DISCUSSION PAPER SERIES

IZA DP No. 218

Parametric and Semiparametric Estimation in Models with Misclassified Categorical Dependent Variables

Christian Dustmann
Arthur van Soest

November 2000

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Parametric and Semiparametric Estimation in Models with Misclassified Categorical Dependent Variables

Christian Dustmann
University College London, IFS, CEPR, London and IZA, Bonn

Arthur van Soest
Department of Econometrics, Tilburg University and IZA, Bonn

Discussion Paper No. 218
November 2000

IZA

P.O. Box 7240
D-53072 Bonn
Germany

Tel.: +49-228-3894-0
Fax: +49-228-3894-510
Email: iza@iza.org

IZA Discussion Paper No. 218
November 2000

ABSTRACT

Parametric and Semiparametric Estimation in Models with Misclassified Categorical Dependent Variables[*]

We consider both a parametric and a semiparametric method to account for classification errors on the dependent variable in an ordered response model. The methods are applied to the analysis of self-reported speaking fluency of male immigrants in Germany. We find that a parametric model which explicitly allows for misclassification performs better than a standard ordered probit model and than a model with random thresholds. We find some substantial differences in parameter estimates and predictions of the different models.

JEL Classification: C14, C35, J15

Keywords: Immigrants, speaking fluency, misclassification error

Christian Dustmann
Department of Economics
University College London
Gower Street
London WC1E 6BT
UK
Tel.: +44 20 7679 5212
Fax: +44 20 7916 2775
Email: c.dustmann@ucl.ac.uk

[*] We are grateful to John Rust, two anonymous referees, Bertrand Melenberg, Frank Windmeijer and Marcel Das for useful comments, and to Joel Horowitz and Marlene Mueller for providing us with their Gauss programs.

1 Introduction

Many empirical studies in economics and other social sciences are concerned with the analysis of ordered categorical dependent variables. Categorical data can be affected by misclassification error. This is especially the case if the categorical assignment is based on subjective self-reported evaluations, as used in many empirical analyses. Examples are studies which analyze data on job satisfaction (see, for example, Clark and Oswald (1994)), satisfaction with health (Kerkhofs and Lindeboom (1995)), or future expectations of household income (Das and Van Soest (1997)).

In applied work, nonlinear parametric limited dependent variable models are typically used for analysis categorical dependent variables. Misclassification can in this case lead to seriously biased parameter estimates, even if the parametric model correctly specifies the unobservable "true" categorical variable. To deal with this problem, estimators have been proposed for parametric binary choice models which correct for misclassification by explicitly incorporating the 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 estimate the two misclassification probabilities that one regime is observed given that the other regime is active. They then use these probabilities to correct the estimates of the price equations in each regime for the misclassification errors. Hausman et al. (1998) estimate binary choice models for job change. In their parametric models, they find significant probabilities of misclassifying in both directions. They also estimate a semiparametric model, and find that the semiparametric estimates are similar to the parametric estimates allowing for misclassification.

In this paper, we focus on self–reported evaluations. The term "misclassification" requires some discussion. We implicitly assume that there is some (unobserved) "true" classification scale, and some (unobserved) "true" cutoff points on this scale which determine what someone's reported evaluation should look like. Misclassification can then be modeled using estimators as those mentioned above, designed for objective outcomes, where the misclassification probability depends on the "true" (discrete) outcome. However, people may deviate from the "true" cut off points. Accordingly, the thresholds may not be objectively determined, but (implicitly) chosen by the respondent. In this case it can be argued that the thresholds in the ordered response model may vary with observed and unobserved characteristics of the respondent. Terza (1985) introduces a model with unobserved thresholds which vary with observed respondent characteristics. Das (1995) extends this model by treating the thresholds as random variables, i.e. to allow them to depend on unobserved characteristics.

In this paper, we consider parametric and semiparametric models for ordered categorical dependent variables with more than two outcomes. Our first parametric model generalizes the binary response models by Lee and Porter (1984) and Hausman et al. (1998) by incorporating probabilities of misclassifying outcomes in a model with more than two categories. Other than in the binary case, the identification of the misclassification probabilities for the intermediate categories relies on distributional assumptions. Accordingly, parametric estimates of the effects of the 'true' outcome may be sensitive to these distributional assumptions. Semiparametric estimation may therefore be a useful alternative. Our second parametric model is an extension which allows for random thresholds (see Das (1995)). We show that this changes the ordered model for three outcomes into a bivariate probit model. Misclassification probabilities can be included in this model in very much the same way as in the model with fixed thresholds.[1]

Under the assumption that misclassification probabilities or random thresholds do not vary with individual characteristics, we show that, if the misclassification probabilities are not too large, the parametric models are special cases of a single index model satisfying a weak monotonicity condition. This model can be estimated using the semiparametric technique of Horowitz and Haerdle (1996), combining average derivatives

[1] We owe the suggestion to use random thresholds to an anonymous referee.

estimation with a GMM type of estimator to take account of discrete regressors.

We apply both parametric and semiparametric estimators to data on self–reported speaking fluency of male immigrants to Germany. A growing literature is concerned with the determinants of language fluency, and its effects economic performance in the host countries (see, for instance, McManus, Gould and Welch (1983), Rivera-Batiz (1990), Chiswick (1991), Chiswick and Miller (1995), and Dustmann (1994)). Nearly all studies are based on self-reported language categorizations. As discussed above, it is likely that this type of data is even more affected by misclassification than objective variables, such as the job change variable in Hausman et al. (1998). Besides errors resulting from, for instance, misunderstanding of survey questions, responses of this type may be misclassified because of heterogeneity in the underlying subjective standards.[2]

The results of our analysis show that allowing for misclassification errors has some effect on the estimates of the parameters in the speaking fluency equation. The parametric model which allows for misclassification is a clear improvement to the standard ordered probit model. The estimated probabilities of misclassification into the extreme categories are large. A likelihood ratio test (accounting for the fact that the null hypothesis fixes parameters on the boundary of the parameter space) shows that the null hypothesis that all misclassification probabilities are zero is clearly rejected. On the other hand, adding random thresholds to the model does not lead to significant improvement.

A formal test of the parametric models against a semiparametric alternative is proposed, based upon uniform confidence bands of the nonparametric regression function of the dependent variable on the parametrically estimated index. We find for both the

[2] Dustmann and van Soest (1998) show that misclassification is substantial in this data. They compare answers to the same survey questions on self-reported speaking fluency given by the same individuals at different points in time. They find that, under the assumption that a deterioration of language capacity is not possible, more than one fourth of the total variance in the language indicator is due to misclassification. This number is a lower bound, since it accounts only for classification errors which are not time persistent.

standard ordered probit model and the model with misclassification probabilities that the parametric regression function is contained in the confidence bands, so that the parametric models cannot be rejected.

The paper is organized as follows. In section 2, we present the models and their estimators. In section 3.1, we briefly describe the data for our empirical application. Parametric and semiparametric estimates are presented in Section 3.2. In Section 3.3, we compare predictions of the two parametric models and the semiparametric model, and discuss some specification tests of the parametric models. Section 4 concludes.

2 Categorical Data and Misclassification

For simplicity, we assume that the dependent variable is observed on an ordinal scale with three levels, coded 1, 2 and 3. Most of our results extend straightforwardly to the case of more than three categories, but the parametric models will lead to more intricate computations for the likelihoods and more auxiliary parameters. Starting point is the standard ordered probit model, not allowing for misclassification errors. It assumes observed categorical information is related to an underlying latent index y^* as follows (the index indicating the individual is suppressed):

$$y^* = x'\beta + u, \tag{1}$$

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

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

Here x is a vector of explanatory variables including a constant term, β is the vector of parameters of interest, and u 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. A normalization of the scale has to be added for identification. This will be discussed below. Throughout, we assume that the observations (y, x) 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 may be substantial if some observations on the endogenous variable suffer from misclassification. They propose a generalization of the binary probit model to take account of misclassification errors. We extend their framework for the binary probit model to the ordered probit model.

Assume that the individual is observed to be in category y, but that the (unobserved) true category is z, which is related to the latent variable y^* as in the ordered probit case:

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

The probabilities of misclassification are given by:

$$\text{Prob}(y = j | z = k) = p_{k,j}, \quad j, k = 1, 2, 3, j \neq k, \tag{5}$$

where $p_{k,j}$ is the probability that an observation which belongs in category k is classified in category j. If $p_{k,j} = 0$ for all j, k with $j \neq k$, then there is no misclassification and the model simplifies to the ordered probit model.

