Institute for Employment Research

The Research Institute of the Federal Employment Agency

IAB

IAB-Discussion Paper 10/2010

Articles on labour market issues

Ethnic Concentration and Language Fluency of Immigrants

Alexander M. Danzer
Firat Yaman

Ethnic Concentration and Language Fluency of Immigrants

Quasi-Experimental Evidence from the Guest-Worker Placement in Germany*

Alexander M. Danzer (Royal Holloway, University of London and IZA Bonn)
Firat Yaman (IAB)

* The authors would like to thank Herbert Brücker, Dan Hamermesh, Victor Lavy, Stephen Trejo, Jonathan Wadsworth, Natalia Weisshaar, Katja Wolf as well as seminar participants at Austin, Royal Holloway, IAB Nürnberg and Essen. Special thank goes to Jan Goebel from the SOEP department at DIW Berlin. All remaining errors are our own. This paper was partly written as part of the Marie Curie Research Training Network on 'Transnationality of Migrants' TOM, which is funded by the European Commission through the Human Resources and Mobility action of its Sixth Framework Programme (EC Contract No. MRTN-CT-2006-035873)

Contents

Abstract

The paper analyses the impact of regional own-ethnic concentration on the language proficiency of immigrants. It solves the endogeneity of immigrants' location choices by exploiting the fact that guest-workers in Germany after WWII were initially placed by firms and labor agencies. We find a robust negative effect of ethnic concentration on immigrants' language ability. Simulation results of a simultaneous location and learning choice model confirm the presence of the effect and show how immigrants with high learning cost select into ethnic enclaves. Under the counterfactual scenario of a regionally equal distribution of immigrants the share of German-speakers increases only modestly.

Zusammenfassung

Dieses Paper analysiert die Wirkung von regionalen ethnischen Konzentrationen auf die Sprachkompetenz von Einwanderern. Wir lösen die Endogenität der Wohnortentscheidungen der Einwanderer in dem wir den Umstand nutzen, dass Einwanderer im Rahmen der Gastarbeiteranwerbung ihren ersten Arbeitsort nicht auswählen konnten. Wir finden belastbare negative Effekte der ethnischen Konzentration auf die Beherrschung der deutschen Sprache. Simulationsergebnisse eines Modells mit gleichzeitigen Lern- und Wohnortentscheidungen unterstützen diesen Befund und zeigen, dass Einwanderer mit hohen Lernkosten gezielt in ethnische Enklaven ziehen. Unter dem Szenario einer Gleichverteilung der Einwanderer über Deutschland würde sich der Anteil der deutsch sprechenden Einwanderer nur geringfügig erhöhen.

JEL classification: J61, R23, F22

Keywords: Enclave, ethnic concentration, language proficiency, immigrants, instrumental variable, random utility mode

1. Introduction

Immigration and the social and economic performance of immigrants have been controversial policy issues for decades, both in North America and Europe. From the immigrants' perspective, leaving behind a familiar social context and adapting to a new environment can be challenging; however, the experience is exacerbated if immigrants do not succeed in integrating into the host country's society, a state often associated with the failure to learn the majority language. The existence of segregated "parallel" societies which are said to be characterised by poverty risk and unemployment, has fueled the debate on the integration of immigrants in many countries—and among them Germany, the country with the largest immigrant population in Europe. An often articulated political concern refers to immigrant groups forming self-sufficient enclaves and thus challenging the life-style as well as formal or informal institutions of native societies. In Germany, those fears may be related to the official denial of being a country of immigration until recently[1], despite the fact that 10 percent of the population is classified as foreign.

The scope of this paper is to analyse the effect of regional ethnic concentration on the language proficiency of immigrants. The empirical analysis is based on the quasi natural experiment generated by the specific nature of guest-worker immigrants to Germany after WWII. By exploiting the fact that immigrants were exogenously placed in firms across West Germany, we estimate the causal effect of own-ethnicity concentration on a basic type of human capital, namely German speaking and writing proficiency. By merging several representative data sets and addressing potential endogeneity bias with an IV approach we provide robust evidence of a small negative effect from ethnic concentration on language fluency. Similar questions have been addressed for other large-scale immigration countries like the USA, Australia and Canada; however, none of these studies can solve the problem of endogenous self-selection of migrants by using a quasi randomized placement of migrants in neighbourhoods.

The economic consequences of immigrants' language command have been studied intensively and for many countries (see for example Chiswick and Miller (2002) and (2005) and Bleakley and Chin (2004) for the USA, Dustmann and Fabbri (2003) for the UK, Dustmann (1994) and Dustmann and van Soest (2001) and (2002) for Germany, and Chiswick and Miller (1995) for Australia). Comparing the "fluency penalty" across the cited articles is complicated by differences in immigration histories (e.g., rates of return migration), and by differences in methodologies including the survey instrument to investigate self-assessed language proficiency (e.g., the U.S. and Australian censuses distinguish between four levels of English proficiency, whereas the German GSOEP data have five levels); however, the entire literature confirms that immigrants with good speaking and writing abilities perform better in the labour market in terms of employment and earnings compared to immigrants who speak and write poorly.

[1] "Germany is not an immigration country." was the leading principle for immigration and "foreigner"-policies in the coalition contract between conservatives and liberal democrats in 1982 (Herbert, 2001, pp. 247-248). See also the essay "Integration ist machbar" by Bade in the daily newspaper Die Welt (2009).

Another strand of the immigration literature has focused on the influence of ethnic enclaves on economic performance and/or language fluency. Theoretical arguments for the influence of ethnic capital and its transmission through neighbourhoods on immigrants' performance have been made by Borjas (1995) and (1998). Most studies that we are aware of find a negative association between ethnic concentration and language proficiency (Cutler et al. (2008), Chiswick and Miller (2005), Lazear (1999) for the USA, Warman (2007) for Canada, and Dustmann and Fabbri (2003) for the UK, and Chiswick and Miller (1996) for Australia). Only the paper by Cutler et al. (2008) attempts to correct for the potential self-selection of immigrants into specific neighbourhoods (ghettos) by using an occupational instrument matrix; however, occupation, location and language choice might be parts of the same decision. For Germany, no study analyses the link between ethnic concentration and language proficiency.[2] The cited papers vary substantially in the size of the regions for which ethnic concentrations are defined, but the negative effect is consistently stronger when the regions are defined on less aggregated levels. If immigrants who are less willing or able to learn a language cluster in local neighbourhoods and counties this finding is not surprising.

Stronger identification attempts have been made in studies estimating the effect of ethnic concentration on earnings. Cutler and Glaeser (1997) use instruments pertaining to the administration and topography of the regions, such as public finance, the number of local governments and rivers in US metropolitan statistical units, and find negative effects of segregation for African-Americans. The authors are the first to seriously address the issue of endogenous selection of migrants; however, it seems at least debatable whether the structure of past public finance is sufficiently exogenous to the destination choice of immigrants in the face of long-term ghettoisation trends. In order to reduce the extent of mobility across metropolitan areas the authors limit their analysis to African-Americans aged 20 to 30—a strategy that might introduce other biases when focusing on outcomes that pertain to the transition period between education and labour market. Policies concerning the exogenous placement of asylum seekers in specific municipalities have provided natural experiments for two small Scandinavian countries. Damm (2009) and Edin et al. (2003) use the initial exogenous placement of refugees in Denmark and Sweden to instrument for current exposure to their own ethnic group. The latter two studies have so far most convincingly solved the problem of self-selection, however, it should be noted that asylum seekers may differ from labour migrants with respect to background and behaviour. Edin et al. find that living in enclaves improves earnings of less skilled refugees while no significant effect pertains for those with more than 10 years of education. Damm finds that higher ethnic concentrations increase earnings irrespective of skill levels.

This paper combines the latter two strands of literature by using an initially exogenous placement policy of labour migrants in order to instrument the effect of regional ethnic composition on language ability. Apart from being the first study that can provide evidence from a natural ex-

[2] Sociological research has dealt with the neighbourhood quality of ethnic clusters in Germany (Drever, 2004).

periment in a large traditional immigration country, our paper adds to the literature a learning and location choice model which yields testable implications for the link between ethnic concentration and language proficiency, and which is able to explain why studies on the effect of enclaves on earnings remain contradictory. Furthermore, the model allows simulating counterfactual outcomes for changes in regional ethnic concentration or average immigrant characteristics. This exercise is informative for gauging the impact of potential future immigration and is illustrated with the example of Germany fully opening her labour market for the Central and Eastern European countries of the EU in 2011.

The paper focuses on the language skills of immigrants for the following reasons:

- Language skills are a crucial part of the human capital endowment of an immigrant and the employment and earnings implications are well documented (see below). By looking at language as an endogenous variable we dissect one of the proximate determinants of labour market outcomes.[3]
- If the costs of learning German are of a non-monetary nature (e.g., effort) while benefits are largely reflected in wages, the assumption of a monotonic and continuous dependency of wages on ethnic concentrations might give rise to misspecifications, as will be shown in the next section.
- The benefits of having a good command of German extend to many areas outside the labour market (e.g., participation in the civil society or use of health care) and have been used to measure successful integration of immigrants.

The remainder of the paper is as follows: In Section 2, we set up our simple learning and location choice model. Section 3 gives a brief overview of the guest-worker programme in Germany and underlines specific characteristics that resulted in exogenous placement of immigrants across German regions. Section 4 provides a detailed overview of the identification strategy used throughout the paper, a description of data sources as well as a discussion of the choice of the regional aggregation level. Section 5 provides the results from the econometric analyses. Section 6 discusses potential explanations for the difference in OLS and IV estimates as well as the potential bias from measurement error. Section 7 contains some brief policy simulations based on our structural model, while Section 8 concludes.

2. Theory

In this section, we turn to the random utility model which derives location and learning choice probabilities through utility maximizing behavior. Suppose learning German is costly, and the cost of immigrant *i* can be described by some observable characteristics X_i, a vector of parame-

[3] It should, however, be noted that ethnic enclaves may provide alternative, but often low paid service sector opportunities.

ters β, and an unobservable component ε, such as ability, which is assumed to be continuous and which we allow, but do not require, to differ across regions *j*. Assuming the cost to be linear in variables we write:

$$c_i^j = X_i'\beta + \varepsilon_i^j \qquad (1)$$

We assume that an immigrant enjoys some benefit from the number or share of people he can interact with. An example is the model by Lazear (1999), in which two people in a region are matched randomly and trade occurs with a fixed payoff if both can interact, that is, speak the same language. In this case the benefit would be the expected payoff before a match occurs and it would be linear in the share of people an agent can interact with. For the moment we just use a generic function $\xi(x^j)$ where x^j equals the fraction of people the immigrant can interact with in location *j*, so that x^j takes the value x_f^j if the immigrant does not speak German (the subscript *f* standing for foreign) and $x_f^j + x_n^j$ if he speaks German (*n* standing for native). The shares of natives and all foreigner groups (denoted by an indexing set *F*) have to sum to 1:

$$x_n^j + \sum_{z \in F} x_z^j = 1 \qquad x_n^j, x_z^j \geq 0. \qquad (2)$$

Different locations are then characterised by:

1. variables differing across locations but equal for all immigrants in that location, W_j,
2. variables characterizing the ethnic concentration, x_f^j, x_n^j which differ across locations and across immigrant groups (but not across immigrants of the same country of origin),
3. an unobservable, continuously distributed component r_i^j.

Denoting by S_i an indicator taking on the value of one, if the immigrant learns German, and zero otherwise, utility of choosing *S* and location *j* for a given immigrant *i* is thus:

$$U_{i,(S,j)} = S_i * (\xi(x_n^j + x_f^j) - X_i'\beta) + (1 - S_i) * \xi(x_f^j) + W_j'\gamma + r_i^j - S_i\varepsilon_i^j \qquad (3)$$

The first term describes the deterministic part (from the point of the econometrician) of utility from learning German. The immigrant can interact with both natives and immigrants of his own group, but incurs the cost $X_i'\beta$. The second term is the utility enjoyed by interacting with other members of the immigrant group only. The term $W_j'\gamma + r_i^j$ describes the utility specific to the region for the immigrant, regardless of whether or not he learns German, and the last term is an unobservable part of the cost of learning German. The choice set consists of all unordered distinct pairs of (S, j) and the chosen alternative is

$$(S^*, j^*) = \underset{s\in\{0,1\},\, j\in J}{\arg\max}\{U_{(s,j)}\} \tag{4}$$

Equation (4) simply states that learning and location decisions are part of the same choice, a fact that was acknowledged but not formalised earlier by Lazear (1999) and Bauer et al. (2005).

