimates, the original unpurified estimates plotted in Figure 6 are used in the remainder of the paper. For those interested in the evidence for DIF, the Appendix documents DIF calculations using the gender/birthplace groupings, which are among the most common groupings in the psychometrics literature. Birthplace is treated as a binary variable (European/Western versus all others), in order to maximize cell size and split respondents roughly based on cultural/linguistic distance from English and French. Questions were then flagged for DIF using the four gender-birthplace groups, with 2 of the 4 NIS questions and 5 of the 6 TeO questions being flagged.

5 Determinants and Effects of Language Skills

To assess the determinants of language skills, studies generally regress skills on demographic characteristics. Comparison across studies is difficult, because language skills are measured in slightly different ways, survey samples are different, and different regressors may be included, not to mention probit coefficients are difficult to interpret if not reported as marginal effects. Nevertheless, for results from other studies, see Chiswick (1991, Tables 3 and 5), Dustmann (1994, Table 3), Chiswick & Miller (1995, Table 1), Lazear (1999, Table 3), Dustmann & van Soest (2004, Tables 2 and 3). All studies agree that language skills increase with time since migration and decrease with age at migration. Married people and females tend to have lower skills, employed people higher, and skills are inversely correlated with geographic and linguistic distance to immigrants' birthplaces or native tongues. Tenure (years since migration) and age at migration are generally found to be most important, both empirically and theoretically (Borjas, 2015; Marrone, 2017).

I replicate previous analyses using the NIS and TeO. The regressions include a vector of demographic characteristics X_i, where the outcomes are fluency or log annual income, as given by equation 4. Where possible, the following different measures of fluency are examined: a binary measure of speaking; an ordinal measure of speaking; the average of speaking, reading, writing, and comprehending; and the latent trait $\hat{L}$. For ease of interpretation, all fluency measures are scaled to have values between 0 and 1.

$$fluency_i = \beta X_i + \epsilon_i \tag{4a}$$

$$\log(income_i) = fluency_i + \beta X_i + \epsilon_i \tag{4b}$$

Table 4 shows regression results for equation 4a. The results agree with past studies on the signs of most variables, including the concavity in tenure and the interaction between tenure and age at migration. The R^2 increases as the measure of fluency becomes more granular; but apart from that it is evident that different measures of skill yield largely similar inferences. The OLS coefficients have the same sign and statistical significance across regression specifications. And, in the case of the ordinal, average, and continuous measures, the magnitudes of the coefficients are similar as well. For the TeO survey in particular, t-tests on the equality of coefficients between regressions (4) and (5) do not find any statistically significant differences for any variables except for AAM^2, the MENA dummy, and the Sub-Saharan dummy. Thus, for practical purposes a simple average of reading, writing, speaking, and comprehending sufficiently approximates the continuous variable $\hat{L}$.

Table 5 shows coefficients for a regression of log-income on demographic characteristics and different measures of fluency. The signs are the same as those found in previous studies: wages are positively correlated with migrating early in life, with longer tenures, with being European, and with education. However, the wage premium for language skills depends on how those skills are measured. The measured language fluency premium increases as the measure of fluency becomes more granular. Even after controlling for education and birthplace, the premium falls overall but still increases across regression measures. Similarly, the R^2 rises, albeit it weakly, in the same way. Note that these patterns are not mechanical; increasing the granularity of the fluency measure does not necessarily imply that the measured wage premium will increase. Rather, it depends on the degree to which the different measures of fluency are distinguishing respondents based on income. If responses to all survey questions were completely correlated then using the ordinal measure of verbal fluency would be just as good as any other measure, and there would be little to no difference between regression coefficients.

In fact, binary or ordinal measures of verbal fluency under-estimate the true returns to full language proficiency because they conflate respondents with verbal-only proficiency only and those with full proficiency in all skills. For the same reason, they also mis-estimate the return to verbal proficiency by itself. To see this most simply, consider the following back-of-the-envelope calculations. Based on the second columns of the panels in Table 5, the binary measure of fluency implies a premium of 37.5% in the NIS and 5.9% in the TeO survey. Now consider the sixth columns, which use the continuous measure $\hat{L}$. Using Table 3, scaling $\hat{L}$ yields average values for Group 1 of 0.41 in NIS and 0.49 in TeO, while Group 3 is scaled to $\bar{L} = 1$. The implied premium for full proficiency relative to low proficiency (conditional on education and birthplace) is therefore 77% in NIS and 10% in TeO. Respondents in Group 2 have average $\bar{L} = 0.85$ in NIS and 0.72 in

TeO, implying premia of 58% and 4.5% for verbal proficiency alone. Thus, the average premium for full proficiency is an additional 50-100% of the premium for verbal proficiency alone.

5.1 Determinants of Verbal-Only versus Full Proficiency

This analysis shows substantive differences in income between immigrants with verbal and those with full proficiency in the language of their destination country. Even more than the actual magnitude of the returns to language skills, the causes of such differences in skills is of interest from policy and scholarly perspectives. Understanding the mechanisms and pathways of non-verbal language skill formation could impact the understanding of immigrants' employment opportunities, job mobility, and the potential policy levers that could encourage labor market integration.

The NIS and TeO surveys can be used for a preliminary analysis of the mechanisms for language learning that differentiate Groups 2 and 3. To eliminate the obvious mechanisms of being a native speaker or having been in the US/French school systems from a young age, the analysis focuses on respondents who migrated later in life and who did not speak English/French as a native language. "Old" is defined based on the biological cutoff for learning easily learning a new language, which occurs roughly at adolescence.

I follow Bleakley & Chin (2004), who popularized the use of a migration threshold of 11 years old. Unfortunately, their causal framework cannot be reliably implemented with either the NIS or TeO surveys. In their analysis of immigrants to the USA, the interaction between arriving young (before age 11) and coming from a non-English speaking country functioned as an instrumental variable to address the endogeneity of language learning. In principle, such an analysis is possible in the NIS and TeO as well. However, in practice, the baseline group of young, native-speaking arrivals is too small to allow for reliable identification; the instrumental variable is effectively perfectly correlated with arriving at a young age.

A full causal analysis is therefore infeasible, but a simple correlational analysis can suggest some of the pathways by which immigrants in Group 3 formed the non-verbal skills lacked by Group 2. The NIS and TeO provide several variables indicating the sources of language learning before and after migration, the use of language in daily life, and characteristics of the respondent's social networks. These variables can be used to identify differences between respondents in Groups 2 and 3, as well as examining income for each of the groups separately.

Table 6 documents the demographic correlates of skills and wages among respondents in the NIS. The first two columns show coefficients from logit regressions where the outcome variable is skill group (Group 3 versus Group 2). The last two columns show OLS coefficients for log-income regressions within each skill group. in general, education, marriage the degree of English usage, and having an employment visa are positively associated with skill and wages; endogamy and gender are negatively correlated. Among the full sample, taking an English class before migration, being married, and having arrived on an employment visa are the most significant predictors of high skill. Among the employed, having an exclusively English-speaking friend group is also significant. These variables imply that formal training in English prior to migration are most important in determining skill, as well as in allowing for the occupational mobility and social connections required by many to migrate.

For skill group 2, having a degree from the USA is highly predictive of higher income, as are employment and refugee/asylee visas. The former indicates that US education may compensate for the fact that those in Group 2 were less likely to have taken English classes prior to migration. For skill group 3, visa type and college education are more predictive of high income. For both groups, visa status is likely a proxy for job mobility and also reflects selection into the labor force: those with employment visas were – almost by definition – more geographically mobile and well-connected to the global labor market. As for refugees, the U.S. refugee settlement system encourages self-sufficiency from the very beginning, and it has been established that refugees have both higher employment and faster income convergence than many other immigrant groups (see Rush, 2015, for a more nuanced analysis).

Table 7 relates job and educational characteristics to skill group and income. Respondents in Group 3 are more likely to have higher-skilled jobs, and to have used social networks to procure a job. However, none of the coefficients are statistically significant. Within skill group, the highest-skill jobs of course predict higher wages. Procuring a job prior to immigration predicts higher wages for those in Group 3, but has no effect for Group 2; procuring a job through a friend or relative has a dampening but statistically insignificant effect on wages.

The TeO survey yields largely the same results. Table 8 documents regression coefficients for demographic and language variables. Those with high skill are more likely to be employed, and education is highly predictive of skill as well. Finishing one's education in France also highly predicts skill group, as does having learned French in school before migration. For income, having procured a job through friends or relatives is highly correlated with lower wages. Education and job type are important as well. In general, this table reinforces the story that non-verbal language skills are most likely to be learned in formal educational settings, rather than being picked up from one's social environment. Unlike the USA, visa type is not predictive, possibly because of France's very different visa structure and because immigrants from the European Union do not need visas at all.

