An Analysis of Speaking Fluency of Immigrants Using Ordered Response Models with Classification Errors and Random Thresholds

Christian Dustmann[†] Arthur van Soest [‡]

January 2003

Key words: Semi-parametric Estimation, Measurement error, Assimilation
JEL codes: C14, C35, J15

[*] We are grateful to an anonymous referee, an associate editor, the editor, Bertrand Melenberg, Frank Windmeijer and Marcel Das for useful comments.

[†] University College London, and Institute for Fiscal Studies, London, e-mail: c.dustmann@ucl.ac.uk

[‡] Tilburg University, Department of Econometrics, P.O. Box 90153, 5000 LE Tilburg, Netherlands. e-mail: avas@kub.nl.

An Analysis of Speaking Fluency of Immigrants Using Ordered Response Models with Classification Errors and Random Thresholds

Abstract

Ordered categorical dependent variables, as frequently analysed in economics and other social sciences, are often affected by misclassification error. In addition, if these variables refer to subjective evaluations, a further source of error is the choice of scale across individuals. An example for a variable that is likely to suffer from both errors is speaking fluency of immigrants as reported in survey data. We develop parametric models that take account of this by incorporating misclassification errors in an ordered response model, and by allowing for subjectively chosen boundaries between the categories. As an alternative, we consider a semi-parametric model that nests all parametric models. We apply these estimators to the analysis of English speaking fluency of immigrants in the UK, focusing on Lazear's theory that due to either learning or self-selection, there is a negative relation between speaking fluency and the ethnic minority concentration in the region. Specification tests show that the model allowing for misclassification errors outperforms the ordered probit model. All models lead to the same qualitative conclusions on the relation between speaking fluency and minority concentration, but there is substantial variation in the size of parameter estimates and marginal effects.

1 Introduction

Many empirical studies in economics and other social sciences are concerned with the analysis of ordered categorical dependent variables, like banded data on earnings, income, or hours worked. This data, often retrieved from surveys, has a true objective underlying scale, but can be affected by misclassification error. Another type of categorical data that has become increasingly popular in applied econometrics is based on subjective evaluations. Examples are data on job satisfaction (see, for example, Clark and Oswald (1996)), satisfaction with health (Kerkhofs and Lindeboom (1995)), future expectations of household income (Das and Van Soest (1997)), or subjective evaluations of English speaking fluency of immigrants in the UK (e.g. Chiswick (1991), Chiswick and Miller (1995) and Dustmann (1994)), which we will analyze in this paper. Such data may suffer from the same misclassification problem. Moreover, the bounds used to distinguish, for example, good from reasonable, reasonable from bad, etc., may be specific to the person who evaluates (the respondent or the interviewer).

In applied work, ordered categorical dependent variables are typically analyzed with ordered probit or ordered logit models. In these non-linear models, misclassification can lead to biased estimates of the parameters of interest. To deal with this problem in the binary choice case, several parametric models have been introduced that explicitly incorporate misclassification probabilities as additional parameters. Lee and Porter (1984) estimate an exogenous switching regression model for market prices of grain, distinguishing regimes where firms are cooperative and noncooperative. They observe an imperfect indicator of the actual regime and extend the standard probit model with two misclassification probabilities for the events that regime A is observed given that regime B is active or vice versa. They estimate these probabilities jointly with the parameters of the price equations in both regimes. Hausman et al. (1998) estimate binary choice models for job changes. Using parametric models, they find significant probabilities of misclassifying in both directions. Using semi-parametric models, they obtain estimates of the slope coefficients of interest that are similar to the estimates in the parametric model allowing for misclassification.

In this paper, we follow Porter and Lee and Hausman et al. and incorporate misclassification errors in an ordered response model. Moreover, we focus on the case where the dependent variable is a subjective evaluation on a discrete ordered scale, with subjectively chosen boundaries (thresholds) between the categories. These thresholds may vary across the observations. This is allowed for in the same way as in Das (1995), who treats the thresholds in the ordered probit model as random variables, depending on observed and unobserved characteristics. To test for misclassification or random thresholds, standard tests cannot be used since the null hypothesis puts the parameters on the boundary of the parameter space. We apply simulation based testing procedures recently developed by Andrews (2001). In addition, we consider a semi-parametric model that nests all parametric models and avoids distributional assumptions on the error terms. Since this is a single–index model, the slope parameters of interest can be estimated using the semi–parametric least squares estimator of Ichimura (1993).

The main issue in the application is the relation between host country language proficiency of immigrant minorities and the regional concentration of the minority group. Understanding the assimilation and adaptation of minority and immigrant groups is an important and growing area of research in economics, becoming more relevant as societies are increasingly characterized by a mix of individuals with different cultural backgrounds. Speaking a common language is a key factor in this process. In an influential recent study, Lazear (1999) has developed a model where trade between different groups requires the ability to communicate with each other. To enhance trading possibilities, minority individuals may learn the language of the majority group. The incentive of learning the language is larger the smaller the relative size of the minority group. Moreover, minority individuals with low proficiency in the majority language may sort themselves into communities where individuals speaking their own minority language are concentrated. As Lazear points out, the two processes both lead to a negative association between minority concentration and fluency in the majority language. If the effect of minority concentration on language is created primarily through learning, then the interaction between minority concentration and years of residence should contribute to explaining language proficiency. On the other hand, if sorting is the only relevant mechanism, then this interaction should not be significant. Comparing data from the U.S. census for 1900 and 1990, Lazear concludes that only sorting matters in 1990, while learning was important in 1900.

We investigate the same issue for the UK, using cross–section data on immigrants from ethnic minority communities drawn in 1994. Our parameter of interest are, as in Lazear's study, the effects of the regional minority concentration and its interaction with years of residence on English language proficiency of immigrants.

In survey data, language proficiency is typically evaluated by the respondent or the interviewer on a four or five point scale, ranging from bad or very bad to very good. It seems likely that evaluators differ in what they think is the threshold between bad and reasonable, reasonable and good, etc. In addition, the reported variable may suffer from the same misclassification error as objective variables, such as the job change variable investigated by Hausman et al. (1998). Dustmann and van Soest (2002) focus on the latter type of error, comparing answers to identical survey questions on self–reported speaking fluency in the host country language by the same immigrants at different points in time. They find that, under the assumption that a decrease of language capacity is not possible, more than one fourth of the total variance in the observed speaking fluency variable is due to random misclassification.

Our main empirical question is whether generalizing the ordered response model to allow for misclassification and random category bounds affects the answers to the economic questions concerning the relation between language proficiency, minority concentration, and years of residence.

The results of our empirical analysis show that allowing for classification errors is a clear improvement to the standard ordered probit model. In particular, the estimated probabilities of misclassification into the extreme categories are large. A formal test based upon Andrews (2001) clearly rejects the null hypothesis that all misclassification probabilities are zero. Allowing for misclassification also leads to substantially different estimates of some of the slope coefficients of the regressors. In our application, allowing for random thresholds is much less important. Andrews tests show that this does not lead to significant improvements in either the ordered probit model or the model with misclassification.

The qualitative conclusions on the effect of minority concentration on speaking fluency do not change if misclassification is allowed for. The effect is significantly negative. This is confirmed by the semi-parametric estimates. The estimates of the size of the marginal effects, however, are biased substantially if misclassification is ignored, particularly at low values of the concentration index. The interaction term between years of residence and minority concentration is significant at the 10% level only in the parametric models and insignificant in the semi-parametric model, suggesting that, for our particular application, self-selection is a better explanation for the negative relation between minority concentration and speaking fluency than learning.

The paper is organized as follows. In section 2, we present the models and their estimators. In section 3, we briefly describe the data. Semi-parametric and parametric estimates are presented in Sections 4 and 5. In Section 6, we compare predictions of the two parametric models and the semi-parametric model and test the parametric specifications. Section 7 concludes.