For notational simplicity, denote the observable part of utility by

$$V_{i,(1,j)} = \xi(x_n^j + x_f^j) - X_i'\beta + W_j'\gamma$$
$$V_{i,(0,j)} = \xi(x_f^j) + W_j'\gamma$$

and the composite error term $r_i^j - S_i\varepsilon_i^j$ by $\omega_{i,(S,j)}$. Omitting the individual index, the probability of learning German and choosing location *j* is given by

$$\begin{aligned} P(S=1, j) &= P(V_{(1,j)} - V_{(s,k)} \geq \omega_{(s,k)} - \omega_{(1,j)} \quad \forall s \in \{0,1\}, k \neq j) \\ &= \Phi_{(1,j)}(-V_{(1,1)}, -V_{(0,1)}, ..., +V_{(1,j)}, -V_{(0,j)}, ..., -V_{(1,J)}, -V_{(0,J)}) \end{aligned} \tag{5}$$

with $\Phi_{(1,j)}(\mathbf{V})$ being the distribution function of $(\omega_{(1,1)}, \omega_{(0,1)}, ..., \omega_{(1,J)}, \omega_{(0,J)})$ for $P(S=1, j)$ at **V**. The second equality is simply saying that the probability of choosing a particular (S, j) is increasing in the associated utility and decreasing in the observed utility of any other alternative. In order to be able to make statements about the reaction of learning probabilities to changes in the immigrant share within a region we need to introduce an assumption concerning the payoff function ξ:

Assumption 1. $\xi(x)$ *is differentiable and strictly increasing in its argument.*

It follows:

Proposition 1. *Let the choice problem of the immigrant be described by equations (2) and (3). Under assumption 1, and holding constant the shares of all immigrant groups other than f:*

1. $\frac{\partial P(S=1 \mid j)}{\partial x_f^j} \leq 0$

2. $\frac{\partial P(S=1, j)}{\partial x_f^j} \leq 0$

3. $\frac{\partial P(k)}{\partial x_f^j} \leq 0, \qquad \frac{\partial P(j)}{\partial x_f^j} \geq 0, \quad k \neq j$

4. $\frac{\partial P(S=1)}{\partial x_f^j} \leq 0$

with strict inequalities if $\omega_{(s,j)}$ *has strictly positive density everywhere.*

Proof: see Appendix.

All of the inequalities above follow the same intuition: An increase of x_f^j increases the observable part of utility of only one choice, which is moving to *j* and not learning German. In particular, it leaves the utility of choice *(S=1,j)* unchanged, since the increased immigrant share just replaces natives and does not change the interaction possibilities for a German-speaker. This effect is captured in the (-1+1) terms in the proof. Thus, all options including learning German are decreased in value relative to *(S=0,j)*. Furthermore, since all choice probabilities other than *P(S,j)* are decreasing in $V_{(S,j)}$, the probability of moving to any location $k \neq j$ is also decreasing.

If we assume ξ to be concave, the condition that a higher share of the own immigrant group *f* replace the respective share of natives can be relaxed. Furthermore, the results can be generalised to the case where x_f^j and x_n^j stand for the absolute number of immigrants and natives in a region. At least in the latter case, concavity would not be an innocuous assumption if there exist externalities in benefits from *x* for some range, such as threshold values for the supply of certain goods and services.

>> Figure 1 about here <<

An interesting implication of the model arises when the costs of learning German are unobserved (time and effort spent learning), but benefits are to some extent reflected in higher earnings (see Figure 1). The earnings (solid line) of immigrants will then be increasing in the ethnic concentration for immigrant *i* for concentrations above a certain threshold value $x > \bar{x}_i$ with $\bar{x}_i$ being the concentration of own group members at which immigrants stop to learn German. For all values below $\bar{x}_i$ the immigrant learns German and her earnings are invariant to $x \in [0, \bar{x}_i]$. If $\bar{x}_i$ is smaller for less educated immigrants (they do not learn German even at low ethnic concentrations), empirical studies might find a positive concentration effect on earnings for less educated and no effect for better educated immigrants (which for example is found by Edin et al.), or might find inconclusive results.

To compare the quantities in proposition 1, we need another assumption:

Assumption 2. *The probability of learning German conditional on location j reaches 1 as* x_f^j *approaches zero:*

$$\lim_{x_f^j \to 0} P(S = 1 \mid j) = 1$$

Proposition 2. *Let the choice problem of the immigrant be described by equations (2) and (3). Under assumptions 1 and 2:*

$$\frac{\partial P(S=1)}{\partial x_f^j} > \frac{\partial P(S=1 \mid j)}{\partial x_f^j}$$

for small x_f^j *(*x_f^j *approaching 0).*

Proof: see Appendix.

Note that the "small" x_f^j condition is sufficient, and less restrictive and/or alternative conditions can be found. For example, the inequality will hold if the marginal utility from contact with other people goes to infinity as the share of people one can interact with goes to 0, $\lim_{x \to 0} \xi'(x) = \infty$, or whenever

$$\sum_{k\neq j}\frac{\partial P(k)}{\partial x_f^j}P(S=1\,|\,k)\geq\sum_{k\neq j}\frac{\partial P(k)}{\partial x_f^j}P(S=1\,|\,j)$$

The intuition of the proposition is that if the immigrant is limited to one location, he cannot "escape" the incentive to learn German by moving to another location. Lowering the immigrant share in a location where it was low initially is not going to change the learning decision of the immigrant, since he can choose from a multitude of locations.

The results are fairly general and do not require any distributional assumptions other than continuity on the ω. In particular, no covariance structure is assumed. To illustrate the working of the model we provide a short example. Let the ω be independently (across choices and individuals) and identically distributed type I extreme value errors, resulting in the well-known multinomial logit model with the choice probabilities given by

$$P(S,j)=\frac{e^{V_{(S,j)}}}{\sum_k\left(e^{V_{(1,k)}}+e^{V_{(0,k)}}\right)} \qquad (6)$$

Let $V_{(S,j)}=S*\left(\xi(x_n^j+x_f^j)\right)+(1-S)*(\xi(x_f^j))$, so that observable utility is given only by the composition of the population. Finally, let $\xi(x)=\ln(x)$. It is easy to verify that assumptions 1 and 2 hold under this specification. We would have:

$$P(S=1\,|\,j)=\frac{1}{1+x_f^j}\in[(1/2),1]$$

$$\frac{\partial P(S=1\,|\,j)}{\partial x_f^j}=-\frac{1}{(1+x_f^j)^2}\in[-1,-(1/4)]$$

$$P(S=1)=\frac{J}{J+\sum_k x_f^k}\in[(1/2),1]$$

$$\frac{\partial P(S=1)}{\partial x_f^j}=-\frac{J}{(J+\sum_k x_f^k)^2}\in[-(1/J),-(1/4J)]$$

It can be verified that

$$\lim_{x_f^j\to 0}\frac{\partial P(S=1\,|\,j)}{\partial x_f^j}<\lim_{x_f^j\to 0}\frac{\partial P(S=1)}{\partial x_f^j}$$

The model is highly stylised to highlight the decision problem and the tradeoffs that immigrants face, and naturally it has some shortcomings. First, we accept the payoff function ξ as a black-box mechanism. Agents benefit from increased communication prospects with other agents, but

we do not link these benefits to any "deep" parameters or structures.[4] A more serious problem might be the omission of moving costs. Here, we are mainly led by data restrictions in our decision not to model moving costs. The bottleneck in the empirical part is the number of immigrants in the German Socio-Economic Panel, with roughly 2,000 observations in 1985 and 1,000 observations in 2001. Very few of those move across regions, as we define them, and we cannot know for what reasons they change their location. Our conjecture is that moving costs would bring the marginal probabilities of learning German conditional and unconditional on location closer to each other, since "escaping" a region becomes more costly.

While the working of the model as summarised above is instructive for understanding the choice situation and the trade-offs each immigrant faces as well as for thinking about counterfactuals, the estimation strategy should depend on the hypothesis to be tested. In the empirical section we aim at estimating

$$\frac{\partial P(S=1|\ j)}{\partial x_f^j}$$

for its intuitive interpretation (as a treatment effect). Identification will rely on our assumption that initial placement and location choices for a certain time period after arrival in Germany were exogenous to immigrants with respect to their willingness/ability to learn German.

A more holistic estimation (allowing for simultaneous learning and location choices of immigrants) of responses to different ethnic concentration counterfactuals will be carried out in the experiments section. Naturally, the latter will require more of the structure outlined above (and consequently will be more restrictive), but can be carried out without the use of instruments. Thus, it provides a robustness check of the direct estimation of the treatment effect.

3. The Guest-Worker Programme in West Germany 1955-1973

The 1950s and 60s in Germany have become known as the time of the „Wirtschaftswunder" (economic miracle), an episode of rapid post-war reconstruction and economic growth. The miracle has been facilitated by an inflow of refugees from East Germany and territories formerly belonging to the German Reich or inhabited by a German-speaking population. As this inflow (8.3 million until 1950) ebbed off, labour shortages became evident, and between 1959 and 1962 the number of vacancies overtook the number of people registered as unemployed. The guest-worker recruitment in Germany began with the German-Italian Recruitment Treaty signed

[4] While this could be done (Lazear (1999) being a possible starting point), it would be only of secondary interest in answering our research question.

in December 1955 to meet the hunger for labour of the German economy.[5] Subsequent treaties were signed with Greece and Spain in 1960, Turkey in 1961, Portugal in 1964, and Yugoslavia in 1968.

>> Figure 2 about here <<

Figure 2 shows the development of the share of the foreign population in Germany, where foreign is defined as not holding German citizenship. Until 1960 the presence of guest-workers was a marginal phenomenon, but we see that recruitment gained momentum in the early 60s and increased steadily until 1967. A dip in the share of foreign employees occurred in 1967 as the result of a brief recession, which however did not affect the further inflow of the foreign population. Within 13 years, the share of foreign employees rose from less than one to twelve percent. Recruitment was halted in 1973 as a consequence of a more severe economic recession; however, the upward trend of the foreign population continued modestly due to family reunification.

The composition of the foreign population has been subject to substantial changes, as seen in Figure 3. While Italians constituted the most numerous group of foreigners in 1969, the Turkish population overtook all other groups in 1971 and has been widening the gap ever since. Notably, the numbers of Turks never decreased after the recruitment stop, as it did for other guest-worker groups.

>> Figure 3 about here <<

Technically, the recruitment was performed by a recruitment commission in the sending country which was jointly set up by the Federal Employment Agency of Germany and the Labour administration of the sending country. German firms requested workers according to their needs and the commission assigned workers from an application pool to specific firms. Workers signed one-year contracts with their first employers at decentralised labour office branches before arriving in Germany. Permits to live in Germany for the duration of one year were issued, but the permission was conditional on employment with the employer of the contract. Accommodation and travel costs were covered by the employer, so that monetary and administrative costs of the application and the move were essentially zero for the guest-worker. The recruitment was designed to attract workers with very low skill requirements. In Germany, most guest-workers became employed in manufacturing, notably in the construction, mining, metal and ferrous industries. As of 1966, 72% of the foreign workforce comprised unskilled workers.

[5] The description of the recruitment history and its technicalities draws on Herbert (2001).

4. Identification

The basic question we attempt to answer in this research is whether the ethnic composition in their neighbourhood negatively impacts on the language fluency of immigrants. We use the quasi-natural experiment of the guest-worker immigration that took mainly place in the 1960s and 1970s in order to establish a causal link between area composition and individual ability to speak and write German. Guest-workers were little educated and generally without any knowledge of the German language upon arrival thus reducing the problem of selective migration.[6] As guest-workers were contracted in their home countries based on the (mostly manual) labour demand of German firms and administered by outlets of the German Labour Office, migrants had no control over their placement in Germany.[7] The idea is then to compare immigrants who were placed in areas with different ethnic compositions and thus with different incentives and costs to learn German. Those confronted with a high regional density of non-Germans will be less likely to require a good command of the host country language for their daily interactions. At the same time, they have fewer opportunities to learn from the interaction with German speakers. The natural counterfactual for a person living in a cluster with a high concentration of own ethnic co-residents is a person of the same ethnicity in a low-concentration area. Comparing persons of the same ethnicity, levels out the potential bias from linguistic distances between languages.

The ideal set-up of our investigation would be to have a data source with objective measures of language speaking and writing fluency for immigrants who were randomly distributed over Germany without ever changing their place of residence. In this case, we could simply estimate the basic OLS model

$$y = \alpha + \beta x + \gamma(ysm) + \kappa + \mu + u \qquad (7)$$

where *y* stands for language ability, *x* stands for ethnic concentration, (*ysm*) stands for exposure to the host country language (years since migration), κ are country of origin fixed effects, μ are regional fixed effects and *u* is a random error term. The estimated coefficient β would report the own-ethnic concentration effect which should carry a negative sign in case we expect ethnic concentration to inhibit learning German, that is if assumption (1) in the theory section holds.

Data

In order to estimate the causal effect of ethnic concentration on language fluency, this paper combines different data sources. As we are interested in language ability of individual immigrants, we make use of the guest-worker sample B of the German Socio-Economic Panel

[6] In a recent study on linguistic integration of immigrants in Germany, still more than 90 percent of Turkish immigrants responded that they had no usable German knowledge upon arrival (Rother, 2008).

[7] Given this procedure, the initial placement was exogenous to the guest-workers. From the perspective of family members moving to Germany in the framework of the family reunification, the location was also exogenous.