6 Conclusion

The main lesson of this study is straightforward: commonly-used measures of language skill may be biased or present an incomplete picture of the relevant aspects of skills. To improve the measurement of language skills and analysis of their economic effects, researchers must incorporate multiple dimensions of language skills. At the very least, they must measure skills using evidence of both verbal and non-verbal aspects of linguistic fluency: speaking as well as reading, writing, or understanding.

There is good news: based on the evidence presented here, systematic differences across groups in the form of differential item functioning are not a major cause for concern. As a result, a simple average of reading, writing, speaking, and comprehending (all measured on the same ordinal scale) may be sufficiently granular as to suffice for many analyses.

Yet the analysis also revealed important differences within groups of immigrants based on verbal skill. By distinguishing different degrees of fluency based on several measures of language competence, the continuous latent trait estimates in this paper showed that verbal fluency and non-verbal fluency are not the same, and that the wage returns to the former are dwarfed by the returns to a comprehensive set of all skills. Although merely suggestive, the evidence provided here indicates that these non-verbal skills are often formed prior to migration, or in formal educational settings, and that they are correlated with labor force integration, geographic mobility, and embeddedness in linguistic social networks.

To the extent that we care about immigrant wage assimilation, then, the formation of nonverbal skills is of policy and scholarly interest. An important goal for future research is to determine the degree to which this suggestive evidence is causal, and the degree to which nonverbal fluency influences selection both on the types of people who migrate and on the types of immigrants who enter the labor force. What policies can encourage nonverbal fluency after migration? How does nonverbal fluency impact occupational mobility?

One starting point may be to improve the availability of linguistic skill data, incorporating multiple measures of skill in the larger-scale surveys that would be necessary to perform the sort of analysis required.

A Evidence for Differential Item Functioning

This Appendix provides evidence for differential item functioning (DIF) in the NIS dataset, based on gender-birthplace groupings. Birthplace groups are a binary categorization of European/non-European, which separates immigrants who are most culturally and linguistically similar to the US or France. When purifying for DIF on the same groupings, the TeO survey yielded qualitatively similar results. In both cases, the correlation between purified and unpurified latent trait estimates exceeded 99%. The same was true when other grouping variables were used. Thus, these results displayed here serve only as an example to document why DIF purification would not affect the main analysis.

Using the method described in Section 2, two of the four questions were flagged for DIF at the 1% significance level: the interviewer assessments of speaking and understanding. As a result, group-specific parameters for the graded response model were calculated for these two questions. Sample-wide parameters were still calculated for the other two questions.

Figure A.1 shows the resulting group-specific category response curves, compared to the original curves from Figure 4. For some groups, DIF purification resulted in large changes in the response curve. Compare the lines in panel (a) corresponding to Europeans with those corresponding to non-Europeans: conditional on the same low latent trait value, interviewers were more likely to categorize Europeans as being in the lowest category, and less likely to classify non-Europeans. But this is offset by the opposite phenomenon at the top end of the distribution: for those with the same high latent trait value, interviewers were more likely to say Europeans spoke "very well." This is called non-uniform DIF, and could be evidence for various forms of bias. For instance, if interviewers expect Europeans to speak better English than immigrants from other regions, then they may be more quick to classify them in the worst category upon finding that they don't speak well. Several other explanations are plausible, for the differences at both ends of the latent trait distribution.

Such differences in interviewers' ratings may be of concern in survey design and implementation, or when individual responses to a specific question are of interest. In this setting, where the latent trait itself is of primary concern, DIF only matters insofar as it affects the final estimate of L_i. Whatever the cause of the differential item functioning, Figure A.2 shows that the overall effect on the latent trait is indeed minimal, partly because of the offsetting differences in response propensities at the upper and lower ends of the distribution. With baseline values of $\hat{L}_i$ ranging between -1.7 and 1.3, the vast majority of individuals had scores change by less than 0.03 after purification. Only a few – primarily European males – had scores change by more than 0.05. Figure A.3 shows that these changes individual values were not sufficient to impact the overall distribution of traits. Correlations between unpurified and purified estimates are greater than 99%. The choice of purified versus unpurified trait does not change the results from the main regression analyses.

References

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

Berman, Eli, Lang, Kevin, & Siniver, Erez. 2003. Language-skill Complementarity: Returns to Immigrant Language Acquisition. *Labour Economics*, **10**(3), 265–290.

Bleakley, Hoyt, & Chin, Aimee. 2004. Language Skills and Earnings: Evidence from Childhood Immigrants. *The Review of Economics and Statistics*, **86**(2), 481–496.

Borjas, George J. 2014. *Immigration Economics*. Harvard University Press.

Borjas, George J. 2015. The Slowdown in the Economic Assimilation of Immigrants: Aging and Cohort Effects Revisited Again. *Journal of Human Capital*, **9**(4), 483–517.

Chiswick, Barry R. 1991. Speaking, Reading, and Earnings Among Low Skilled Immigrants. *Journal of Labor Economics*, **9**(2), 149–170.

Chiswick, Barry R., & Miller, Paul W. 1995. The Endogeneity Between Language and Earnings: International Analyses. *Journal of Labor Economics*, **13**(2), 246–288.

Choi, Seung W., Gibbons, Laura E., & Crane, Paul K. 2011. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *Journal of Statistical Software*, **39**(8), 1–30.

Crane, Paul K., Gibbons, Laura E., Jolley, Lance, & van Belle, Gerald. 2006. Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar. *Medical Care*, **44**(11), S115–S123.

Dustmann, Christian. 1994. Speaking Fluency, Writing Fluency and Earnings of Migrants. *Journal of Population Economics*, **7**(2), 133–156.

Dustmann, Christian, & Fabbri, Francesca. 2003. Language Proficiency and Labour Market Performance of Immigrants in the UK. *The Economic Journal*, **113**(489), 695–717.

Dustmann, Christian, & Görlach, Joseph-Simon. 2016. The Economics of Temporary Migrations. *Journal of Economic Literature*, **54**(1), 98–136.

Dustmann, Christian, & van Soest, Arthur. 2001. Language Fluency and Earnings: Estimation with Misclassified Language Indicators. *The Review of Economics and Statistics*, **83**(4), 663–674.

Dustmann, Christian, & van Soest, Arthur. 2004. An Analysis of Speaking Fluency of Immigrants Using Ordered Response Models with Classification Errors. *Journal of Business and Economic Statistics*, **22**(3), 312–321.

Hausman, Jerry A., Abrevaya, Jason, & Scott-Morton, Fiona M. 1998. Misclassification of the Dependent Variable in a Discrete Response Setting. *Journal of Econometrics*, **87**(2), 239–269.

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.

Lazear, Edward P. 1999. Culture and Language. *Journal of Political Economy*, **107**(6), S95–S126.

Lee, Lung-fei, & Porter, Robert H. 1984. Switching Regression Models with Imperfect Sample Separation Information: With an Application on Cartel Stability. *Econometrica*, **52**(2), 391–418.

Marrone, James V. 2017. *Linguistic and Cultural Change as a Human Capital Process*. Working Paper.

McManus, Walter, Gould, William, & Welch, Finis. 1983. Earnings of Hispanic Men: The Role of English Language Proficiency. *Journal of Labor Economics*, **1**(2), 101–130.

McManus, Walter S. 1985. Labor Market Costs of Language Disparity: An Interpretation of Hispanic Earnings Differences. *The American Economic Review*, **75**(4), 818–827.

Muraki, Eiji, & Carlson, James E. 1995. Full-Information Factor Analysis for Polytomous Item Responses. *Applied Psychological Measurement*, **19**(1), 73–90.

Reckase, Mark D. 2009. *Multidimensional Item Response Theory*. New York: Springer.

Rush, Nayla. 2015 (November). *Fact-Checking a Fact Sheet on Refugee Resettlement*. Tech. rept. Center for Immigration Studies.

Samejima, Fumiko. 1969. *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Monograph Supplement, 34 Psychometric Monograph No. 17. Psychometric Society.

Schoellman, Todd. 2010. The Occupations and Human Capital of U.S. Immigrants. *Journal of Human Capital*, **4**(1), 1–34.

Swaminathan, Hariharan, & Rogers, H. Jane. 1990. Detecting Differential Item Functioning Using Logistici Regression Procedures. *Journal of Educational Measurement*, **27**(4), 361–370.

Zumbo, Bruno D. 2007. Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, **4**(2), 223–233.

