Christelis, Dimitris; Sanz-de-Galdeano, Anna

Working Paper

Smoking persistence across countries: an analysis using semi-parametric dynamic panel data models with selectivity

IZA Discussion Papers, No. 4336

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Christelis, Dimitris; Sanz-de-Galdeano, Anna (2009) : Smoking persistence across countries: an analysis using semi-parametric dynamic panel data models with selectivity, IZA Discussion Papers, No. 4336, Institute for the Study of Labor (IZA), Bonn, https://nbn-resolving.de/urn:nbn:de:101:1-20090909344

This Version is available at:
https://hdl.handle.net/10419/36062

DISCUSSION PAPER SERIES

IZA DP No. 4336

Smoking Persistence Across Countries: An Analysis Using Semi-Parametric Dynamic Panel Data Models with Selectivity

Dimitris Christelis
Anna Sanz-de-Galdeano

July 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Smoking Persistence Across Countries: An Analysis Using Semi-Parametric Dynamic Panel Data Models with Selectivity

Dimitris Christelis
SHARE and CSEF, University of Naples Federico II

Anna Sanz-de-Galdeano
Universitat Autònoma de Barcelona and IZA

Discussion Paper No. 4336
July 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4336
July 2009

ABSTRACT

Smoking Persistence Across Countries: An Analysis Using Semi-Parametric Dynamic Panel Data Models with Selectivity*

We study smoking persistence in ten countries using data from the European Community Household Panel. Such persistence may be due to true state dependence but may also reflect individual unobserved heterogeneity. We distinguish between the two by using semi-parametric dynamic panel data methods applied to both the decision to smoke or not and to the decision on the number of cigarettes smoked. Our model allows for correlation of the two time-varying error terms, i.e. for selectivity. We find that for both smoking decisions true state dependence is in general much smaller when unobserved individual heterogeneity is taken into account, and we also uncover large differences in true state dependence across countries. Finally, we find that taking into account heaping in the reported number of cigarettes smoked considerably improves the fit of our model.

JEL Classification: C33, C34, D12, I10, I12

Keywords: smoking, panel data, selectivity

Corresponding author:

Anna Sanz-de-Galdeano
Departament d'Economia i Historia Economica - Edifici B
Universitat Autonoma de Barcelona
Bellaterra 08193
Barcelona
Spain
E-mail: galdeano76@gmail.com

* We are grateful to Matilde P. Machado, Pierre-Carl Michaud and Arthur van Soest for useful discussions and comments. We would also like to thank the participants to the 4th CSEF-IGIER Symposium and the 15th International Conference on Panel Data. Anna Sanz-de-Galdeano acknowledges financial support from the Spanish Ministry of Science and Technology grant SEJ2007-62500.

I. Introduction

One of the most obvious patterns in individual smoking behaviour is its persistence over time: individuals who smoked in the past are more likely to be current smokers. Such persistence may be due to true state dependence but it may reflect individual unobserved heterogeneity as well. For example, individuals' persistent smoking habits may stem from unobserved factors such as their degrees of risk aversion, rates of time preference, abilities to acquire and process relevant information and their health attitudes. The distinction between state dependence and unobserved heterogeneity is very important from a policy perspective: if smoking persistence were mostly reflecting individual time-invariant unobserved heterogeneity, the effectiveness of many policies aimed at reducing smoking rates as well as their long-run compositional effects would be seriously called into question. On the other hand, there would be scope for policies aimed at influencing some of the personal characteristics that crystallize unobserved heterogeneity at young age, e.g. health attitudes.

Our paper adds to the existing literature in several important ways. First, we study the dynamics of smoking behaviour across European countries from a comparative perspective. Previous studies addressing this issue and analysing the dynamics of smoking behaviour are country-specific and generally rely on US data (Gilleskie and Strumpf, 2004; Chaloupka, 1991), with the exception of Labeaga (1999) who uses data for Spain. Our study on the other hand uses data from the European Community Household Panel, a longitudinal micro-level database that provides cross-country comparable information on both smoking behaviour and socioeconomic variables that are important in predicting smoking.

It is well known that smoking rates are generally higher in Europe than in the US (Cutler and Glaeser, 2006). There are also important differences, however, within European countries, not only in smoking rates and persistence but also in factors such as smoking regulations (taxation, limits on smoking in public places and restrictions on youth access to tobacco products), social norms on and tolerance towards smoking behaviour and awareness of the health risks associated with smoking.

We will relate our country-specific estimates of true state dependence to some of these factors by relying on additional data from the World Health Organization and the Eurobarometer surveys.

Second, we isolate the effect of state dependence net of unobserved heterogeneity by exploiting the panel nature of the ECHP and using nonlinear panel data methods applied to both the decision to smoke or not and to the decision of how many cigarettes to smoke. Several papers that have used individual-level panel data on consumption expenditures have highlighted the importance of accounting for time invariant heterogeneity across households when investigating habit formation in consumption behaviour (Browning and Collado, 2007; Carrasco, Labeaga and López-Salido, 2005). To the best of our knowledge, however, no previous study has simultaneously modelled the smoking participation and consumption (conditional on participation) decisions within a dynamic framework, allowing for the time-varying error terms in the two decisions to be correlated and accounting for unobserved fixed effects in a semi-parametric fashion in order to distinguish between true and spurious state dependence.

Gilleskie and Strumpf (2004) analyze, as we do in this study, the dynamics of smoking participation and consumption decisions. They do not, however, incorporate selectivity into the model, thus assuming that the time variant heterogeneity terms in the participation and consumption equations are uncorrelated. This assumption is not trivial. For example, if this correlation is significant (as we find in our study), then it implies that factors affecting participation have an effect on cigarette consumption as well. Ignoring this effect can lead to inconsistent estimates. Harris and Zhao (2007, henceforth HZ) propose a zero-inflated ordered probit model to allow for both a double hurdle specification and for correlation of the time varying unobservables from the participation and cigarette consumption equations. Unlike Gilleskie and Strumpf (2004) however, HZ (2007) cannot model the dynamics of smoking behaviour because they use cross-sectional survey data. We extend the framework of HZ by using a dynamic panel data specification and accounting in a semi-parametric fashion for the presence of unobserved time invariant heterogeneity potentially correlated with the regressors.

We find that our estimates of true state dependence from our panel models are substantially smaller than the ones obtained when individual unobserved heterogeneity is not taken into account. This holds for both males and females and in all countries considered. However, even after accounting for unobserved individual heterogeneity, the effect of lagged smoking behavior remains economically and statistically significant in most cases. We also find that both allowing for selectivity turns out to be an empirically relevant issue.

In addition, given that the data on the number of cigarettes smoked is characterized not only by an excess of zero observations but also by heaping, we experiment with both a linear model and an ordered probit panel model for the cigarette consumption equation. We find that the latter specification gives a better fit than the former one, and thus conclude that it is important that researchers account for heaping when studying cigarette consumption.

Finally, we investigate the sources of the observed cross-country variation in true state dependence. To this purpose, we rely on additional information on individuals' attitudes and beliefs about the effects of smoking and its social acceptance (taken from several Eurobarometer surveys) as well as on indicators of country-specific smoking restrictions (taken from the World Health Organization Tobacco Control Database). We find suggestive evidence that cross-country differences in true state dependence may be related to country-specific smoking regulations. On the other hand, cross-country variation in attitudes toward smoking seems to reflect unobserved heterogeneity and, in consequence, it cannot explain differences in true state dependence across countries once this heterogeneity has been taken into account.

In Section II we provide information on our data and the variables we use in our empirical models, which are discussed in Section III. Section IV presents the results from our estimation, while Section V highlights the importance of true state dependence by running simulation exercises across the countries in our sample. Section VI concludes.

II. Data sources, variable definitions and descriptive analysis

The principal individual-level dataset used in this analysis is the European Community Household (ECHP), a standardized multi-purpose annual longitudinal survey carried out in all 15 countries of the European Union between 1994 and 2001. The ECHP not only contains a wide range of economic and socio-demographic information both at the household and the individual level, but it also includes questions related to the health status and smoking behaviour of European adults (16+). Moreover, given that it was centrally designed and coordinated by the Statistical Office of the European Commission (Eurostat) a good level of comparability across countries and over time is ensured.[1]

We restrict our study to waves 5-8 (years 1998-2001) from the ECHP because no questions on smoking are asked in the first four waves of the survey in any of the countries. Additionally, we are forced to exclude some countries from the analyses because of insufficient or unavailable smoking information and focus instead in the following subset of countries: Finland, Denmark, UK, Ireland, Belgium, Austria, Italy, Spain, Portugal and Greece.[2]

The smoking information available from the ECHP is based on two questions. First, respondents are asked "do you smoke or did you ever smoke?" and these are the five possible answers: "smoke daily", "smoke occasionally", "do not smoke, used to smoke daily", "do not smoke, used to smoke occasionally" or "never smoked". Second, individuals who have declared to smoke daily currently or in the past, are asked to report the number of cigarettes they smoke per day (currently or in the past).