(GSOEP) which was started in 1984 and which provides detailed information on individual and household characteristics. This sample initially comprised 1,393 households with either a Greek, Italian, Spanish, Turkish, or Yugoslavian household head. Due to the limited sample size of the GSOEP we have to use administrative data in order to generate regional concentration measures of guest-workers. Unfortunately, the 1984 wave of the GSOEP does not allow sufficiently detailed regional merging with other data sources: instead we use the 1985 wave comprising 2,346 immigrants with full information from the five most important guest-worker countries.

The main outcome of interest is language knowledge. Like the previous literature we use indicators of self-assessed language fluency and writing ability which are measured on a five-category ordinal Likert-scale ranging from "not speaking at all" (lowest category) to "speaking very well" (highest category).[8] For most of the analysis, we use a binary variable for speaking and writing ability which takes on the value one for the two highest scores on the Likert scale and zero otherwise. As can be seen from Table 1, less than half of the sample claimed to speak or write German at least at a good level in 1985.

>> Table 1 about here <<

Demographic information comprises gender, marital status, country of origin, age at migration, years since migration, years of schooling, a dummy variable for education abroad and a dummy indicating the presence of children in the household. Table 1 further reveals that the average immigrant entered Germany at relatively young age (23 years) and had spent almost 15 years in the country. Educational attainments are rather low (at nine years of schooling) which is consistent with the fact that the vast majority of educational degrees was attained in the home country. The gender mix as well as the common presence of children in immigrant households reflect the migration for family unification, which became dominant after the recruitment stop in 1973.

Given the scope of the guest-worker programme it might be surprising that the German government never collected detailed information on where guest-workers moved and for how long they stayed, leaving us with general data sources. To generate ethnic concentration measures, we use the IAB[9] Beschäftigtenstichprobe of 1975, a two percent sample of all persons with social security insurance in Germany. This employee-sample comprises 2% of the entire employee population plus recipients of certain social transfers like unemployment benefits. Employers mandatorily register employees for the payment of social security taxes, so that employees can be tracked through their social security numbers until dropping out of the labour force.

[8] Objective language measures are to date unavailable in Germany. The federal office for Migration and Integration initiated an "integration panel" which started in 2007 with a focus on the effect of language course participation on language ability. Again, no objective language evaluation was possible due to legal uncertainties and the absence of a coherent test scheme (Rother, 2008).

[9] The research institute affiliated with the Federal Employment Agency of Germany.

Instrumental variable approach

Although the placement of immigrants in Germany was exogenous to them, they were in reality allowed to move after one year of work (including being allowed to return). However, until the economic recession in mid 1970s, guest-workers would move only to follow labour demand, and normally only short distances (i.e., within region). These moves must be understood as steps towards settling down in Germany, after many guest-workers had spent the first time in employer-provided accommodation. The fact that immigrants moved across regions might imply that the propensity to move into ethnically homogeneous regions (enclaves) is correlated with some unobservable characteristics of migrants. For instance, migrants who are less able or willing to learn German could self-select into ethnic clusters in order to reduce the costs adherent to absent language skills. If this was the case we would expect naive OLS estimates of the enclave effect to be biased away from zero.

To overcome this bias, we use an instrumental variable approach where we estimate the following system of equations:

$$\begin{aligned} y &= \alpha + \beta x + \gamma(ysm) + \kappa + \mu + u \\ x &= \gamma + \lambda z + e \end{aligned} \qquad (8)$$

where *z* is the instrument which satisfies the assumptions *Cov(z, x) ≠ 0* and *Cov(z, u) = 0*. The IV estimator

$$\hat{\beta}^{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

can be interpreted as the ratio between the reduced form relationship between *y* and *z* over the first stage relationship of *y* and *x*. The main instrument used in this application is the ethnic composition of regions in 1975, thus ten years before our language ability measures were taken and at the time when the guest-worker programme had just come to an end. At this time, the placement of guest-workers was still predominantly exogenous to them, implying that z is uncorrelated with any unobservable factors that are accumulated in *u*.

We identify the effect of ethnic concentration on language acquisition through the use of the exogenous 1975 ethnic composition as an instrument for ethnic concentrations in 1985, the earliest year for which we have a sample of immigrants with German proficiency and residence county information. The identifying assumption is that until 1975 the guest-workers have not changed their locations according to characteristics that are correlated with the ability or willingness to learn German. We do not need to assume that guest-workers never moved, but that whatever influenced their moving decision (if they moved) was not correlated with unobservable characteristics influencing the learning decision. Given the economic boom until 1973 and the pervasively low levels of education and skills among the guest-workers, we do not expect much

sorting across regions until 1975; since transitions from one employer to another were most likely to happen within rather than across regions, it helps that regions are defined at a fairly aggregate level (discussed in the next section).

Our instrument might contain measurement error, as we instrument the ethnic concentration exposure in 1985 with the regional concentration of 1975, although we cannot observe individual places of residence in 1975. This is potentially problematic if guest-workers have moved within this ten year period, and in consequence ethnic concentrations have become stronger for those who were not likely to learn German in the first place. Using data from the employee-sample of the IAB we investigate whether there are systematic differences in the exposure to ethnic concentration before and after moving across regions.

We base our calculations on all guest-workers who were present in the sample in 1975 and in 1984, and who "moved"; those are migrants who were registered for work (or benefit receipt) in different regions in 1975 and 1984. We base our analysis on the workplace location rather than residence, because of higher non-responses for the latter.[10] We observe that 17% of the guest-workers have moved across regions between 1975 and 1984, as compared to 14% of German nationals.[11] Some but not the entire differential is due to the fact that the immigrants who moved were predominantly younger and male. Fifty two percent of those who moved turned to neighbouring regions.

We also construct a variable DIFF defined on guest-workers who moved between 1975 and 1984, which is the difference in the regional ethnic concentration that a guest-worker experienced between 1975 and 1984. For example, if a Turk lived in Munich in 1975 and in Berlin in 1984, DIFF would be the concentration of Turks in Berlin in 1984 minus the concentration of Turks in Munich in 1984. Thus, the change in concentration after moving cannot be attributed to differential trends in the overall population of different immigrant groups. Figure 4 plots the density distribution of changes in ethnic concentration (DIFF) of 2,523 guest-workers in the IAB sample who moved between 1975 and 1984. The distribution peaks around zero (the mean of DIFF being 0.0036) and has somewhat more mass to the right. For 54% of the movers the concentration of their own ethnic group changed by less than one percentage point and for 70% by less than 1.5 percentage points.

>> Figure 4 about here <<

It could still be the case though, that guest-workers who moved to regions with lower concentrations differ systematically from those who moved to higher concentration locations. We thus regress the variable DIFF on educational attainment, age, and nationality dummies (all as reported in 1984). If systematic sorting was present, we would expect educational attainment and age to correlate (albeit imperfectly) with ability or willingness to learn German. For this test we group

the educational information in four categories, educ1 being education less than high-school (Gymnasium, qualifying for college) without vocational training, educ2 high-school degree or vocational training (but not both), educ3 high-school degree and vocational training, and educ4 college degree.

>> Table 2 about here <<

Table 2 reports these OLS results, with educ4 and the Greek dummy being the omitted categories. If anything, lower educational attainments show some weak correlation with a positive change in ethnic concentration, although none of the dummies is significant at conventional levels. The size of coefficients is very small, and the mean change in concentrations for guest-workers with the lowest education remains below two tenths of a percentage point when compared to workers with college degrees. Age seems to be irrelevant. Turks and Yugoslavs were more likely to move to regions with higher concentrations, but here, too, the magnitudes of the coefficients are modest. When we include interactions between ethnic and educational category dummies (column 2), even the ethnic dummies lose their significance (none of the interactions comes close to significance). In general, the variable DIFF is explained very poorly by the regression, with R^2 not even reaching 0.02. We conclude that sorting of guest-workers along any observable characteristics has been absent or very modest between 1975 and 1985.

To further test the validity of our instrument we also employ a second variable, the regional election result of the Social Democratic Party of Germany (SPD) in the national elections of 1976. This instrument is sufficiently correlated with the 1985 ethnic composition of regions, as guest-workers were predominantly placed in regions with dominant mining and heavy industry sectors, which were traditional strongholds of the SPD. Beyond the link through ethnic composition, the instrument is not correlated with individual language ability, as guest-workers were not entitled to vote in the national election unless having adopted German citizenship. At that time, this was true only for a negligible fraction of guest-workers and language knowledge was no criterion for the admission to German citizenship. Additionally, the political landscape in Germany largely ignored the fact that guest-workers were starting to settle down and that the intended „rotation principle“ of the migration flows (guest-workers should return after a first employment spell) never came into effect. Consequently, none of the political parties broached the issue of integration or language policy at that time.

Choice of regional level of aggregation

Conditional on data availability, ethnic concentrations can be measured at several levels of aggregation. Ideally, the effect of ethnic concentration on language proficiency should be measured within geographic units containing people's daily life context. Generally, there is a qualita-

[10] Reporting residence was not mandatory.

[11] When looking at inter-regional moves, the level of mobility seems low when compared to the USA.

tive trade-off between small units of aggregation that closely reflect the idea of ethnic neighbourhoods (e.g., census tracts in the US ghettoisation literature with an average size of three to five thousand inhabitants; Cutler and Glaeser, 1997; or municipalities in Sweden with a median population size of 16,000 inhabitants; Edin, Frederiksson and Aslund, 2003) and larger units, that circumvent the potential bias from self-selection into neighbourhoods (e.g., through the use of metropolitan level data (CMA); Warman, 2007; Cutler, Glaeser and Vigdor, 2008). The latter approach assumes that the problematic self-selection of individuals into ethnic enclaves mainly takes place within cities rather than across. In Germany, a reasonable level of aggregation contains cities with their economically integrated suburban areas, or counties in rural areas, amounting to average populations of 100,000 to 250,000. Unlike within American cities, the degree of ethnic and social segregation is much lower in Germany (see Musterd (2005)). Census tract units used in the American neighbourhood effects literature can thus be expected to reflect individuals' space of interaction too narrowly for the German case. Our analysis is based on so-called Anpassungsschichten, which are regional units comprising a larger city and the economically linked hinterland. In West Germany including West Berlin, there were 111 Anpassungsschichten in 2001 with an average population size between 135 and 500 thousand inhabitants, respectively. These units are broader than preferred, however, they are the lowest level of aggregation for which the Microcensus remains fully representative for minority nationalities. Since early waves of the Microcensus do not identify the Anpassungsschicht of the household, we use a 2% employee sample (IAB employee sample, see below) to project population shares for the same regional units for the years 1975 and 1985. Thus, the broad regional aggregation has also the advantage of decreasing the degree of measurement error resulting from projecting for the population. We wish to emphasize that this degree of aggregation will deliver more conservative estimates than when based on smaller regional units. Consequently, we provide lower bound estimates of the true effect of ethnic concentration on the language proficiency of immigrants. By including Anpassungsschicht and ethnicity fixed effects, we exploit only variation in ethnic concentrations that is not systematic across ethnicities or across regions. If the chosen level of aggregation effectively reduced the bias from sorting, our OLS estimates should be very close to the true effect of own-ethnic concentration.

5. Results

In the following we provide empirical evidence of a German language penalty from living among members of the same ethnicity which is robust when accounting for the endogeneity of immigrants' post-initial-placement location choice. Figure 5 gives an initial idea of the correlation between ethnic concentration in the location of immigrants (here the log of the normalised frequency) and their average language fluency in German (as a share of immigrants who speak German well or very well). The correlation between the two variables of interest is negative, with the variance across regions being substantial. It becomes evident that larger regions contain higher ethnic concentrations.

>> Figure 5 about here <<

Main results

Table 3 indicates that there is a significantly negative return to language fluency from living in an area with higher own-ethnic concentration. When including control variables, the coefficient becomes more pronounced and is -0.037.[12] Equivalently, if the ethnic concentration increases by one standard deviation, the probability that a person is fluent in German decreases by 2.6 percent. Although the effect of own-ethnic concentration might seem small at first, one has to consider the high level of aggregation it refers to. Other authors have found similar effects at high levels of aggregation for the USA (Chiswick and Miller, 2005) or Canada (Warman, 2007). The table further reports results from specification (2) which comprises an instrumental variable approach. The use of the instrument (in columns 3 and 4) returns a very similar coefficient, modestly further from zero than our OLS estimate.[13]

>> Table 3 about here <<

Table 4 reports results from the same estimations using writing fluency as the dependent variable. Interestingly, the concentration effects are equally precisely measured when compared to Table 3, however, the effects are substantially closer to zero and significantly different thereof only in the IV estimation (columns 3 and 4). In Table 5, we repeat the analysis of speaking fluency and test the robustness of our results by using alternative measures of ethnic concentration. Columns 1 and 4 use the absolute number of own-ethnic minority members, which simply reflects a transformation of our initial results (see also theory section). The remaining columns use alternative measures of ethnic concentrations. The dissimilarity index ranges between zero and one with the corner solution representing the state of perfectly equal distribution across space and the state of perfect concentration of all minority members in one region.[14] The isolation index is a measure ranging between zero and one which reflects the degree of isolation which an average member of an ethnicity faces on top of the equal distribution of this ethnicity across space.[15] As can be seen from Table 5, our results are robust to the use of alternative measures of ethnic concentration or segregation; two-stage-least-squares estimators are consistently more negative than the OLS estimates.