2 Categorical Data and Misclassification

We assume that the dependent variable is observed on an ordinal scale with three levels, coded 1, 2 and 3. In our application, this corresponds to speaking English slightly or not at all, reasonably well, or very well, respectively. The models we discuss extend straightforwardly to the case of more than three categories, but the parametric models will lead to more auxiliary parameters and more intricate expressions for the likelihood function. Starting point is the ordered probit model, not allowing for classification errors. It relates observed categorical information for respondent i to an underlying latent index y_i^* as follows:

$$y_i^* = x_i'\beta + u_i, \tag{1}$$

$$y_i = j \quad \text{if} \quad m_{j-1} < y_i^* < m_j, \quad j = 1, 2, 3\,, \tag{2}$$

$$u_i | x_i \sim N(0, \sigma^2)\,. \tag{3}$$

Here x_i is a vector of explanatory variables including a constant term, β is the vector of parameters of interest, and u_i is the error term. We assume $m_0 = -\infty$, $m_1 = 0$, $m_3 = \infty$. The variance σ^2 and the bound m_2 can be seen as nuisance parameters. We will fix σ^2 to 100 to identify the scale. Throughout, we assume that the observations (y_i, x_i) are a random sample from the population of interest.

2.1 A Parametric Misclassification Model

For the binary choice case, Hausman et al. (1998) show that the bias in estimates of β can be substantial if some observations on the endogenous variable are misclassified. They propose a generalization of the binary probit model to take account of classification errors. We extend this model to the ordered probit case.

We assume that the reported category is y_i, but the (unobserved) true category is z_i, which is related to the latent variable y_i^* as in the ordered probit model:

$$z_i = j \quad \text{if} \quad m_{j-1} < y_i^* < m_j\,, \quad j = 1, 2, 3\,. \tag{4}$$

The probabilities of misclassification are given by:

$$\mathrm{P}(y_i = j | z_i = k, x_i) = p_{k,j}, \quad j, k = 1, 2, 3, j \neq k. \tag{5}$$

Thus $p_{k,j}$ is the probability that an observation belonging to category k is classified in category j. If $p_{k,j} = 0$ for all j, k with $j \neq k$, there is no misclassification and the model simplifies to the ordered probit model. The model with three categories has six misclassification probabilities $p_{k,j}$.

In this model, the latent variable y_i^* can be seen as a perfect indicator of speaking fluency on a continuous scale, something like the score on the ideal objective speaking fluency test. The "true" category z_i is the categorical outcome based upon this score. Misclassification means that the wrong outcome is reported. It should be acknowledged that this is only one way to model misclassification. For example, another source of misclassification would be measurement error in y_i^*, but a normally distributed measurement error would be captured in u_i and would not be identified. A third source would be individual variation in cut-off points. This is discussed in the next subsection.

The main identifying assumption in the model is that $p_{k,j}$ does not depend upon x_i (except through z_i). This is the common identifying assumption in this literature, used by Hausman et al. (1998), Lee and Porter (1984), and in other applications such as Douglas et al. (1995). Such an assumption can only be avoided if a different measurement can be used as a benchmark, such as, in our empirical example, objective measurement of language proficiency (see Charette and Meng (1994)).

For the binary choice case (with categories denoted 0 and 1), Hausman et al. (1998) show that identification of $p_{k,j}$, $j, k = 0, 1$ does not rely on the normality assumption, as long the support of $x_i'\beta$ is the whole real line, i.e., as long as there are enough observations with very low and very high values of $x_i'\beta$. The probabilities of misclassification are then given by:

$$p_{1,0} = \lim_{x_i'\beta \to \infty} \mathrm{P}(y_i = 0|x_i) \text{ and } p_{0,1} = \lim_{x_i'\beta \to -\infty} \mathrm{P}(y_i = 1|x_i)\,.$$

Hausman et al. (1998) show that their model satisfies the single index property that $E\{y_i|x_i\}$ depends on x_i via $x_i'\beta$ only. Therefore, β is identified up to scale and sign. The additional condition required for identification is that $p_{0,1}$ and $p_{1,0}$ are not too large:

$$p_{1,0} + p_{0,1} < 1. \tag{6}$$

This guarantees that $E\{y_i|x_i\}$ increases with $x_i'\beta$. Accordingly, the sign of β is also identified, and (5) implies that the $p_{0,1}$ and $p_{1,0}$ are non-parametrically identified.

For the ordered probit case with categories 1, 2 and 3 and six misclassification probabilities, we get

$$\begin{aligned} E\{y_i|x_i\} &= 2 - p_{2,1} + p_{2,3} - \Phi((m_1 - x_i'\beta)/\sigma)(1 - p_{1,2} - p_{2,1} + p_{2,3} - 2p_{1,3}) \quad (7) \\ &+ [1 - \Phi((m_2 - x_i'\beta)/\sigma)](1 - p_{3,2} - p_{2,3} + p_{2,1} - 2p_{3,1}) \, . \end{aligned}$$

Thus the condition that $E\{y_i|x_i\}$ increases with $x_i'\beta$ for every value of $x_i'\beta$ implies (instead of (6) for the binary choice case):

$$p_{1,2} + p_{2,1} - p_{2,3} + 2p_{1,3} < 1 \text{ and } p_{2,3} + p_{3,2} - p_{2,1} + 2p_{3,1} < 1 \, . \quad (8)$$

This condition is satisfied for small enough values of the misclassification probabilities. A sufficient condition for (8) is given by Abrevaya and Hausman (1999):

$$p_{1,1} > p_{2,1} > p_{3,1} \text{ and } p_{3,3} > p_{2,3} > p_{1,3} \, . \quad (9)$$

This condition is stronger than (8) but easier to understand intuitively.

The argument for nonparametric identification in the binary choice case applies to $p_{1,j}$ and $p_{3,j}$, but not to $p_{2,1}$ or $p_{2,3}$. Identification of these is achieved in this parametric model by imposing normality of the error terms. The model can straightforwardly be estimated by Maximum Likelihood (ML), where the $p_{k,j}$ are estimated jointly with the slope parameters β. The ML estimates are consistent, asymptotically normal, and asymptotically efficient if the assumptions (including normality of the errors) are satisfied.

2.2 Random Threshold Variation across Respondents

Evaluators (typically the respondent or the interviewer) are usually not precisely instructed how to construct their subjective score y_i^* on a continuous scale or which cut–off points to use for the discrete outcomes. This suggests that there will be (unobserved) heterogeneity in y_i^* and the cut-off points. Unobserved heterogeneity in y_i^* is picked up by the error term u_i in (3). (To identify β, it has to be assumed that such heterogeneity is independent of the regressors.) In this subsection we discuss how heterogeneity in the cut-off points m_1 and m_2 can be incorporated.

Extending the ordered probit model (with or without misclassification probabilities) to allow for heterogeneity in the threshold values is intuitively attractive, since it implies that two evaluators who perceive the same latent value y_i^* may still give different answers on the ordinal scale, using their own interpretation of what is, for instance, good, reasonable, or bad speaking fluency.

Ordered probit models with category bounds that vary across respondents have been introduced by Terza (1985) and Das (1995). While Terza (1985) only allows for variation of the category bounds with observed (exogenous) respondent characteristics, Das (1995) also allows for unobserved heterogeneity in the bounds. Here we follow Das (1995). We first discuss the model without classification errors. Its specification is as follows.

$$\begin{aligned}
y_i^* &= x_i'\beta + u_i, \qquad (10)\\
m_{ji}^* &= w_i'\gamma_j + v_{ji} \quad j = 1, 2\\
y_i &= 1 \quad \text{if} \quad y_i^* \leq \min(m_{1i}^*, m_{2i}^*)\\
y_i &= 2 \quad \text{if} \quad \min(m_{1i}^*, m_{2i}^*) < y_i^* \leq \max(m_{1i}^*, m_{2i}^*)\\
y_i &= 3 \quad \text{if} \quad y_i^* > \max(m_{1i}^*, m_{2i}^*)\\
u_i &\sim N(0, \sigma^2), \;\; v_{ji} \sim N(0, \sigma_j^2), \;\; j = 1, 2
\end{aligned}$$

$$u_i, v_{1i} \text{ and } v_{2i} \text{ are independent of each other and of } x_i \text{ and } w_i \qquad (11)$$

The ordering in the thresholds cannot be determined a priori: with positive probability, m_{1i}^* exceeds m_{2i}^*, and the model with categories $(-\infty, m_{1i}^*)$, (m_{1i}^*, m_{2i}^*), and