We focus on current daily smoking behaviour and define a dichotomous indicator that is equal to one if individuals declare to smoke daily and value zero otherwise. As for the amount of

[1] For further details on the ECHP, see Peracchi (2002).

[2] No questions on smoking are asked in any of the waves in Luxembourg, France and the Netherlands. Smoking information is unavailable in Germany (2000) and Sweden (1998) and incomplete in Sweden (1999-2001), where there is no information on the number of cigarettes smoked.

cigarettes smoked per day, we set it equal to zero for those who are not daily smokers.[3] For those respondents who declare to be daily smokers and report zero number of cigarettes, we allow for the possibility that they might be non-participants as well. These cases, however, represent only 0.44% of the observations in our sample.[4]

A common problem faced by all empirical analyses of self-reported cigarette consumption is the bunching or heaping of reported values. Figure 1 displays the frequency distribution of cigarettes consumed by daily smokers in our sample. Not surprisingly, there is substantial heaping of cigarette counts at multiples of 5, with the largest heap at 20 (the amount of cigarettes typically contained in one pack). As Wang and Heitjan (2008) indicate, if these were true cigarette counts, using typical count data models would not be appropriate, while if they were not, heaping could lead to biased estimates. In Section III below, we will discuss how we address the heaping problem.

Table 1 displays descriptive statistics on smoking transitions and persistence by gender. Column 1 reports smoking rates among adult Europeans, which are always remarkably higher for males than for females with the exception of Ireland and the UK. while this gender difference is particularly large in Mediterranean countries. Columns 2 and 3 report the rates of yearly transitions into and out of smoking that occur during the estimation period. In Column 4 we report the mobility index proposed by Shorrocks (1978), which takes the value 0 if there are no transitions and the value 1 if there is no persistence. It is well known that most smokers start when they are very young.[5] Hence, it is not surprising that, given that individuals' average age in our sample is around 41 years, the percentage of inflows is always lower than the percentage of outflows. Regarding persistence, the Spearman rank correlation between the smoking rate and the mobility index is

[3] This choice is due to the fact that when individuals answered "yes, I smoke occasionally", the question on the number of cigarettes was not asked at all.

[4] We also experimented with considering them as smokers (i.e. participants who don't smoke any cigarettes) in the context of the double-hurdle model discussed below, and our results were practically identical.

[5] See, for instance, Gruber and Zinman (2000).

positive (0.35), but only statistically significant at the 12% level (p-value=0.124); that is, there are more smoking transitions in countries with higher prevalence of smoking. We also find that this result is mainly driven by inflows or transitions from non-smoking to smoking status: inflows are significantly more frequent in countries with higher smoking rates (correlation=0.665, p-value=0.001), whereas the correlation between smoking rates and outflows is far from being statistically significant at standard levels of testing.

The multi-purpose nature of the ECHP allows us to incorporate in our analysis a rich set of covariates, summarized in Table 2, that are used to explain variations in individual smoking behaviour. These include individual and household socioeconomic characteristics such as age, education, marital status, labour market status, home ownership, presence of children under 12 years of age in the household, household size and real household income as well as health-related variables indicating whether the individual spent at least one night in the hospital and consulted a specialist at least once during the previous 12 months.[6]

Regarding tobacco prices (which are set at the national level in the countries that we examine), the short length of our panel prevents us from estimating price elasticities because, as can be seen in Figure 2, price variability during the period under study is very small. This could be due to problems we faced when constructing price data that are both intra- and intertemporally consistent for the countries in our sample.[7] Hence, we include time dummies in all our analyses in order to partially capture price effects.

On top of relying on the observations for which non-missing information is available for both the smoking indicators and the covariate variables previously outlined, we further restrict our analyses to the balanced panel sample (as in Wooldridge, 2005). It is important to note, however, that the pattern of both the outcome and predictor variables is essentially invariant across the

[6] Given that these last two variables refer to the year previous to the interview, there should be no reverse causality between them and current smoking.

[7] A fuller discussion of the construction of our price data is provided in Appendix A.1

balanced and unbalanced samples, as can be seen from Table 2. This suggests that attrition and non-response do not seem to be associated with any of the variables considered in the analyses (including smoking participation and intensity), and that, in principle, they should not be correlated with any of the unobserved characteristics that affect individual smoking decisions (sample sizes by country and gender for the balanced and unbalanced panels are displayed in Appendix Table A.1).

A different type of attrition potentially related with smoking behaviour can be caused by the increased probability of death for smokers when old age is reached. We focus on individuals who are at most 65 years old in order to minimize this problem.

Finally, we will also rely on data from the Eurobarometer surveys and on official figures from Eurostat and the World Health Organization. While we mostly use this information in a descriptive manner, we believe that being aware of the wide variety of institutional setups, beliefs and attitudes related to smoking habits across Europe can provide useful insights when attempting to account for cross-country differences.

III. Empirical model

We estimate a double-hurdle model of cigarette consumption (see Jones 1989, Labeaga 1999, HZ), which postulates that individuals do not smoke any cigarettes because either they are not smokers (non-participation) or they are at a corner solution while being smokers. The double hurdle model has been shown by HZ (using cross-sectional data from an Australian smoking survey) to provide a better fit than a simpler single hurdle model. We allow for different specifications for the decision to smoke or not and for how many cigarettes to smoke, while including the lagged decision as a determinant in both cases.

In order to model the decision to smoke or not, we define a binary variable y^1 that takes the value one if an underlying latent variable y^{*1} is larger than zero, while it is equal to zero otherwise. The equation for the latent variable y^{*1} for individual i at time t is as follows

$$y_{i,t}^{*1} = \mathbf{X}_{i,t}^{1}\boldsymbol{\beta}_1 + \theta_1 y_{i,t-1}^{1} + c_i^1 + \varepsilon_{i,t}^{1} \quad (1)$$

where $\mathbf{X}_{i,t}^{1}$ is the matrix of our explanatory variables (assumed to be strictly exogenous) and c_i^1 the unobserved heterogeneity affecting the decision to smoke or not (we will specify it further below, when we discuss the corresponding heterogeneity for the decision about how many cigarettes to smoke). The parameter θ_1 captures the true state dependence due to smoking in the previous period, while the time varying error term $\varepsilon_{i,t}^{1}$ is assumed to follow a standard normal distribution, conditional on $\mathbf{X}_{i,t}^{1}, y_{i,t-1}^{1}, c_i^1$. Denoting by φ the normal density function and with Φ its associated cumulative distribution, the probability of smoking participation is therefore is given by

$$\text{Prob}(y_{i,t}^{1} = 1 \mid \mathbf{X}_{i,t}^{1}, y_{i,t-1}^{1}, c_i^1) = \Phi\left(\mathbf{X}_{i,t}^{1}\boldsymbol{\beta}_1 + \theta_1 y_{i,t-1}^{1} + c_i^1\right) \quad (2)$$

With respect to the number of cigarettes smoked, we assume that there is a second latent variable y^{*2} that is related to observed number of cigarettes y^2 in the following way:

$$y_{i,t}^{2} = \begin{cases} 0 \text{ if } y_{i,t}^{*1} <= 0 \\ 0 \text{ if } y_{i,t}^{*1} > 0 \text{ and } \mathrm{y}_{i,t}^{*2} = 0 \\ y_{i,t}^{*2} \text{ if } y_{i,t}^{*1} > 0 \text{ and } y_{i,t}^{*2} > 0 \end{cases} \quad (3)$$

The second possibility in (3) is a feature of the double hurdle model that differentiates it from conventional sample selection models, because it allows for no smoking even if one has overcome the participation threshold. The latent variable y^{*2} is assumed to have the following linear specification:

$$y_{i,t}^{*2} = \mathbf{X}_{i,t}^{2}\boldsymbol{\beta}_2 + \theta_2 y_{i,t-1}^{2} + c_i^2 + \varepsilon_{i,t}^{2} \quad (4)$$

As was the case with the decision to smoke or not, we account for true state dependence in the number of cigarettes smoked through the parameter θ_2. The term c_i^2 represents the unobserved heterogeneity affecting the number of cigarettes smoked. The time-varying error term $\varepsilon_{i,t}^{2}$ is assumed to be normally distributed with a zero mean and a standard deviation equal to σ_2.