>> Table 4 about here <<

---

[12] The largest part of the effect stems from variation across regions (50 percent). Thirty-nine percent of the effect is due to variation across ethnicities.

[13] We also use a variety of transformations of this instrument (e.g., ranks) yielding qualitatively the same results.

[14] The formula for the Dissimilarity index is $\frac{1}{2}\sum_{j}^{N}\left|\frac{\#ethnicity_{ij}}{\#ethnicity_{i,total}}-\frac{\#non-ethnicity_{ij}}{\#non-ethnicity_{i,total}}\right|$ for ethnicity i in region j.

[15] The formula for the Isolation index is $\frac{\sum_{j}^{N}\frac{\#ethnicity_{ij}}{\#ethnicity_{i,total}}\times\frac{\#ethnicity_{ij}}{population_{ij}}-\frac{\#ethnicity_{i,total}}{population_{i,total}}}{\min\left(1,\frac{\#ethnicity_{i,total}}{population_{smallest}}\right)-\frac{\#ethnicity_{i,total}}{population_{i,total}}}$ for ethnicity in region j.

>> Table 5 about here <<

In Table 6 we add further robustness concerning our dependent variable. So far, we have used a binary indicator for speaking and writing fluency. These variables are, however, generated from ordinal rankings of five answer categories. Columns 1 and 2 report basic results for OLS regressions that use the full information of the language self-assessment. Although the coefficients are hard to interpret, one can infer the robustness of our results from them.[16] Columns 3 to 6 use a transformed binary concentration measure that takes the value one if the ethnic concentration of an ethnicity in a region lies above the 75th percentile of the entire ethnic concentration distribution, and zero otherwise. Due to the loss of information, the precision of the estimation in column 3 is lower compared to the one with continuous concentration measures. As column 4 shows the contact rate with natives might matter more for language acquisition than simply living in own-ethnicity enclaves. Although living with fewer Germans outside enclaves might be beneficial for language acquisition,[17] the absence of native speakers inside enclaves has a strongly negative impact on language knowledge. Columns 5 and 6 report differential effects for older and younger migrants according to their age at migration. The comparison of both columns shows that older immigrants bear most of the negative impact from enclaves while those who immigrated at younger age have no disadvantage from living in an enclave; these results reconfirm findings for other countries (Warman, 2007). The joint coefficient of a young immigrant in an enclave is significantly positive 0.092 (s.e. 0.049).

>> Table 6 about here <<

Table 7 shows further instrumental variable estimation results. Given the relatively small sample size, we prefer the use of only one instrument. However, we have a second instrument at hand with which to test for over identification of the equation. Using only the second instrument—the election outcome for the Social Democratic Party of Germany (SPD) in 1976—the 2SLS estimator becomes even more negative. Employing both instruments at once we produce an over identified model: the estimated coefficient moves very close to our initial result and the Hanson test statistics confirms that our instruments satisfy the orthogonality condition. Even when introducing a number of interactions (column 4) we cannot clearly reject the null hypothesis that the instruments are invalid. Columns 5 and 6 produce the reduced form results for both instruments.

>> Table 7 about here <<

>> Table 8 about here <<

[16] The results are also robust to the use of ordered probit estimation.

[17] Generally, immigrants tend to have more social contacts with other immigrants irrespective of ethnicity. As a result, German might be the language of communication among immigrants from diverse ethnic backgrounds.

Table 8 adds evidence from a non-parametric perspective. We perform nearest neighbour propensity score matching to generate the closest counterfactuals of our observations artificially.[18] For two different binary treatment variables, the first line reports the result without matching. The remaining rows are different versions of the matching estimator employing different numbers of nearest neighbours. As evidenced in the table, applying the matching estimator increases the language ability of the control group, i.e. in the unmatched sample we underestimate the language ability of those residing outside enclaves. Also, the average treatment effects are clearly significant, lending further robustness to our earlier results.

6. Measurement Error

The model estimated in (7) has several sources of potential measurement error which will be discussed in this section. More specifically, we wish to explain the fact that 2SLS estimates are more negative than standard OLS results.

Models using language ability as an explanatory variable (e.g., in wage regressions) have discussed the measurement error inherent to self-assessed language knowledge (Dustmann and van Soest, 2001; Bleakley and Chin, 2004). Survey respondents might generally misjudge their language ability, and the deviation of self-assessed from "objective" fluency might be correlated with level of education (i.e., better educated might have a better idea of their true language ability) and level of language ability (i.e., those in the upper part of the fluency distribution have less room for over-estimating their ability with the reverse being true for the other extreme of the fluency distribution).[19] In our application, language fluency is, however, the dependent variable and measurement error herein reduces precision while it does not introduce any bias into the estimates.

More serious than in the dependent variable is measurement error in independent variables as it may bias the estimated coefficients. As such, this type of error might potentially drive OLS estimates closer to zero and explain our finding of more negative 2SLS estimates. Our ethnic concentration measures are computed for five ethnicities (Greek, Italian, Spanish, Turkish, and Yugoslav) from the IAB Employee sample 1985 which comprises two percent of all individuals with social security insurance in Germany. It seems reasonable to assume that these densities suffer from measurement error, especially in regions which comprise a generally low share of foreign population or few individuals of one single ethnicity. In support of these measures, it should be noted that social security insurance was compulsory in Germany at that time (and still

[18] The matching estimators are well-fitted with full support.

[19] In our sample, there is a strong central tendency in the five category Likert scale with only 15 percent of respondents claiming to have no (category 1) or very good (category 5) language ability.

is) and that unemployed individuals are also included in the sample. Further, due to the demand-driven nature of the guest-worker programme, ethnic minorities were more equally distributed across German regions than one would expect under more labour supply driven arrangements. As such, the extent of measurement error is probably not correlated with characteristics of the region other than size and thus should be of little concern in our estimation.

Attenuation bias towards zero could be shown in our data when the instrument has better measurement properties than the original density measures. In our case, this seems rather unlikely, as the instrument (ethnic concentration measured from IAB sample 1975) comes from the same data source as our original variable that is potentially plagued by measurement error. In order to show that this type of error is of less importance here, we use another instrument that does not suffer from the problem: The regional election outcomes of the Social Democratic Party (SPD) of Germany in the year 1976.

>> Table 9 about here <<

As both sources of measurement error seem not responsible for the observed outcome, we turn to a last potential solution. Given that the 2SLS estimates exploit only variation in ethnic concentration across space and ethnicities that was present in 1975, while OLS estimates rely on the respective variation for 1985, a change in this variation might result in different outcomes. In other words, if the fluency penalty from ethnic concentration differs in the 1975 sample from the 1985 population, OLS and 2SLS will differ for qualitative reasons (Angrist and Imbens, 1995). If immigrants moved across regions between 1975 and 1985 in a non-random fashion, our instrumental variable approach will only estimate a Local Average Treatment Effect (LATE) for those who did not intend to move. A useful check whether "stayers" suffer stronger from ethnic concentration can be performed by exploiting information on the year when the immigrant household moved to the current place of residence. The full retrospective information is only available in the first wave of the GSOEP (1984) and that is why we lose some observations. As Table 9 suggests, migrants who live longer at their current place of residence have much more negative coefficients on the ethnic concentration measures. It should be noted, that although this piece of evidence indicates that the language penalty differs with the propensity to have moved, it cannot answer the question whether people moved voluntarily (i.e., sorting) and whether they moved across regional units. However, if we did not account for sorting of immigrants who were less willing or able (omitted variable x_2) to learn German into ethnic enclaves (x_1), then the ethnic concentration measure will overestimate the true penalty on learning German since x_2 is expected to be negative for language fluency and $Corr(x_1, x_2) > 0$. When estimating a reduced form regression with the instrument rather than the ethnic concentration measure of 1985, we indeed

find a slightly lower coefficient of -0.030 instead of -0.037. While being present in our data, sorting over time accounts for less than 20 percent of the coefficient.

7. Policy Simulation

The regressions reported in Section 5 and 6 aimed at identifying and quantifying the effect of ethnic concentrations on language proficiency. As we have discussed in the theoretical part of the paper, the effect of a rise in ethnic concentration on an immobile immigrant's propensity to learn German (the case in the previous sections) will typically be different from the impact if the immigrant can move across regions. From a policy perspective the latter case is of more relevance. Furthermore, we showed that changes in the concentration impact stronger on the location conditioned probability of learning German as compared to the unconditional probability, at least for low x_f.

It is thus interesting to study how learning and location choices would have behaved under different scenarios of ethnic concentrations and individual characteristics. To this avail, we estimate the model outlined in Section 2 as a multinomial choice model and use the estimated parameters to perform some model simulations on counterfactual distributions of immigrants across regions and educational attainments.

We conduct this exercise on a more recent sample of immigrants (the 2001 wave of the GSOEP). Above we had used the 1985 wave of the panel to reduce as much measurement error in our instrument as possible. The experiments, however, are more relevant for recent data, because we now observe immigrants whose decisions to live in Germany have become permanent and who arguably had the chance to settle in a region of their own choice.

Multinomial Choice Model

Recall that a choice alternative is given by a pair of learning and location decisions. If the ω_i^j in equation (5) are distributed type I extreme value, the choice probabilities are the ones given in equation (6), resulting in the well-known multinomial logit model. The properties of this model are discussed at length in McFadden (1974). The model is consistent with a globally concave likelihood function. Importantly, consistency is preserved when the estimation is performed on a subset of choice alternatives, a pivotal property when the choice-set for optimizing agents is very large. In our case, the choice-set consists of all possible learning-location decisions (with approximately 90 observed locations there are 180 alternatives). For the analysis we have conducted estimations on all chosen plus four additional randomly selected alternative locations (without replacement), amounting to ten distinct (S, j) choice pairs. We have also performed estimation on two, three, and four locations to test the robustness of our estimates in depend-

ence of the choice size, with only negligible differences in the results. The preservation of consistency is guaranteed by the Irrelevance of Independent Alternatives feature of the multinomial logit model. At the same time, this is an important limitation of the model, since we would expect that "shocks" to the same learning decisions (in different locations) are correlated. In other words, if an immigrant is likely to learn German in location *j*, we would expect him to be likely to learn German in all other locations, too. The multinomial logit does not allow for such a correlation structure.

To relax this restriction we also estimate a multinomial probit model on the full choice set where the unobservable shocks may be be correlated for the same learning and location decisions. The respective covariance structure yields two desirable features: First, an immigrant receiving a high shock to learning German in one location is also likely to receive the same shock in other locations (the immigrant being of a specific learner-type). Second, an immigrant who has a high shock to "not learning in location *j*" is likely to have a high shock for "learning in location *j*" (with reasons drawing the immigrant to the location regardless of ethnic concentrations and learning costs). Estimation is slightly more complicated than in the multinomial logit case, since the choice probabilities in equation (5) do no longer have a closed form solution. Instead, the probabilities are approximated via simulation. For details of the estimation algorithm see the Appendix.

Simulation Results

Since estimation of the full multinomial probit model is very time-intensive, we decided to use a parsimonious specification of the choice model. The *X* in equation (1) thus include a constant, age at migration, years since migration, years of education, and a dummy for having obtained the highest educational degree in the country of origin, all of which were significant predictors of language proficiency in an OLS framework. The *W* in equation (3) contains two variables: total regional population which is precisely projected from the German Microcensus of the year 2000 and normalised to one for the least populous region, and the regional unemployment rate which is aggregated over county data from the German Federal Employment Services. Finally, the payoff function in equation (3), ξ, is specified as a quadratic function in its argument, not including a constant. As is standard, the variance of the unobservable part of utility is normalised to $\frac{\pi^2}{6}$ in the multinomial logit case. For the probit model we normalise the variance of ε to one, and implicitly estimate the variance of *r*. Thus, we estimate nine parameters for the logit, and ten parameters for the probit case. Our preferred model is the probit, since it allows for a richer (and more realistic) covariance structure, but for comparison and because of the computational burden of finding standard errors in the probit model we report results of both specifications.

Estimates are reported in Table 10. Both the multinomial logit and probit yield identical signs on all coefficients. The payoff function ξ is concave and attains its maximum at an ethnic concentration of approximately 75% in both cases. A higher unemployment rate reduces the probability of living in the corresponding region, whereas a larger total population size increases it. Years since migration and years of schooling reduce the cost of language learning, while a higher age at migration and a foreign educational degree increase the cost. We do not report standard errors for our probit model, since we have no closed form solution for the derivative of the likelihood-function, and bootstrap-methods would be too computation-intensive. However, standard errors from the logit model should provide some guidance for the relative importance of the variables. First, the effects of own ethnic concentration is estimated with considerable precision. Second, the coefficients on regional characteristics have low standard errors, too. Third, the coefficient on years of schooling is most precisely estimated among the learning-cost variables; its value suggests a prominent role of education in determining the cost of learning German. Both models suggest that one additional year of schooling reduces learning costs in a magnitude comparable to 10 to 12 additional years of residence in Germany.