Table 1: Potential Misclassification in the New Immigrant Survey

(a) Two Assessments of Respondent's English Fluency

		Self			
		1 (Not Well)	2	3	4 (Very Well)
Interviewer	1 (Not at All)	1318	1640	471	141
	2	59	212	31	7
	3	24	271	236	34
	4	30	182	717	246
	5 (Very Well)	92	145	695	1887

(b) Self-Assessment: Speaking vs. Understanding

		Speak			
		1 (Not Well)	2	3	4 (Very Well)
Understand	1 (Not Well)	1213	65	1	0
	2	301	1782	79	5
	3	10	565	1548	70
	4 (Very Well)	0	36	525	2240

Note.– Each panel shows the frequency for various pairs of responses to different questions on English language skill.

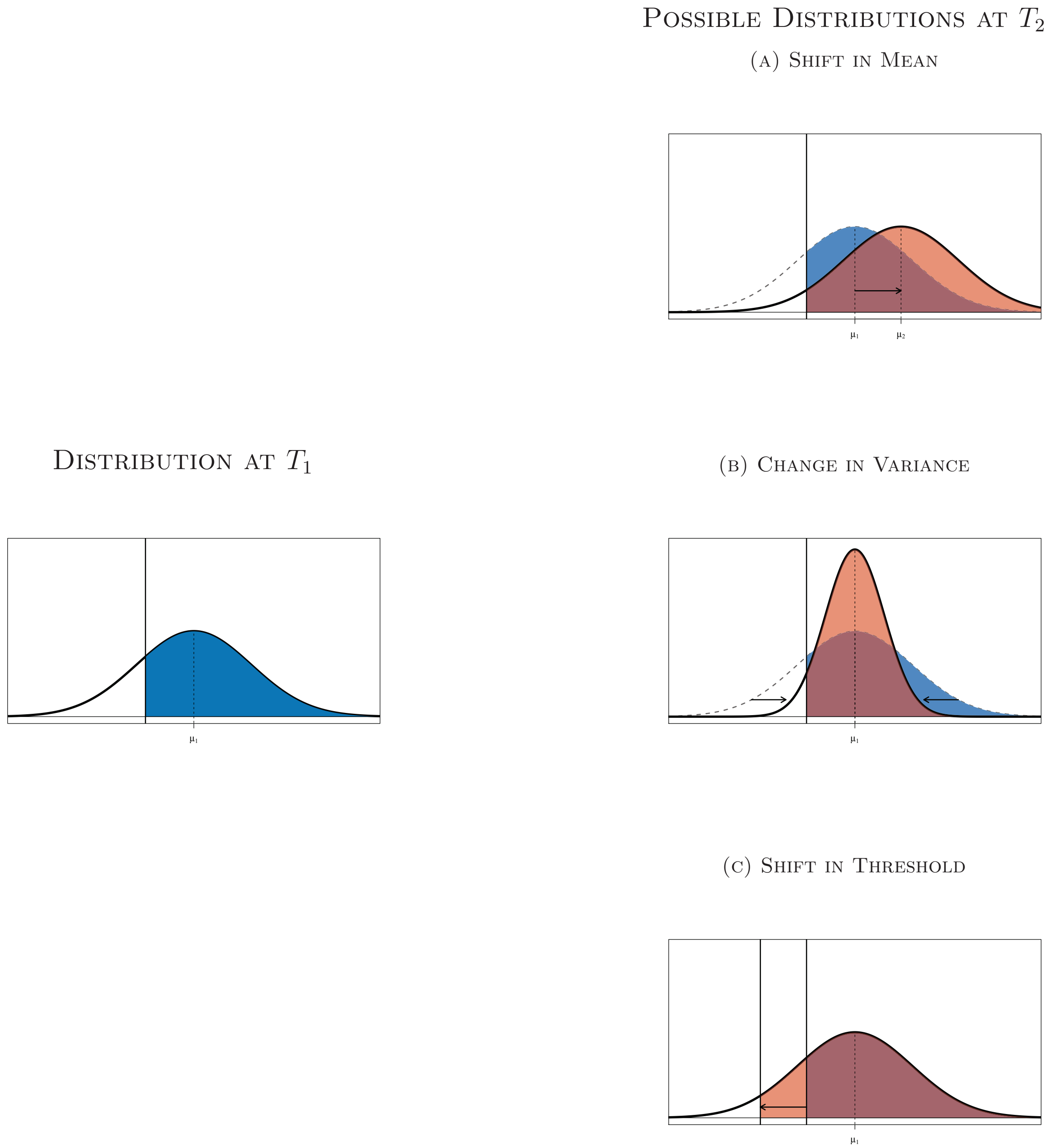


Figure 1: Lack of identification for changes in Likert scale variables.

Left panel: hypothetical distribution of language fluency at initial survey date T_1. 80% of the sample is above the threshold. Right panels: alternative ways in which the distribution or threshold could shift at second survey date T_2, all of which yield 95% of the distribution above the threshold.

Table 2

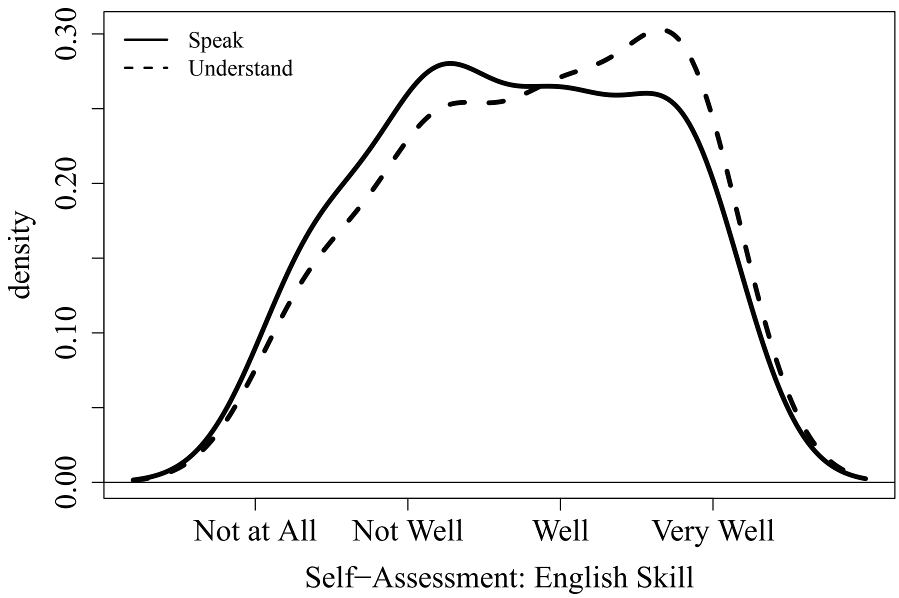
Summary Statistics

	NIS			TeO		
	Full Sample	Speak Well	Do Not Speak Well	Full Sample	Speak Well	Do Not Speak Well
N	8426	2314	6112	8303	5527	2776
Language† (%)						
Speaks well	25.7	1	0	73.6	100	0
Understands well	31.5	96.9	8.4	76.1	98.1	14.6
Reads well	N/A	N/A	N/A	60.6	92.2	7.8
Writes well	N/A	N/A	N/A	64.9	87.4	2.2
Ever used as child	15.0	40.7	5.9	42.7	55.0	8.3
Native-speaking	4.6	17.3	0	28.6	38.9	0
Spoke well on arrival	N/A	N/A	N/A	26.6	36.2	0
Took class before arrival	34.7	43.6	31.5	N/A	N/A	N/A
Took class after arrival	16.8	3.8	21.3	20.1	11.2	45.1
Migration						
Mean Age at Migration	33.5	28.7	35.1	17.3	14.2	26.0
Median Age at Migration	30	27	31	18	12	24
Mean Tenure	6.0	6.8	5.8			
Median Tenure	3	5	3	23	27	13
Demographics (%)						
European	12.5	12.6	12.5	29.5	29.5	29.3
Sub-Saharan	N/A	N/A	N/A	13.4	15.2	10.4
MENA	N/A	N/A	N/A	38.6	42.3	28.4
Africa & Mid. East	11.0	15.8	9.3	N/A	N/A	N/A
Asian	29.0	26.2	30.0	14.1	9.3	27.6
Latin American	42.7	27.9	47.7	N/A	N/A	N/A
Female	56.4	49.8	58.7	54.0	53.7	55.0
Married	74.0	72.4	74.5	61.7	57.4	73.7
Endogamous	58.2	41.1	64.1	34.8	25.6	60.5
Has children	70.1	56.4	74.8	74.2	71.5	81.7
Education/Employment (%)						
H.S.	18.3	19.3	17.9	10.9	11.6	9.2
2-yr degree	13.9	12.6	14.3	9.1	11.3	2.9
4-yr college or more	30.2	49.9	23.4	21.3	25.1	10.7
Finished school after migration	13.3	28	8.3	52.1	66.9	10.8
Employed	55.6	68.1	51.2	66.7	70.4	56.2
Mean Income	$25658	$39061	$19441	€21763	€23302	€16446
Low-skilled worker	32.8	23.9	35.9	6.4	4.3	12.3
Skilled worker	10.9	16.3	9.1	11.8	10.5	15.4
Professional	11.9	27.6	6.4	43.2	50.2	23.7
Job through relative/friend	18.0	9.2	22.0	18.6	16.1	25.7
Job found before arrival	7.5	11.6	5.6	7.4	8.0	6.5
Visa Type (%)						
Employment	9.7	19.6	6.3	14.9	11.7	23.7
Student	N/A	N/A	N/A	9.9	11.9	4.3
Family	58.2	56.5	58.8	27.0	23.3	37.5
Refugee	6.7	4.1	7.5	6.3	4.3	11.9

† Language skills refer to English for the NIS and French for TeO. Doing "well" means rating 4 out of 4 on self-reported Likert scale.