In a model with three categories, there are six misclassification probabilities $p_{k,j}$. Thus compared to the standard ordered probit for three categories, this model has six additional parameters.

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 $x'\beta$ is the whole real line, i.e. as long as observations with very low and very high values of $x'\beta$ occur with nonzero probability. The probabilities of misclassification are then given by:

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

If $p_{0,1}$ and $p_{1,0}$ do not depend on x and if u is independent of x, the model satisfies the single index property: $E\{y|x\}$ depends on x via $x'\beta$ 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 condition guarantees that $E\{y|x\}$ increases with $x'\beta$. Accordingly, the sign of β is also identified, and it then follows from (5) that the $p_{0,1}$ and $p_{1,0}$ are nonparametrically identified.

For the ordered probit case with three categories coded 1, 2 and 3, and with misclassification probabilities, we obtain

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

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

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

This condition is satisfied for small enough values of the misclassification probabilities. A sufficient condition for (8) is given by Abrevaya and Hausman (1998):

$$p_{1,1} > p_{2,1} > p_{3,1} \text{ and } p_{3,3} > p_{2,3} > p_{1,3} \tag{9}$$

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 on 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. They will generally be inconsistent if the errors are not normally distributed.

2.2 Threshold Variation across Respondents

With self-assessed categorical survey data, no information is given to the respondents on how to construct their score y^* on a continuous scale, or which cut-off points to choose between the discrete outcomes. Thus it may well be that both the underlying latent score and the cut-off points contain individual heterogeneity. Our data requires us to assume that deviations between true scores and scores used by the respondents do not vary systematically with observed characteristics. Unobserved heterogeneity in the way respondents form their scores is picked up by the error term u in (1).

An extension of the model with explicit misclassification errors is to allow for heterogeneity in the threshold values used by the respondents. This is intuitively attractive, since it allows that even if two respondents have the same latent speaking fluency y^*, and are perfectly well aware of that fluency, they still might give different answers on the ordinal scale, since the survey questions leave room for the respondent's own interpretation of what is good, reasonable, or bad speaking fluency.

Ordered probit models with category bounds varying 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. The subjective nature of the data suggests that the latter framework is more relevant, and we therefore set up the model for three categories, following Das (1995).

We first discuss the model without misclassification errors. The model specification is then as follows.

$$y^* = x'\beta + u, \tag{10-a}$$

$$m_j^* = z'\gamma_j + v_j \quad j = 1, 2 \tag{10-b}$$

$$y = 1 \quad \text{if} \quad y^* \leq \min(m_1^*, m_2^*) \tag{10-c}$$

$$y = 2 \quad \text{if} \quad \min(m_1^*, m_2^*) < y^* \leq \max(m_1^*, m_2^*) \tag{10-d}$$

$$y = 3 \quad \text{if} \quad y^* > \max(m_1^*, m_2^*) \tag{10-e}$$

$$u, v_1 \text{ and } v_2 \text{ are independent of each other and of } x \tag{10-f}$$

$$u \sim N(0, \sigma^2), \;\; v_j \sim N(0, \sigma_j^2), \;\; j = 1, 2. \tag{10-g}$$

The problem with random thresholds in the ordered response model is that the ordering in the thresholds cannot be determined a priori: with positive probability, m_1^* exceeds m_2^*, and the model with categories $(-\infty, m_1^*)$, (m_1^*, m_2^*), and (m_2^*, ∞) is not well-defined. Das(1995) solves this problem by using the ordered thresholds instead of the original ones, and we follow his approach. In the case with three categories, this boils down to replacing m_1^* by $\min(m_1^*, m_2^*)$ and m_2^* by $\max(m_1^*, m_2^*)$. The probabilities of three outcomes ($y = 1$, $y = 2$ or $y = 3$) for this model can be rewritten as follows.

$$P(y = 1) = P(u - v_1 < z'\gamma_1 - x'\beta \text{ and } u - v_2 < z'\gamma_2 - x'\beta) \tag{11-a}$$

$$\begin{aligned} P(y = 2) \;\; = \;\; & P(u - v_1 < z'\gamma_1 - x'\beta \text{ and } u - v_2 > z'\gamma_2 - x'\beta) + \\ & + P(u - v_1 < z'\gamma_1 - x'\beta \text{ and } u - v_2 > z'\gamma_2 - x'\beta) \end{aligned} \tag{11-b}$$

$$P(y = 3) = P(u - v_1 > z'\gamma_1 - x'\beta \text{ and } u - v_2 > z'\gamma_2 - x'\beta) \tag{11-c}$$

This shows that this model is a bivariate probit model which does not distinguish between the two regimes leading to outcome $y = 2$. It also makes clear that some normalizations are needed to identify the model. The scale is normalized in the same way as in the other models, by setting one of the slope coefficients in β to 1 or -1. To identify the location, we set $\gamma_1 = -\gamma_2$. While this is numerically equivalent to several other normalizations, we chose this normalization because it treats good and bad fluency symmetrically. Moreover, it implies that an increase of $|z'\gamma_1|$ induces an increase in the probability of giving the intermediate answer, allowing to interpret $|z'\gamma_1|$

as the tendency to respond in the intermediate category.[3]

The covariance structure of the bivariate probit model is given by $V(u - v_1) = \sigma^2 + \sigma_1^2$, $V(u - v_2) = \sigma^2 + \sigma_2^2$, and $Cov(u - v_1, u - v_2) = \sigma^2$. Thus the three variances of u, v_1 and v_2 are exactly identified. Relaxing 10-f and allowing for correlations between the three error terms, would lead to an unidentified model.

Note that this model can also be interpreted somewhat differently. The respondent can be seen as evaluating his own speaking fluency twice (once based upon $-z'\gamma_1 - v_1 + x'\beta + u$, and once upon $-z'\gamma_2 - v_2 + x'\beta + u$. If both evaluations are positive, the answer $y = 3$ (good or very good) is given. If both are negative, $y = 1$ (bad or very bad) is reported. If one evaluation is positive and the other is negative, $y = 2$.

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' outcomes z are given by (11-a), (11-b) and (11-c), (with y replaced by z). The probabilities of the reported outcomes given the true outcomes are then again given by (5).

2.3 A Semiparametric 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, semiparametric estimation seems a good alternative.

Let us consider the Hausman et al. (1998) model with fixed thresholds and misclassification probabilities. The conditional mean of the observed categorical variable y in model (1) - (5) is given by (7). Accordingly, the mean of y conditional on x depends on x only through the index $x'\beta$. Therefore, (1)-(5) is a special case of the single index model given by

[3] Replacing γ_1 by $-\gamma_1$ (i.e., interchanging γ_1 and γ_2,) does not change the probabilities. The sign of γ_1 can be identified by imposing the additional constraint that $z'\gamma_1$ is more often the lower bound than the upper bound, i.e. $z'\gamma_1 \leq 0$ for at least half of the observations.

$$E\{y|x\} = H(x'\beta)\,, \tag{12}$$

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

$$u \text{ is independent of } x\,, \tag{13}$$

we get the following expression instead of (7):

$$\begin{aligned} E\{y|x\} \;&=\; 2 - p_{2,1} + p_{2,3} - G(m_1 - x'\beta)(1 - p_{1,2} - p_{2,1} + p_{2,3} - 2p_{1,3}) + \\ &\quad\; [1 - G(m_2 - x'\beta)](1 - p_{3,2} - p_{2,3} + p_{2,1} - 2p_{3,1})\,, \end{aligned} \tag{14}$$

where G is the distribution function of the error term u ($G(t) = P[u \le t]$).

Again, the right-hand side depends on x only through $x'\beta$, so that (1), (2), (4), (5) and (13) lead to the same single index model (12). The link function H is then given by G in (14). The crucial assumption here is that the misclassification probabilities in (4)- (5) do not depend on x. This is the typical identifying assumption in this type of literature, used by Hausman et al. (1998), Lee and Porter (1984), but also in other applications such as Douglas et al. (1995). Such an assumption can only be avoided if a completely different measurement can be used as a benchmark, such as, in our empirical example, objective measurement of language proficiency (see Charette and Meng (1994)).

It is straightforward to extend (14) to the following results:

Proposition:

(a) If (1), (2), (5) and (6) are satisfied, then $E\{y|x\} = H(x'\beta)$ for some function H, i.e. the model is a single index model.