>> Table 10 about here <<

Figure 6 compares the concentration effects on the probability to learn German for the multinomial choice results (as given in the first part of proposition 1, that is, conditional on location) and an OLS regression of language proficiency (as given in our benchmark regressions reported earlier). It should be noted that the OLS coefficient from the 2001 sample (-0.045) is higher in absolute terms than the coefficient obtained from the 1985 wave (-0.037). This is consistent with sorting of high-learning cost immigrants into regions with high ethnic concentrations between 1985 and 2001. As expected, the treatment effect over a wide range of concentrations is smaller in both choice models. For our preferred probit specification, the derivative of $P(S=1|\ j)$ at an own-ethnicity concentration of 3% is -3.9. Furthermore, the probit model exhibits a smaller treatment effect at all concentrations, whereas the logit model has higher marginal effects of concentration on German proficiency for lower concentrations, as can be seen by the steeper slope of the logit curve at low concentrations.

>> Figure 6 about here <<

Counterfactuals

Which level of language proficiency would prevail, if Germany had been able to place immigrants in specific regions to equalise their distribution across German regions? Or, if Germany had been screening guest-worker applicants by their level of education? To answer these questions we simulate four different scenarios with the help of our probit estimates: The first is the

"real" world. We simulate the learning and location model 500 times and compare four simulated moments to the actual data. The second scenario is an equal distribution of all ethnic groups over the country; that is replacing actual ethnic concentrations in the regions by the West-German average (as a placement policy might have done). In the third case we increase each immigrant's education by one year of schooling. Finally, we simulate a one percentage point increase of own-ethnic concentrations in each region.

The main outcome of interest is the fraction of first-generation immigrants deciding to learn German across the different scenarios. The other three moments are plausibility checks of our model: we report the fraction of our sample deciding to live in the region with the largest population, which is Berlin. Berlin is an "outlier" among all regions, with its population at least doubling the population of 86 out of the 87 other regions. Consequently, the capital is chosen most often in our benchmark simulations and the fraction of immigrants deciding to live there indicates the degree of clustering. A further indicator for clustering is the number of regions chosen by at least one observation of our sample, with 88 being the maximum. Finally, we also look at the fraction of immigrants in Berlin who decide to learn German.

>> Table 11 about here <<

Results of our experiments are reported in Table 11 together with the real data moments. The reported numbers are averaged over 500 simulations, with standard deviations in parentheses. For example, 46.6% of the sample decide to learn German in our model simulation (column 1), coinciding almost exactly with the true fraction of German-speakers. 7.4% decide to live in Berlin, out of which 44.1% learn German. The actual share of immigrants residing in Berlin in all immigrants in West Germany is 5.3%, so that our model slightly overestimates the attraction of this region. All or almost all regions are chosen by at least one immigrant in every simulation.

When moving from those results to an equal distribution of immigrants throughout Germany, the effect on language proficiency is positive, but small. In other words, immigrants who are treated with a lower immigrant share than in the benchmark scenario are those, who had high learning costs to begin with. Even though the incentive has increased after equalizing the concentrations, only few of them are induced to learn German.

To the opposite, the increased education scenario leads to a considerable improvement in German proficiency. The fraction of German learners increases by 6 percentage points compared to the benchmark, without effecting the distribution of immigrants across Germany much. Regional factors (population and unemployment) largely determine the preference over regions. Berlin and other populous regions are chosen most often, and within those regions the share of speakers increases. Of those immigrants choosing to live in Berlin, 49.7% learn German now as compared to the benchmark share of 44.1%.

Finally, an increase in own-ethnic concentrations by one percentage point in all regions leads to a decrease in language proficiency by 3.8 percentage points, which is just about the change in the probability of learning German conditional on location. Given the near-linearity of the concentration effect (see Figure 6) this is not surprising: For the learning decision a common increase of concentrations across regions should yield the same effect as an equivalent increase in the actual location while being locked in. A closer look at the choices of immigrants under this scenario reveals that clustering in higher-concentration areas becomes now more pronounced. Berlin is preferred only second most often despite its large population. More immigrants decide to move to a region around the city of Stuttgart[20]. This region comprises a population of 2 million inhabitants and is characterized by low unemployment and above-Berlin concentration levels for all ethnic groups except for Turks. The concentrations of the Turkish population in Berlin and the region around Stuttgart are about the same. Some immigrants who found it optimal to learn German are now induced not to learn, and thus the relative importance of the ethnic concentration in the settlement choice increases, making high concentration areas more of a drawing magnet.

8. Conclusion

This paper is the first attempt to investigate the effect of own-ethnic regional concentration on the language ability of labour immigrants theoretically and empirically. Using the example of the guest-workers who were paired with German firms exogenously, we find small but negative causal effects from living among immigrants from the same country of origin in the mid 1980s in Germany. The effect becomes slightly larger after addressing potential sorting into enclaves with an instrumental variable approach. We discuss several sources of measurement error and conclude that the instrument produces a local average treatment effect (LATE). Given the level of aggregation used in the analysis, our ethnic concentration effects are lower bound estimates. Research on more disaggregated ethnic enclaves might be desirable in order to better reflect immigrants' daily life context; however, the lack of highly disaggregated data in Germany prevents more profound investigations.

The paper provides a simple random utility model which allows for simultaneous learning and location choices. Using estimated parameters to simulate the effect of own-ethnic concentration on language ability suggests a lower impact than estimated by a simple OLS strategy. Applying the model on more recent data from Germany, we find an increased tendency for immigrants to

[20] The region consists of the counties Böblingen, Esslingen, Göppingen, Ludwigsburg, and Rems-Murr-Kreis.

sort into regions with co-nationals. Generally, an additional year of education increases the propensity that immigrants learn German much stronger than would a placement policy that produces equal ethnic distributions of immigrants across Germany. Despite finding a negative effect from own-ethnic concentration on language ability, we conclude that public policy might achieve better integration outcomes by targeting education levels rather than location choices.

References

Angrist, J.D., Imbens, G.W., 1995. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. Journal of the American Statistical Association 90(43), 431-442.

Bade, K. J., 2009. Integration ist machbar. In: Die Welt, Nov 7th 2009.

Bauer, T.K., Dietz, B., Zimmermann, K.F., Zwintz, E., 2005. German Migration: Development, Assimilation, and Labour Market Effects. In: Zimmermann, K.F. (Ed.), European Migration, What Do We Know? Oxford University Press, Oxford, pp. 197-261.

Bauer, T.K., Epstein, G.S., Gang, I.N., 2005. Enclaves, Language, and the Location Choice of Migrants. Journal of Population Economics 18(4), 649-662.

Bleakley, H., Chin, A., 2004. Language Skills and Earnings: Evidence from Childhood Immigrants. Review of Economics and Statistics 86(2), 481-496.

Borjas, G.J., 1995. Ethnicity, Neighbourhoods, and Human-Capital Externalities. American Economic Review 85(3), 365-390.

Borjas, G.J., 1998. To Ghetto or not to Ghetto: Ethnicity and Residential Segregation. Journal of Urban Economics 44(2), 228-253.

Chiswick, B.R., Miller, P.W., 1995. The Endogeneity between Language and Earnings: International Analyses. Journal of Labor Economics 13(2), 246-288.

Chiswick, B.R., Miller, P.W., 1996. Ethnic Networks and Language Proficiency among Immigrants. Journal of Population Economics 9(1), 19-35.

Chiswick, B.R., Miller, P.W., 2002. Immigrant Earnings: Language Skills, Linguistic Concentrations and the Business Cycle. Journal of Population Economics 15(1), 31-57.

Chiswick, B.R., Miller, P.W., 2005. Do Enclaves Matter in Immigrant Adjustment? City & Community, 4(1), 5-36.

Cutler, D.M., Glaeser, E.L., 1997. Are Ghettos Good or Bad? Quarterly Journal of Economics 112(3), 827-872.

Cutler, D.M., Glaeser, E.L., Vigdor, J.L., 2008. When are Ghettos Bad? Lessons from Immigrant Segregation in the United States. Journal of Urban Economics 63(3), 759-774.

Damm, A.P., 2009. Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence. Journal of Labor Economics 27(2), 281-314.

Drever, A.I., 2004. Separate Spaces, Separate Outcomes? Neighbourhood Impacts on Minorities in Germany. Urban Studies 41(8), 1423-1439.

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

Dustmann, C., van Soest, A., 2001. Language Fluency and Earnings: Estimation with Misclassified Language Indicators. Review of Economics and Statistics 83(4), 663-674.

Dustmann, C., van Soest, A., 2002. Language and the Earnings of Immigrants. Industrial and Labor Relations Review 55(3), 473-492.

Dustmann, C., Fabbri, F., 2003. Language Proficiency and Labour Market Performance of Immigrants in the UK. Economic Journal 113(489), 695-717.

Edin, P.A., Fredriksson, P., Aslund, O., 2003. Ethnic Enclaves and the Economic Success of Immigrants - Evidence from a Natural Experiment. Quarterly Journal of Economics 118(1), 329-357.

Herbert, U., 2001. Geschichte der Ausländerpolitik in Deutschland – Saisonarbeiter, Zwangsarbeiter, Gastarbeiter, Flüchtlinge. C.H.Beck, Munich.

Lazear, E.P., 1999. Culture and Language. Journal of Political Economy 107(6), S95-S126.

McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105-142.

Musterd, S., 2005. Social and Ethnic Segregation in Europe: Levels, Causes, and Effects. Journal of Urban Affairs 27(3), 331-348.

Rother, N., 2008. Das Integrationspanel. Ergebnisse zur Integration von Teilnehmern zu Beginn ihres Integrationskurses. Bundesamt für Migration und Flüchtlinge, Working paper 19, Nürnberg.

Train, K.E., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, New York.

Warman, C., 2007. Ethnic Enclaves and Immigrant Earnings Growth. Canadian Journal of Economics 40(2), 401-422.

Figures and Tables

Figure 1: Earnings and regional ethnic concentration *(x)*

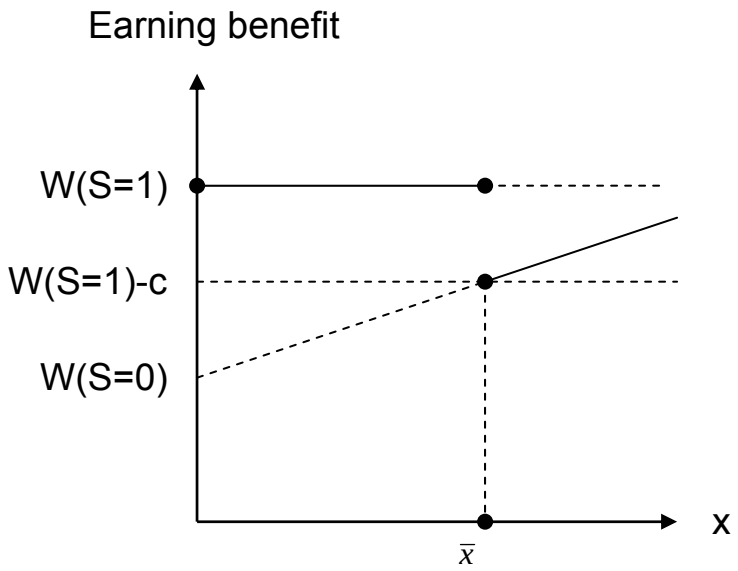
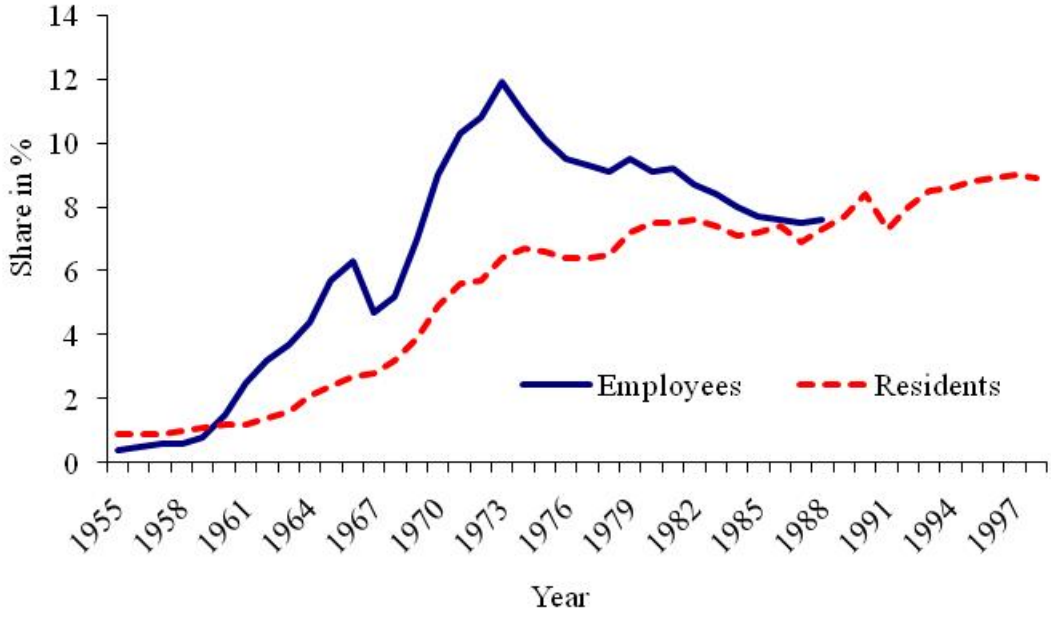
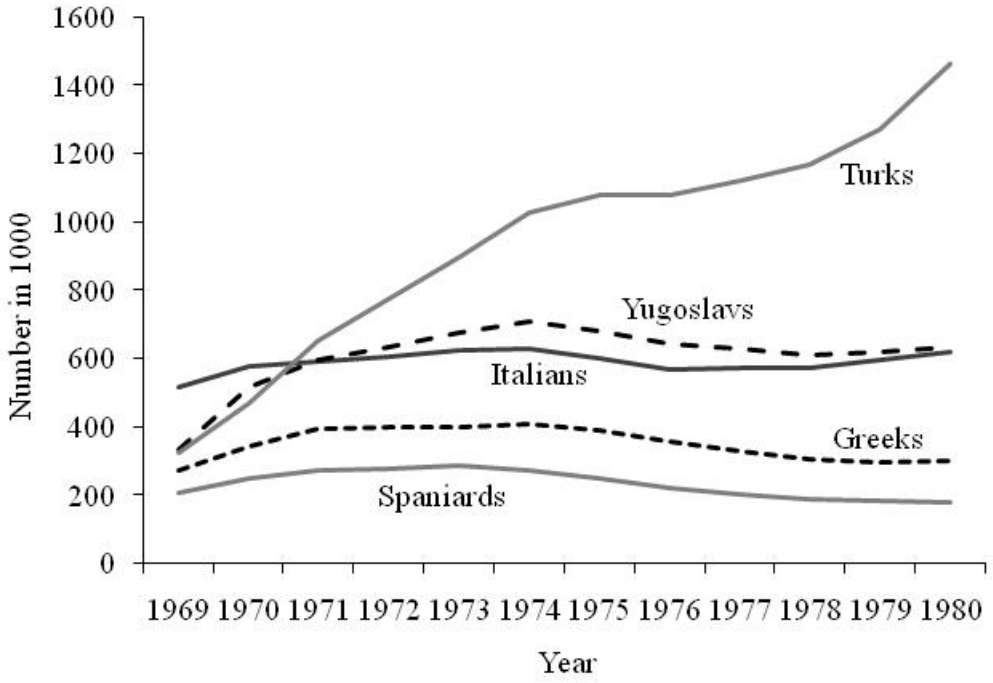