Furthermore, it is assumed to be correlated with $\varepsilon_{i,t}^{1}$, with a correlation coefficient equal to ρ. In other words, we allow for selectivity in the decision on the number of cigarettes smoked. This selectivity reflects the influence of time-varying factors on both decisions, e.g. the introduction of a heavily advertised new brand of cigarettes might affect both whether one becomes a smoker and the number of cigarettes smoked. Importantly, selectivity also implies that we have to jointly estimate models for the two decisions.[8]

The formulation of the smoking decision in (2) and (4) does not encompass the rational addiction model of Becker and Murphy (1988), in which the current period utility of smoking is influenced by previous smoking behaviour. This assumption, together with another one stating that individuals are forward-looking and consider the influence of today's smoking choice on future smoking behaviour, result in a current period demand equation that contains both the lagged smoking decision and the expected future one, together with lagged, current and and future tobacco prices[9]. As Gilleskie and Strumpf (2005) remark, however, it is the optimal expected value of consumption based on currently available information rather than the actual ex-post value that matters and it is hard to argue that future prices are known with perfect foresight;[10] rather, they are likely to be forecast as a function of currently available information. Therefore, we follow the formulation of Gilleskie and Strumpf (2000, 2005) and make the expected value of cigarette consumption a function of past and present consumption.

[8] Semykina and Wooldridge (2008) and Labeaga et al. (2009) also use panel models that allow for sample selectivity; their approaches, however, are different from ours, as they do not use a full maximum likelihood estimation framework and make parametric assumptions about the random effect terms.

[9] See, for instance, Chaloupka (1991) and Becker, Grossman, and Murphy(1991, 1994).

[10] Coppejans et al. (2007) actually develop a demand model for goods that are subject to habit formation and show in their application for the market for cigarettes that it indeed depends on individual beliefs about the evolution of future prices.

In addition, incorporating future information on cigarette consumption and prices is impractical in our context for two further practical reasons: i) as already discussed, the time variability of our price data is quite limited, and thus the inclusion of future prices as regressors would add very little additional information; ii) we would be forced to use two ECHP waves instead of three, and this would lead to considerably smaller sample sizes.

As Wooldridge (2005) points out, the assumption that $\mathbf{y}_{i,t} = (y_{i,t}^1, y_{i,t}^2)$ depends on its once lagged value $\mathbf{y}_{i,t-1}$ but not on any of its other lags implies that the joint distribution of $(\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, ..., \mathbf{y}_{i,T})$ conditional on $\mathbf{y}_{i,0}, \mathbf{c}_i = (c_i, c_i)$, $\mathbf{X}_{i,t} = (\mathbf{X}_{i,t}^1, \mathbf{X}_{i,t}^2)$ can be written as

$$\prod_{t=1}^{T} f(\mathbf{y}_{i,t} \mid \mathbf{y}_{i,t-1}, \mathbf{X}_{i,t}, \mathbf{c}_i) \tag{5}$$

The presence of the unobserved heterogeneity vector $\mathbf{c}_i$ complicates our estimation because one needs to address the ensuing incidental parameters problem. Furthermore, given that our specification for both smoking decisions is dynamic, we have to take into account the endogeneity of the decisions in the initial period. To address both issues, we adapt in our two-equation context the conditional maximum likelihood approach of Wooldridge (2005), who breaks the unobserved heterogeneity terms in two parts: i) one that is correlated with our regressors through the use of the Mundlak-Chamberlain specification (Mundlak 1978, Chamberlain 1980, 1984) and with the initial value of the corresponding outcome variable; ii) another that is uncorrelated with our regressors. Hence, the specification for the unobserved heterogeneity terms is as follows:

$$c_i^g = \mathbf{Z}_i^g \boldsymbol{\gamma}_g + \delta_g y_{i,0}^g + v_i^g \tag{6}$$

with $g = 1,2$. The matrix $\mathbf{Z}^g$ denotes the means of the time-varying regressors, y_0^g the decisions in the initial period and v_i^g random terms uncorrelated with $\mathbf{Z}^g$.

Following Heckman and Singer (1984), we assume that v_i^g has a K-point nonparametric distribution with support $\{v_1^g, v_2^g, ..., v_K^g\}$. In order to facilitate convergence of our likelihood

function, we assume that the elements of $\mathbf{v}_k = (v_k^1, v_k^2)$ occur with a common probability p_k, $(k = 1,...,K)$.[11]

As a result of the above, the likelihood, conditional on the initial choices $\mathbf{y}_{i,0}$ and $\mathbf{W}_i = (\mathbf{X}_i^1, \mathbf{X}_i^2, \mathbf{Z}_i^1, \mathbf{Z}_i^2)$ becomes

$$f(\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, ..., \mathbf{y}_{i,T} \mid \mathbf{W}_i, \mathbf{y}_{i,0}) = \int \prod_{t=1}^{T} [f(\mathbf{y}_{i,t} \mid \mathbf{y}_{i,t-1}, \mathbf{X}_{i,t}, \mathbf{c}_i)] dG(\mathbf{c}_i \mid \mathbf{y}_{i,0}, \mathbf{Z}_i) = \prod_{t=1}^{T} \prod_{k=1}^{K} p_k \left[h(\mathbf{y}_{i,t} \mid \mathbf{y}_{i,t-1}, \mathbf{y}_{i,0}, \mathbf{W}_{i,t}, \mathbf{v}_k) \right] \tag{7}$$

The final step needed to make the likelihood operational is to specify the form of the joint distribution of the two decisions, while taking into account the selectivity induced by the correlation between the time-varying unobservables $\varepsilon_{i,t}^1, \varepsilon_{i,t}^2$. To this effect, we adapt the formulation proposed by Zabel (1992) and Greene (2007) so as to accommodate the non-parametric distribution of $\mathbf{v} = (v^1, v^2)$, and also allow for a double hurdle model specification. If we denote by L the number of cigarettes smoked, and by $\varepsilon_{i,t}^2(L,k) = L - (\mathbf{X}_{i,t}^2 \boldsymbol{\beta}_2 + \theta_2 y_{i,t-1}^2 + \mathbf{Z}_i^2 \boldsymbol{\gamma}_2 + \delta_2 y_{i,0}^2 + v_k^2)$ the time-varying error of the second stage decision, expressed as a function of L and the k[th] value of the purely random part of the unobserved heterogeneity, then the joint probability of the two smoking decisions can be written as

$$h(\mathbf{y}_{i,t} \mid \mathbf{y}_{i,t-1}, \mathbf{y}_{i,0}, \mathbf{W}_{i,t}, \mathbf{v}_k) = \begin{cases} \text{Prob}(y_{i,t}^1 = 0) + \text{Prob}(y_{i,t}^1 = 1, y_{i,t}^{*2} = 0) = \\ \Phi(-\mathbf{X}_{i,t}^1 \boldsymbol{\beta}_1 - \theta_1 y_{i,t-1}^1 - \mathbf{Z}_i^1 \boldsymbol{\gamma}_1 - \delta_1 y_{i,0}^1 - v_k^1) + \\ \Phi\left(\dfrac{\mathbf{X}_{i,t}^1 \boldsymbol{\beta}_1 + \theta_1 y_{i,t-1}^1 + \mathbf{Z}_i^1 \boldsymbol{\gamma}_1 + \delta_1 y_{i,0}^1 + v_k^1 + (\rho / \sigma_2) \varepsilon_{i,t}^2(0,k)}{\sqrt{1-\rho^2}} \right) \dfrac{1}{\sigma_2} \phi\left(\dfrac{\varepsilon_{i,t}^2(0,k)}{\sigma_2} \right) \\ \text{if not smoking any cigarettes (L = 0)} \\ \Phi\left(\dfrac{\mathbf{X}_{i,t}^1 \boldsymbol{\beta}_1 + \theta_1 y_{i,t-1}^1 + \mathbf{Z}_i^1 \boldsymbol{\gamma}_1 + \delta_1 y_{i,0}^1 + v_k^1 + (\rho / \sigma_2) \varepsilon_{i,t}^2(L,k)}{\sqrt{1-\rho^2}} \right) \dfrac{1}{\sigma_2} \phi\left(\dfrac{\varepsilon_{i,t}^2(L,k)}{\sigma_2} \right), \\ \text{if smoking one or more cigarettes (L > 0)} \end{cases} \tag{8}$$

[11] Other recent papers that use this non-parametric specification of heterogeneity are Halliday (2008) and Michaud and Tatsiramos (2008).

The log likelihood of all our sample observations, which needs to be maximized with respect to $\boldsymbol{\alpha} = (\theta_1, \theta_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \delta_1, \delta_2, \sigma_2, \rho, \mathbf{v}_2, .., \mathbf{v}_K, p_2, .., p_K)$ [12], is hence equal to

$$\ln L = \sum_{i=1}^{N} \left(\log \prod_{t=1}^{T} \prod_{k=1}^{K} p_k \left[h(\mathbf{y}_{i,t} \mid \mathbf{y}_{i,t-1}, \mathbf{y}_{i,0}, \mathbf{W}_{i,t}, \mathbf{v}_k) \right] \right) \tag{9}$$

Clearly, the likelihood function described in (8)-(9) is extremely complicated; therefore one needs to exercise care when choosing the initial values of the parameters in order to facilitate convergence. For each country/gender combination, we have experimented with a number of initial parameter values and report the results from those that led to the largest value of the likelihood.[13]

We also estimated our model with various numbers of non-parametric distributions points. We present results with four distribution points (the maximum number for which we managed to obtain convergence of the likelihood in general), given that the fit from the model estimated using four points was better than the one using three points, as determined by the value of the Akaike Information Criterion (AIC). The results for our magnitude of interest, however, namely state dependence in both smoking decisions (as measured by the average partial effects (APEs) of lagged smoking), did not in general change much when moving from three to four points. Hence, we surmise that the marginal benefit of adding more distribution points is quite small compared to the ensuing considerable additional computational cost.