(m^*_{2i}, ∞) is not well-defined. Das (1995) solves this problem by using the ordered thresholds instead of the original ones. In the case with three categories, this means that m^*_{1i} is replaced by $\min(m^*_{1i}, m^*_{2i})$ and m^*_{2i} by $\max(m^*_{1i}, m^*_{2i})$. The probabilities of the three outcomes ($y_i = 1$, $y_i = 2$ or $y_i = 3$) for this model can be rewritten as follows:

$$\mathrm{P}(y_i = 1|x_i, w_i) = \mathrm{P}(u_i - v_{1i} < w_i'\gamma_1 - x_i'\beta \text{ and } u_i - v_{2i} < w_i'\gamma_2 - x_i'\beta)\,, \tag{12-a}$$

$$\begin{aligned}\mathrm{P}(y_i = 2|x_i, w_i) \;=\;& \mathrm{P}(u_i - v_{1i} < w_i'\gamma_1 - x_i'\beta \text{ and } u_i - v_{2i} > w_i'\gamma_2 - x_i'\beta) \\ &+\mathrm{P}(u_i - v_{1i} < w_i'\gamma_1 - x_i'\beta \text{ and } u_i - v_{2i} > w_i'\gamma_2 - x_i'\beta)\,,\end{aligned} \tag{12-b}$$

$$\mathrm{P}(y_i = 3|x_i, w_i) = \mathrm{P}(u_i - v_{1i} > w_i'\gamma_1 - x_i'\beta \text{ and } u_i - v_{2i} > w_i'\gamma_2 - x_i'\beta)\,. \tag{12-c}$$

This is a bivariate probit model that does not distinguish between the two regimes leading to outcome $y_i = 2$. It is clear that scale and location need to be fixed to identify the model. The scale is set by choosing $\sigma^2 = 100$, as in the other models. To identify the location, we set $\gamma_1 = -\gamma_2$. This is equivalent to several other normalizations but has the advantage of symmetry. It implies that an increase of $|w'\gamma_1|$ induces an increase in the probability of giving the intermediate answer. The sign of γ_1 is identified by imposing that $w'\gamma_1$ is more often the lower bound than the upper bound (i.e., $w'\gamma_1 \leq 0$ for at least 50% of the observations).

The covariance structure of the bivariate probit model is given by $V(u_i - v_{1i}) = \sigma^2 + \sigma_1^2$, $V(u_i - v_{2i}) = \sigma^2 + \sigma_2^2$, and $Cov(u_i - v_{1i}, u_i - v_{2i}) = \sigma^2$. Thus the variances of u_i, v_{1i} and v_{2i} are identified. Relaxing (11) by allowing for non–zero correlations between the three error terms would lead to an unidentified model.

This model can also be interpreted as follows. Two evaluations are reported: one based upon $-w_i'\gamma_1 - v_{1i} + x_i'\beta + u_i$, and one upon $-w_i'\gamma_2 - v_{2i} + x_i'\beta + u_i$. If both evaluations are positive, the answer $y_i = 3$ (good or very good) is given. If both are negative, $y_i = 1$ (bad or very bad) is reported. If one evaluation is positive and the other is negative, $y_i = 2$.

In the empirical application, speaking fluency is evaluated by the interviewer. The data provide no information on the interviewer so that interviewer characteristics can enter only through v_{1i} and v_{2i}. Including respondent characteristics in w_i seems less natural here. We experimented with this but found no significant results. In the results that we will report, w_i consists of a constant term and threshold heterogeneity comes through v_{1i} and v_{2i} only.

Explicitly allowing for misclassification in this model is possible in the same way as in the standard ordered probit model. The probabilities for the true categorical outcomes z_i are given by (12-a), (12-b) and (12-c), with y_i replaced by z_i. The probabilities of the reported outcomes given the true outcomes are again given by (5).

2.3 A Semi-parametric Approach

The parametric ML estimates of the slope parameters β in the models introduced above require distributional assumptions and may not be robust to misspecification. If we are interested in β only and consider the $p_{k,j}$ as nuisance parameters, semi-parametric estimation seems a good alternative.

Consider the model with fixed thresholds and misclassification probabilities. The conditional mean of the observed categorical variable y_i in model (1) - (5) given x_i is given by (7). It depends on x_i only through the index $x_i'\beta$. Thus (1)-(5) is a special case of the single index model given by

$$E\{y_i|x_i\} = H(x_i'\beta)\,, \tag{13}$$

where H is an unknown link function. If we relax the normality assumption (3) and replace it by the assumption

$$u_i \text{ is independent of } x_i\,, \tag{14}$$

we get the following expression instead of (7):

$$\begin{aligned} E\{y_i|x_i\} &= 2 - p_{2,1} + p_{2,3} - G(m_1 - x_i'\beta)(1 - p_{1,2} - p_{2,1} + p_{2,3} - 2p_{1,3}) + \\ &\quad [1 - G(m_2 - x_i'\beta)](1 - p_{3,2} - p_{2,3} + p_{2,1} - 2p_{3,1})\,, \end{aligned} \tag{15}$$

where G is the distribution function of the error term u_i ($G(t) = \mathrm{P}[u_i \leq t]$).

Again, the right-hand side depends on x_i only through $x_i'\beta$, so that (1), (2), (4), (5) and (14) lead to the single index model (13) with link function H given by (15). As stated before, the crucial assumption here is that the misclassification probabilities in (4)- (5) do not depend on x_i.

Moreover, under the same assumptions, it is straightforward to show that the conditional variance $V\{y_i|x_i\}$ also depends on x_i through the same index $x_i'\beta$ only. This implies that the model for y_i is heteroskedastic but the heteroskedasticity has a special form. Finally, it is easy to show that the inequalities in (8) imply that H can be chosen non–decreasing.

An expression similar to (15) can be derived from the extension of the model which allows for random cut–off points. Under the additional assumption that the variation in the cut–off points is independent of observed characteristics x_i, the model with random cut–off points is also a single index model and the statements above remain valid.

Thus the models discussed above are all special cases of the general single index model (13) for some (unknown) link function H. In this model, the vector β of slope parameters is identified up to scale; the constant term is not identified. A number of asymptotically normal root n consistent estimators for β in this model have been discussed in the literature, requiring various assumptions on the distribution of the explanatory variables x_i and regularity conditions on the link function H. Ichimura (1993) uses nonlinear least squares combined with nonparametric estimation of H. This estimator requires numerical minimization of a non–convex objective function. Hausman et al. (1998) use the maximum rank correlation estimator of Han (1987). This also requires numerical optimization. We experimented with applying this estimator, but ran into convergence problems with the Han estimator, possibly due to the relatively large number of explanatory variables.

Attractive from a computational point of view is the class of (weighted or unweighted) average derivative estimators (see, for example, Powell et al. 1989). They require that the distribution of x is absolutely continuous and are therefore not directly applicable to our empirical example. Horowitz and Haerdle (1996) have developed an estimator which allows for discrete variables, but not for interaction terms of continuous variables. Since interaction terms are important in our particular application, the Horowitz and Haerdle (1996) estimator cannot be applied. We will therefore focus on Ichimura's semi–parametric least squares (SLS) estimator.

Ichimura's SLS estimator minimizes the sum of squares $S_n(\beta)$ over β, where

$$S_n(\beta) = 1/n \sum (y_i - \hat{E}[y_i|x_i'\beta])^2. \tag{16}$$

Here $\hat{E}[y_i|x_i'\beta]$ is a univariate kernel regression estimate of y_i on the index $x_i'\beta$ (for given β). Finding the β at which (16) is minimized requires an iterative procedure. If smooth kernel weights are used, the function to be minimized is smooth in β and a Newton-Raphson technique can be used to find the optimal β, i.e., $\hat{\beta}_{SLS}$. Ichimura (1993) shows that, under appropriate regularity conditions, this yields a $\sqrt{n}$ consistent asymptotically normal estimator of β_0. He also derives the asymptotic covariance matrix of this estimator and shows how it can be estimated consistently.

Ichimura (1993) also indicates how to design an asymptotically efficient weighted semi-parametric least squares (WSLS) estimator that uses SLS as the first step. For the general case, this requires nonparametric regression of the squared SLS residuals on x and leads to problems if x contains interaction terms or discrete variables. In our case, however, we have seen above that the natural generalization of the parametric models implies that $V[y_i|x_i]$ depends on x_i only through $x_i'\beta$, and for this special case Ichimura shows that the efficient WSLS estimator requires weighting with $\hat{V}[y_i|x_i'\hat{\beta}_{SLS}]^{-1}$, obtained by a non-parametric regression of the squared SLS residuals on the index $x_i'\hat{\beta}_{SLS}$.