(b) If, in addition,

$$p_{1,2} + p_{2,1} - p_{2,3} + 2p_{1,3} \le 1$$

$$\text{and} \tag{15}$$

$$p_{2,3} + p_{3,2} - p_{2,1} + 2p_{3,1} \leq 1,$$

then H can be chosen non decreasing.

(c) If, moreover, (15) holds with strict inequalities and u has a continuous distribution with support (a, b), then H can be chosen strictly increasing and differentiable with positive derivative on the interval $(-b + m_1, -a + m_2)$.

A similar expression to (14) can still be derived from the extension of the model which allows for random cutoff points. Under the additional assumption that the variation in the cutoff points is independent of observed respondent characteristics, the model with random cutoff points is a single index model and the proposition remains valid.

We have shown that the models discussed above are all special cases of the general single index model (12) 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 estimators for β in this model have been discussed in the literature, under varying assumptions on the distribution of the explanatory variables x and regularity conditions on the link function H. Ichimura (1993) derives an asymptotically normal root n consistent estimator based upon nonlinear least squares combined with nonparametric estimation of H. This estimator has the drawback that it is computationally burdensome, since it 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. Hausman et al. (1998) report that no convergence problems occurred in their Monte Carlo experiments or their empirical work. Our experience, however, is different: we ran into convergence problems, possibly due to the comparatively large number of explanatory variables.

Attractive from a computational point of view is the class of average derivative estimators or weighted average derivative estimators (see, for example, Powell et al.

(1989)). These estimators require the distribution of x to be absolutely continuous, and are therefore not directly applicable to our empirical example. Horowitz and Haerdle (1996), however, have recently developed an estimator which builds upon a weighted average derivative estimator for the slope parameters (up to a scale normalization) of the continuous explanatory variables as a first step. The parameters of the discrete explanatory variables are estimated in a second step. Their estimator is consistent and asymptotically normal for all slope parameters (up to a normalizing scale parameter). Horowitz and Haerdle also show how to estimate the asymptotic covariance matrix consistently. The estimator does not require numerical optimization and is computationally very convenient. On the other hand, it requires the choice of a kernel (in the first step) and several smoothness parameters (in both steps). Procedures for choosing the smoothness parameters in the second step in an optimal way are not available, so that some ad hoc choices cannot be avoided. We will apply the Horowitz and Haerdle estimator using their kernel but using various values of the smoothness parameters.

One of the regularity conditions for the Horowitz and Haerdle estimator is a weak monotonicity condition on the link function H: H has to be monotonically increasing on some nonempty interval with a priori specified range. In the (parametric) models with misclassification probabilities, this condition is satisfied if the misclassification probabilities satisfy (15) with strict inequalities. Thus this regularity condition does not invalidate the claim that the model is more general than the parametric single index models.

We briefly sketch the idea of the weighted average derivative estimator and the extension of Horowitz and Haerdle. Details can be found in Powell et al. (1989) and Horowitz and Haerdle (1996). For continuous x, the weighted average derivative is given by

$$\delta = E\{f(x)\frac{\partial E\{y|x\}}{\partial x}\} = -2E\{y\frac{\partial f}{\partial x}\}\,. \tag{16}$$

Here $f(x)$ is the density of x. It can be estimated by a differentiable nonparametric kernel regression estimator $\hat{f}(x)$. The derivative of this function estimate is a nonparametric estimator of $\frac{\partial f}{\partial x}$. According to (16), an estimate of δ is obtained as -2 times the sample mean of the $y_i \frac{\partial \hat{f}(x_i)}{\partial x}$. Powell et al. (1989) show that this weighted average derivative estimator is consistent for δ, and derive its limit distribution (under appropriate regularity conditions).

In the single index model (12), we have

$$\delta = E\{f(x)G'(x'\beta)\}\beta \,. \tag{17}$$

Hence, the vectors δ and β are identical up to a scale factor. The weighted average derivative estimator can therefore be used to estimate β up to scale. By means of normalization, one of the slope coefficients is set to 1 or -1, so that the others are identified.

Now consider the case with both continuous and discrete regressors. Denote them by x and z, respectively, and write the single index model as

$$E\{y|x,z\} = H(x'\beta + z'\alpha) \,. \tag{18}$$

Horowitz and Haerdle (1996) partition the sample into subsamples with given values of z. Within each subsample, $z'\alpha$ is constant, and the model is a single index model in the continuous variables x only. This gives a consistent weighted average derivative estimator for β (which does not include the constant, and has one coefficient normalized to 1 or -1) for each subsample. Horowitz and Haerdle obtain a consistent but more efficient estimator for β by combining these estimates, using minimum distance (i.e., they take the weighted average of the separate estimates, using the inverse of their estimated covariance matrices as weights).

To derive an estimator for α, let z_1 and z_2 be two different values of z corresponding to cells 1 and 2 in the partition, and let $H_1(x'\beta)$ and $H_2(x'\beta)$ be the within cell link functions. Thus $H_i(x'\beta) = H(x'\beta + z_i'\alpha)$, $i = 1, 2$. This gives a relation between H_1 and H_2 which is used by Horowitz and Haerdle (1996) to derive a condition which should be satisfied by α. Assume H is monotonically increasing on an interval with range $[c_0, c_1]$ (this is the weak monotonicity condition referred to above). Define $h_i(t) = max[c_0, min[H_i(t), c_1]]$ $(i = 1, 2)$. Then it is easy to see from a graph of h_1 and h_2, and straightforward to prove using some algebra, that

$$\int_{-\infty}^{\infty} [h_2(t) - h_1(t)]dt = (c_1 - c_0)(z_2 - z_1)'\alpha\,. \tag{19}$$

This yields a moment restriction on α. Plugging in estimates of the link functions H_1 and H_2 yields an estimate of the left hand side of (19) and transfers (19) into a sample moment condition. Horowitz and Haerdle combine such sample moment conditions for different pairs z_1 and z_2, and thus derive a GMM type estimator for α. The estimator not only depends on the nonparametric estimators used for estimating β and the link functions, but also on the choice of c_0 and c_1. It is clear that these have to be chosen such that $[c_0, c_1]$ is contained in the range of H, but it is not clear what the optimal choice is, since the gains of using a larger interval $[c_0, c_1]$ and thus more observations, should be compared to the loss due to inaccurate estimation of the tails of H.

A final remark concerns the chosen cardinal scale of our observed dependent variable y, the outcomes of which we coded by 1, 2 and 3. If we change the coding to e.g. 1, 2 and 4, this leads to a different link function, and to a different single index estimator. The link with the parametric model through the monotonicity condition also changes somewhat, since (15) will change. All single index estimators obtained with different coding will be consistent (under the appropriate assumptions, including the monotonicity condition), but it is not clear which one is most efficient. We do not pursue this issue and only consider the coding 1, 2 and 3.

3 Application

3.1 Data and Variables

We apply both the parametric and the semiparametric estimator to data on speaking fluency of male immigrants in West–Germany. The data are drawn from the first (1984) wave of the German Socio-Economic Panel (GSOEP). The (GSOEP) contains a boost sample of households with a foreign born head from Turkey, Yugoslavia, Italy, Greece and Spain, which were the typical emigration countries to Germany in the 1950s - 1970s. In the first wave (1984), this boost sample contains 1500 households, which are representative for the targeted immigrant population. The foreign born individuals are asked a number of specific questions regarding their economic and social circumstances, as well as their language proficiency. For our analysis, we use only males who where older than 15 years at immigration. All survey questions are asked in the immigrant's home country language (see Dustmann (1994) for more details).

Our dependent variable is a self–reported indicator of speaking fluency, reported on a five point scale. Due to the small number of observations in the extreme categories, we have transformed this information into a three level variable: $y_i = 3$ if individual i's fluency in the host country language is good or very good; $y_i = 2$ if fluency is reported to be on an intermediate level; $y_i = 1$ if fluency is bad or very bad.

The choice of explanatory variables is motivated by human capital theory. We use years since migration (YSM), age at entry (AGEENT), schooling (SCH), after school education (EDU), and dummy variables indicating the immigrant's nationality (T, Y, I, G) as regressors. The time of residence in the host country is a measure of exposure to the host country language, and we would expect individuals to improve their language fluency with residence. Age at entry is expected to affect language fluency negatively for two reasons: individuals who are older at entry may have a shorter pay off period for investments into language capital; and individuals' ability to learn a new language may decrease with age. Individuals with higher levels of education should find it more easy to learn a new language, since higher education may reflect higher ability, and since education increases the productivity of accumulating language capital.[4]

[4] For instance, individuals who know how to read and to write learn a new language in a more systematic way than individuals who lack these skills. Also, the better educated may be more efficient in the acquisition of further knowledge.