Our estimation if further complicated by the fact that, as already discussed, the data on number of cigarettes smoked is plagued by heaping, which can cause severe biases in estimation (Heitjan and Rubin, 1990; Wang and Heitjan, 2008; Lillard, Bar and Wang, 2008). Ideally, one would like to have a separate model for the heaping process, which would allow for various heaping

[12] Given that the probabilities sum up to 1, one needs to estimate only K-1 probability parameters. In addition, as in Michaud and Tatsiramos (2008), one element in the support of each of the two non-parametric shocks is normalized to zero. See also Appendix A.2 for further details on the parameter vector $\boldsymbol{\alpha}$.

[13] For the case of females in Belgium, we show results from an estimation that converged to a local maximum, i.e. we have obtained slightly higher values of the likelihood but without convergence. The coefficient estimates in both cases, however, are very close to each other, so we conjecture that APEs should be similar as well.

possibilities for any given reported number of cigarettes smoked. This approach, however, proved impractical in our case given the already considerable complexity of our semi-parametric dynamic panel sample selection model. Therefore, we had take into account heaping in a simple fashion, namely by coarsening the data around the possible heaping points. To this effect we created an ordered indicator variable s that took the same value when the observed number of cigarettes was reported to be between two successive thresholds μ_{j-1} and μ_j. By inspecting Fig. 1, we observe that the heaping points are multiples of five and hence we chose the following values for our thresholds: 7.5, 12.5, 17.5, 22.5, 27.5 and 32.5. There is an additional threshold of zero that serves as a determinant of the second hurdle, as in HZ. The mapping between s and y^{*1}, y^{*2} is as follows

$$s_{i,t} = \begin{cases} 0 \text{ if } y_{i,t}^{*1} <= 0 \text{ or } (y_{i,t}^{*1} > 0 \text{ and } y_{i,t}^{*2} <= 0 = \mu_0) \\ j \text{ if } y_{i,t}^{*1} > 0 \text{ and } y_{i,t}^{*2} > \mu_{j-1} \text{ and } \mathrm{y}_{i,t}^{*2} <= \mu_j \quad (j = 1,..., J-1) \\ J \text{ if } y_{i,t}^{*1} > 0 \text{ and } y_{i,t}^{*2} > \mu_{J-1} \ \ (J = 7) \end{cases} \tag{10}$$

Since s is an ordinal indicator, we can use the zero-inflated ordered probit with selection of HZ, suitably adapted to our dynamic panel data context. The use of a discrete variable to denote the number of cigarettes implies that the equation for the latent variable y^{*2} is now as follows:

$$y_{i,t}^{*2} = \mathbf{X}_{i,t}^2 \boldsymbol{\beta}_2 + \sum_{j=1}^{J} \theta_2^j ds_{i,t-1}^j + c_i^2 + \varepsilon_{i,t}^2 \tag{11}$$

where $ds_{i,t}^j$ is a dummy variable that is equal to one if $s_{i,t} = j$, and zero otherwise $(j = 1,..,J)$. The equation for c_i^2 is modified in an analogous fashion

$$c_i^2 = \mathbf{Z}_i^2 \boldsymbol{\gamma}_2 + \sum_{j=1}^{J} \delta_2^j ds_{i,0}^j + v_i^2 \tag{12}$$

All the above imply that the joint decision probability of $\mathbf{t}_{i,t} = (y_{i,t}^1, s_{i,t})$, denoted by $h(\mathbf{t}_{i,t} \mid \mathbf{t}_{i,t-1}, \mathbf{t}_{i,0}, \mathbf{W}_{i,t}, \mathbf{v}_k)$, is as follows:

$$
h(\mathbf{t}_{i,t} \mid \mathbf{t}_{i,t-1}, \mathbf{t}_{i,0}, \mathbf{W}_{i,t}, \mathbf{v}_k) = \begin{cases}
\text{Prob}(y^1_{i,t} = 0) + \text{Prob}(y^1_{i,t} = 1, y^{*2}_{i,t} = 0) = \\
\Phi(-\mathbf{X}^1_{i,t}\boldsymbol{\beta}_1 - \theta_1 y^1_{i,t-1} - \mathbf{Z}^1_i\boldsymbol{\gamma}_1 - \delta_1 y^1_{i,0} - v^1_k) \; + \\
\Phi_2\left(\mathbf{X}^1_{i,t}\boldsymbol{\beta}_1 + \theta_1 y^1_{i,t-1} + \mathbf{Z}^1_i\boldsymbol{\gamma}_1 + \delta_1 y^1_{i,0} + v^1_k, \dfrac{\mu_0 - (\mathbf{X}^2_{i,t}\boldsymbol{\beta}_2 + \sum_{j=1}^{J}\theta_2^j ds^j_{i,t-1} + \mathbf{Z}^2_i\boldsymbol{\gamma}_2 + \sum_{j=1}^{J}\delta_2^j ds^j_{i,0} + v^2_k)}{\sigma_2}, -\rho \right) \\
\text{if not smoking any cigarettes } (s_{i,t} = 0, \mu_0 = 0) \\
\Phi_2\left(\mathbf{X}^1_{i,t}\boldsymbol{\beta}_1 + \theta_1 y^1_{i,t-1} + \mathbf{Z}^1_i\boldsymbol{\gamma}_1 + \delta_1 y^1_{i,0} + v^1_k, \dfrac{\mu_j - (\mathbf{X}^2_{i,t}\boldsymbol{\beta}_2 + \sum_{j=1}^{J}\theta_2^j ds^j_{i,t-1} + \mathbf{Z}^2_i\boldsymbol{\gamma}_2 + \sum_{j=1}^{J}\delta_2^j ds^j_{i,0} + v^2_k)}{\sigma_2}, -\rho \right) - \\
\Phi_2\left(\mathbf{X}^1_{i,t}\boldsymbol{\beta}_1 + \theta_1 y^1_{i,t-1} + \mathbf{Z}^1_i\boldsymbol{\gamma}_1 + \delta_1 y^1_{i,0} + v^1_k, \dfrac{\mu_{j-1} - (\mathbf{X}^2_{i,t}\boldsymbol{\beta}_2 + \sum_{j=1}^{J}\theta_2^j ds^j_{i,t-1} + \mathbf{Z}^2_i\boldsymbol{\gamma}_2 + \sum_{j=1}^{J}\delta_2^j ds^j_{i,0} + v^2_k)}{\sigma_2}, -\rho \right) \\
\text{if the number of cigarettes smoked is larger than } \mu_{j-1} \text{ and less or equal to } \mu_j \; (s_{i,t} > j-1 \\
\text{and } s_{i,t} <= j), \; j = 1,...,J-1 \\
\Phi_2\left(\mathbf{X}^1_{i,t}\boldsymbol{\beta}_1 + \theta_1 y^1_{i,t-1} + \mathbf{Z}^1_i\boldsymbol{\gamma}_1 + \delta_1 y^1_{i,0} + v^1_k, \dfrac{(\mathbf{X}^2_{i,t}\boldsymbol{\beta}_2 + \sum_{j=1}^{J}\theta_2^j ds^j_{i,t-1} + \mathbf{Z}^2_i\boldsymbol{\gamma}_2 + \sum_{j=1}^{J}\delta_2^j ds^j_{i,0} + v^2_k) - \mu_{J-1}}{\sigma_2}, \rho \right) \\
\text{if the number of cigarettes smoked is larger than } \mu_{J-1} \; (s_{i,t} = J)
\end{cases}
\tag{13}
$$

where Φ_2 denotes the bivariate normal distribution. The log likelihood function in (9) is modified accordingly.

It is important to note that the threshold values μ_j are typically unknown in a standard ordered probit model. Obviously, this is not true in our case, because we choose these values in order to address heaping in the number of cigarettes smoked. Knowing μ_j has important practical implications. Given that the ordered probit estimates the ratio between μ_j and σ_2 (the standard deviation of ε^2), our knowledge of μ_j allows us to identify σ_2. As a result, and given that our model also estimates the ratio of the coefficients of the equation for y^{*2} to σ_2 (as is apparent from (13)), we can identify the parameters $\boldsymbol{\beta}_2, \boldsymbol{\theta}_2 = (\theta_2^1,...,\theta_2^J), \boldsymbol{\gamma}_2$ and $\boldsymbol{\delta}_2 = (\delta_2^1,...,\delta_2^J)$ as well.