Figure 2: Share of foreign population in Germany



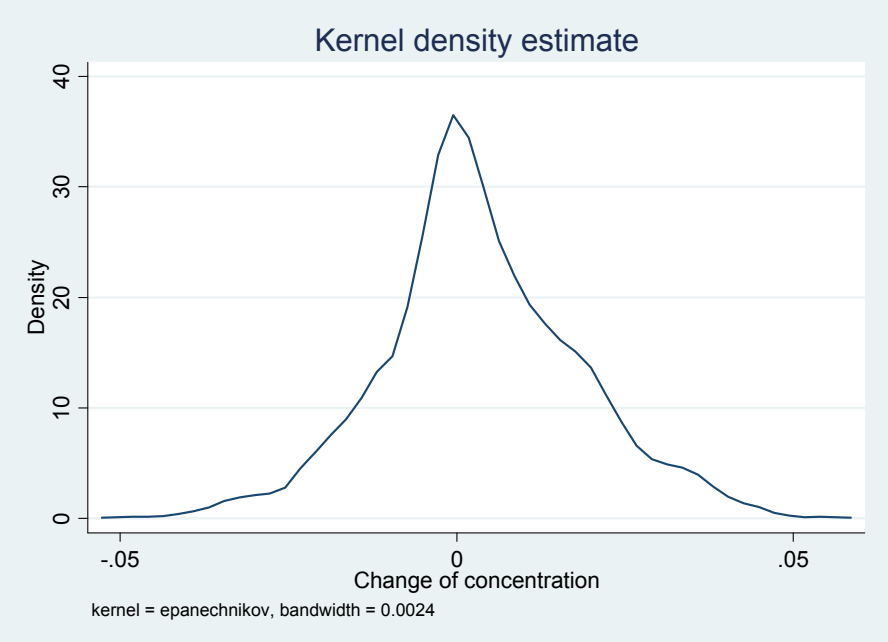
Source: Herbert (2001), pp. 198-199, and Bauer, Dietz et al. (2005).

Figure 3: Absolute number of foreign population by source country



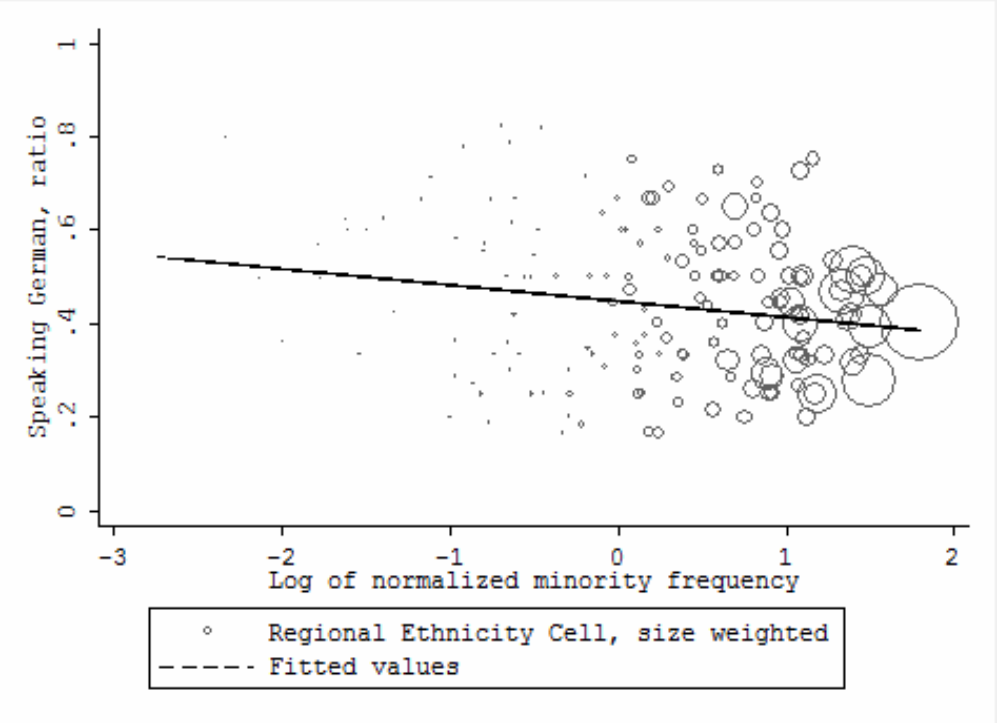
Source: Herbert (2001), pp. 198-199, and Bauer, Dietz et al. (2005).

Figure 4: Kernel density estimate of ethnic concentration change over time (DIFF)



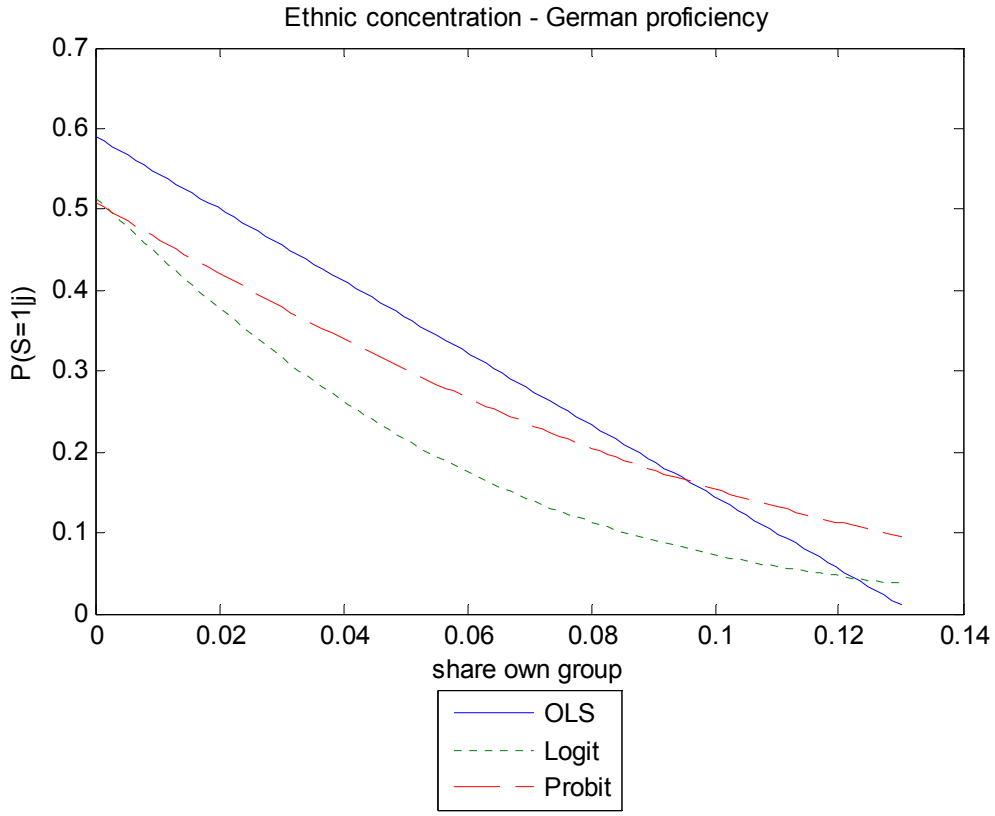
Source: IAB employee sample 1975 and 1984; authors' calculations.

Figure 5: Correlation between own ethnic concentration and average speaking ability



Source: GSOEP 1985, IAB 1985; authors' calculations.

Figure 6: Estimated learning probabilities across estimators



Source: GSOEP 1985, IAB 1985; authors' calculations.

Table 1: Descriptive statistics

	Number of observations	Mean
Speak German	2346	42.3%
Write German	2339	45.0%
Male	2346	55.6%
Age at migration	2346	23.39
Years since migration	2346	14.62
Years of schooling	2346	9.08
Schooling abroad	2346	83.6%
Married	2346	78.8%
Children in household	2346	64.4%
Turkish	2346	34.3%
Yugoslav	2346	19.0%
Italian	2346	19.6%
Spanish	2346	12.3%
Greek	2346	14.8%

Source: GSOEP 1985; authors' calculations.

Table 2: Determinants of DIFF

	(1) OLS	(2) OLS
Educ1	0.0013	0.0027
	(0.0008)	(0.0058)
Educ2	-0.0001	0.0012
	(0.0009)	(0.0049)
Educ3	0.0016	0.0094
	(0.0037)	(0.0152)
Age	-2.04E-06	2.10E-06
	(3.85E-05)	(4.00E-05)
Turkish	0.0029**	0.0051
	(0.0012)	(0.0057)
Italian	-0.0002	-0.0013
	(0.0014)	(0.0019)
Yugoslav	0.0037***	0.0032
	(0.0012)	(0.0057)
Spanish	-0.0008	-0.0001
	(0.0019)	(0.0038)
Interactions	no	yes
Constant	0.0005	-0.0001
	(0.002)	(0.0057)
Observations	2523	2523
R-squared	0.012	0.016

Note: Dependent variable: DIFF = difference in ethnic concentration of individual specific region of residence between 1975 and 1985. Omitted categories: educ4 and Greek nationals. *** p<0.01, ** p<0.05, * p<0.1; robust standard errors in parentheses. Source: IAB 1975/1985; authors' calculations.

Table 3: Determinants of speaking ability

	(1) OLS	(2) OLS	(3) First stage	(4) 2SLS
Frequency of own ethnicity	-0.025***	-0.037***		-0.042***
	(0.009)	(0.014)		(0.015)
Male		0.099***	0.001	0.099***
		(0.018)	(0.011)	(0.017)
Age at migration		-0.035***	0.005**	-0.035***
		(0.003)	(0.003)	(0.003)
Age at migration squ.		0.036***	-0.009*	0.036***
		(0.006)	(0.005)	(0.006)
Years since migration		0.009***	-0.001	0.009***
		(0.002)	(0.001)	(0.002)
Years of schooling		0.052***	-0.002	0.052***
		(0.005)	(0.003)	(0.005)
Schooling abroad		-0.040*	-0.042***	-0.040*
		(0.023)	(0.016)	(0.023)
Married		-0.119***	-0.027	-0.119***
		(0.028)	(0.018)	(0.027)
Children in household		0.013	-0.005	0.012
		(0.020)	(0.013)	(0.019)
1975 Frequency of own ethnic.			0.718***	
			(0.009)	
Constant	0.460***	0.200**	0.459***	0.226***
	(0.056)	(0.080)	(0.054)	(0.081)
Observations	2346	2346	2346	2346
R-squared	0.258	0.360	0.968	0.360

Note: Dependent variable: Binary variable of speaking ability (Speaking very good or good = 1, speaking fair, poor or not at all = 0). Regressions control for ethnicity fixed effects and fixed effects for 85 regional *Anpassungsschichten* in West Germany. *** p<0.01, ** p<0.05, * p<0.1; robust standard errors in parentheses. Source: GSOEP 1985 and IAB 1975/1985; authors' calculations.