Table 1: Variable Definitions and Sample Statistics

Variable	Variable	Mean	Std Dev
Speaking Fluency bad/very bad	SPF=1	0.208	0.406
Speaking Fluency intermediate	SPF=2	0.408	0.492
Speaking Fluency good/very good	SPF=3	0.384	0.487
Age at Entry	AGEENT	27.67	7.06
Years since Migration	YSM	15.16	5.40
Years of Schooling[2]	SCH	1.29	2.68
Years of job specific education[2]	EDU	1.28	2.31
Country of Birth: Turkey	T	0.30	0.46
Country of Birth: Yugoslavia	Y	0.21	0.41
Country of Birth: Greece	G	0.14	0.35
Country of Birth: Italy	I	0.21	0.41
Country of Birth: Spain	S	0.14	0.35

[2] after the age of 14

Source: German Socio-Economic Panel 1984; 1185 observations

We have also included country of origin dummies. Immigrants may be a self selected group. Since selection is determined by the economic conditions in home and host country, country of origin dummies may pick up level effects in the average ability level (see Borjas (1987)). Also, the relation between language proficiency and return migration may vary across the origin countries. Moreover, these dummies may reflect language distance and cultural differences, which affect the acquisition of language capital. Finally, origin dummies may capture enclave effects, if individuals from different origins have different propensities to live in ethnic communities.

Definitions and summary statistics of all the variables can be found in Table 1. The first four explanatory variables are measured in years and can be interpreted as continuous variables. The four dummy variables for nationalities, however, are obviously discrete.

3.2 Results Parametric Models

The estimation results for the parametric models are presented in Table 2. As explained above, one of the slope parameters has to be normalized to 1 or -1 in the semiparametric model. To make the parametric models comparable with the semiparametric model, we have used the same normalization in the parametric models. We have set the coefficient of AGEENT equal to -1. This variable has a significant negative effect and the largest absolute t-value if the parametric models are estimated with the usual normalization $\sigma = 1$.

The first columns give the results of the standard ordered probit model. They are in accordance with other studies on language fluency. Years since migration, years of schooling and years of job specific education all have the expected positive effect on speaking fluency. The country dummies indicate that both the Spanish base group and Turkish workers have significantly lower probabilities to be fluent in German than the other groups.

In the second set of estimates, the misclassification probabilities are explicitly included. Since these probabilities are by definition nonnegative, standard t-tests or likelihood ratio tests on $p_{k,j} = 0$ are inappropriate (see Shapiro (1985), for example). 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 three of them. This suggests that adding the probabilities of misclassification is indeed 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 (1998b). The LR test statistic does not have the usual chi squared distribution under the null, due to the one sided nature of the test and due to the fact that under the null, the parameter vector is not in the interior of the parameter space. Andrews (1998b) 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

Table 2: Parametric Models

	Ordered Probit		Misclass. Model		Random Thresholds		Misclass. Random Th.	
	Coef	St er	Coef	St er	Coef	St er	Coef	St er
constant	34.318	3.013	21.453	6.339	20.615	3.062	15.940	4.503
T	-2.137	2.214	-1.663	2.104	-2.789	2.227	-2.874	2.108
Y	13.123	2.749	12.284	2.628	13.428	2.744	10.834	2.300
G	5.326	2.635	6.648	2.470	5.582	2.518	5.192	2.206
I	4.709	2.401	5.484	2.334	4.456	2.390	4.251	2.079
YSM	0.353	0.135	0.433	0.130	0.355	0.134	0.413	0.117
AGEENT	-1.000	—-	-1.000	—-	-1.000	—-	-1.000	—-
SCH	1.121	0.234	1.731	0.419	1.214	0.309	1.642	0.340
EDU	0.793	0.249	1.914	0.463	0.888	0.323	0.994	0.261
σ	19.952	1.887	10.340	4.466	17.846	1.906	4.170	11.845
m_2	25.011	2.450	12.467	11.484	12.009	1.251	8.647	3.381
$p_{1,2}$			0.225	0.124			0.253	0.142
$p_{1,3}$			0.156	0.063			0.095	0.110
$p_{2,1}$			0.069	0.267			0.057	0.339
$p_{2,3}$			0.119	0.455			0.363	0.168
$p_{3,1}$			0.036	0.023			0.031	0.019
$p_{3,2}$			0.227	0.069			0.236	0.052
σ_1					0.283	9.538	10.738	10.906
σ_2					14.709	4.600	0.000	—
Log-Likelihood	-1145.25		-1134.62		-1142.56		-1132.92	

explain in detail how this can be applied in our case. We find a critical 5% value of 8.6. Since the realization of the LR test statistic is 21.3, the null hypothesis is rejected. Thus, also a formal test confirms that allowing explicitly for misclassification errors improves the fit of the model significantly compared to the ordered probit model.

The estimates of $p_{2,1}$ and $p_{2,3}$ have particularly large standard errors, reflecting the problem that these are harder to identify. The estimated probabilities are small enough to satisfy the inequality conditions in (15). This implies monotonicity of the link function if the parametric model is written as a single index model, so that the monotonicity assumption required for the semiparametric estimator is fulfilled.

The qualitative effects of the regressors have not changed in this more general specification. Still, some of the estimated slope coefficients in the second model differ substantially from those in the ordered probit model. In particular, the effects of the educational variables have increased considerably. The standard deviation of the error term u has decreased by almost 50 percent. This is because part of the unsystematic variation in observed speaking fluency is now explained by classification errors. The estimate of the bound m_2 has changed accordingly.

The third panel is the model with random thresholds, without misclassification probabilities. We only present the results for the model where the variation in the thresholds is independent of the observed characteristics. A (standard) likelihood ratio test does not reject this model against the more general model at the 5% or the 10% level (LR test statistic 9.7; 10% critical value 13.4). The final two parameters are the estimated standard deviations of the thresholds. These estimates are rather inaccurate. One of them is virtually equal to zero, but the other one is not. A likelihood ratio test similar to the one discussed above (following Andrews, 1998b) rejects the ordered probit model against the model with random thresholds at the 5% level (LR test statistic 5.4; 5% critical value 5.0). The estimates of the slope parameters are close to those in the ordered probit model.

In the final columns, both misclassification probabilities and random thresholds are incorporated. The estimates of the misclassification probabilities are similar to those in the model with fixed thresholds, except for $p_{2,3}$. This estimate is implausibly large. As in the model with fixed thresholds, this estimate and the estimate of $p_{2,1}$ are very inaccurate. The estimates of the other $p_{j,k}$, however, remain quite accurate and the point estimates are close to their values in the model with fixed thresholds. This finding is in line with the fact that these are nonparametrically identified, while the other two are not. The estimates of the misclassification probabilities again satisfy the monotonicity conditions (15). A likelihood ratio test (following Andrews, 1998b) of this model against the previous one again rejects the hypothesis that all $p_{j,k}$ are zero at all conventional significance levels (test statistic 19.3; 5% critical value; 8.7). Most of the estimates of the slope parameters are very close to those in the model with fixed thresholds. The exception is the effect of job specific education, the estimate of which is closer to the ordered probit estimate.

The parametric results can be summarized as follows. Allowing for misclassification probabilities significantly improves the fit of the model, and changes some of the estimates of the slope parameters. On the other hand, it is hard to estimate some of the misclassification probabilities accurately, particularly those which we cannot identify nonparametrically. Allowing for random thresholds has less important consequences than allowing for misclassification probabilities. The estimates of the standard deviations of the thresholds are inaccurate, and the slope parameter estimates are generally close to those in the fixed threshold model. The hypothesis that the thresholds do not vary with observed regressors is not rejected, which gives some support in favor of a single index model.[5]

[5] Following a suggestion by one of the referees, we have also estimated a model with fixed thresholds and with nonzero misclassification probabilities for adjacent categories only. This is the second model with the restrictions $p_{1,3} = p_{3,1} = 0$. These restrictions are not rejected by an Andrews LR test (test statistic 4.8; 5% critical value 9.8). Like the model with six misclassification probabilities, this model significantly out performs ordered probit (Andrews LR test statistic 16.4; 5% critical value 7.9). It gives very similar slope coefficients as the unrestricted misclassification model, but some of

Given the data at hand, it will not be possible to disentangle misclassification errors from random thresholds or other sources of misspecification of the ordered probit model without making specific assumptions. If we are not interested in misclassification errors as such, but only in the effects of the regressors on speaking fluency, we can avoid these assumptions and use a semiparametric single index model.