This result is important because in our model the latent variable y^{*2} has a precise and economically relevant meaning, which is not typically the case in ordered variable models. Namely,

y^{*2} denotes the true, but not accurately measured (due to heaping), number of cigarettes smoked. This fact, combined with the fact that the coefficients in (11) and (12) are all identified, implies that we can recover from our estimated ordered probit model the predicted mean of the true number of cigarettes smoked. Therefore, there is no need to examine the probability that y^{*2} lies in a given range, as would typically happen in an ordered variable model in which the latent variable has neither an economic meaning nor an identified predicted mean; instead, we can calculate the average partial effect (APE) of any variable of interest directly on the estimated conditional mean of cigarettes smoked.

In order to compare the fit between the linear panel model in (8) and the ordered probit panel model in (13), we computed the AIC values for both models. They are shown in Table 3 and it is obvious that for all country/gender combinations the ordered probit panel model gives a much better fit than the linear panel one. This result implies that heaping needs to be addressed, even in a simple manner as the one we use in this paper, when estimating models that use information on the reported number of cigarettes smoked.

To check the robustness of our results, and given that the number of cigarettes is a count variable, we substituted a Poisson specification for the linear one in the smoking intensity equation, while maintaining the non-parametric specification for $\mathbf{v}$ and modifying (8) appropriately. Unfortunately, convergence of the likelihood function proved very difficult to obtain.[14]. The AIC values are in most cases modestly better (in terms of fit) than the ones from the semi-parametric linear model, but again very inferior to those from the semi-parametric ordered probit model. The relative similarity of the fit from the Poisson model to the one from the linear one is not surprising, given that the conditional median number of cigarettes smoked is relatively large (20 for males and 15 for females). Hence a linear model could be an appropriate alternative to a count data one.

[14] Results from both the fully parametric linear model and the semi-parametric Poisson model are available upon request.

As an additional robustness check, we experimented with the fully parametric specification of unobserved heterogeneity proposed by Greene (2007) for the linear panel model. Hence we substituted a bivariate normal distribution for the nonparametric one for $\mathbf{v}$, and estimated the resulting linear model through simulated maximum likelihood. The estimation turned out to be extremely time consuming (it took roughly 15 times longer on average than when using the nonparametric specification for $\mathbf{v}$), and the likelihood function converged with great difficulty. For the cases in which we managed to obtain convergence, the AIC values were slightly worse than the ones obtained from the semi-parametric linear model.

To sum up, we find that our semi-parametric ordered probit model is by far superior in terms of fit to both a linear and a Poisson model, and convergence of its associated likelihood function is not prohibitively time consuming.[15] Furthermore, in the case of the linear model the non-parametric distributional specification for the random effect $\mathbf{v}$ is preferable in terms of fit to the fully parametric specification. Even more importantly, results from the non-parametric distribution of $\mathbf{v}$ should be considerably more robust to misspecification than the ones obtained using the bivariate normal distribution (see Mroz, 1999). Therefore, we will base our subsequent analyses of smoking behaviour in our sample on the semi-parametric ordered probit model.

IV. Empirical Results

As already discussed, our aim is to estimate the true state dependence in both the decision to smoke and in the decision on the number of cigarettes smoked. For the decision to smoke, this true state dependence is given by the APE of the lagged smoking decision $y^1_{i,t-1}$ on the probability to

[15] The model takes on average approximately 1.5 hours to converge for a single country/gender combination, on a 2.7 GHz quad-core PC running the multi-processor version of Stata. Roughly the same time is required for the linear semi-parametric model.

smoke in the current period. This probability is equal to the probability that one has overcome both the participation hurdle and is not at a corner solution, that is

$$1 - \left[\text{Prob}(y_{i,t}^1 = 0) + \text{Prob}(y_{i,t}^1 = 1, y_{i,t}^{*2} = 0)\right] \tag{14}$$

The probability of non-smoking (and thus its complement) can be easily computed from (8) and (13), after integrating out the non-parametric terms $\mathbf{v}$ using their associated estimated probabilities. We have also examined the APEs on just the probability of participation, that is on $\text{Prob}(y_{i,t}^1 = 1)$, and they were virtually identical to the APEs on the probability of smoking shown in (14).[16]

With respect to the number of cigarettes smoked, we are interested in the APE of the lagged number of cigarettes smoked on the current predicted mean number of cigarettes *conditional* on smoking, that is *conditional* on the number of cigarettes smoked being larger than zero. Due to selectivity, this conditional mean is not equal to the linear index for the current number of cigarettes smoked (shown in (4) for the linear model. and in (11) for the ordered probit one). Rather, it is equal to

$$E(y_{i,t}^2 \mid y_{i,t}^2 > 0) = \sum_{k=1}^{K} p_k \left(\mathbf{X}_{i,t}^2 \boldsymbol{\beta}_2 + \sum_{j=1}^{J} \theta_2^j ds_{i,t-1}^j + \mathbf{Z}_i^2 \boldsymbol{\gamma}_2 + \sum_{j=1}^{J} \delta_2^j ds_{i,0}^j + v_k^2 + \rho\sigma_2 \frac{\varphi(\mathbf{X}_{i,t}^1 \boldsymbol{\beta}_1 + \theta_1 y_{i,t-1}^1 + \mathbf{Z}_i^1 \boldsymbol{\gamma}_1 + \delta_1 y_{i,0}^1 + v_k^1)}{\Phi(\mathbf{X}_{i,t}^1 \boldsymbol{\beta}_1 + \theta_1 y_{i,t-1}^1 + \mathbf{Z}_i^1 \boldsymbol{\gamma}_1 + \delta_1 y_{i,0}^1 + v_k^1)} \right) \tag{15}$$

The last term in (15) denotes the inverse Mills ratio and is due to the correlation between the time varying unobservables $\varepsilon_{i,t}^1, \varepsilon_{i,t}^2$ of the two decisions. While (15) holds for the semi-parametric ordered probit model, for the linear model the lagged dummies of the ordinal variable should be replaced by the lagged number of cigarettes smoked $y_{i,t-1}^2$ as in (4), and an analogous change is needed for cigarette consumption in the initial period. As discussed in Section IV, in the case of the ordered probit model our knowledge of the threshold values allows us to identify the conditional mean in (15).

[16] We do not examine the estimates of θ_1, the coefficient of the lagged dependent variable $y_{i,t-1}^1$, because in binary models coefficients are identified only up to scale.

We compute APEs via Monte Carlo simulation (details can be found in Appendix A.2). For the sake of brevity, we only report the APE of smoking between 17.5 and 22.5 cigarettes in period $t-1$ (compared to not smoking at all) on the number of cigarettes currently smoked, for the ordered probit model. The corresponding APE for the linear model is computed as the change in the predicted number of cigarettes smoked due to a change from not smoking to smoking 20 cigarettes in period $t-1$. In both cases, switching from non-smoking to smoking in period $t-1$ affects the conditional mean of cigarettes smoked in period t not only through the change in the lagged cigarette consumption but also through the change in the lagged discrete smoking indicator $y_{i,t-1}^{1}$ that enters into the inverse Mills ratio term.[17] In order to illustrate the importance of accounting for unobserved heterogeneity when estimating true state dependence, the APEs for both smoking decisions derived from the two semi-parametric panel models are compared to the APEs derived from the corresponding two simple models, where the unobserved heterogeneity terms c^1, c^2 are set to zero (in what follows, we refer to these models as pooled models).

When ρ is sizeable, that is when selectivity is present, the APE on the conditional mean will be quite different from the coefficient of the lagged consumption variable. We indeed find that our estimates of ρ are typically large and statistically significant (they are shown in Table A.2)[18], which implies that ignoring selectivity can lead to serious inconsistencies in our results, both for coefficient estimates and for APEs.

Tables 4A and 4B display, for males and females respectively, country-specific APEs of lagged smoking participation and lagged number of cigarettes smoked from the four aforementioned models. It is worth noting that females and males share several common features.

[17] As a result, the APE of the lagged cigarette consumption on the conditional mean of the current one will not be in general equal to the corresponding coefficient.

[18] Due to space constraints, we do not display results on any other estimated regression parameters. They are available upon request from the authors.

The estimated APEs always decrease when unobserved individual heterogeneity is accounted for, which confirms our prior that persistent smoking behaviour does not reflect only true state dependence. For instance, the probability of smoking for Danish females (males) who smoked in the previous period is on average 88.0 (87.3) percentage points (pp) higher than for their non-smoking counterparts according to the pooled ordered probit estimates, while the corresponding APEs from the semi-parametric ordered probit model (26.2 and 26.9 pp for females and males, respectively) are reduced by more than 60%. The relative magnitude of this reduction is generally similar in all countries for females and males for both the linear and the ordered probit model; the APE of lagged smoking participation, however, remains always strongly statistically significant at standard levels of testing. The results associated with the smoking intensity equation display a similar pattern: the pooled ordered probit APE of smoking between 20 cigarettes in the previous period is generally at least twice as big as the corresponding APE derived from the semi-parametric ordered probit estimation. Following on with the Danish case, the pooled ordered probit APE amounts to 9.4 (6.5) cigarettes for females (males), while the semi-parametric ordered probit APE is reduced by more than 50%: 3.2 cigarettes for females and 1.6 cigarettes for males. Moreover, there are a few cases in which, after accounting for unobserved individual heterogeneity, statistically significant results are no longer obtained in the intensity equation: this is the case for Portugal and Italy for females and Portugal and Ireland for males. The APE for Greek males in the intensity equation was not statistically significant either in the pooled or in the semi-parametric panel ordered probit model.