Implementing the SLS and WSLS estimators in practice requires a choice of kernel and bandwidth. We will work with the Gaussian kernel. For consistency, the bandwidth should tend to zero if $n \to \infty$ at a slow enough rate. Although a large literature on the optimal bandwidth choice exists for the non–parametric regression problem itself, it is not clear how to determine the optimal bandwidth for estimating β. Theoretical results for similar problems suggest that under-smoothing will be optimal, i.e., the optimal bandwidth will be smaller than the optimal bandwidth for the non–parametric regression of y_i on $x_i'\beta$. The common approach for choosing a bandwidth in a situation like this is to experiment with the bandwidth which would be optimal for the non-parametric regression problem (given a value of β) and with smaller bandwidth values (to under-smooth). We will present results for several values of the bandwidth.

Once β_{SLS} (or β_{WSLS}) is obtained, the link function H can be estimated by a non–parametric (kernel) regression of y_i on the estimated index $x_i'\hat{\beta}_{SLS}$. The usual asymptotic properties of a kernel estimator apply since $\hat{\beta}_{SLS}$ converges at a faster rate than the non-parametric estimator.

3 Data

We apply the models and techniques discussed above to analyze the effect of minority concentration on immigrants' proficiency in the host country language. The empirical analysis is based on the Fourth National Survey on Ethnic Minorities (FNSEM), a cross- sectional survey carried out in the UK in 1993 and 1994. Individuals included are aged 16 or more. There are 5196 observations in the minority sample. We focus on a homogeneous sample of 1471 men of Indian ethnicity (from India, Bangladesh, Pakistan or Uganda). The FNSEM contains information on the concentration of the individual's own minority group at ward level, which has been matched to the survey from the

1991 Census.[1] The language information in the survey is based on the interviewer's evaluation of the individual's language ability in English, with categorical answers (speaks English) *very well*, *fairly well*, *slightly*, and *not at all*. For the empirical analysis, we have combined the categories *slightly* and *not at all* and recoded the three categories to 3 (*very well*), 2 (*fairly well*) and 1 (*slightly or not at al*).

Summary statistics on the resulting categorical speaking fluency variable and on other individual characteristics are presented in Table 1. About 47 percent of the 1471 men in the survey data are reported to speak English very well. For only 4.3%, the interviewer reports *not at all*; this group is merged with the 22.6% in the category *slightly*.

On average, concentration of minorities of the same ethnicity as the respondent is about 16.2%, with substantial variation in the sample and a sample standard deviation of 15.2%. There is a clear negative correlation between language proficiency and minority concentration. Average minority concentration in the subsample of people with low speaking fluency is about 20.8%, in the subsample of the most fluent speakers it is only 13.7%. The rank correlation coefficient is -0.215.

4 Semi-parametric Estimates

Some SLS and WSLS estimates explained in section 2.3 are presented in Table 2. In the first column, SLS estimates are presented with the bandwidth set equal to $1.06\hat{\sigma}(x'\hat{\beta})n^{-0.2}$, where n is the number of observations and $\hat{\sigma}(x'\hat{\beta})$ is the estimated standard deviation of the single index. This is the rule of thumb estimate for the optimal bandwidth in the kernel regression (Silverman, 1986). Since under-smoothing typically gives more efficient estimates for the single index (Powell, 1994), we also present the results for a bandwidth that is half as large (third column). The differences

[1] In the UK, a ward is the smallest geographical area identified in the Population Census. According to the 1991 census, the mean population within a ward is 5459 individuals, and the median is 4518.

between the two sets of estimates or their standard errors are small, confirming the general finding in this literature that the SLS results are not sensitive to the choice of the bandwidth (see, for example, Bellemare et al., 2002). The second column presents the WSLS estimates, using the same bandwidth as column I. These estimates are very similar to those in column I. Estimated standard errors are somewhat smaller in most cases, in line with the fact that WSLS is asymptotically efficient while SLS is not, but there are also two parameters for which the estimated standard error is slightly larger for WSLS than for SLS. Results for the smaller bandwidth (not presented) tell the same story.

Standard errors are based upon the asymptotic distribution of the estimator. Bootstrapped standard errors give the same economic conclusions and are therefore not presented. They are larger than the asymptotic standard errors for some parameters and smaller for others.

The constant term is not estimated. The coefficient of YSM (years since migration) is normalized to 0.9634, its estimate in the ordered probit model (see below). This normalization makes it easy to compare semi-parametric and parametric results. The variable YSM has a significant positive effect with a large absolute t-value in all parametric models, which justifies the assumption that the coefficient is nonzero, the (only) necessary condition for using this normalization.

The estimation results are qualitatively in line with Lazear (1999). Since not only YSM itself but also YSM squared and YSM interacted with the minority concentration index are included among the regressors, the effect of an increase of YSM on expected speaking fluency varies across observations. Still, the marginal effect of increasing years since migration on expected fluency is positive at almost all observations. The negative sign of YSM squared implies that this effect is smaller for those with longer years of residence. Conditional on years since migration, older immigrants are less fluent in English than younger immigrants. The country of origin dummies indicate that, keeping other characteristics constant, immigrants from Pakistan and Bangladesh are significantly less fluent than immigrants from India, whereas the individuals of afro-asian origin are the most fluent.

Speaking fluency falls with minority concentration at a declining rate, confirming Lazear's finding for the U.S. One explanation for this is that individuals who live in areas with high concentrations of residents of their own minority have lower incentives to learn the majority language. Another explanation is that individuals select their area of residence according to their language proficiency. As Lazear points out, a significant negative effect of the concentration variable on speaking fluency is consistent with both explanations. In both cases, the individual's (location or learning) choice is determined by the objective to maximize interaction with individuals with whom they share a common language.

To distinguish between the two explanations, Lazear adds an interaction term between minority concentration and years of residence (YSM). An insignificant interaction term favors the self selection hypothesis, since the learning argument would imply a negative interaction effects (a larger learning rate, i.e., a higher effect of YSM, when learning pays off more, i.e., when minority concentration is lower). In Table 2, the coefficient on the interaction term of years since migration and minority concentration is negative but insignificant and close to zero, favoring the self selection hypothesis. Interestingly, this is similar to what Lazear finds for the 1990 U.S. census.

In figure 1, we have drawn the estimated link function H in (13) for the first set of results in Table 2. For the other results, the figure looks very similar.[2] The figure also contains 95 percent uniform confidence bounds (based upon Haerdle and Linton (1994)). The estimated link function is increasing on its full domain, except at very low values of the index, for which the estimates are imprecise due to the small number of observations in that region. In an ordered response model without misclassification, the value of the link function should tend to 1 if the index value tends to $-\infty$. The figure suggests that this is not the case. This could be due to misclassification of some

[2] We use the quartic kernel. The bandwidth is chosen by visual inspection.

respondents with low speaking fluency ($y_i = 1$).

5 Parametric Estimates

Estimates for several parametric models are presented in Table 3. The first columns give the results of the standard ordered probit model. In the second column, misclassification probabilities are incorporated (see section 2.1). The third column allows for random cut–off points (see section 2.2) but not for misclassification. Results in the fourth column allow for non-zero misclassification probabilities as well as random thresholds.

The four sets of parametric estimates of the slope coefficients are generally in line with each other in terms of signs and significance levels, but there are substantial differences in magnitude. We will discuss the magnitude of the most important estimates below when we look at predicted marginal effects on the probabilities of good and reasonable speaking fluency. The coefficients all have the same sign as in the semi-parametric model. Fluency increases with years since migration at a decreasing rate. Immigrants from Pakistan and Bangladesh are less fluent than immigrants from India, while afro-asian immigrants have the highest fluency, *ceteris paribus*. Speaking fluency is lower in regions where the concentration of immigrants from the same country of origin is larger.