3.3 Semiparametric Estimates

In Table 3, the semiparametric estimates using the estimator of Horowitz and Haerdle (1996) are presented. The constant term is not estimated and, as before, the coefficient of AGEENT is normalized to -1. Note that the sign of this coefficient is identified, due to the assumption that the link function is increasing. We find the same sign as in the parametric models.

There are no guidelines for choosing the optimal bandwidth h in the kernel regressions or for choosing the parameters c_0 and c_1 in estimating the parameters of the discrete variables. Starting from the Horowitz and Haerdle (1996) choices (after rescaling the dependent variable so that it has range in [0,1]), we performed some Monte Carlo simulations with several values.[6] In general, we found that the results were more sensitive to the choice of c_0 and c_1 than to the choice of h. To give the reader some idea about the sensitivity for these choices, we present two sets of results. Other choices led to similar conclusions, though sometimes with much larger standard errors.

The standard errors of the semiparametric estimates can be computed in two ways. First, the asymptotic distribution of the estimator as derived by Horowitz and Haerdle (1996) can be used. The second option is to use bootstrapped standard errors. Our Monte Carlo simulations suggest that the bootstrapped standard errors are closer to the true standard errors than the asymptotic standard errors. We therefore present

the misclassification probability estimates are rather different.

[6] Detailed results of these are available upon request from the authors.

Table 3: Semiparametric Estimation Results

	$h = 5$; $c_0 = 0.3$; $c_1 = 0.7$		$h = 3$; $c_0 = 0.2$; $c_1 = 0.8$	
	Coeff.	Bootst. s.e.	Coeff.	Bootst. s.e.
T	-4.05	2.11	-4.31	1.82
Y	8.92	2.16	7.00	1.95
G	5.32	2.10	3.47	1.92
I	3.58	2.33	0.73	2.04
YSM	0.212	0.118	0.232	0.147
AGEENT	-1.00	—	-1.00	—
SCH	0.849	0.205	0.853	0.259
EDU	1.258	0.374	1.541	0.442

Note: the Dependent variable is rescaled to values 0, 0.5 and 1 to make the scale comparable to the scale of a 0-1 variable used in Horowitz and Haerdle (1996); this is needed to make comparable choices of smoothness parameters h, c_0 and c_1.

the bootstrapped standard errors in Table 3.[7] The bootstrapped standard errors on the coefficients of the continuous variables are typically larger than the asymptotic standard errors, suggesting that the asymptotic standard errors underestimate the true ones. This seems a rather common finding in the semiparametric literature (cf. Horowitz (1993), for example). For the nationality dummies, however, some of the asymptotic standard errors are larger than the bootstrapped standard errors.

The coefficients have the same sign as in the parametric models. Their order of magnitude is in most cases also similar to that in the previous models. There are some differences, but in most cases, confidence intervals overlap. The effect of general schooling (compared to the effect of age at entry) is not as strong as in the parametric models with misclassification errors. The same holds for years since migration, which is no longer significant at the 5% level. On the other hand, the differences between semi-

[7] A straightforward bootstrapping procedure is used, resampling 500 new data sets with replacement from the original data set; the new data sets have the same number of observations (1185) as the original data set.

Table 4: Monte Carlo Results Semiparametric Estimator

	$h = 5$; $c_0 = 0.3$; $c_1 = 0.7$		$h = 3$; $c_0 = 0.2$; $c_1 = 0.8$	
	Bias	St. Deviation	Bias	St. Deviation
T	-0.411	2.130	0.588	1.874
Y	-4.571	2.489	-5.237	2.281
G	-1.353	2.526	-1.813	2.130
I	-1.172	2.301	-1.688	1.954
YSM	-0.151	0.128	-0.060	0.178
AGEENT	—-	—	—-	—-
SCH	-0.534	0.189	-0.539	0.292
EDU	-0.209	0.281	-0.093	0.428

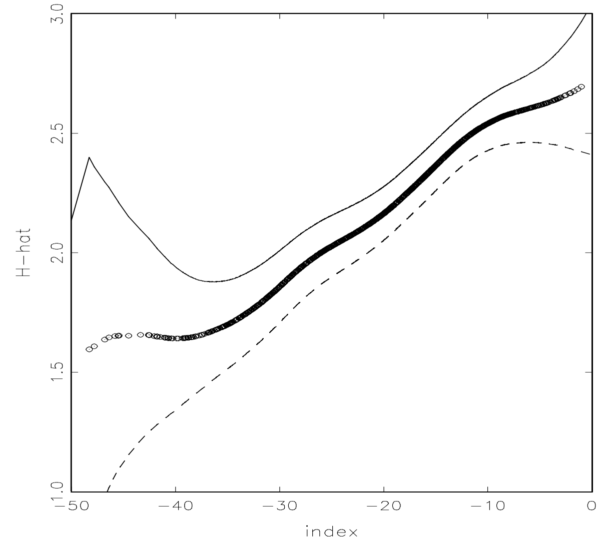
Note: dependent variable generated using ordered probit results in Table 2; 500 Monte Carlo replications; smoothness parameters chosen as in Table 3.

parametric and parametric estimates of the coefficients on the home country dummies are not very large, compared to their standard errors. The semiparametric estimates imply that Turkish immigrants are less fluent than the reference group of Spanish immigrants, with a difference which is on the edge of being significant. Greek and Italian immigrants are no longer significantly different from the Spanish immigrants.

Neither the bootstrapped, nor the asymptotic standard errors are uniformly larger than in the parametric models. All estimators converge at the same rate, but if (one of) the parametric model(s) is not misspecified, the parametric ML estimator would be asymptotically more efficient. Smaller estimated standard errors for the semiparametric estimates can be due to finite sampling error in estimating the standard errors, or due to misspecification of the parametric models.

In general, most of the semiparametric parameter estimates are closer to zero than in the parametric models. To investigate whether this is due to misspecification of the parametric models, or due to finite sample bias in the semiparametric estimates, we present a brief summary of our Monte Carlo results in Table 4. Here the exogenous variables are the observed data (1185 observations), but the dependent variable is generated using the ordered probit model estimates in Table 2. The results are based upon 500 independent Monte Carlo draws. We find that the differences between semiparametric and parametric estimates can to a large extent be attributed to the finite sample bias in the semiparametric estimates. This bias appears to be substantial for the data at hand. For example, the finite sample bias towards zero on the coefficient of years since migration is 20% to 40% of the true value. If the semiparametric estimate in Table 3 would be corrected for this, the semiparametric estimate would come much closer to its parametric counterparts. Similar results hold for most other parameters.

Figure 1. Link Function Semiparametric Model



In figure 1, we have drawn the estimated link function H in (12) for the first set of smoothness parameters in Table 3. For the other set of smoothness parameters, the figure looks very similar.[8] 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 not

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

very precise 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 those with low speaking fluency ($y = 1$).

3.4 Comparison of the Three Models

In this subsection we compare two parametric models and the semiparametric model. We do not consider the models with random thresholds, since their predictions are very close to those with fixed thresholds. Similarly, we only look at the semiparametric model with the first set of smoothness parameters in Table 3, since the other set gives very similar results. First, we look at predictions, i.e. the estimated (conditional) probabilities of bad, intermediate and good speaking fluency (given x), or the conditional mean of the outcome y or z (coded 1, 2 or 3), which summarizes these three probabilities in a linear combination. In the ordered probit model, observed and true speaking fluency (y and z) coincide, but in the model with misclassification errors they do not. Comparing predictions of observed and true speaking fluency should tell us how different the implications of the two parametric models are. The semiparametric model only identifies the observed speaking fluency probabilities, and we compare these with those of the two parametric models. Finally, we formally test for misspecification of the parametric models, using a graphical test against a semiparametric single index alternative.

The means and standard deviations of the predictions of the observed outcomes are similar in the three models.[9] The mean predictions are also similar to the sample means of the outcomes. Larger differences are found with the predictions of the true outcomes

[9] For the parametric models, the predictions are straightforward functions of the estimated parameters. For the semiparametric model, predicted probabilities can be obtained by nonparametric regression of the dummies for good (including very good) or bad (including very bad) on the index. This is similar to the nonparametric estimator of H discussed above.

according to the misclassification model. In particular, the average predictions of bad or good true fluency are larger than the corresponding predictions for observed fluency. Accordingly, the sample dispersion of the predictions for z, the speaking fluency variable free of misclassification error, is larger than that for the predictions of y, the observed speaking fluency indicator.