Another interesting feature shared by both females and males is that the magnitude of the pooled APEs is not only large but remarkably homogeneous across countries. For example, the pooled ordered APEs in the participation equation range between 87.8 pp for British males and 74.0 pp for Irish or Greek males, while the smallest APE for females (74.6 pp in Spain) is not too different from the largest one either (93.8 in Austria). Instead, when we turn to the semi-parametric ordered probit model, there is much larger cross-country variability in the magnitude of the APEs of

interest, which now range between 9.7 pp (Ireland) and 54.4 pp (Italy) for males, and between 4.3 and 52.1 pp for females in the same countries. These results are in line with those from the linear models and those associated with the intensity equation, which indicates that there are different patterns of unobserved heterogeneity across countries.

Finally, it is also worth mentioning that our results do not exhibit a clear geographical pattern. For females, state dependence appears to be stronger in the North than in the South of Europe, with the exception of Italy, where state dependence is relatively high. However, this is not the case for males, for whom the impact of lagged smoking behaviour is relatively small in both Spain and Greece.

An interesting but difficult question is what are the forces driving these cross-country differences in the magnitude of true state dependence. We provide some tentative evidence on this by relying on country measures of the beliefs about and social acceptance of smoking (constructed from various Eurobarometer surveys,[19] as well as on indicators of smoking regulations (from the World Health Organization Tobacco Control Database.[20] In particular, we correlate these indicators with different measures of inflows into and outflows from smoking status in an attempt to uncover patterns in the magnitude of smoking transitions across countries.

Several Eurobarometer surveys ask respondents whether they believe that smoking causes cancer and death, whether they would be encouraged to quit if they got scientific proof that smoking causes serious illnesses, if they have heard about passive smoking, and if they believe that smoking can cause health problems to non smokers. The first and the second questions are asked in Eurobarometer 41.0 (1994), the third in Eurobarometer 38.0 (1992) and the fourth question is asked in Eurobarometers 38.0 (1992) and 64.1 (2005). These questions inform us about the extent to

[19] The Eurobarometer is conducted on behalf of the European Commission in order to monitor public opinion in the European Union. For detailed information on the Standard and Special Eurobarometer Surveys, see http://www.gesis.org/en/data%5Fservice/eurobarometer/index.htm

[20] See http://data.euro.who.int/tobacco/

which adult individuals in European countries are aware of the health consequences of smoking both for smokers and for non smokers. Other questions from the Eurobarometer surveys allow us to assess the degree of social acceptance that smoking enjoys. In particular, individuals are asked whether they are in favour of having smoking bans in any indoor public space and whether they believe that advertisement for cigarettes should not be regulated in any way. These questions are asked in Eurobarometer 64.1 (2005), and in Eurobarometer 41.0 (1994), respectively.

As for smoking regulations, we consider smoking restrictions in government facilities, indoor workplaces and offices and restaurants. The World Health Organization (WHO) Tobacco Control Database provides information, for all the countries in our sample, on whether there is a ban, a partial restriction, a voluntary agreement or no regulation at all in each of these contexts.[21]

In Table 5 we report Spearman rank correlations between the indicators outlined above and two different measures of outflows from and inflows into smoking: those computed from the data (displayed in Table 1) and those derived from the semi-parametric panel ordered probit model. We choose to report correlations with these two sets of transition indicators because the former are influenced by both true state dependence and unobserved individual heterogeneity while the latter should be less affected by unobserved individual heterogeneity due to our panel estimation procedures. Hence, given that individuals' attitudes towards smoking as well as their beliefs about the hazards of smoking are part of individuals' unobserved heterogeneity, one would expect these attitudes and beliefs to be significantly correlated with at least some of the descriptive flows. On the other hand, they should not to be significantly correlated with the transition rates derived from the semi-parametric panel ordered probit model, the coefficients of which should be unaffected by unobserved individual heterogeneity.

[21] Other anti-smoking policies such as taxes or the minimum age for buying tobacco products are not considered in this analysis because they do not display sufficient country-variation within the group of countries and data period under study.

This expectation is generally confirmed by the results displayed in Table 5. For instance, outflows computed from the data (that is, the percentage of individual smokers who quit between consecutive years) are significantly higher in countries with a higher percentage of individuals who are in favour of having smoking bans in indoor public spaces and who believe that smoking causes cancer and death. Along the same lines, quitting is significantly less frequent in countries where more individuals believe that advertisement for cigarettes should not be regulated in any way. The correlations for inflows and those pertaining to variables capturing the beliefs about the harms of smoking, however, are not statistically significant at standard levels of testing.

The regulatory indicators, unlike the previous variables capturing beliefs and attitudes, are country-specific rather than individual-specific. Therefore, it may be the case that cross-country differences in the magnitude of the flows derived from the semi-parametric panel ordered probit model can partially be explained by differences in regulations. Actually, this seems to be the case with the existence of a smoking ban in government facilities: countries with such a ban also have significantly higher outflows as measured after accounting for unobserved individual heterogeneity.

V. Discussion

There are two important messages from the results presented in Section IV: i) accounting for unobserved individual heterogeneity leads to smaller estimates of true state dependence in smoking; ii) even after taking unobserved heterogeneity into account, the estimated true state dependence is very large in the vast majority of cases.

The goal of this Section is twofold. First, we want to quantify the implications for smoking behaviour of our accounting for unobserved individual heterogeneity. Second, we want to assess the effect of state dependence in smoking behaviour on smoking prevalence and transitions. To this purpose, we perform two simulation exercises (more details on the simulations can be found in Appendix A.3).

In our first simulation, we compare the impact that lagged smoking behaviour has on smoking prevalence when unobserved individual heterogeneity is accounted for and when it is not. First, we simulate individuals' smoking decisions using all the semi-parametric panel ordered probit coefficient estimates and compare them with the analogous figures when the coefficient estimate on the lagged smoking variable is set to zero. This difference and its associated standard errors are displayed in columns 1 and 2 of Table 6, respectively.[22] We then make the same comparison but instead we rely on the pooled ordered probit model, and report the corresponding results in columns 3 and 4 of Table 6. The difference between these two differences, displayed in column 5 of Table 6, is our measure of the change in the effect of state dependence on smoking prevalence when we do address the issue of unobserved individual heterogeneity and when we do not.

Our results indicate that our double difference is statistically significant for both genders and in all countries considered. Moreover, its magnitude is generally large: not accounting for unobserved individual heterogeneity would lead to a much lower prevalence of smoking (10 pp in all countries, except for Italy for males and Italy and Portugal for females) when putting the coefficient of state dependence to zero, compared to when we account for such heterogeneity. Given that the estimated state dependence is much stronger when unobserved heterogeneity is not taken into account, putting it to zero induces many more transitions out of smoking (and thus a much lower overall smoking prevalence) in the pooled model than in the panel model.

The goal of our second simulation exercise is to illustrate how much smoking prevalence and transitions would be changed if the magnitude of the state dependence parameters were modified to be equal to the smallest estimated ones found among the countries in our sample. We observe that Ireland has the smallest state dependence parameters for both genders overall; therefore we take Ireland's estimates and use them to simulate counterfactual prevalence of smoking and associated transitions for each of the other countries using our semi-parametric panel

[22] All the magnitudes of interest and their standard errors are computed via Monte Carlo simulation, conducted in a way analogous to the one use for the computation of APEs, as described in Appendix A.2.

ordered probit model. These counterfactual outcomes are then compared to the corresponding results obtained when using the originally estimated country-specific state dependence parameters. In both cases, coefficients of all other covariates remain equal to their original estimates for each country, which allows us to focus our attention only on the consequences of altering the state dependence parameters.

One would expect that simulated transitions in each country are lower than their counterfactual counterparts obtained when using Ireland's coefficients, since lower state dependence induces more transitions. The results shown in Table 7 confirm this conjecture, and the differences are always statistically significant (transitions into smoking are displayed in Column 3 while those out of smoking in column 5). As an example, transitions for Italian males are significantly less likely than they would have been had this country had the lower Irish coefficient of true state dependence: inflows into and outflows out of smoking are reduced by 3.5 and 37.8 percentage points, respectively. Given that transitions out of smoking are quantitatively much more important than transitions into smoking, the overall effect on the smoking rate is negative, that is the counterfactual smoking rate computed using Ireland's coefficients is smaller than the originally estimated rate. For example, the average smoking rate for males in Belgium during the estimation period is 9.3 pp higher than the corresponding counterfactual rate, and this difference is important for all country/gender combinations.