The estimated coefficient on the interaction term of minority concentration and years since migration is always negative and significant at approximately the two-sided 10% level in the first two models, and at a somewhat higher level in the models with random thresholds. This is different from the semi-parametric estimates, which were negative but smaller in magnitude and not significant at all. While the semi-parametric evidence suggested that the negative effect of minority concentration on speaking fluency is due to self selection into local areas and not due to the effort in learning the language, the parametric results suggest that learning could play a role as well. Still, t-values are not high enough to draw any final conclusions on this. For those with zero years of residence, the estimated pattern of speaking fluency as a function of minority concentration is decreasing up to about the 88th percentile of minority concentration according to the model with misclassification only, up to about the 92nd percentile for the model with random thresholds only, and up to the 94th percentile for the semi-parametric models. This suggests that already shortly after entry, immigrants in low minority concentration areas speak better English, something which can only be explained by self selection.

The misclassification probabilities in column 2 are by definition nonnegative, implying that standard t-tests or likelihood ratio tests on $p_{k,j} = 0$ are inappropriate (see, e.g., Shapiro (1985)). Still, the estimates of the $p_{k,j}$ and their standard errors imply that 0 is not contained in the one-sided 95% confidence intervals of four of them, suggesting that adding the probabilities of misclassification is an improvement compared to the standard ordered probit model. A formal test of the hypothesis $p_{k,j} = 0$ for all $j \neq k$ can be based upon the likelihood ratio, using the method proposed by Andrews (2001). The LR test statistic does not have the usual chi–squared distribution under the null, since the test is one–sided and since under the null, the parameter vector is not in the interior of the parameter space. Andrews (2001) demonstrates that the LR test statistic can still be used and shows how to compute the appropriate asymptotic critical values, using a quadratic approximation to the likelihood. In the appendix we give the algorithm that is used for our case. We find a 5% critical value of 9.04 and a 1% critical value of 12.88. Since the realization of the LR test statistic is 16.72, the null hypothesis is rejected at the 1% level. This confirms that allowing for misclassification errors improves the fit of the model significantly.

The estimates of the misclassification probabilities amply satisfy the inequalities in (9) that are sufficient for identification and imply monotonicity of the link function. The estimates of $p_{2,1}$ and $p_{2,3}$ have the largest standard errors, reflecting the problem that these are harder to identify. Compared to the ordered probit model, most slope coefficients and the estimate of the category bound m_2 have increased by approximately a factor 2. Due to the normalization, this can also be seen as a reduction of the standard deviation of the error term u by about 50 percent. The interpretation is that part of the unsystematic variation in observed speaking fluency is now explained by classification errors.

The third specification presented in Table 3 is the model with random thresholds but without misclassification probabilities. The final two parameters are the estimated standard deviations of the thresholds. One of them is equal to zero, but the other one is not, although its standard error is as large as the point estimate. A likelihood ratio test similar to the one discussed above (following Andrews 2001) does not reject the ordered probit model against the model with random thresholds at the 10% level (LR test statistic 3.7; 5% and 10% critical values 5.09 and 3.77, respectively). The estimates of the slope parameters are close to those in the ordered probit model.

In the final columns of Table 3, both misclassification probabilities and random thresholds are incorporated. The estimates of the misclassification probabilities $p_{2,1}$ and $p_{2,3}$ are extremely inaccurate now, suggesting a serious identification problem in this rich parametric model. The other misclassification probabilities, which are nonparametrically identified, are estimated more accurately and the estimates are close to those in the model with fixed thresholds. The estimates of the misclassification probabilities satisfy the monotonicity conditions (8) but not the stronger conditions in (9). An Andrews (2001) LR test of this model against the previous one again rejects the hypothesis that all $p_{j,k}$ are zero at conventional significance levels (test statistic 15.16; 1% critical value 12.44). On the other hand, an Andrews test of the model with misclassification probabilities and fixed thresholds (specification 2) cannot be rejected against the more general model with random thresholds (LR test statistic 2.12; 10% critical value 2.90). Thus all Andrews tests taken together lead to the unambiguous conclusion that misclassification is significant but random variation in thresholds is not, supporting specification 2, with misclassification probabilities and fixed thresholds.

The results of the parametric models can be used to analyze the size of the effects of concentration of immigrants of a certain language minority on true speaking fluency, not affected by misclassification error or variance in thresholds. Table 4 summarizes the results. It presents the estimated marginal effects of minority concentration on the probabilities of at least slight fluency and very good fluency according to each of the models in Table 3 at the first, second and third quartile of the sample distribution of the concentration index. Other regressors have been set to their sample means. The estimated marginal effects are functions of the estimates of β and m_2 (models 1 and 2) or γ (models 3 and 4). Misclassification probabilities are discarded; the marginal effects refer to the true classification, not to the reported classification. In models 3 and 4, random variation of the thresholds is also discarded, and the mean threshold values are used.

The table shows some substantial differences in the estimated marginal effects. For example, let us compare two otherwise identical immigrants in a region with approximately median ethnic concentration. If the area of the one has a 1%-point higher ethnic concentration than the area of the other immigrant, the ordered probit model predicts a 1.33%-point higher probability of speaking English very well for the immigrant in the lower concentration area. According to the misclassification model, the difference has the same sign but is much larger, about 2.27 %-points (with standard error 0.44%-points).

Model 2 allows for misclassification and significantly outperforms the ordered probit model. On the other hand, it leads to much larger standard errors on the estimated marginal effects. As an intermediate case, we also estimated a model that allows for misclassification in an adjacent category, but not in non-adjacent categories. In other words, we imposed $p_{1,3} = p_{3,1} = 0$ in model 2 (without threshold variation). We do not present detailed results for this model, since this model if formally rejected against model 2. Still, most of the estimation results are similar to those in model 2. The estimates of the misclassification probabilities are, for example, $\hat{p}_{1,2} = 0$ (the lower bound), $\hat{p}_{2,1} = 0.2528$ (standard error 0.0468), $\hat{p}_{2,3} = 0.3023$ (standard error 0.0384), and $\hat{p}_{3,2} = 0.0877$ (standard error 0.0379), values which are similar to those in Table 3. The estimated marginal effects are also similar to those of model 2, but with standard errors that are about 20% smaller, on average.

Comparing Two Parametric Models

In Figures 2 and 3,**Arthur: Figures need relabeling - there are now figures 1, 2a, 2b, and 4** we compare the predictions of two parametric models: ordered probit and the misclassification model. We do not consider the models with random thresholds, since we found no support for these. We look at the estimated probabilities that true fluency is (at least) good and that reported fluency is good. In the ordered probit model, observed and true speaking fluency (y and z) coincide, but in the model with misclassification errors they do not.

Figure 2 presents a scatter plot of the predicted probabilities of good speaking fluency according to the two parametric models. For the misclassification model (vertical axis), the figure shows the predictions of the true speaking fluency variable z. For the ordered probit model (horizontal axis and 45 degree line), predictions of y and z coincide. We find that the misclassification model leads to more probability estimates close to zero or one than the ordered probit model, leading to a larger dispersion in $\hat{P}[z = 3|x]$ according to the misclassification model than according to ordered probit. Still, the correlation between the two sets of predictions is quite large (the sample correlation coefficient is 0.97).

In Figure 3, we compare predictions of the probability that individuals *report* good or very good speaking fluency. In the misclassification model, the probability of reporting good or very good fluency is never close to one or zero. For most observations with predicted probabilities not close to one or zero, the predictions according to ordered probit and misclassification models are similar. The correlation coefficient is almost 0.99.

The substantial differences between true and reported fluency in the misclassification model confirm the conclusion from the misclassification probabilities in Table 4: generalizing the ordered probit model by incorporating misclassification probabilities is useful in this empirical example. The same conclusion is obtained for the probability of bad or very bad speaking fluency (figures not reported).

Mis-specification Tests of Parametric Models

In principle, the parametric models could be tested against the semi-parametric model using a Hausman test. Under the null that the parametric model is correct, the parametric ML estimates are asymptotically efficient and the SLS estimates are consistent. Under the alternative that the semi-parametric model is correctly specified but the parametric model is not, only the SLS estimates are consistent. Thus a chi-squared test can be based on the difference between parametric and semi-parametric estimates. Unfortunately, however, the estimated standard errors of the SLS estimates are not always larger than those of the parametric ML estimates. This implies that the Hausman test statistic cannot be computed. This problem remains if bootstrapped standard errors are used for the semi-parametric model. The procedure of Newey (1985) can not be used as it does not apply to the semi-parametric estimator.