Figure 2. P(good or very good); Parametric Models

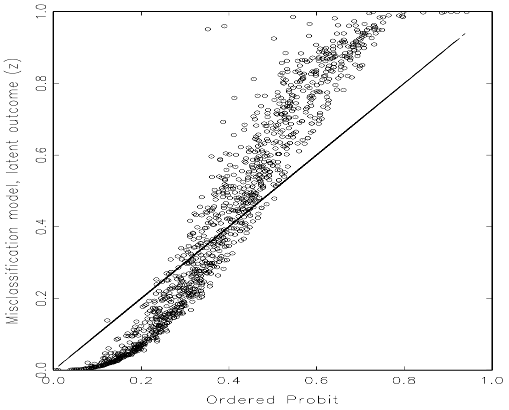
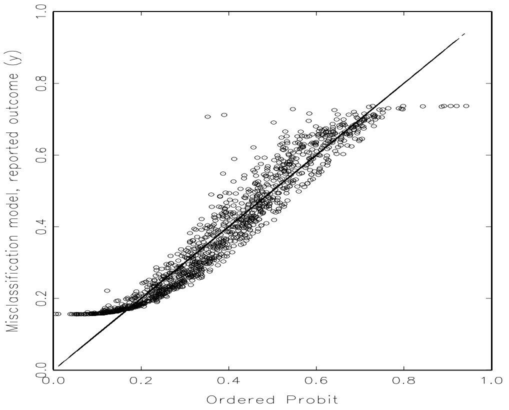


Figure 3. P(good or very good); Parametric Models



In Figure 2, we present a scatter plot of the predicted probabilities of speaking the language well or very well according to the two parametric models. For the misclassification model (vertical axis), the figure shows the predictions of the latent 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, explaining the large dispersion in $\hat{P}[z = 3|x]$ according to the misclassification model. Still, the correlation between the two sets of predictions is quite large (the sample correlation coefficient is 0.96).

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. Again, the correlation coefficient is about 0.96.

The substantial differences between latent and observed outcomes in the misclassification model confirm the conclusion from the misclassification probabilities in Table 2: generalizing the ordered probit model by incorporating misclassification probabilities is useful in this empirical example. While the predictions for the reported variable y are similar for ordered probit and misclassification model, except for observations in the tails, the predictions for the latent variable z are not.[10]

In a similar way, we have compared the predictions of the reported outcomes according to the semiparametric model with those of the ordered probit model and those of the misclassification model. In most cases, all three predictions are similar. This is confirmed by the sample correlation coefficients, which all exceed 0.9. Thus in spite of the apparent differences between the parametric and semiparametric estimates which we saw in the previous subsection, most predictions are similar.

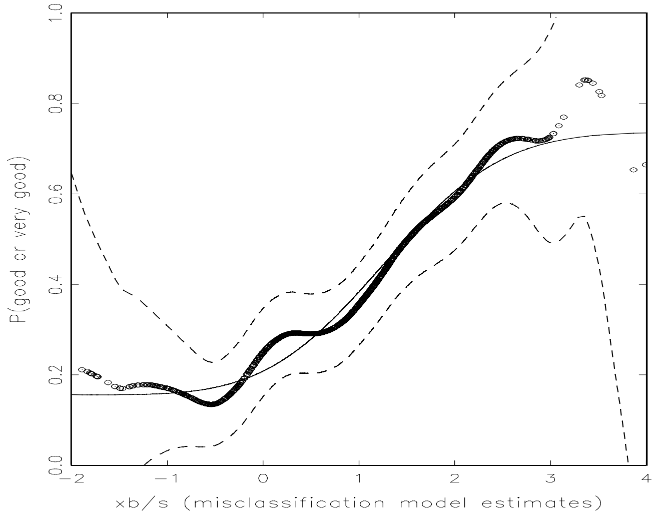
Misspecification Tests of Parametric Models

In Figure 4, we present a graphical tests of the parametric model with misclassification against a semiparametric single index alternative. This test is similar to the test proposed by Horowitz (1993) for the parametric binary choice model. The null hypothesis is that the parametric model is correctly specified; the test should have some power in the direction of the semiparametric alternative which we discussed, but it is not clear whether it has power in other directions of misspecification (such as other than single index models).

The figure presents two functions of the index estimate $x'b/s$, where b and s are the parametric estimates of β and σ in Table 2. The solid line reflects the predicted probabilities $\hat{P}[y_i = 3|x_i] = \hat{P}[y_i = 3|x_i'b/s]$ according to the parametric model, as a function of $x_i'b/s$. The circles are nonparametric kernel regression estimates of the

[10] We come to the same conclusions when we draw figures of the probability of bad or very bad speaking fluency. These figures are not reported.

Figure 4. Test of Parametric Misclassification Model



observed dummy indicator variable $I(y_i = 3)$ on the same index $x_i'b/s$. The dashed lines are nonparametric uniform 95% confidence bands around these kernel estimates.[11]

Under the null hypothesis that the parametric model is specified correctly, b/s is consistent for β/σ. In that case, the parametric formula 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/s 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 thus 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 (asymptotic) test can be based on the uniform confidence bands around the nonparametric curve.

[11] Since the estimator b/s converges to β/σ at rate root n, which is faster rate than the rate of convergence of the nonparametric estimator, the standard errors of b and s are asymptotically negligible, and the uniform confidence bands are calculated as if b/s were known.

The result is that the solid curve is everywhere between the uniform confidence bands, so that the parametric model cannot be rejected.[12] This can be seen as support in favor 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 this model is rejected against the model with misclassification errors. This casts doubt on the power of this type of test. In particular, most of the difference between ordered probit and misclassification model predictions is for values of the index in the tails (see Figure 3). In that region, the width between the uniform confidence bands shows that nonparametric estimates are not very accurate.

4 Summary and Conclusions

In models with ordered categorical dependent variables where the categorical assignment is based on subjective self-reported evaluations, misclassification is likely to be considerable, and may lead to seriously biased parameter estimates and predictions. Parametric estimators which incorporate and estimate misclassification probabilities, as well as semiparametric estimators, are an alternative to standard parametric models. Extending the work of Lee and Porter (1984) and Hausman et al. (1998), we have introduced a parametric model which incorporates misclassification probabilities for the case of more than two ordered categories. We show that this model is a special case of a semiparametric single index model which, if misclassification probabilities are not too large, satisfies some monotonicity condition. It can therefore be estimated with a recently developed estimator of Horowitz and Haerdle (1996).

We analyze the determinants of immigrants' language proficiency proficiency, and compare the results of the standard model with those of the parametric model with misclassification and with the semiparametric results. In all models, the signs of the estimated slope coefficients are the same. Magnitudes and significance levels of the

[12] The same conclusion is obtained if $P[y_i = 1|x_i'b]$ is used instead of $P[y_i = 3|x_i]$.

effects vary, however. Monte Carlo simulations and tests suggest that the parametric model allowing for misclassification errors performs well in describing the data, and that the differences between the estimates of that model and the semiparametric estimates may be due to finite sample bias in the semiparametric estimates. We also consider a model with random category thresholds, but this does not add much to the fixed thresholds case.

When analysis categorical variables which are likely to suffer from misclassification, the misclassification model and the semiparametric estimator we have suggested appear to be a substantial improvement. The parametric misclassification model is also easy to implement, and it gives predictions of the true categorization. It also provides estimates of the misclassification probabilities, which may be also of interest. 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.

References

- Abrevaya, J. and J. Hausman (1999), Semiparametric estimation with mismeasured dependent variables: an application to duration models for unemployment spells, mimeo, University of Chicago.

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

- Andrews, D. (1998b), Testing when a parameter is on the boundary of the maintained hypothesis, mimeo, Yale University.

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

- 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 (1994), Unhappiness and unemployment, *Economic Journal*, 104, 648-659.

- 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 (1998), Language and the earnings of immigrants, CEPR discussion paper series No. 2012.

- 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), Semiparametric estimation of a work-trip mode choice model, *Journal of Econometrics*, 58, 49-70.

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

- Ichimura, H. (1993), Semiparametric 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.

- 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.

- McManus, W., W. Gould, and F. Welch (1983), Earnings of Hispanic men: the role of English language proficiency, *Journal of Labor Economics*, 1, 101-130.