All in all, both our simulation exercises show that changes in state dependence, induced by taking into account unobserved heterogeneity or by counterfactually varying the magnitude of the true state dependence coefficients, have very important consequences for smoking rates and the associated smoking transitions. Therefore, our simulation results provide further evidence that reducing true state dependence, even when such dependence is quite smaller than the observed one, is a worthwhile target for policy makers who are interested in reducing the prevalence of smoking.

VI. Conclusion

We study smoking persistence in Europe from a comparative perspective using internationally comparable data from the ECHP for ten countries. The longitudinal nature of our dataset allows us to analyze the dynamics of smoking behavior and to investigate the extent to which its observed persistence reflects true state dependence as opposed to individual unobserved heterogeneity. To this purpose, we use semi-parametric nonlinear panel data methods and consider both smoking participation and smoking intensity decisions, allowing for correlation in the two time-varying error terms.

From a methodological perspective, we depart from and complement previous related studies in several important ways. As in Gilleskie and Strumpf (2004), we use longitudinal data to explore the dynamics of smoking participation and intensity decisions; however, we additionally incorporate selectivity into our model. This point is empirically relevant, as we find that the time varying unobservables in the smoking participation and cigarette consumption equations are indeed significantly correlated with each other in a number of cases. In addition, like HZ, we use a double hurdle specification and incorporate selectivity into the model; however, we do so within a dynamic framework and use a semi-parametric approach to deal with the fixed effects that are potentially correlated with the regressors.

We find statistically significant and economically relevant estimates of true state dependence in most countries. For both males and females, however, we show that accounting for individual unobserved heterogeneity leads to a large reduction in the magnitude of the impact of lagged smoking behavior. In other words, a non negligible fraction of the observed persistence in smoking reflects unobserved individual heterogeneity rather than true state dependence. Our results also indicate that taking into account of heaping in self-reported cigarette consumption data is important, given that our ordinal specification yields a substantially better fit than linear or Poisson models that do not take such heaping into account.

The reduced size of our estimated state dependence is an important finding from a policy perspective because it suggests that the cumulative impact of policies aimed at reducing smoking

prevalence is likely to be smaller in the long run than one would expect when departing from the spurious estimates of the impact of lagged smoking behavior resulting from pooled models. In the light of this evidence, interventions at early ages that effectively impact the development of the personal traits that form individual unobserved heterogeneity should be a useful tool in the policy mix. We also show via our simulations, however, that our estimated true state dependence, even if smaller in magnitude after correcting for the presence of unobserved individual heterogeneity, has important implications for smoking prevalence across all countries in our sample. Therefore, it should remain a target for policymakers who have a reduced smoking rate as an objective.

References

Becker, G.S., Murphy, K.M., 1988. “A theory of rational addiction.” *Journal of Political Economy* 96(4), 675-700.

Becker, G. S., Grossman, M., Murphy, K. M., 1991. "Rational Addiction and the Effect of Price on Consumption." *American Economic Review* 81(2), 237-241

Becker, G. S., Grossman, M., Murphy, K. M., 1994. "An Empirical Analysis of Cigarette Addiction." *American Economic Review* 84(3), 396-418.

Browning, M., Collado, M.D., 2007. “Habits and heterogeneity in demands: a panel data analysis.” *Journal of Applied Econometrics* (22), 625-640.

Carrasco, R., Labeaga, J.M., López-Salido, J.D., 2005. “Consumption and habits: evidence from panel data.” *Economic Journal* (115), 144-165.

Chaloupka, F., 1991. “Rational addictive behavior and cigarette smoking.” *Journal of Political Economy* 99(4), 722-742.

Chamberlain, G., 1980. “Analysis of covariance with qualitative data.” *Review of Economic Studies* 47, 225–238.

Chamberlain, G., 1984. “Panel Data.” In *Handbook of Econometrics, Vol. 2*, Z. Griliches and M.D. Intriligator (eds). Elsevier Science.

Coppejans, M., Gilleskie, D., Sieg, H., Strumpf, K., 2007. ”Consumer Demand under Price Uncertainty: Empirical Evidence from the Market for Cigarettes.” *Review of Economics and Statistics* 89 (3), 510-521.

Cutler, D.M, Glaeser, E.L., 2006. “Why do Europeans smoke more than Americans?” NBER Working Paper No. 12124.

Gilleskie, D.B., Strumpf, K.S., 2000. “The behavioral dynamics of youth smoking.” NBER Working Paper No. 7838.

Gilleskie, D.B., Strumpf, K.S., 2005. “The behavioral dynamics of youth smoking.” *Journal of Human Resources* 40(4), 822-866.

Greene, W., 2007. "Censored data and truncated distributions." In *The Palgrave Handbook of Econometrics Volume 1: Econometric Theory*, T.C. Mills. and C. Patterson, (eds.), 695-736. Palgrave Macmillan.

Gruber, J., Zinman, J., 2000. "Youth smoking in the US: evidence and implications." NBER Working paper No. 7780.

Halliday, T.J., 2008. "Heterogeneity, state dependence and health." IZA Discussion Paper No. 3463.

Harris, M.N., Zhao, X., 2007. "A zero-inflated ordered probit model with an application to modelling tobacco consumption."*Journal of Econometrics* 141, 1073-1099.

Heckman, J., Singer, B., 1984. "A method for minimizing the impact of distributional assumptions in econometric models for duration data." *Econometrica* 52(2), 271-320.

Heitjan, D.F., Rubin, D., 1990. "Inference from coarse data via multiple imputation with an application to age heaping." *Journal of the American Statistical Association* 84 (410), 304-314

Jones, A.M., 1989. "A double-hurdle model of cigarette consumption." *Journal of Applied Econometrics* 4, 23-29.

Lillard, D.R., Bar, H., Wang, H, 2008. "A heap of trouble? Accounting for mismatch bias in retrospectively reported data (with application to smoking cessation and (non) employment)." Mimeo.

Labeaga, J.M., 1999. "A double-hurdle rational addiction model with heterogeneity: Estimating the demand for tobacco." *Journal of Econometrics* 93, 49-72.

Jiménez-Martín, S., Labeaga, J. M., Rochina-Barrachina, M.E. 2009. "Some estimators for dynamic panel data sample selection and switching models." Mimeo.

Michaud, P.C., Tatsiramos, K., 2008. "Fertility and female employment dynamics in Europe: the effect of using alternative econometric modelling assumptions." IZA Discussion Paper No. 3853.

Mroz, T. A, 1999. “Discrete factor approximations in simultaneous equation models: Estimating the impact of a dummy endogenous variable on a continuous outcome.” *Journal of Econometrics* (92): 233-274.

Mundlak, Y., 1978. “On the pooling of time series and cross sectional data.” *Econometrica* 56, 69–86.

Semykina, A., Wooldridge, J.M., 2007. “Estimation of dynamic panel data models with sample selection.” Mimeo.

Train, K., 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press: Cambridge.

Wooldridge, J.M., 2005. “Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity.” *Journal of Applied Econometrics*, 20, 39-54.

Wang, H., Heitjan, D.F., 2008. “Modelling heaping in self-reported cigarette counts.” *Statistics in Medicine* 27, 3789–3804.

Zabel, J. 1992. “Estimating Fixed and Random Effects Models with Selectivity,” *Economics Letters*, 40, 269-272.

Appendix A.1 Construction of Relative Tobacco Prices

The relative tobacco prices plotted in Figure are both comparable across countries and over time. In order to construct these indicators we depart from Eurostat's data on prices[23] and follow several steps. First, we apply the rate of change of the harmonized index of consumer prices (HICP) for tobacco (which is comparable across time but not across countries) to the price level index for tobacco (which is cross country comparable but not intertertemporally consistent). This procedure yields a measure of tobacco prices that is both comparable across countries and over time. Second, we apply the rate of change of the HICP to the price level index of household consumption expenditure (HFCE), which yields a general price indicator that is comparable both across countries and over time. Finally, we divide the first calculated price variable by the second one in order to obtain our measure of relative tobacco prices.