Note.– Means are calculated using survey weights. Variables are described in Section 3. See data dictionary for complete details on constructing variables from raw survey data. N/A = Not Available

(a) NIS



(b) TeO

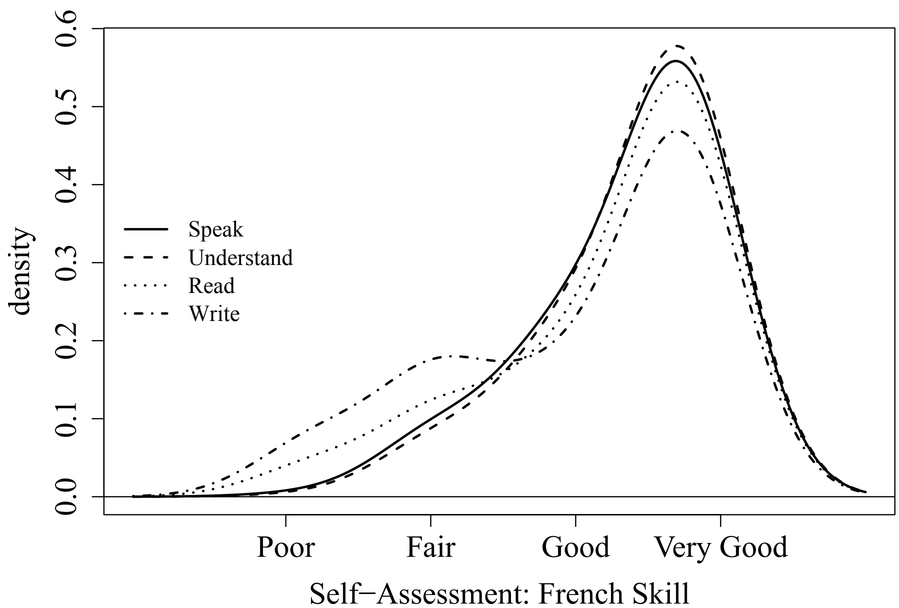


Figure 2: Implied distributions of language skills

Densities are calculated from survey questions each with four possible response categories.

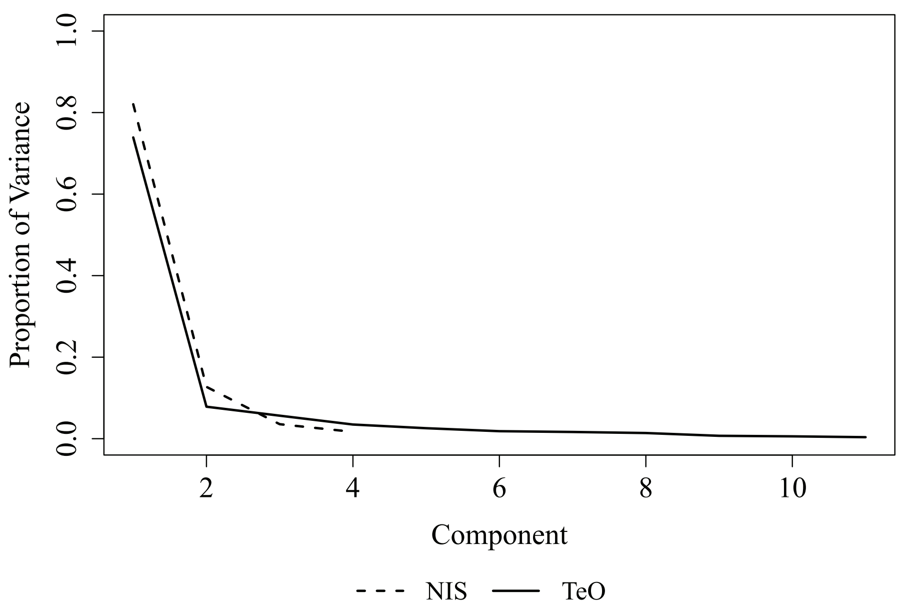
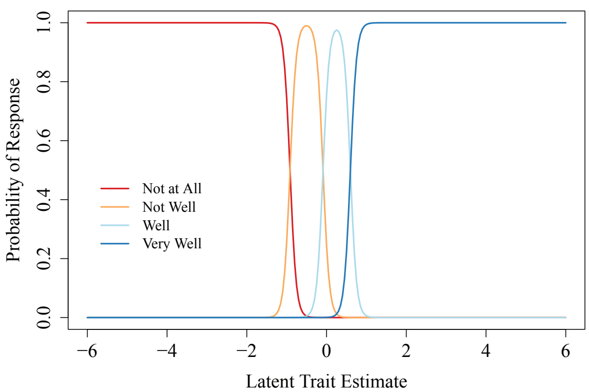
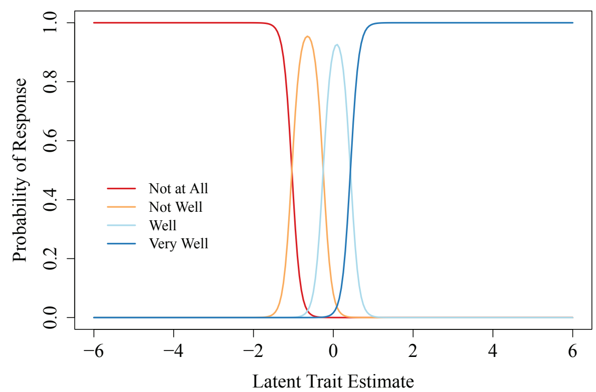


Figure 3: Principal component analysis of survey questions

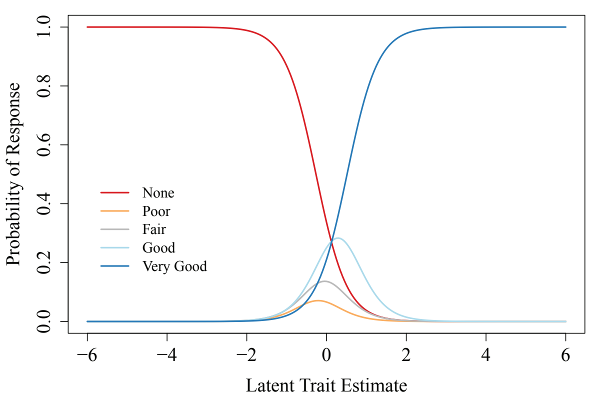
(a) Self-Assessment: Speaking



(b) Self-Assessment: Comprehension



(c) Interviewer Assessment: Speaking



(d) Interviewer Assessment: Comprehension

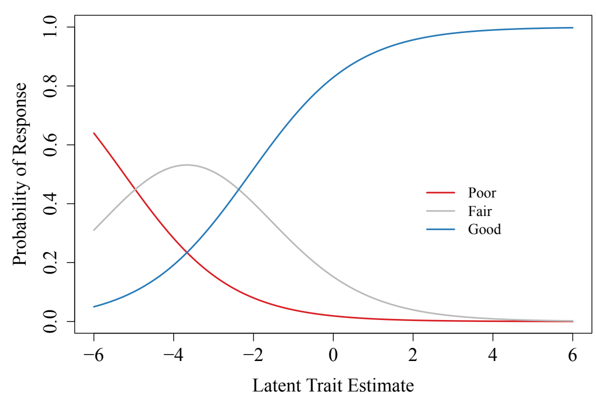
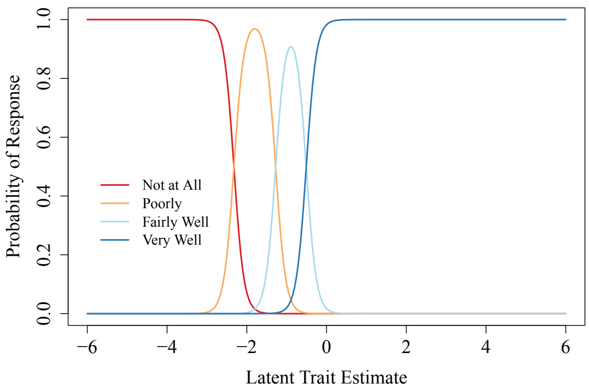