An alternative, graphical, specification test of parametric models is introduced by Horowitz (1993). The null hypothesis is that the parametric model is correctly specified. The result for the parametric model with misclassification is given in Figure 4. It presents two functions of the index estimate $x'b/s$, where b and s are the parametric estimates of β and σ in Table 3. The solid curve gives the predicted probabilities $\hat{P}[y_i = 3|x_i] = \hat{P}[y_i = 3|x_i'b]$ according to the parametric model, as a function of $x_i'b$. The dashed curves gives nonparametric kernel regression estimates of the observed dummy indicator variable $I(y_i = 3)$ on the same index $x_i'b$ with uniform 95% confidence bands. Since the estimator b converges to β at rate root n, which is faster rate than the rate of convergence of the nonparametric estimator, the standard errors of b are asymptotically negligible, and confidence bands are calculated as if b was known.

Under the null that the parametric model is specified correctly, b is consistent for β and the parametric expression for the predicted probability $\hat{P}[y_i = 3|x_i]$ is consistent for $P[y_i = 3|x_i]$. The null hypothesis, however, also implies that $P[y_i = 3|x_i]$ is a single index function of $x_i'\beta$, and b is a consistent estimate of this single index (up to scale). The nonparametric curve is the estimated link function, and it will also be consistent for $P[y_i = 3|x_i]$. Thus under the null both curves are consistent for the same function, and should be similar. The null hypothesis will be rejected if the nonparametric (circled) curve is significantly different from the parametric (solid) curve. Since the parametric curve is based upon estimates which converge at rate $\sqrt{n}$, while the nonparametric curve converges at the lower rate $n^{0.4}$, the imprecision in the former curve can be neglected compared to that in the latter, and the test can be based on the uniform confidence bands around the nonparametric curve.

The result is that the solid curve is everywhere between the uniform confidence bands, so that the parametric model cannot be rejected. This can be seen as support in favour of the parametric misclassification model. It should be admitted, however, that the same test cannot reject the ordered probit model either, while we already saw that the Andrews test rejects this model against the model with misclassification errors. This casts some doubt on the power of this type of test. The same conclusions are obtained if $P[y_i = 1|x_i'b]$ is used instead of $P[y_i = 3|x_i]$.

6 Summary and Conclusions

In models with ordered categorical dependent variables where the categorical assignment is based on subjective evaluations, misclassification may have two sources: Classical misclassification due to simple reporting errors, and misclassification due to a subjective choice of scale. Both sources can lead to seriously biased parameter estimates and predictions. Parametric estimators which incorporate and estimate misclassification probabilities, as well as semi-parametric estimators, are an alternative to standard parametric models. Extending the work of Lee and Porter (1984) and Hausman et al. (1998), we introduce a parametric model that incorporates misclassification probabilities for the case of more than two ordered categories, and that allows for scale heterogeneity. We show that this model is a special case of a semi-parametric single index model that can be estimated with semi-parametric least squares.

Using these models, we analyze the association between minority concentration and speaking fluency of immigrants, using data for the UK. We find that the misclassification model is a significant improvement compared to the standard probit model. Allowing for random thresholds in addition does not lead to further improvements. The qualitative effects of minority concentration are similar, supporting Lazear's finding for the US that speaking fluency falls with minority concentration. Marginal effects show, however, that the size of the correlation and the shape of the relationship between fluency and minority concentration are quite different according to the two models. The models both give weak evidence in favor of a learning effect, reflected by a negative interaction effect of minority concentration and years since migration that is significant at the one sided 10% level. The evidence in favor of self selection of more fluent immigrants into areas with lower minority density is much stronger and insensitive to the chosen model. Semi-parametric estimates in a model that nests all parametric models considered confirm the qualitative conclusions, although the evidence of a learning effect is even weaker.

A shortcoming of the model is that probabilities of misclassification in intermediate categories are not precisely estimated, since their identification relies on parametric assumptions. Better estimates of all misclassification probabilities would require additional data, for example alternative measurements (Charette and Meng (1994)), or panel data. This is on our research agenda.

Appendix: Andrews Test

This appendix explains how to test the null hypothesis H_0: $p_{jk} = 0$, $j, k = 1, 2, 3, j \neq k$ against the alternative $p_{jk} > 0$ for at least one pair $j \neq k$.[3] Since the model is not defined for $p_{jk} < 0$, the parameter vector is not an internal point of the parameter space under the null hypothesis, implying that standard asymptotic theory of the ML estimator does not apply.[4] Andrews (1999) derives the asymptotic distribution of a class of a general class of estimators including ML when the true parameter value is on the boundary of the parameter space. Andrews (2001) applies the results in Andrews (1999) to derive the asymptotic distribution of the quasi-likelihood ratio test statistic, which is what we need. (Andrews (2001) also allows for nuisance parameters which play a role under the alternative only; such parameters do not appear in our case.) See Theorem 4 in Andrews (2001). (The special case without nuisance parameters that are not identified under the null also follows from Theorem 3 in Andrews (1999).) It is straightforward to check that the regularity assumptions required for this theorem are satisfied in our example, since observations are i.i.d., ML estimation is used, the log likelihood has continuous right partial derivatives of second order, and the parameter space has the form of a convex cone. Checking the regularity conditions is basically the same as for the example of a random coefficients model in Andrews (1999).

Let LR be the likelihood ratio test statistic: $2(\ln L_1 - \ln L_0)$, where L_1 is the unrestricted maximum of the likelihood (allowing for all $p_{j,k} \geq 0$) and L_0 is the restricted maximum (imposing $p_{j,k} = 0$ for all $j, k = 1, 2, 3$. The parameter vector can be written as $\theta = (\theta_1', \theta_2')'$, where θ_2 contains the six misclassification probabilities $p_{1,2}, \ldots, p_{3,2}$ and θ_1 contains the other 12 (unrestricted) parameters of the model. The parameter space can be written as $V = (-\infty, \infty)^{12}$ x $[0, \infty)^6)$, and the null hypothesis is $\theta \in V_0 = (-\infty, \infty)^{12}$ x $\{0\}^6$. (We ignore the obvious lower bound on the threshold

[3] Tests for random thresholds against fixed thresholds are constructed in the same way.

[4] It also imples that alternative tests for inequality constraints such as those of Andrews (1998) or Szroeter (1997) cannot be applied.

m_2), since it is not binding and irrelevant for the local approximations.) Let J be minus the expected value of the Hessian of the log likelihood contribution of a random observation at the true parameter values, which, under the null, can be consistently estimated in the usual way by $\hat{J}$, the sample mean of the matrix of second order partial derivatives at each observation, evaluated at the restricted ML estimates. Similarly, let I be the expected value of the outer product of the gradient of the log likelihood contribution of a random observation, and $\hat{I}$ its natural estimate under the null. The only difference with the usual case of an internal point of the parameter space is that right partial derivatives are used for the parameters $p_{j,k}$.

Theorem 4 in Andrews (2001) now implies that LR has the same asymptotic distribution as

$$Inf_{[\theta \in V_0]}\ q(\theta) - Inf_{[\theta \in V]}\ q(\theta) \tag{17}$$

with

$$q(\theta) = (\theta - Z)' J (\theta - Z), \ \ Z \sim N(0, J^{-1} I J^{-1}) \tag{18}$$

The asymptotic distribution of LR is thus be obtained by the following simulation procedure:

- plugging in the estimates $\hat{J}$ for J and $\hat{I}$ for I. (As in the usual ML case, I and J coincide under the null, so an asymptotically equivalent procedure is to use an estimate for only one of them.)
- generating multivariate normal draws of Z,
- solving the two quadratic programming problems in (17)for each draw,
- considering the thus obtained simulated distribution of the difference between the two minimum values.

References

- Abrevaya, J. and J. Hausman (1999), Semi-parametric estimation with mismeasured dependent variables: an application to duration models for unemployment spells, *Annales d'Economie et de Statistique*, 55-56, 243-276.

- Andrews, D. (1998), Hypothesis testing with a restricted parameter space, *Journal of Econometrics*, 84, 155-199.

- Andrews, D. (1999), Estimation when a parameter is on a boundary *Econometrica*, 67, 1341-1384.

- Andrews, D. (2001), Testing when a parameter is on the boundary of the maintained hypothesis, *Econometrica*, 69, 683-734.