Table 4: Determinants of writing ability

	(1) OLS	(2) OLS	(3) First stage	(4) 2SLS
Frequency of own ethnicity	-0.007	-0.021		-0.036**
	(0.009)	(0.014)		(0.015)
Male		0.093***	0.001	0.093***
		(0.018)	(0.011)	(0.017)
Age at migration		-0.035***	0.005**	-0.035***
		(0.003)	(0.003)	(0.003)
Age at migration squ.		0.036***	-0.009*	0.036***
		(0.006)	(0.005)	(0.006)
Years since migration		0.010***	-0.001	0.010***
		(0.002)	(0.001)	(0.002)
Years of schooling		0.066***	-0.002	0.066***
		(0.005)	(0.003)	(0.005)
Schooling abroad		-0.046**	-0.042***	-0.046**
		(0.023)	(0.016)	(0.022)
Married		-0.122***	-0.027	-0.122***
		(0.027)	(0.018)	(0.026)
Children in household		0.044**	-0.005	0.043**
		(0.020)	(0.013)	(0.020)
1975 Frequency of own ethnic.			0.718***	
			(0.009)	
Constant	0.962**	0.325	0.459***	0.327
	(0.415)	(0.356)	(0.054)	(0.349)
Observations	2339	2339	2339	2339
R-squared	0.257	0.378	0.968	0.377

Note: Dependent variable: Binary variable of writing ability (Writing very good or good = 1, writing fair, poor or not at all = 0). Regressions control for ethnicity fixed effects and fixed effects for 85 regional Anpassungsschichten in West Germany. *** p<0.01, ** p<0.05, * p<0.1; robust standard errors in parentheses. Source: GSOEP 1985 and IAB 1975/1985; authors' calculations.

Table 5: Determinants of speaking ability; robustness check with alternative enclave measures

	(1) OLS	(2) OLS	(3) OLS	(4) 2SLS	(5) 2SLS	(6) 2SLS
Number of members of ethnicity	-0.036***			-0.053***		
	(0.012)			(0.016)		
Dissimilarity index		-1.471***			-2.776**	
		(0.498)			(1.137)	
Log of Isolation index			-0.015*			-0.085**
			(0.008)			(0.035)
Male	0.099***	0.100***	0.099***	0.100***	0.098***	0.096***
	(0.018)	(0.018)	(0.018)	(0.017)	(0.018)	(0.019)
Age at migration	-0.035***	-0.036***	-0.035***	-0.035***	-0.036***	-0.036***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Age at migration squ.	0.036***	0.036***	0.036***	0.036***	0.037***	0.037***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Years since migration	0.009***	0.009***	0.009***	0.009***	0.010***	0.008***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Years of schooling	0.052***	0.049***	0.050***	0.051***	0.049***	0.052***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Schooling abroad	-0.040*	-0.035	-0.036	-0.040*	-0.033	-0.042*
	(0.023)	(0.024)	(0.024)	(0.023)	(0.024)	(0.025)
Married	-0.121***	-0.120***	-0.121***	-0.121***	-0.119***	-0.117***
	(0.028)	(0.028)	(0.028)	(0.027)	(0.028)	(0.029)
Children in household	0.012	-0.002	-0.002	0.009	-0.002	0.012
	(0.020)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)
Constant	0.743***	-0.989*	-0.948*	0.846***	-1.147**	-2.058***
	(0.112)	(0.573)	(0.576)	(0.137)	(0.585)	(0.786)
Observations	2346	2282	2278	2346	2282	2278
Instrument	no	no	no	yes	yes	yes
R-squared	0.361	0.309	0.307	0.360	0.306	0.281

Note: Dependent variable: Binary variable of speaking ability (Speaking very good or good = 1, speaking fair, poor or not at all = 0). Instrument: 1975 Frequency of own ethnicity. All regressions control for ethnicity fixed effects. Regressions (1), (4) and (5) control for fixed effects for 85 regional Anpassungsschichten in West Germany, all other regressions control for regional unemployment rate and log of regional income level. *** p<0.01, ** p<0.05, * p<0.1; robust standard errors in parentheses.
Source: GSOEP 1985 and IAB 1975/1985; authors' calculations.

Table 6: Determinants of speaking ability; robustness checks and extensions

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	OLS	OLS	OLS
	Five-scale ordinal variable		Binary variable: Speaking			
Frequency of own ethnicity	-0.078***	-0.064**				
	(0.019)	(0.027)				
Enclave (Own ethnic concentration above p75)			-0.040*	0.002	-0.060**	0.032
			(0.022)	(0.031)	(0.026)	(0.038)
Frequency of Germans below p75				0.086***		
				(0.028)		
Enclave*Freq. of Germans below p75				-0.131***		
				(0.044)		
Immigrated at young age (16 or below)					0.066	
					(0.043)	
Enclave*Immigrated at young age (16 or below)					0.085**	
					(0.044)	
Immigrated as adult						-0.060
						(0.042)
Enclave*Immigrated as adult						-0.099**
						(0.043)

Male		0.214***	0.101***	0.106***	0.103***	0.102***
		(0.033)	(0.018)	(0.018)	(0.018)	(0.018)
Age at migration		-0.058***	-0.035***	-0.035***	-0.028***	-0.028***
		(0.007)	(0.003)	(0.003)	(0.004)	(0.004)
Age at migration squ.		0.031**	0.036***	0.036***	0.026***	0.027***
		(0.013)	(0.006)	(0.006)	(0.007)	(0.007)
Years since migration		0.031***	0.009***	0.008***	0.009***	0.009***
		(0.004)	(0.002)	(0.002)	(0.002)	(0.002)
Years of schooling		0.118***	0.049***	0.049***	0.050***	0.051***
		(0.009)	(0.005)	(0.005)	(0.005)	(0.005)
Schooling abroad		-0.046	-0.040*	-0.043*	-0.055**	-0.057**
		(0.048)	(0.024)	(0.023)	(0.023)	(0.023)
Married		-0.309***	-0.120***	-0.120***	-0.109***	-0.111***
		(0.052)	(0.028)	(0.028)	(0.028)	(0.028)
Children in household		-0.010	-0.004	0.000	0.003	0.004
		(0.038)	(0.020)	(0.019)	(0.020)	(0.020)
Constant	4.086***	3.063***	-1.031*	-0.984*	-1.141**	-1.049*
	(0.132)	(0.156)	(0.571)	(0.562)	(0.565)	(0.559)
Observations	2346	2346	2282	2340	2340	2340
R-squared	0.324	0.454	0.308	0.316	0.315	0.316

Note: Dependent variable: Binary variable of speaking ability (Speaking very good or good = 1, speaking fair, poor or not at all = 0). Regressions control for ethnicity fixed effects and fixed effects for 85 regional Anpassungsschichten in West Germany. *** p<0.01, ** p<0.05, * p<0.1; robust standard errors in parentheses. Source: GSOEP 1985 and IAB 1985; authors' calculations.

Table 7: Results with alternative and multiple instruments

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) 2SLS	(5) Reduced form	(6) Reduced form
Frequency of own ethnicity (instrumented)	-0.039***	-0.068**	-0.040***	-0.034**		
	(0.009)	(0.030)	(0.012)	(0.014)		
1975 Frequency of own ethnicity; (1)					-0.028***	
					(0.011)	
SPD election result; (2)						-0.025**
						(0.011)
Male	0.099***	0.100***	0.099***	0.099***	0.099***	0.099***
	(0.016)	(0.016)	(0.016)	(0.017)	(0.018)	(0.018)
Age at migration	-0.035***	-0.035***	-0.035***	-0.035***	-0.036***	-0.035***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
Age at migration squ.	0.036***	0.036***	0.036***	0.036***	0.036***	0.036***
	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)
Years since migration	0.009***	0.009***	0.009***	0.009***	0.009***	0.009***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
Years of schooling	0.051***	0.051***	0.051***	0.052***	0.052***	0.052***
	(0.011)	(0.010)	(0.011)	(0.005)	(0.005)	(0.005)
Schooling abroad	-0.039***	-0.040*	-0.039***	-0.039*	-0.037	-0.038*
	(0.015)	(0.020)	(0.015)	(0.023)	(0.023)	(0.023)
Married	-0.119***	-0.120***	-0.119***	-0.119***	-0.118***	-0.118***

	(0.015)	(0.015)	(0.015)	(0.027)	(0.028)	(0.028)
Children in household	0.011*	0.009	0.011*	0.011	0.011	0.014
	(0.006)	(0.018)	(0.006)	(0.019)	(0.020)	(0.020)
Constant	-1.949***	-1.852**	-1.947***	-1.967**	-1.319	-0.679
	(0.514)	(0.800)	(0.514)	(1.000)	(1.091)	(1.042)
Instruments	(1)	(2)	rank(1), rank(2)	(2) and interaction rank(2)* pop size		
Number of instruments	1	1	2	60		
Hanson overidentification test, p-value	—	—	0.959	0.163		
Observations	2340	2340	2340	2340	2340	2340
R-squared	0.364	0.362	0.364	0.364	0.364	0.362

Note: Dependent variable: Binary variable of speaking ability (Speaking very good or good = 1, speaking fair, poor or not at all = 0). Instrument (1) is the own-ethnic concentration of the year 1975. Instrument (2) is the election outcome of the Social Democratic Party of Germany (SPD) in the 1976 national elections. Note that there are six missing observations for instrument (2). Therefore all regressions are performed on a slightly smaller sample than in Table 3, which explains the small differences in estimates. Regressions control for ethnicity fixed effects and fixed effects for 85 regional *Anpassungsschichten* in West Germany. *** p<0.01, ** p<0.05, * p<0.1; robust standard errors in parentheses. Source: GSOEP 1985, IAB 1975/1985 and official results of the *Bundestag* elections 1976; authors' calculations.

Table 8: Propensity score matching results

Treatment Variable (1)	Sample	Treated	Controls	Difference	s.e.	T-stat
Frequency > 75p	Unmatched	0.393	0.431	-0.038	(0.025)	-1.51
	One neighbour	0.391	0.454	-0.063	(0.043)	-1.45
	Two neighbours	0.391	0.485	-0.093	(0.038)	-2.47
	Three neighbours	0.391	0.479	-0.087	(0.036)	-2.42
	Four neighbours	0.391	0.462	-0.071	(0.035)	-2.01
Treatment Variable (2)	**Sample**	**Treated**	**Controls**	**Difference**	**s.e.**	**T-stat**
Dissimilarity index > p75	Unmatched	0.378	0.428	-0.051	(0.033)	-1.53
	One neighbour	0.378	0.486	-0.108	(0.052)	-2.08
	Two neighbours	0.378	0.454	-0.076	(0.046)	-1.67
	Three neighbours	0.378	0.474	-0.096	(0.043)	-2.24
	Four neighbours	0.378	0.458	-0.080	(0.041)	-1.95

Note: Nearest neighbour matching using propensity score matching, probit estimation of propensity score. Propensity score estimation includes standard covariates (see Table 1). Number of observations is 2,346. For first treatment variable, three observations are off-support, for second treatment all observations are on support. Source: GSOEP 1985, IAB 1985; authors' calculations.

Table 9: Heterogeneous effect by year of moving to current place of residence

	(1) With full information on year of move 1985	(2) Move before 1982	(3) Move before 1979	(4) Move before 1975	(5) Move before 1970
Frequency of own ethnicity	-0.036**	-0.057***	-0.093***	-0.068**	-0.191***
	(0.014)	(0.015)	(0.021)	(0.031)	(0.069)
Male	0.097***	0.100***	0.105***	0.138***	0.075
	(0.018)	(0.019)	(0.023)	(0.031)	(0.068)
Age at migration	-0.035***	-0.038***	-0.035***	-0.033***	-0.017
	(0.003)	(0.004)	(0.005)	(0.006)	(0.015)
Age at migration squ.	0.036***	0.040***	0.037***	0.035***	0.002
	(0.006)	(0.007)	(0.008)	(0.010)	(0.028)
Years since migration	0.009***	0.009***	0.010***	0.007**	0.022***
	(0.002)	(0.002)	(0.003)	(0.004)	(0.007)
Years of schooling	0.052***	0.052***	0.050***	0.045***	0.030*
	(0.005)	(0.005)	(0.007)	(0.009)	(0.017)
Schooling abroad	-0.043*	-0.042*	-0.023	-0.040	0.057
	(0.024)	(0.025)	(0.032)	(0.044)	(0.105)
Married	-0.125***	-0.096***	-0.134***	-0.123**	-0.253*
	(0.029)	(0.031)	(0.041)	(0.053)	(0.131)
Children in household	0.012	0.005	-0.019	-0.014	0.034
	(0.020)	(0.021)	(0.026)	(0.035)	(0.090)
Constant	0.163*	0.222*	0.534***	0.637***	1.327***
	(0.098)	(0.133)	(0.097)	(0.153)	(0.473)
Observations	2289	2000	1362	784	222
R-squared	0.361	0.375	0.391	0.428	0.507

Note: Dependent variable: Binary variable of speaking ability (Speaking very good or good = 1, speaking fair, poor or not at all = 0). Regressions control for ethnicity fixed effects and fixed effects for 85 regional Anpassungsschichten in West Germany. *** p<0.01, ** p<0.05, * p<0.1; robust standard errors in parentheses. Source: GSOEP 1984-85 and IAB 1975/1985; authors' calculations.