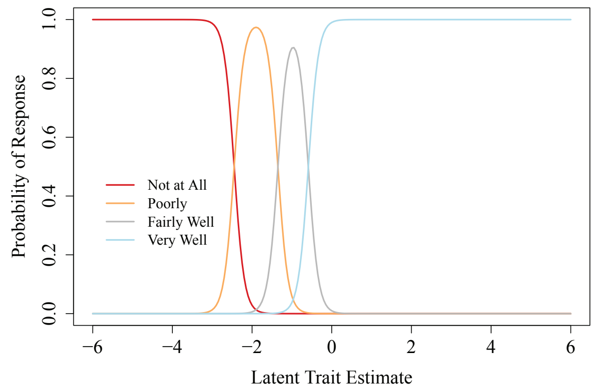
Figure 4: Category Response Curves for NIS Questions

Category response curves show the probability of providing a given response to a question as a function of latent trait L, parametrized as a Graded Response Model as in Section 2.

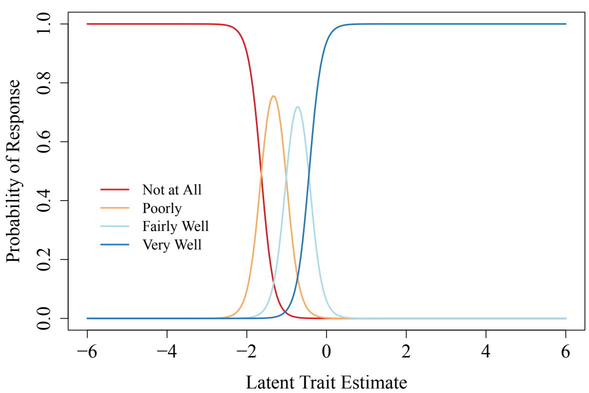
(a) Self-Assessment: Speaking



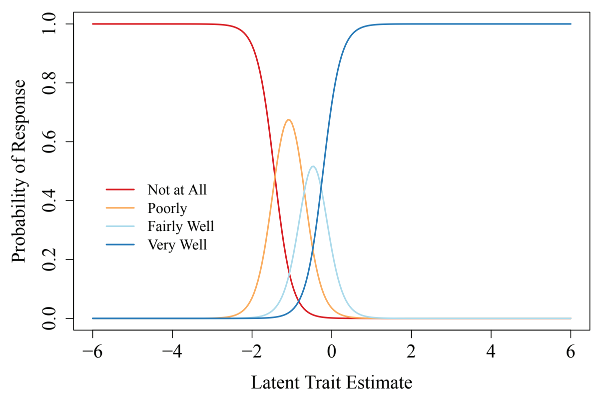
(b) Self-Assessment: Comprehension



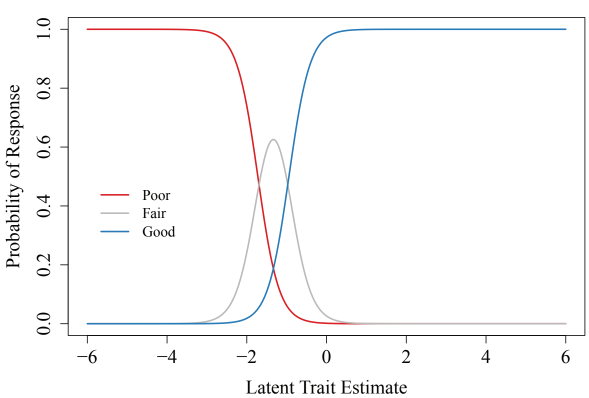
(c) Self-Assessment: Reading



(d) Self-Assessment: Writing



(e) Interviewer Assessment: Comprehension



(f) Extent of Interview Translation

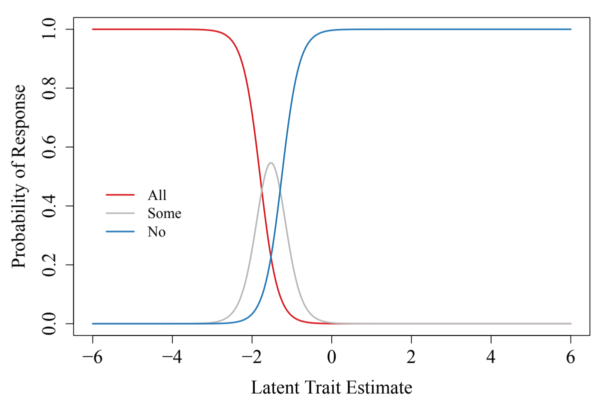


Figure 5: Category Response Curves for TeO Questions

Category response curves show the probability of providing a given response to a question as a function of latent trait L, parametrized as a Graded Response Model as in Section 2.

(a) NIS

(b) TeO

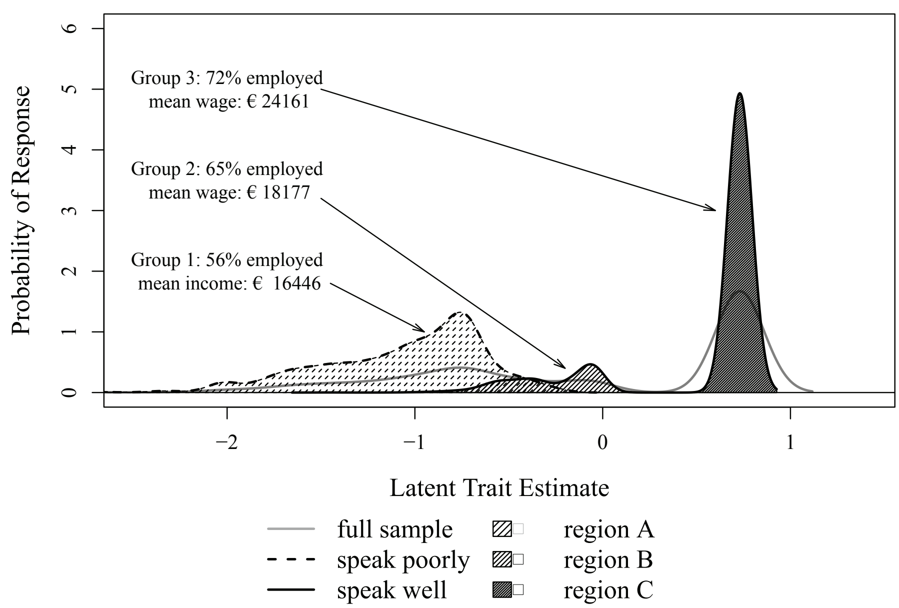


Figure 6: Latent trait distributions

Baseline GRM latent trait distributions using questions listed in Section 3.

Table 3

LATENT TRAIT ESTIMATES BY SKILL GROUP

		NIS		TeO	
		ORIGINAL	SCALED	ORIGINAL	SCALED
Group 1	Mean	-0.46	0.42	-1.05	0.48
	Median	-0.50	0.41	-0.95	0.52
Group 2	Mean	0.86	0.85	-0.29	0.71
	Median	0.86	0.85	-0.17	0.74
Group 2/Group 3 Threshold		1.15	0.94	0	0.79
Group 3	Mean	1.32	1	0.73	1
	Median	1.32	1	0.73	1

NOTE.– Table shows summary statistics of latent trait estimates $\hat{L}$, by groups illustrated in Figure 6.