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

- Rivera-Batiz, F. (1990), English language proficiency and the economic progress of immigrants, *Economics Letters* , 34, 295-300.

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

- 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.

Appendix: Testing the null hypothesis of no misclassification errors

In this appendix we discuss 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 $j \neq k$. Since the model is not defined for $p_{jk} < 0$, the parameter vector cannot be an internal point of the parameter space under the null hypothesis, which implies that standard asymptotic theory of the ML estimator does not apply. Andrews (1999) explains how to derive the asymptotic distribution of an estimator when the true parameter value is on the boundary of the parameter space for a very general class of estimators and allowing for non i.i.d. observations, and our maximum likelihood framework with i.i.d. observations is a special case of this. In Andrews (1998b), it is explained how the results in Andrews (1999) can be used to determine the asymptotic distribution of the quasi-likelihood ratio test statistic. This is what we will apply here.[13]

A first important feature of the test is that inequality constraints are tested rather than equality restrictions. In line with the literature going back to Chernoff (1954), this implies that the likelihood ratio test statistic is not asymptotically chi-squared, but some mixture of chi-squared distributions. For testing inequality constraints where the true parameter vector is in the interior of the parameter space, several procedures have been designed which are relatively easy to apply in practice. See, for example, Andrews (1998a) or Szroeter (1997). None of these can be applied in our case however, since - as discussed above - the true parameter vector is not in the interior of the parameter space under the null.

The basic idea of the Andrews (1999, 1998b) approach to deal with this problem is that the log likelihood near the true parameter values can still be approximated by a quadratic function, using first and second order partial left or right derivatives. Since

[13] Andrews (1998b) also allows for nuisance parameters which play a role under the alternative only (and are thus not identified under the null). Such parameters do not appear in our case.

$p_{j,k} < 0$ is not feasible but right partial derivatives are. Their continuity properties imply that the second order approximation remains useful for deriving the asymptotic distribution of the likelihood ratio test statistic.

Theorem 4 in Andrews (1998b) provides the required results.[14] It is straightforward to check that the regularity assumptions required for this theorem are all satisfied in our example, using that 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 given 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 value of the likelihood (allowing for all values $p_{j,k} \geq 0$) and L_0 is the restricted maximum (imposing $p_{j,k} = 0$ for all $j, k = 1, 2, 3,\ j$. 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 10 (unrestricted) parameters of the model. The parameter space can be written as $V = (-\infty, \infty)^{10}$ x $[0, \infty)^6)$, and the null hypothesis is $\theta \in V_0 = (-\infty, \infty)^{10}$ x $\{0\}^6$.[15] 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 $p_{j,k}$.

[14] We only need the special case without nuisance parameters which are unidentified under the null. The result for this special case follows also from Theorem 3 in Andrews (1999).

[15] We ignore the obvious lower bounds on some of the other parameters (the standard deviation of the error term σ and the threshold m_2), since these are not binding and do not matter for the local approximations.

Theorem 4 in Andrews (1998b) now implies that LR has the same asymptotic distribution as

$$Inf_{[\theta \in V_0]}\ q(\theta) - Inf_{[\theta \in V]}\ q(\theta) \tag{20}$$

with

$$q(\theta) = (\theta - Z)'J(\theta - Z), \ \ Z \sim N(0, J^{-1}IJ^{-1}) \tag{21}$$

The asymptotic distribution of LR can thus be obtained by

- plugging in the estimates $\hat{J}$ for J and $\hat{I}$ for I,[16]
- generating multivariate normal draws of Z,
- solving the two quadratic programming problems for each draw,
- considering the thus obtained simulated distribution of the difference between the two minimum values.

With our estimates, this procedure gave a 5% critical value of about 8.6 for the LR statistic.[17] The realization of the LR statistic is 21.3 (see Table 2), so that the null is rejected. The conclusion is that the misclassification probabilities are jointly significantly different from zero.

[16] As in the usual ML case, I and J coincide under the null, so an asymptotically equivalent procedure would be to use an estimate for only one of them.

[17] As expected, this is smaller than the corresponding chi-square critical value with six degrees of freedom, which would be obtained in the standard case with equality restrictions at an internal point.

IZA Discussion Papers

No	Author(s)	Titel	Area	Date
121	J. C. van Ours	Do Active Labor Market Policies Help Unemployed Workers to Find and Keep Regular Jobs?	4/6	3/00
122	D. Munich J. Svejnar K. Terrell	Returns to Human Capital under the Communist Wage Grid and During the Transition to a Market Economy	4	3/00
123	J. Hunt	Why Do People Still Live in East Germany?	1	3/00
124	R. T. Riphahn	Rational Poverty or Poor Rationality? The Take-up of Social Assistance Benefits	3	3/00
125	F. Büchel J. R. Frick	The Income Portfolio of Immigrants in Germany - Effects of Ethnic Origin and Assimilation. Or: Who Gains from Income Re-Distribution?	1/3	3/00
126	J. Fersterer R. Winter-Ebmer	Smoking, Discount Rates, and Returns to Education	5	3/00
127	M. Karanassou D. J. Snower	Characteristics of Unemployment Dynamics: The Chain Reaction Approach	3	3/00
128	O. Ashenfelter D. Ashmore O. Deschênes	Do Unemployment Insurance Recipients Actively Seek Work? Evidence From Randomized Trials in Four U.S. States	6	3/00
129	B. R. Chiswick M. E. Hurst	The Employment, Unemployment and Unemployment Compensation Benefits of Immigrants	1/3	3/00
130	G. Brunello S. Comi C. Lucifora	The Returns to Education in Italy: A New Look at the Evidence	5	3/00
131	B. R. Chiswick	Are Immigrants Favorably Self-Selected? An Economic Analysis	1	3/00
132	R. A. Hart	Hours and Wages in the Depression: British Engineering, 1926-1938	7	3/00
133	D. N. F. Bell R. A. Hart O. Hübler W. Schwerdt	Paid and Unpaid Overtime Working in Germany and the UK	1	3/00
134	A. D. Kugler G. Saint-Paul	Hiring and Firing Costs, Adverse Selection and Long-term Unemployment	3	3/00
135	A. Barrett P. J. O'Connell	Is There a Wage Premium for Returning Irish Migrants?	1	3/00
136	M. Bräuninger M. Pannenberg	Unemployment and Productivity Growth: An Empirical Analysis within the Augmented Solow Model	3	3/00

137	J.-St. Pischke	Continuous Training in Germany	5	3/00
138	J. Zweimüller R. Winter-Ebmer	Firm-specific Training: Consequences for Job Mobility	1	3/00
139	R. A. Hart Y. Ma	Wages, Hours and Human Capital over the Life Cycle	1	3/00
140	G. Brunello S. Comi	Education and Earnings Growth: Evidence from 11 European Countries	2/5	4/00
141	R. Hujer M. Wellner	The Effects of Public Sector Sponsored Training on Individual Employment Performance in East Germany	6	4/00
142	J. J. Dolado F. Felgueroso J. F. Jimeno	Explaining Youth Labor Market Problems in Spain: Crowding-Out, Institutions, or Technology Shifts?	3	4/00
143	P. J. Luke M. E. Schaffer	Wage Determination in Russia: An Econometric Investigation	4	4/00
144	G. Saint-Paul	Flexibility vs. Rigidity: Does Spain have the worst of both Worlds?	1	4/00
145	M.-S. Yun	Decomposition Analysis for a Binary Choice Model	7	4/00
146	T. K. Bauer J. P. Haisken-DeNew	Employer Learning and the Returns to Schooling	5	4/00
147	M. Belot J. C. van Ours	Does the Recent Success of Some OECD Countries in Lowering their Unemployment Rates Lie in the Clever Design of their Labour Market Reforms?	3	4/00
148	L. Goerke	Employment Effects of Labour Taxation in an Efficiency Wage Model with Alternative Budget Constraints and Time Horizons	3	5/00
149	R. Lalive J. C. van Ours J. Zweimüller	The Impact of Active Labor Market Programs and Benefit Entitlement Rules on the Duration of Unemployment	3/6	5/00
150	J. DiNardo K. F. Hallock J.-St. Pischke	Unions and the Labor Market for Managers	7	5/00
151	M. Ward	Gender, Salary and Promotion in the Academic Profession	5	5/00
152	J. J. Dolado F. Felgueroso J. F. Jimeno	The Role of the Minimum Wage in the Welfare State: An Appraisal	3	5/00
153	A. S. Kalwij M. Gregory	Overtime Hours in Great Britain over the Period 1975-1999: A Panel Data Analysis	3	5/00
154	M. Gerfin M. Lechner	Microeconometric Evaluation of the Active Labour Market Policy in Switzerland	6	5/00