Table 10: Multinomial Choice estimates of payoff function ξ, regional pull and push factors, and cost of learning German

	(1) Logit	(2) Probit
Quadratic	-19.1728	-7.4106
	(1.401)	
Linear	27.9283	11.1724
	(1.3668)	
Unemployment	-0.1139	-0.0551
	(0.0135)	
Population	0.1966	0.1084
	(0.0168)	
Constant	10.0109	4.7767
	(0.8761)	
Age at migraton	0.0847	0.0493
	(0.0576)	
Years since migration	-0.0241	-0.0188
	(0.0781)	
Years of schooling	-0.297	-0.1904
	(0.1341)	
Schooling abroad	0.5863	0.3608
	(1.8063)	
LL	-2061.3	-4781.0
LL ratio index	0.1206	0.0917

Note: Estimated by Maximum Likelihood. Sandwich standard errors in parentheses. The first two estimates refer to parameters *a* and *b* in the payoff function $\xi(x) = a * x^2 + b * x$.
Source: GSOEP 2001; authors' calculations

Table 11: Counterfactual simulations

	(1) Benchmark	(2) Equal distribution	(3) More education	(4) Higher concentration	(5) Data
German speakers (%)	46.6%	48.2%	52.6%	42.8%	46.8%
	(1.4)	(1.4)	(1.4)	(1.3)	
Living in Berlin (%)	7.4%	7.3%	7.4%	7.4%	5.3%
	(0.85)	(0.86)	(0.84)	(0.83)	
German speakers in Berlin (%)	44.1%	47.4%	49.7%	40.2%	71.4%
	(5.7)	(5.8)	(5.8)	(5.7)	
Regions with at least one immigrant	87.4	87.0	87.4	87.4	88
	(0.8)	(0.9)	(0.8)	(0.8)	

Note: Means from 500 simulations on a sample of 1,018 immigrants. Standard deviations in parentheses.
Source: GSOEP 2001; authors' calculations.

Appendix

Proof of proposition 1:

1. The probability of learning German conditional on location *j* is:

$$P(S=1 \mid j) = P\left(U_{(1,j)} > U_{(0,j)}\right) = P\left(V_{(1,j)} - V_{(0,j)} > \omega_{(0,j)} - \omega_{(1,j)}\right)$$

Taking the derivative with respect to x_f^j gives:

$$\frac{\partial P(S=1 \mid j)}{\partial x_f^j} = g_j * (\xi'(x_f^j + x_n^j)(-1+1) - \xi'(x_f^j)) \leq 0$$

where g_j denotes the density of $\omega_{(0,j)} - \omega_{(1,j)}$

2. Taking the derivative of $P(S=1, j)$ with respect to x_f^j gives:

$$\frac{\partial P(S=1, j)}{\partial x_f^j} = \frac{\partial \Phi_{(1,j)}}{\partial V_{(1,j)}} * \xi'(x_f^j + x_n^j) * (-1+1) + \frac{\partial \Phi_{(1,j)}}{\partial (-V_{(0,j)})} * (-\xi'(x_f^j)) \leq 0$$

3. The probability of moving to *k* is:

$$P(k) = P(S=1,k) + P(S+0,k) = \Phi_{(1,k)} + \Phi_{(0,k)}$$

Then $\frac{\partial P(k)}{\partial x_f^j} = \frac{\partial \Phi_{(1,k)}}{\partial (-V_{(0,j)})} * (-\xi'(x_f^j)) + \frac{\partial \Phi_{(0,k)}}{\partial (-V_{(0,j)})} * (-\xi'(x_f^j)) \leq 0$

Since $\sum_{k \neq j} \frac{\partial P(k)}{\partial x_f^j} + \frac{\partial P(j)}{\partial x_f^j} = 0,$, it follows that $\frac{\partial P(j)}{\partial x_f^j} \geq 0$.

4. The unconditional probability of speaking German is:

$$P(S=1) = \sum_k P(S=1 \mid k) P(k)$$

with the derivative

$$\begin{aligned} \frac{\partial P(S=1)}{\partial x_f^j} &= \frac{\partial P(S=1 \mid j)}{\partial x_f^j} P(j) + \sum_k P(S=1 \mid k) \frac{\partial P(k)}{\partial x_f^j} \\ &= \frac{\partial P(S=1, j)}{\partial x_f^j} + \sum_{k \neq j} P(S=1 \mid k) \frac{\partial P(k)}{\partial x_f^j} \leq 0 \end{aligned}$$

Q.E.D.

Proof of proposition 2:

We need to show:

$$\lim_{x_f^j \to 0}\left(\frac{\partial P(S=1|\, j)}{\partial x_f^j} P(j) + \lim_{x_f^j \to 0}\sum_k P(S=1|\, k)\frac{\partial P(k)}{\partial x_f^j}\right) > \lim_{x_f^j \to 0}\left(\frac{\partial P(S=1|\, j)}{\partial x_f^j}\right)$$

$$\Leftrightarrow \lim_{x_f^j \to 0} P(S=1, j) \lim_{x_f^j \to 0}\left(\frac{\partial P(S=1|\, j)}{x_f^j}\right) + \lim_{x_f^j \to 0}\sum_{k \neq j} P(S=1|\, k)\frac{\partial P(k)}{\partial x_f^j}$$

$$+ \lim_{x_f^j \to 0} P(S=1|\, j)\frac{\partial P(j)}{\partial x_f^j} > \lim_{x_f^j \to 0}\left(\frac{\partial P(S=1|\, j)}{\partial x_f^j}\right)$$

since $\lim_{x_f^j \to 0} P(S=0, j) = 0$ is implied by assumption 2. With $\lim_{x_f^j \to 0} P(S=1|\, j) = 1$ and $\frac{\partial P(j)}{\partial x_f^j} = -\sum_{k \neq j}\frac{\partial P(k)}{\partial x_f^j}$:

$$\Leftrightarrow \lim_{x_f^j \to 0} P(S=1, j) \lim_{x_f^j \to 0}\left(\frac{\partial P(S=1|\, j)}{x_f^j}\right) + \lim_{x_f^j \to 0}\sum_{k \neq j}\left(\frac{\partial P(k)}{\partial x_f^j}(P(S=1|\, k)-1)\right) > \lim_{x_f^j \to 0}\left(\frac{\partial P(S=1|\, j)}{\partial x_f^j}\right)$$

which holds since $\lim_{x_f^j \to 0} P(S=1, j) \lim_{x_f^j \to 0}\left(\frac{\partial P(S=1|\, j)}{x_f^j}\right) > \lim_{x_f^j \to 0}\left(\frac{\partial P(S=1|\, j)}{\partial x_f^j}\right)$, $\frac{\partial P(k)}{\partial x_f^j} < 0$ and $P(S=1|\, k) - 1 < 0$.

Since ξ is differentiable, there is a neighbourhood around x_f^j for which the inequality holds.

Q.E.D.

Estimation of the simulated multinomial probit model

If we assume the random variables (ε_i^j, r_i^j) to be i.i.d. normal with standard deviations $\sigma_\varepsilon, \sigma_r$, the model to be estimated is a multinomial probit. The mean vector can be set to zero without loss of generality, since the *X* in equation (1) contain a constant and the choice of one location over another is not affected by a level shift of utilities. To understand our estimation routine, consider a choice-set with two locations: we can stack the alternatives as “location 1, not learn”, “location 1, learn”, “location 2, not learn”, and “location 2, learn”. Suppressing the individual index and letting the unobserved learning cost ε be location-independent, the corresponding random vector of ω and its variance-covariance matrix are:

$$\Omega = \begin{pmatrix} r^1 \\ r^1 - \varepsilon \\ r^2 \\ r^2 - \varepsilon \end{pmatrix}, \qquad V(\Omega) = \begin{pmatrix} \sigma_r^2 & \sigma_r^2 & 0 & 0 \\ \cdot & \sigma_r^2 + \sigma_\varepsilon^2 & 0 & \sigma_\varepsilon^2 \\ \cdot & \cdot & \sigma_r^2 & \sigma_r^2 \\ \cdot & \cdot & \cdot & \sigma_r^2 + \sigma_\varepsilon^2 \end{pmatrix}$$

The estimation algorithm for the two times *k* choices (for *k* regions) consists of the following steps (see Train (2003) for a discussion of simulation-based estimations of multinomial choice models):

1. Construct the $(2k \times 2k)$ matrix *L* such that $LL^T = V(\Omega)$. This ensures positive definiteness of the variance matrix. *L* consists of two distinct elements (apart from the zeros). One of them is normalised such that $\sigma_\varepsilon = 1$, since σ_r and σ_ε are not separately identified.

2. For every observation, draw a vector of two times *k* random numbers from the joint normal distribution $N(0, V(\Omega))$. Calculate all $U_{(S,j)}$ from equation (3).

3. To have a smooth probability (rather than a step-function by just counting the number of times an alternative is chosen), calculate $R = \dfrac{\exp(U_{(S,j)} / \lambda)}{\sum_{s=\{0,1\}} \sum_k \exp(U_{(s,k)} / \lambda)}$ where λ can be any number between zero and one.

4. Repeat steps 2 and 3 *N* times and average the *N* "probabilities" *R* to obtain $\hat{P} = (1/N) \sum R$, an approximation to equation (5).

Importantly, the random numbers drawn for each individual should remain constant over all iterations of the maximization routine. We have set the number of simulation steps *N* to 20,000. The higher we set *N*, the closer we approximate the "true" probabilities, at the cost of longer computation time. All estimations are performed with maximum likelihood on the (simulated) probabilities.

Recently published

No.	Author(s)	Title	Date
20/2009	Hohmeyer, K.	Effectiveness of One-Euro-Jobs: Do programme characteristics matter?	8/09
21/2009	Drasch, K. Matthes, B.	Improving Retrospective life course data by combining modularized self-reports and event history calendars: Experiences from a large scale survey	9/09
22/2009	Litzel, N. Möller, J.	Industrial clusters and economic integration: Theoretic concepts and an application to the European Metropolitan Region Nuremberg	9/09
23/2009	Bauer, Th. Bender, S. Paloyo, A.R. Schmidt, Ch.M.	Evaluating the labor-market effects of compulsory military service	11/09
24/2009	Hohendanner, C.	Arbeitsgelegenheiten mit Mehraufwandsentschädigung: Eine Analyse potenzieller Substitutionseffekte mit Daten des IAB-Betriebspanels	12/09
25/2009	Dlugosz St. Stephan, G. Wilke, R.A.	Fixing the leak: Unemployment incidence before and after the 2006 reform of unemployment benefits in Germany	12/09
1/2010	Schmieder J.F. von Wachter, T. Bender, S.	The long-term impact of job displacement in Germany during the 1982 recession on earnings, income, and employment	1/10
2/2010	Heckmann, M. Noll, S. Rebien, M.	Stellenbesetzungen mit Hindernissen: Auf der Suche nach Bestimmungsfaktoren für den Suchverlauf	1/10
3/2010	Schmillen, A. Möller, J.	Determinants of lifetime unemployment: A micro data analysis with censored quantile regressions	1/10
4/2010	Schmieder, J.F. von Wachter, T. Bender, S.	The effects of unemployment insurance on labour supply and search outcomes: Regression discontinuity estimates from Germany	2/10
5/2010	Rebien, M.	The use of social networks in recruiting processes from a firms perspective	2/10
6/2010	Drechsler, J.	Multiple imputation of missing values in the wave 2007 of the IAB establishment panel	2/10
7/2010	Dauth, W.	Agglomeration and regional employment growth	2/10
8/2010	Lietzmann, T.	Zur Dauer der Bedürftigkeit von Müttern : Dauer des Leistungsbezugs im SGB II und Ausstiegschancen	3/10
9/2010	Jahn, E. Rosholm, M.	Looking beyond the bridge: How temporary agency employment affects labor market outcomes	6/10

As per: June 10, 2010

For a full list, consult the IAB website http://www.iab.de/de/publikationen/discussionpaper.aspx

Imprint

IAB-Discussion Paper 10/2010

Editorial address
Institute for Employment Research
of the Federal Employment Agency
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Regina Stoll, Jutta Palm-Nowak

Technical completion
Jutta Sebald

Website
http://www.iab.de

Download of this Discussion Paper
http://doku.iab.de/discussionpapers/2010/dp1010.pdf

For further inquiries contact the author:

Firat Yaman
Phone +49.911.179 3756
E-mail firat.yaman@iab.de