Table 4

Determinants of Language Fluency

(a) NIS

	Measure of Fluency			
	(1)	(2)	(3)	(4)
	Binary (Logit)	Binary (OLS)	Ordinal (OLS)	$\hat{L}$ (OLS)
Age at Migraton (AaM)	-0.00889***	-0.0108***	-0.0109***	-0.00930***
	(0.00236)	(0.00214)	(0.00174)	(0.00145)
$(\text{AaM})^2/100$	0.00997***	0.0127***	0.00857***	0.00661***
	(0.00244)	(0.00217)	(0.00177)	(0.00149)
Yrs Since Migration (YSM)	0.0184***	0.0195***	0.0199***	0.0175***
	(0.00415)	(0.00335)	(0.00287)	(0.00237)
$(\text{YSM})^2/100$	-0.0293**	-0.0284***	-0.0359***	-0.0330***
	(0.0111)	(0.00732)	(0.00632)	(0.00573)
(YSM)×(AaM)	-0.000273**	-0.000271***	-0.000194**	-0.000155**
	(0.0000940)	(0.0000721)	(0.0000679)	(0.0000557)
Female	-0.0174	-0.0114	-0.0256***	-0.0249***
	(0.0103)	(0.0107)	(0.00751)	(0.00627)
Employed	0.0516***	0.0534***	0.0571***	0.0486***
	(0.0115)	(0.0116)	(0.00841)	(0.00699)
Married	0.00988	0.0116	0.0217*	0.0245**
	(0.0129)	(0.0130)	(0.00925)	(0.00777)
Endogamous	-0.0823***	-0.0872***	-0.0877***	-0.0800***
	(0.0125)	(0.0136)	(0.00937)	(0.00776)
Has Children	-0.0343**	-0.0450**	-0.0285**	-0.0168*
	(0.0133)	(0.0152)	(0.0101)	(0.00838)
H.S. Degree	0.104***	0.0941***	0.145***	0.123***
	(0.0148)	(0.0142)	(0.0107)	(0.00896)
Assoc./Vocational Deg.	0.101***	0.0835***	0.160***	0.143***
	(0.0164)	(0.0152)	(0.0119)	(0.00968)
4-yr College or More	0.238***	0.247***	0.298***	0.259***
	(0.0127)	(0.0136)	(0.00976)	(0.00816)
Asia	-0.564***	-0.641***	-0.356***	-0.327***
	(0.0335)	(0.0213)	(0.0153)	(0.0132)
Latin America	-0.629***	-0.715***	-0.493***	-0.439***
	(0.0326)	(0.0215)	(0.0157)	(0.0134)
Africa/Middle East	-0.467***	-0.525***	-0.256***	-0.243***
	(0.0353)	(0.0259)	(0.0169)	(0.0145)
Europe	-0.583***	-0.665***	-0.401***	-0.354***
	(0.0349)	(0.0246)	(0.0168)	(0.0144)
Constant		0.971***	1.001***	0.982***
		(0.0503)	(0.0392)	(0.0326)
N	7797	7797	7797	7797
R^2		0.262	0.429	0.457

* $p < 0.05$
** $p < 0.01$
$* * * \ p < 0.001$

(b) TeO

	Measure of Fluency				
	(1)	(2)	(3)	(4)	(5)
	Binary (Logit)	Binary (OLS)	Ordinal (OLS)	Average (OLS)	$\hat{L}$ (OLS)
Age at Migraton (AaM)	-0.182***	-0.0173**	-0.00625	-0.00991**	-0.0112**
	(0.0508)	(0.00642)	(0.00330)	(0.00361)	(0.00367)
$(\text{AaM})^2/100$	0.260***	0.0121	-0.000401	0.00626	0.00907
	(0.0789)	(0.0108)	(0.00569)	(0.00619)	(0.00629)
Yrs Since Migration (YSM)	0.167***	0.00850	0.00363	0.00242	0.00422
	(0.0359)	(0.00491)	(0.00247)	(0.00272)	(0.00285)
$(\text{YSM})^2/100$	-0.207***	-0.0127*	-0.00579	-0.00389	-0.00672
	(0.0413)	(0.00610)	(0.00304)	(0.00342)	(0.00371)
(YSM)×(AaM)	-0.00292**	-0.0000148	0.0000165	0.00000330	-0.00000860
	(0.00100)	(0.000129)	(0.0000656)	(0.0000726)	(0.0000738)
Female	0.0533	-0.00833	-0.00868	-0.00265	-0.00508
	(0.0879)	(0.0125)	(0.00606)	(0.00688)	(0.00694)
Employed	0.528***	0.0573***	0.0394***	0.0457***	0.0418***
	(0.100)	(0.0151)	(0.00751)	(0.00839)	(0.00858)
Married	0.0777	0.0132	0.00345	0.00169	0.00780
	(0.0997)	(0.0146)	(0.00699)	(0.00798)	(0.00827)
Endogamous	-0.790***	-0.128***	-0.0590***	-0.0695***	-0.0689***
	(0.0836)	(0.0138)	(0.00638)	(0.00715)	(0.00728)
Has Children	-0.118	-0.0175	-0.00934	-0.00751	-0.00810
	(0.118)	(0.0164)	(0.00815)	(0.00889)	(0.00882)
H.S. Degree	1.058***	0.125***	0.0668***	0.0959***	0.0943***
	(0.127)	(0.0184)	(0.00866)	(0.00985)	(0.00989)
2-year college	2.128***	0.212***	0.0965***	0.130***	0.130***
	(0.161)	(0.0203)	(0.00998)	(0.0111)	(0.0115)
More than 2-year college	1.997***	0.201***	0.0957***	0.130***	0.129***
	(0.109)	(0.0153)	(0.00721)	(0.00832)	(0.00855)
Sub-Sahara	1.120***	0.131***	0.0568***	0.0748***	0.0862***
	(0.123)	(0.0180)	(0.00840)	(0.00980)	(0.00963)
Asia	-1.233***	-0.196***	-0.121***	-0.122***	-0.123***
	(0.109)	(0.0169)	(0.00945)	(0.00934)	(0.0104)
MENA	0.847***	0.0858***	0.0315***	0.0478***	0.0554***
	(0.110)	(0.0157)	(0.00755)	(0.00862)	(0.00855)
Other	0.104	0.0204	0.0205	0.0349**	0.0277*
	(0.199)	(0.0293)	(0.0116)	(0.0129)	(0.0141)
Constant	1.506*	0.800***	0.912***	0.907***	0.876***
	(0.708)	(0.0869)	(0.0442)	(0.0485)	(0.0490)
N	8297	8297	8297	8297	8297
R^2		0.331	0.342	0.389	0.422

* $p < 0.05$
** $p < 0.01$
$* * *$ $p < 0.001$

Table 5

DETERMINANTS OF LOG-INCOME

(a) NIS

	MEASURE OF FLUENCY					
	BINARY		ORDINAL		$\hat{L}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Fluency	0.763***	0.456***	1.362***	0.981***	1.725***	1.306***
	(0.0656)	(0.0733)	(0.0938)	(0.119)	(0.109)	(0.141)
Age at Migraton (AaM)	0.114***	0.0921***	0.119***	0.101***	0.120***	0.105***
	(0.0209)	(0.0203)	(0.0191)	(0.0189)	(0.0192)	(0.0190)
$(\text{AaM})^2/100$	-0.148***	-0.120***	-0.146***	-0.126***	-0.145***	-0.128***
	(0.0235)	(0.0229)	(0.0218)	(0.0215)	(0.0218)	(0.0216)
Yrs Since Migration (YSM)	0.303***	0.314***	0.302***	0.305***	0.302***	0.303***
	(0.0371)	(0.0367)	(0.0325)	(0.0332)	(0.0328)	(0.0332)
$(\text{YSM})^2/100$	-0.814***	-0.809***	-0.788***	-0.780***	-0.777***	-0.769***
	(0.0924)	(0.0895)	(0.0834)	(0.0832)	(0.0836)	(0.0831)
(YSM)×(AaM)	-0.000702	-0.000707	-0.000831	-0.000772	-0.000911	-0.000836
	(0.000818)	(0.000803)	(0.000713)	(0.000714)	(0.000721)	(0.000716)
Female	-0.473***	-0.487***	-0.441***	-0.469***	-0.433***	-0.461***
	(0.0606)	(0.0601)	(0.0595)	(0.0595)	(0.0590)	(0.0592)
Constant	5.953***	6.342***	5.204***	5.752***	4.897***	5.410***
	(0.443)	(0.456)	(0.410)	(0.431)	(0.414)	(0.441)
EDUCATION		X		X		X
BIRTHPLACE		X		X		X
N	3121	3121	3121	3121	3121	3121
R^2	0.293	0.338	0.317	0.349	0.326	0.354