155	J. Hansen	The Duration of Immigrants' Unemployment Spells: Evidence from Sweden	1/3	5/00
156	C. Dustmann F. Fabbri	Language Proficiency and Labour Market Performance of Immigrants in the UK	1	5/00
157	P. Apps R. Rees	Household Production, Full Consumption and the Costs of Children	7	5/00
158	A. Björklund T. Eriksson M. Jäntti O. Raaum E. Österbacka	Brother Correlations in Earnings in Denmark, Finland, Norway and Sweden Compared to the United States	5	5/00
159	P.- J. Jost M. Kräkel	Preemptive Behavior in Sequential Tournaments	5	5/00
160	M. Lofstrom	A Comparison of the Human Capital and Signaling Models: The Case of the Self-Employed and the Increase in the Schooling Premium in the 1980's	5	6/00
161	V. Gimpelson D. Treisman G. Monusova	Public Employment and Redistributive Politics: Evidence from Russia's Regions	4	6/00
162	C. Dustmann M. E. Rochina-Barrachina	Selection Correction in Panel Data Models: An Application to Labour Supply and Wages	6	6/00
163	R. A. Hart Y. Ma	Why do Firms Pay an Overtime Premium?	5	6/00
164	M. A. Shields S. Wheatley Price	Racial Harassment, Job Satisfaction and Intentions to Quit: Evidence from the British Nursing Profession	5	6/00
165	P. J. Pedersen	Immigration in a High Unemployment Economy: The Recent Danish Experience	1	6/00
166	Z. MacDonald M. A. Shields	The Impact of Alcohol Consumption on Occupational Attainment in England	5	6/00
167	A. Barrett J. FitzGerald B. Nolan	Earnings Inequality, Returns to Education and Immigration into Ireland	5	6/00
168	G. S. Epstein A. L. Hillman	Social Harmony at the Boundaries of the Welfare State: Immigrants and Social Transfers	3	6/00
169	R. Winkelmann	Immigration Policies and their Impact: The Case of New Zealand and Australia	1	7/00
170	T. K. Bauer K. F. Zimmermann	Immigration Policy in Integrated National Economies	1	7/00
171	C. Dustmann F. Windmeijer	Wages and the Demand for Health – A Life Cycle Analysis	5	7/00
172	D. Card	Reforming the Financial Incentives of the Welfare System	3	7/00

173	D. S. Hamermesh	Timing, Togetherness and Time Windfalls	5	7/00
174	E. Fehr J.-R. Tyran	Does Money Illusion Matter? An Experimental Approach	7	7/00
175	M. Lofstrom	Self-Employment and Earnings among High-Skilled Immigrants in the United States	1	7/00
176	O. Hübler W. Meyer	Industrial Relations and the Wage Differentials between Skilled and Unskilled Blue-Collar Workers within Establishments: An Empirical Analysis with Data of Manufacturing Firms	5	7/00
177	B. R. Chiswick G. Repetto	Immigrant Adjustment in Israel: Literacy and Fluency in Hebrew and Earnings	1	7/00
178	R. Euwals M. Ward	The Renumeration of British Academics	5	7/00
179	E. Wasmer P. Weil	The Macroeconomics of Labor and Credit Market Imperfections	2	8/00
180	T. K. Bauer I. N. Gang	Sibling Rivalry in Educational Attainment: The German Case	5	8/00
181	E. Wasmer Y. Zenou	Space, Search and Efficiency	2	8/00
182	M. Fertig C. M. Schmidt	Discretionary Measures of Active Labor Market Policy: The German Employment Promotion Reform in Perspective	6	8/00
183	M. Fertig C. M. Schmidt	Aggregate-Level Migration Studies as a Tool for Forecasting Future Migration Streams	1	8/00
184	M. Corak B. Gustafsson T. Österberg	Intergenerational Influences on the Receipt of Unemployment Insurance in Canada and Sweden	3	8/00
185	H. Bonin K. F. Zimmermann	The Post-Unification German Labor Market	4	8/00
186	C. Dustmann	Temporary Migration and Economic Assimilation	1	8/00
187	T. K. Bauer M. Lofstrom K. F. Zimmermann	Immigration Policy, Assimilation of Immigrants and Natives' Sentiments towards Immigrants: Evidence from 12 OECD-Countries	1	8/00
188	A. Kapteyn A. S. Kalwij A. Zaidi	The Myth of Worksharing	5	8/00
189	W. Arulampalam	Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages	3	8/00

190	C. Dustmann I. Preston	Racial and Economic Factors in Attitudes to Immigration	1	8/00
191	G. C. Giannelli C. Monfardini	Joint Decisions on Household Membership and Human Capital Accumulation of Youths: The role of expected earnings and local markets	5	8/00
192	G. Brunello	Absolute Risk Aversion and the Returns to Education	5	8/00
193	A. Kunze	The Determination of Wages and the Gender Wage Gap: A Survey	5	8/00
194	A. Newell F. Pastore	Regional Unemployment and Industrial Restructuring in Poland	4	8/00
195	F. Büchel A. Mertens	Overeducation, Undereducation, and the Theory of Career Mobility	5	9/00
196	J. S. Earle K. Z. Sabirianova	Equilibrium Wage Arrears: A Theoretical and Empirical Analysis of Institutional Lock-In	4	9/00
197	G. A. Pfann	Options to Quit	1	9/00
198	M. Kreyenfeld C. K. Spiess G. G. Wagner	A Forgotten Issue: Distributional Effects of Day Care Subsidies in Germany	3	9/00
199	H. Entorf	Rational Migration Policy Should Tolerate Non-Zero Illegal Migration Flows: Lessons from Modelling the Market for Illegal Migration	1	9/00
200	T. Bauer G. S. Epstein I. N. Gang	What are Migration Networks?	1	9/00
201	T. J. Dohmen G. A. Pfann	Worker Separations in a Nonstationary Corporate Environment	1	9/00
202	P. Francois J. C. van Ours	Gender Wage Differentials in a Competitive Labor Market: The Household Interaction Effect	5	9/00
203	J. M. Abowd F. Kramarz D. N. Margolis T. Philippon	The Tail of Two Countries: Minimum Wages and Employment in France and the United States	5	9/00
204	G. S. Epstein	Labor Market Interactions Between Legal and Illegal Immigrants	1	10/00
205	A. L. Booth M. Francesconi J. Frank	Temporary Jobs: Stepping Stones or Dead Ends?	1	10/00
206	C. M. Schmidt R. Baltussen R. Sauerborn	The Evaluation of Community-Based Interventions: Group-Randomization, Limits and Alternatives	6	10/00

207	C. M. Schmidt	Arbeitsmarktpolitische Maßnahmen und ihre Evaluierung: eine Bestandsaufnahme	6	10/00
208	J. Hartog R. Winkelmann	Dutch Migrants in New Zealand: Did they Fare Well?	1	10/00
209	M. Barbie M. Hagedorn A. Kaul	Dynamic Effciency and Pareto Optimality in a Stochastic OLG Model with Production and Social Security	3	10/00
210	T. J. Dohmen	Housing, Mobility and Unemployment	1	11/00
211	A. van Soest M. Das X. Gong	A Structural Labour Supply Model with Nonparametric Preferences	5	11/00
212	X. Gong A. van Soest P. Zhang	Sexual Bias and Household Consumption: A Semiparametric Analysis of Engel Curves in Rural China	5	11/00
213	X. Gong A. van Soest E. Villagomez	Mobility in the Urban Labor Market: A Panel Data Analysis for Mexico	1	11/00
214	X. Gong A. van Soest	Family Structure and Female Labour Supply in Mexico City	5	11/00
215	J. Ermisch M. Francesconi	The Effect of Parents' Employment on Children's Educational Attainment	5	11/00
216	F. Büchel	The Effects of Overeducation on Productivity in Germany— The Firms' Viewpoint	5	11/00
217	J. Hansen R. Wahlberg	Occupational Gender Composition and Wages in Sweden	5	11/00
218	C. Dustmann A. van Soest	Parametric and Semiparametric Estimation in Models with Misclassified Categorical Dependent Variables	1	11/00

An updated list of IZA Discussion Papers is available on the center's homepage *www.iza.org*.