- Bellemare, C., B. Melenberg and A. van Soest (2002), Semi-parametric models for satisfaction with income, *Portuguese Economic Journal*, 1, 181-203.

- Borjas, G.J. (1987), Self-Selection and the Earnings of Immigrants, *American Economic Review*, 77, 531-553.

- Charette, M. and R. Meng (1994), Explaining Language Proficiency, *Economics Letters*, 44, 313-321.

- Chernoff, H. (1954), On the distribution of the likelihood ratio *Annals of Mathematical Statistics*, 54, 573-578.

- Chiswick, B. (1991), Reading, speaking, and earnings among low-skilled immigrants, *Journal of Labor Economics*, 9, 149-170.

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

- Clark, A. and A. Oswald (1996), Satisfaction and comparison income, *Journal of Public Economics*, 61, 359-381.

- Das, M. (1995), Extensions of the ordered response model applied to consumer valuation of new products, CentER DP series No. 9515, Tilburg University.

- Das, M. and A. van Soest (1997), Expected and realized income changes: Evidence from the Dutch socio-economic panel, *Journal of Economic Behavior and Organization*, 32, 137-154.

- Douglas, S., K. Smith Conway and G. Ferrier (1995), A switching frontier model for imperfect sample separation information: with an application to labor supply, *International Economic Review*, 36, 503-527.

- Dustmann, C. (1994), Speaking fluency, writing fluency and earnings of migrants, *Journal of Population Economics*, 7, 133-156.

- Dustmann, C. and A. van Soest (2001), Language and the earnings of immigrants, *Industrial and Labor Relations Review*, forthcoming.

- Haerdle, W. and O. Linton (1994), Applied nonparametric methods, in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Volume IV, North-Holland, Amsterdam.

- Han, A.K. (1987), Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator, *Journal of Econometrics*, 35, 303-316.

- Hausman, J., J. Abrevaya and F. Scott-Morton (1998), Misclassification of a dependent variable in a discrete response setting, *Journal of Econometrics*, 87, 239-269.

- Horowitz, J. (1993), Semi-parametric estimation of a work-trip mode choice model, *Journal of Econometrics*, 58, 49-70.

- Horowitz, J. and W. Haerdle (1996), Direct semi-parametric estimation of single index models with discrete covariates, *Journal of the American Statistical Association*, 91, 1632-1640.

- Ichimura, H. (1993), Semi-parametric least squares (SLS) and weighted SLS estimation of single index models, *Journal of Econometrics*, 58, 71-120.

- Kerkhofs, M. and M. Lindeboom (1995), Subjective health measures and state dependent reporting errors, *Health Economics*, 4, 221-235.

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

- Lee, L.F. and R.H. Porter (1984), Switching regression models with imperfect sample information with an application on cartel stability, *Econometrica*, 52, 391-418.

- Powell, J. (1994), Estimation of semiparametric models, in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Volume IV, North-Holland, Amsterdam, 2444-2523.

- Powell, J., J. Stock, and T. Stoker (1989), Semi-parametric estimation of index coefficients, *Econometrica*, 57, 1403-1430.

- Shapiro, A. (1985), Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints, *Biometrica*, 72, 133-144.

- Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

- Szroeter, J. (1997), Standard normal tests of multiple inequality constraints: first-order large sample theory, mimeo, University College London.

- Terza, J. (1985), Ordered probit: a generalization, *Communications in Statistics – Theory and Methods*, 14, 1-11.

Table 1: Variable Definitions and Sample Statistics

Variable	Code	Mean	Std Dev
Speaks English slightly or not at all	SPF=1	26.85	–
Speaks English fairly well	SPF=2	26.24	–
Speaks English very well	SPF=3	46.91	–
Age (in years)	age	42.38	14.27
Years since Migration	ysm	19.58	9.35
Country of Birth: African	afroas	22.71	–
Country of Birth: Bangladesh	bangladesh	17.88	–
Country of Birth: Indian	indian	29.98	–
Country of Birth: Pakistan	pakistan	29.44	–
Minority Concentration (%)	conc index	16.20	15.20

Source: Fourth National Survey on Ethnic Minorities (FNSEM), 1471 observations

Table 2: Semi-parametric Estimation Results

bandwidth	SLS; $h = 1.5470^1$		WSLS; $h = 0.1.5470^1$		SLS; $h = 0.7563^2$	
	coeff.	st. error	coeff.	st. error	coeff.	st. error
ysm	0 .9634	—	0.9634	—	0.9634	—
age	-0 .9617	0.1071	-0.9923	0.1201	-0.9840	0.1094
conc. index	-25.1826	6.4955	-26.5351	6.4125	-20.2509	6.1338
afroas	4.2826	0.9689	4.1344	0.9171	4.1996	0.9296
pakistan	-4.2943	0.8178	-4.4190	0.7726	-3.7830	0.7825
bangla desh	-4.3825	0.8960	-4.6807	0.8785	-4.1162	0.8716
age sq	0.0061	0.0010	0.0063	0.0010	0.0064	0.0010
ysm sq	-0.0140	0.0011	-0.0139	0.0010	-0.0147	0.0010
conc ind sq	32.3767	9.2184	34.0762	8.8466	24.2073	8.6987
ysm * conc. in	-0.0655	0.1420	-0.0656	0.1524	-0.0872	0.1460

Notes:

1: $1.06\sqrt{\hat{V}(x'\hat{\beta})}n^{-0.2}$ (Silverman's rule of thumb)

1: $0.53\sqrt{\hat{V}(x'\hat{\beta})}n^{-0.2}$

Table 3: Estimation Results Parametric Models

	Ordered		Misclass.		Random		Misclass.	
	Probit		Model		Thresholds		Random Th.	
	coeff.	st. err.	coeff.	st. err.	coeff.	st. err.	coeff.	st. err.
Constant	28.4750	3.3542	55.6333	13.2788	25.5461	3.7574	84.6781	24.8929
ysm	0.9634	0.1342	2.0196	0.4168	1.0323	0.1429	3.5985	1.1871
age	-0.8258	0.1411	-1.6342	0.4246	-0.8817	0.1555	-3.1134	0.9219
conc. index	-34.0386	7.7247	-64.1300	17.8579	-36.6716	8.5027	-119.8281	42.0837
afroas	3.7520	0.9326	7.9896	2.5771	4.1600	1.0014	15.5292	6.0728
pakistan	-6.0868	0.8292	-9.6401	2.2333	-6.2717	0.9171	-18.0508	6.0543
bangla desh	-6.0094	0.9649	-10.0340	2.4232	-6.1796	1.0397	-18.7859	6.3406
age sq	0.0041	0.0015	0.0082	0.0034	0.0043	0.0015	0.0169	0.0060
ysm sq	-0.0152	0.0032	-0.0314	0.0079	-0.0164	0.0032	-0.0548	0.0205
conc ind sq	48.2251	10.8978	93.2072	24.1164	50.5726	11.7550	170.1757	60.1355
ysm * conc. in	-0.3918	0.2307	-0.6923	0.4164	-0.3764	0.2509	-1.0383	0.7712
m_2	8.7001	0.3913	23.2845	5.3590	-4.4034	0.2440	16.8234	7.9191
σ_1					5.2893	2.1503	10.2876	8.1447
σ_2					0	—	23.4378	10.6779
Prob 2 if 1			0	—			0	—
Prob 3 if 1			0.1029	0.0458			0.13891	0.0505
Prob 1 if 2			0.2725	0.0473			0.37238	0.5401
Prob 3 if 2			0.2450	0.0570			0.07365	1.2068
Prob 1 if 3			0.0095	0.0146			0	—
Prob 2 if 3			0.1042	0.0381			0.09293	0.0335
Log-Likelihood	-1317.646		-1309.332		-1315.858		-1308.278	