* $p < 0.05$
** $p < 0.01$
$* * *\ p < 0.001$

(b) TeO

	Measure of Fluency							
	Binary		Ordinal		Average		$\hat{L}$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fluency	0.227***	0.0593	0.492***	0.110	0.544***	0.137	0.592***	0.198*
	(0.0404)	(0.0385)	(0.0872)	(0.0865)	(0.0765)	(0.0834)	(0.0820)	(0.0867)
Age at Migraton (AaM)	0.0676**	0.0516**	0.0667**	0.0512**	0.0686**	0.0520**	0.0698**	0.0529**
	(0.0233)	(0.0184)	(0.0235)	(0.0184)	(0.0236)	(0.0183)	(0.0237)	(0.0184)
$(\text{AaM})^2/100$	-0.0971*	-0.0816*	-0.0947*	-0.0810*	-0.0981*	-0.0820*	-0.100*	-0.0832*
	(0.0418)	(0.0332)	(0.0422)	(0.0334)	(0.0424)	(0.0332)	(0.0428)	(0.0335)
Yrs Since Migration (YSM)	0.0504**	0.0411**	0.0505**	0.0412**	0.0504**	0.0412**	0.0491**	0.0406**
	(0.0172)	(0.0139)	(0.0173)	(0.0140)	(0.0173)	(0.0140)	(0.0174)	(0.0140)
$(\text{YSM})^2/100$	-0.0366	-0.0266	-0.0366	-0.0267	-0.0366	-0.0267	-0.0345	-0.0257
	(0.0224)	(0.0185)	(0.0225)	(0.0185)	(0.0225)	(0.0185)	(0.0224)	(0.0185)
(YSM)×(AaM)	-0.00148**	-0.00104**	-0.00149**	-0.00104**	-0.00147**	-0.00104**	-0.00146**	-0.00104**
	(0.000479)	(0.000376)	(0.000480)	(0.000375)	(0.000485)	(0.000376)	(0.000489)	(0.000379)
Female	-0.294***	-0.346***	-0.295***	-0.346***	-0.300***	-0.347***	-0.299***	-0.347***
	(0.0400)	(0.0353)	(0.0400)	(0.0352)	(0.0400)	(0.0352)	(0.0398)	(0.0353)
Constant	8.493***	8.734***	8.229***	8.680***	8.179***	8.657***	8.148***	8.608***
	(0.322)	(0.256)	(0.300)	(0.234)	(0.304)	(0.234)	(0.304)	(0.235)
Education		X		X		X		X
Birthplace		X		X		X		X
N	4956	4956	4956	4956	4956	4956	4956	4956
R^2	0.133	0.242	0.132	0.242	0.139	0.242	0.144	0.243

* $p < 0.05$
** $p < 0.01$
$* * * \; p < 0.001$

Table 6

DEMOGRAPHIC/SKILL CORRELATES OF VERBAL-ONLY VS. FULL PROFICIENCY: NIS

	LOGIT: GROUP 2 VS. 3		OLS: LOG-INCOME	
	ALL	EMPLOYED	GROUP 2	GROUP 3
	(1)	(2)	(3)	(4)
Used English at age 10	0.354	0.294	0.358	0.164
	(0.184)	(0.234)	(0.388)	(0.216)
Uses only English with friends	0.179	0.544*	-0.929**	-0.0201
	(0.173)	(0.220)	(0.323)	(0.158)
Took Eng. class before migration	0.427**	0.450*	0.268	0.250
	(0.161)	(0.197)	(0.279)	(0.139)
Took Eng. class in last yr.	-0.310	-0.0530	-0.142	0.194
	(0.348)	(0.418)	(0.347)	(0.219)
Employed	0.0715			
	(0.172)			
Finished educ. in USA	0.198	0.277	0.610*	0.0763
	(0.211)	(0.255)	(0.267)	(0.152)
H.S. Degree	-0.0416	-0.119	0.315	0.00632
	(0.249)	(0.332)	(0.446)	(0.280)
Assoc./vocational deg.	-0.0510	-0.125	0.721	0.447
	(0.277)	(0.360)	(0.439)	(0.311)
4-yr College or More	0.152	0.0842	0.565	0.669*
	(0.227)	(0.309)	(0.351)	(0.294)
Married	0.584**	0.389	0.339	0.356
	(0.192)	(0.254)	(0.334)	(0.183)
Endogamous	-0.332	-0.176	-0.195	0.0102
	(0.208)	(0.263)	(0.285)	(0.198)
Female	-0.203	-0.141	-0.237	-0.477***
	(0.150)	(0.192)	(0.235)	(0.134)
VISA TYPE				
Employment	0.583*	0.440	1.254**	1.368***
	(0.280)	(0.352)	(0.431)	(0.388)
Family	-0.282	-0.141	0.626	0.790*
	(0.246)	(0.340)	(0.389)	(0.360)
Diversity	-0.447	-0.760	-0.0919	0.479
	(0.310)	(0.399)	(0.450)	(0.402)
Refugee/Asylee	-0.372	-0.460	1.261**	0.740*
	(0.390)	(0.445)	(0.440)	(0.375)
Legalization	-0.759	-0.783	0.151	0.0433
	(0.419)	(0.503)	(0.571)	(0.520)
Constant	1.747***	1.570*	8.269***	7.918***
	(0.501)	(0.638)	(0.871)	(0.582)
N	1663	1217	140	573

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7

EMPLOYMENT VARIABLES AND VERBAL-ONLY VS. FULL PROFICIENCY: NIS

	LOGIT	OLS: LOG-INCOME	
	GROUP 2 VS. 3	GROUP 2	GROUP 3
	(1)	(2)	(3)
Uses only Eng. at work	0.225	0.319	0.204
	(0.229)	(0.298)	(0.156)
H.S. Degree	-0.235	0.460	0.103
	(0.332)	(0.508)	(0.307)
Assoc./Vocational Deg.	0.0418	0.771	0.482
	(0.371)	(0.431)	(0.334)
4-yr College or More	0.187	0.666	0.534
	(0.321)	(0.399)	(0.286)
Finished educ. in USA	0.0662	0.129	0.113
	(0.263)	(0.285)	(0.132)
Procured job via friend/relative	0.380	-0.739	-0.670
	(0.333)	(0.425)	(0.369)
Procured job prior to migration	0.338	0.00307	0.399**
	(0.333)	(0.570)	(0.144)
JOB SECTOR			
Management	0.640	1.431***	0.869**
	(0.428)	(0.416)	(0.331)
Sciences	0.507	1.340**	1.280***
	(0.406)	(0.413)	(0.268)
Education/Law	0.324	0.531	0.267
	(0.469)	(0.640)	(0.363)
Healthcare	0.317	-0.0749	0.413
	(0.434)	(0.526)	(0.305)
Services	0.0151	0.0820	0.0788
	(0.301)	(0.377)	(0.267)
Ag./Manufacturing	0.0430	1.052*	0.275
	(0.450)	(0.433)	(0.386)
Constant	1.812**	7.112***	8.083***
	(0.610)	(0.762)	(0.501)
N	1151	127	546

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

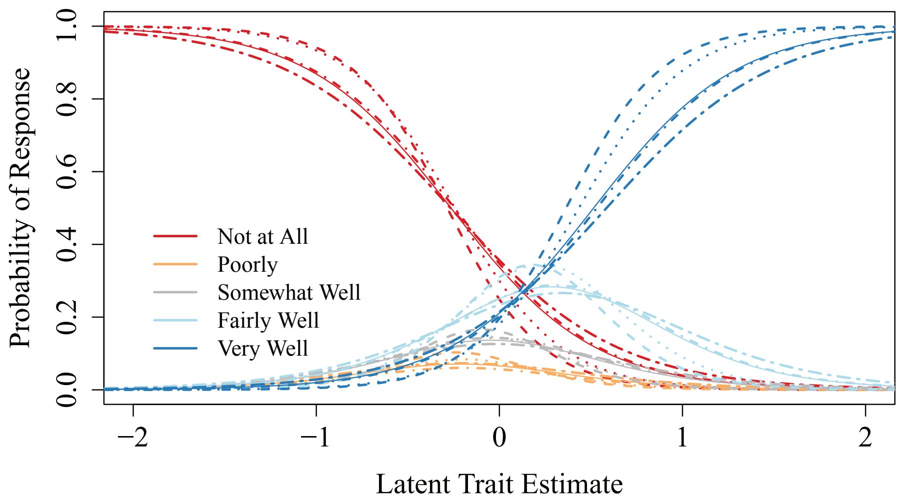
Table 8

Demographic/Skill Correlates of Verbal-Only vs. Full Proficiency: TeO

	Logit: Group 2 vs. 3		OLS: Log-Income	
	All (1)	Employed (2)	Group 2 (3)	Group 3 (4)
Uses only French with friends	0.0708	0.0727	-0.0388	0.0145
	(0.141)	(0.175)	(0.0545)	(0.0466)
Uses French with children	0.258	0.365	-0.174**	0.0264
	(0.171)	(0.217)	(0.0632)	(0.0607)
Uses French with spouse	-0.272	-0.216	-0.0304	-0.0945
	(0.172)	(0.218)	(0.0760)	(0.0569)
Employed	0.587***			
	(0.136)			
Procured job via friend/family		-0.183	-0.124*	-0.145**
		(0.174)	(0.0521)	(0.0532)
Skilled laborer		0.103	0.0549	0.104
		(0.308)	(0.0631)	(0.0888)
Professional		0.603*	0.105	0.340***
		(0.273)	(0.0648)	(0.0783)
H.S. Degree	1.239***	1.028***	-0.0973	0.0653
	(0.223)	(0.300)	(0.100)	(0.0724)
2-year college	1.452***	1.074**	0.284**	0.180**
	(0.282)	(0.356)	(0.109)	(0.0676)
More than 2-year college	1.140***	1.040***	0.203	0.470***
	(0.216)	(0.297)	(0.189)	(0.0684)
Finished Educ in France	0.913***	0.908***	-0.0583	0.0231
	(0.193)	(0.238)	(0.0788)	(0.0658)
Married	0.275	-0.0754	-0.0248	0.0897
	(0.188)	(0.224)	(0.0763)	(0.0702)
Endogamous	-0.138	0.0593	0.100	0.0650
	(0.140)	(0.176)	(0.0608)	(0.0476)
Female	0.156	-0.0235	-0.511***	-0.370***
	(0.147)	(0.198)	(0.0611)	(0.0511)
French Learning Before Migration				
Nat. lang. in country	0.388*	0.154	0.113	-0.0376
	(0.198)	(0.249)	(0.0591)	(0.0623)
Learned from family	-0.0193	-0.0432	-0.00481	0.00397
	(0.182)	(0.249)	(0.0608)	(0.0665)
Learned in school	1.077***	0.986***	-0.0919	-0.0602
	(0.136)	(0.176)	(0.0706)	(0.0478)
Learned from TV/media	0.239	-0.0429	-0.0352	-0.0223
	(0.228)	(0.281)	(0.0711)	(0.0724)
Learned from tapes	0.299	0.451	0.0901	-0.139
	(0.406)	(0.558)	(0.144)	(0.110)
Learned at work	-0.593*	-0.668	0.238**	0.184
	(0.287)	(0.383)	(0.0836)	(0.128)
Visa Type				
Employment	-0.267	-0.296	-0.0401	0.0257
	(0.212)	(0.261)	(0.0838)	(0.0648)
Student	0.180	0.255	0.0353	-0.0953
	(0.266)	(0.356)	(0.170)	(0.0927)
Family	-0.256	-0.243	-0.0917	-0.109
	(0.180)	(0.245)	(0.0907)	(0.0571)
Refugee	0.0610	0.0136	-0.0248	-0.0231
	(0.253)	(0.327)	(0.0931)	(0.0839)
Constant	-1.858***	-1.911**	10.16***	9.563***
	(0.515)	(0.733)	(0.192)	(0.157)
N	1814	1124	476	582

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(a) Interviewer Assessment: Speaking



(b) Interviewer Assessment: Comprehension

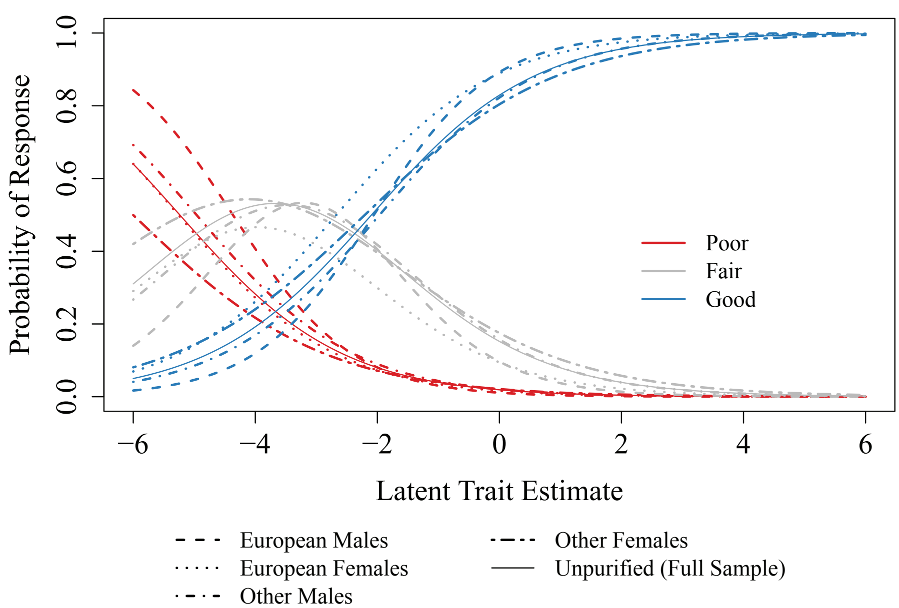


Figure A.1: Category Response Curves After DIF Purification (NIS)

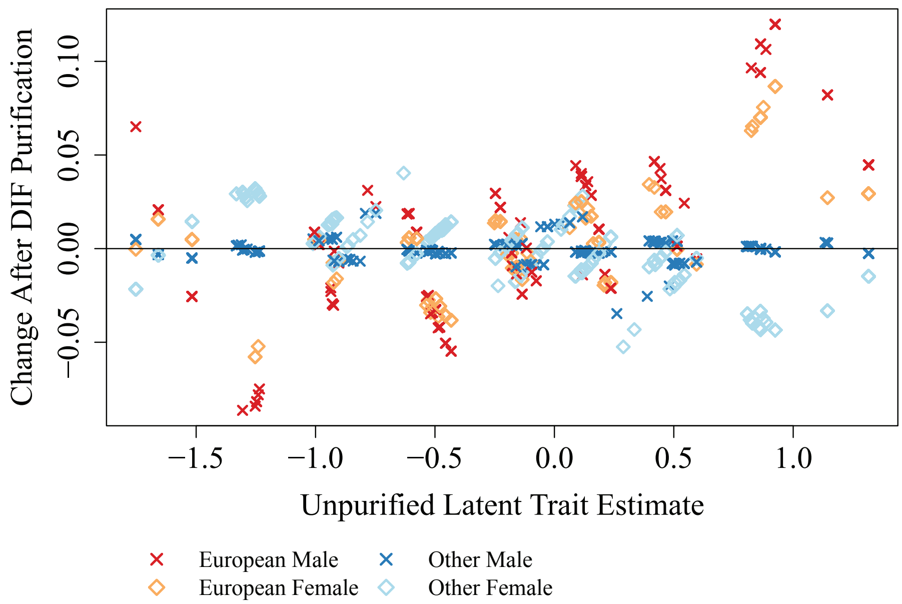


Figure A.2: Changes in Latent Trait Values After DIF Purification

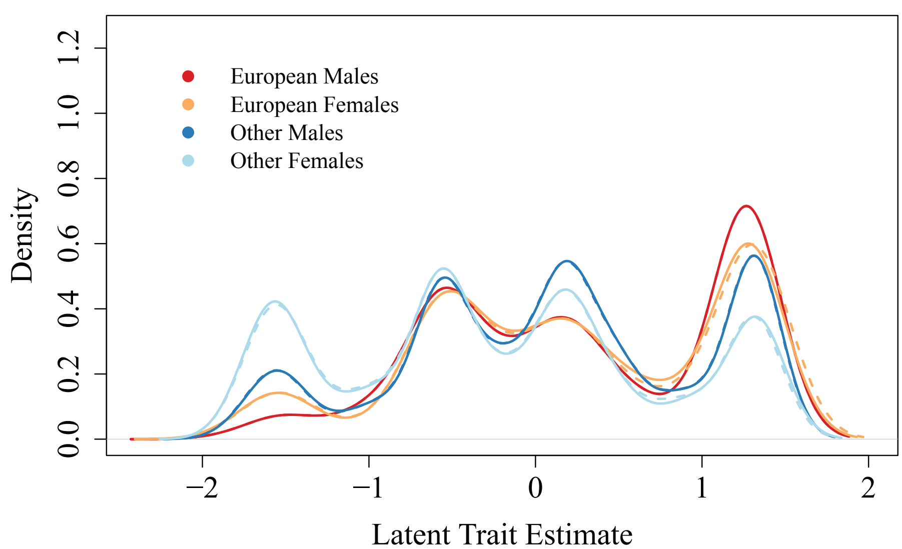


Figure A.3: Latent Trait Distributions Before and After DIF Purification