Table 4: Marginal Effects of Minority Concentration; Parametric Models

Quantile of Minority	Ordered		Misclass.		Random		Misclass.	
Concentration	Probit		Model		Thresholds		Random Th.	
	Effect	st. err.	Effect	st. err.	Effect	st. err.	Effect	st. err.
P(fairly or very fluent)								
at 1st quartile	-0.8811	0.0972	-0.1857	0.1878	-0.8982	0.1055	-0.2915	0.5199
at median	-0.9255	0.1140	-0.3733	0.2614	-0.9610	0.1287	-1.0777	0.9529
at 3rd quartile	-0.7895	0.0928	-0.6286	0.2422	-0.8349	0.1099	-2.3014	0.9506
P(very fluent)								
at 1st quartile	-1.5310	0.1983	-2.8962	0.5999	-1.6076	0.2158	-4.0193	2.8528
at median	-1.3293	0.1638	-2.2688	0.4407	-1.4033	0.1830	-4.3820	1.5748
at 3rd quartile	-0.8728	0.0860	-1.0649	0.1985	-0.9214	0.0990	-1.8027	1.6470

Explanation: marginal effect of an increase of ethnic concentration by 1 %-point on the probability (in %-points) of speaking English fairly or very well (top panel) or very well (bottom panel)

Figure 1. Link Function Semiparametric Model

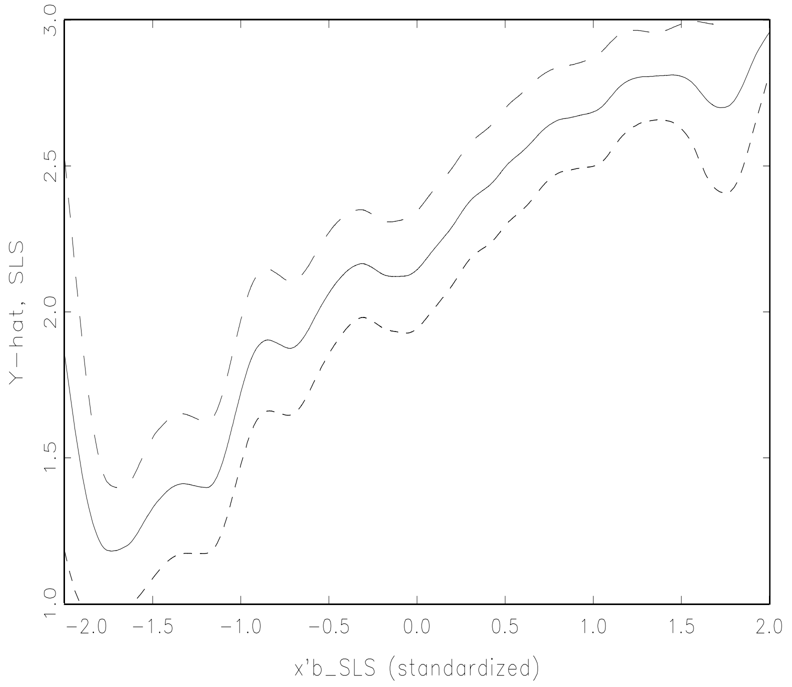


Figure 2. P(Z=3|X) Misclassification Model and Ordered Probit

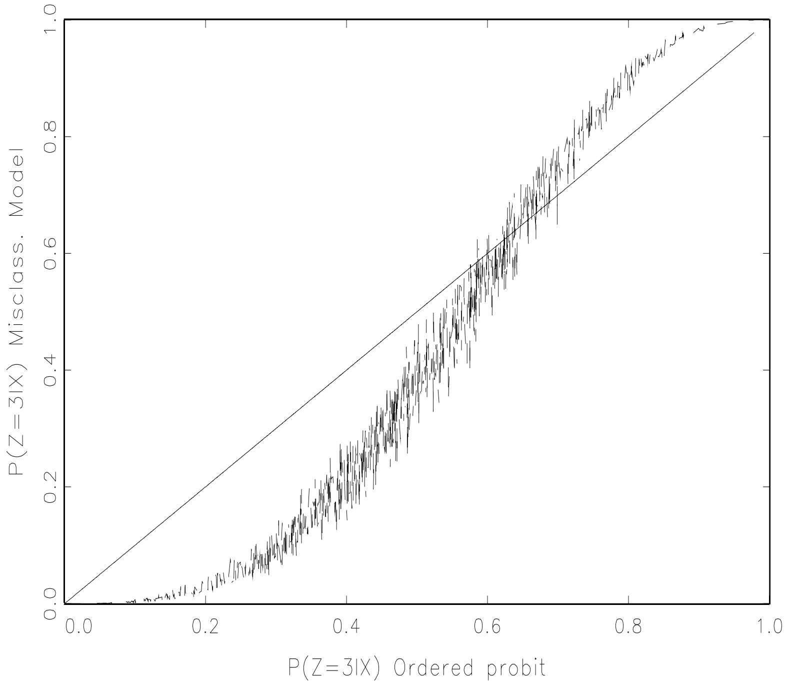


Figure 3. P(Y=3IX) Misclassification Model and Ordered Probit

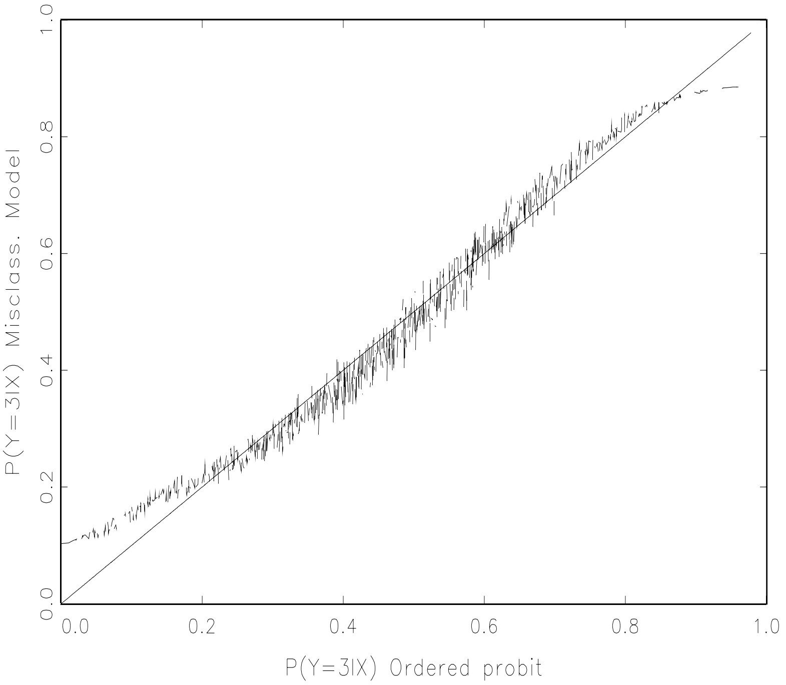
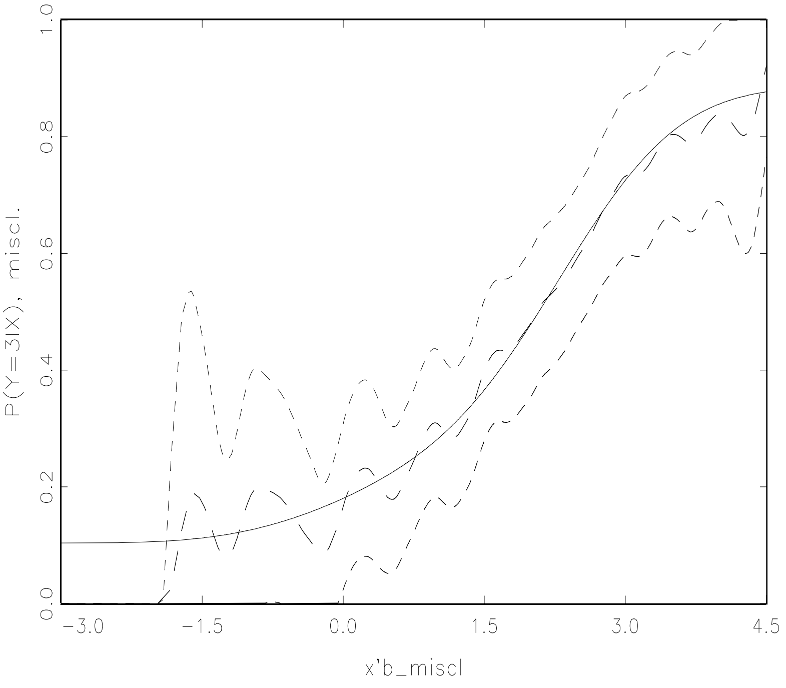


Figure 4. Specification Test Misclassification Model


P(Y=3IX), miscl.
x'b_miscl