Appendix A.2 Calculation of Magnitudes of Interest via Monte Carlo Simulation

The magnitudes σ_2, ρ and $\mathbf{p} = (p_2,...,p_K)$ must all satisfy constraints: σ_2 must be greater than zero, ρ must lie between minus one and one and $p_2,...,p_K$ must be between zero and one. These constraints make convergence of our already complicated likelihood function even more difficult. Therefore we estimate σ_2, ρ and $\mathbf{p}$ as functions of the unconstrained parameters τ, ψ, and $\boldsymbol{\omega} = (\omega_2,...,\omega_K)$. These new parameters thus replace σ_2, ρ and $\mathbf{p}$ in the parameter vector $\boldsymbol{\alpha}$ (shown in Section III), with respect to which the likelihood function is maximized. The mapping between the new parameters and σ_2, ρ and $\mathbf{p}$ is as follows:

[23] See http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1996,45323734&_dad=portal&_schema=PORTAL&screen=welcomeref&open=/prc/prc_ppp&language=en&product=EU_MASTER_prices&root=EU_MASTER_prices&scrollto=150

$$\sigma_2 = e^{\tau}$$
$$\rho = \frac{e^{\psi}-1}{e^{\psi}+1} \tag{16}$$
$$p_k = \frac{e^{\omega_k}}{\sum_{k=1}^{K} e^{\omega_k}} \text{ with } \omega_1 = 0, k = 1,...,K$$

Given that APEs, σ_2, ρ and $\mathbf{p}$ all represent magnitudes that are nonlinear functions of the estimated parameters $\hat{\boldsymbol{\alpha}}^* = (\hat{\theta}_1, \hat{\theta}_2, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \hat{\delta}_1, \hat{\delta}_2, \hat{\tau}, \hat{\psi}, \hat{\mathbf{v}}_2, ..., \hat{\mathbf{v}}_K, \hat{\omega}_2, ..., \hat{\omega}_K)$, we compute their point estimates and standard errors via Monte Carlo simulation (Train, 2003), that is by using the formula

$$E(g(\alpha)) = \int g(\alpha) f(\alpha) d\alpha \tag{17}$$

where $g(\alpha)$ denotes the magnitude of interest and $f(\alpha)$ the joint distribution of all the elements in α. We implement this simulation estimator by drawing 1,000 times from the joint distribution of the vector of parameters $\hat{\alpha}^*$ under the assumption that it is asymptotically normal with mean and variance-covariance matrix equal to the maximum likelihood estimates. For a given parameter draw j we generate the magnitude of interest $g(\hat{\alpha}^{*j})$. For APEs in particular, we first calculate the partial effect corresponding to each individual in our sample and then calculate the APE $g(\hat{\alpha}^{*j})$ as the weighted average (using sample weights) of the effect across individuals.[24] We then estimate $E(g(\alpha))$ and its standard error as the mean and standard deviation respectively of the distribution of $g(\hat{\alpha}^{*j})$ over all parameter draws.

Appendix A.3 Details on the Simulations

For each of the simulations discussed in Section V, and for a given j draw of the parameters $\hat{\alpha}^{*j}$, we draw fifty times from a [0,1] uniform distribution u. Individuals in our sample are considered to be smokers if their predicted probability of smoking, which is equal to $1 - \left[\text{Prob}(y^1_{i,t} = 0) + \text{Prob}(y^1_{i,t} = 1, y^{*2}_{i,t} = 0)\right]$, is larger than the uniform draw under the assumptions

[24] We do not evaluate marginal effects at sample means since this practice can lead to severely misleading results (see Train, 2003, pp. 33-34).

characterizing the given simulation of interest. The opposite happens for a uniform draw that lies above the predicted probability of smoking. This assignment of smoking status in period $t-1$ gives the value of the lagged smoking decision $y^1_{i,t-1}$ in period t. This process is repeated for all uniform draws and for all draws from the distribution of the parameter vector $\hat{\alpha}^*$ (assumed to be normal as described in Appendix A.2). For each combination of the two draws we compute our magnitudes of interest, that is the percentage of smokers, the rate of smoking transitions, and their differences across models and for different assumptions about state dependence as needed. The final estimate of every magnitude of interest and its standard error are then calculated, as described in Appendix A.2, as the average and the standard deviation, respectively, of its distribution over all combinations of the two draws.

Table 1. Smoking Rates and Transitions

Country	(1) Smoking Rate	(2) Rate of Transitions into Smoking	(3) Rate of Transitions out of Smoking	(4) Mobility Index
		Panel A. Males		
Finland	0.274	0.036	0.092	0.129
Denmark	0.302	0.033	0.105	0.138
United Kingdom	0.265	0.032	0.102	0.134
Ireland	0.264	0.061	0.190	0.251
Belgium	0.286	0.043	0.107	0.150
Austria	0.342	0.046	0.108	0.154
Italy	0.323	0.062	0.144	0.207
Spain	0.372	0.098	0.162	0.260
Portugal	0.319	0.050	0.096	0.146
Greece	0.499	0.128	0.128	0.256
		Panel B. Females		
Finland	0.187	0.026	0.105	0.132
Denmark	0.313	0.032	0.095	0.127
United Kingdom	0.263	0.028	0.097	0.125
Ireland	0.281	0.065	0.166	0.230
Belgium	0.212	0.025	0.095	0.120
Austria	0.215	0.039	0.148	0.187
Italy	0.151	0.031	0.166	0.197
Spain	0.223	0.058	0.198	0.256
Portugal	0.092	0.019	0.109	0.128
Greece	0.222	0.051	0.176	0.227

Note: The rate of transitions into smoking denotes the percentage of non-smokers in year t-1 who smoke in year t. The rate of transitions out of smoking is defined analogously.
Source: 1998 – 2001 waves of ECHP

Table 2. Sample Statistics

Variable	(1)	(2)	(3)	(4)
	Males		Females	
	Unbalanced Panel	Balanced Panel	Unbalanced Panel	Balanced Panel
Age	0.418	0.408	0.413	0.422
Primary School Education	0.455	0.454	0.475	0.482
Secondary School Education	0.333	0.339	0.328	0.319
Tertiary Education	0.212	0.207	0.198	0.199
Married	0.630	0.599	0.619	0.645
Divorced or Separated	0.035	0.035	0.055	0.055
Widowed	0.008	0.008	0.043	0.044
Never Married	0.327	0.357	0.283	0.256
Household Size	3.67	3.68	3.60	3.60
Young children at home	0.291	0.284	0.297	0.301
Home Owner	0.795	0.790	0.780	0.786
Spent Night at the Hospital during the Previous Year	0.062	0.063	0.073	0.072
Visited Specialist during the Previous Year	0.342	0.337	0.525	0.532
Employee	0.579	0.572	0.446	0.449
Self-Employed	0.186	0.179	0.086	0.090
Unemployed	0.054	0.058	0.056	0.053
Inactive	0.181	0.191	0.413	0.409
Currently a smoker	0.335	0.335	0.202	0.199
Number of cigarettes smoked (conditional median)	20	20	15	15
No. of observations	88,666	73,260	93,816	78,861

Source: 1998 – 2001 waves of ECHP

Table 3. Akaike Criterion for the Semi-parametric Panel Ordered Probit and Linear Models

Country	(1)	(2)	(3)	(4)
	Males		Females	
	Semi-parametric Ordered Probit Model	Semi-parametric Linear Model	Semi-parametric Ordered Probit Model	Semi-parametric Linear Model
Finland	5,801.7	10,669.2	3,978.6	7,497.3
Denmark	4,136.4	7,975.0	4,196.9	8,522.0
United Kingdom	8,293.8	15,761.8	8,832.0	17,176.8
Ireland	4,988.9	8,601.3	5,147.8	8,892.2
Belgium	4,496.0	8,283.9	3,713.8	7,348.1
Austria	7,165.9	13,682.0	4,774.8	8,440.3
Italy	20,070.8	35,447.7	10,195.0	17,404.9
Spain	20,158.9	35,960.2	13,666.5	23,431.8
Portugal	14,394.3	28,247.1	3,745.2	6,491.4
Greece	17,945.8	35,576.5	9,548.7	16,114.4

Note: Lower values indicate a better fit.

Table 4A. Average Partial Effects of Lagged Smoking Decisions, Males

Country	(1)	(2)		(3)	(4)		(5)	(6)		(7)	(8)	
	Pooled Ordered Probit Model			Pooled Linear Model			Semi-parametric Panel Ordered Probit Model			Semi-parametric Panel Linear Model		
	M. Eff.	Std. Error		M. Eff.	Std. Error		M. Eff.	Std. Error		M. Eff.	Std. Error	
Finland												
Probability of Smoking	0.863	0.009	***	0.847	0.009	***	0.197	0.054	***	0.359	0.045	***
Conditional mean of cigarettes smoked	6.39	0.511	***	6.92	0.420	***	2.80	0.735	***	1.53	0.499	***
Denmark												
Probability of Smoking	0.873	0.011	***	0.855	0.011	***	0.269	0.035	***	0.414	0.050	***
Conditional mean of cigarettes smoked	6.47	1.401	***	8.38	0.536	***	1.68	0.745	**	2.42	0.659	***
United Kingdom												
Probability of Smoking	0.878	0.008	***	0.851	0.008	***	0.330	0.032	***	0.199	0.056	***
Conditional mean of cigarettes smoked	10.31	1.030	***	10.42	0.403	***	2.95	0.575	***	4.23	0.512	***
Ireland												
Probability of Smoking	0.740	0.014	***	0.714	0.015	***	0.097	0.023	***	0.120	0.024	***
Conditional mean of cigarettes smoked	4.76	0.628	***	4.91	0.467	***	1.14	0.749		5.33	0.715	***
Belgium												
Probability of Smoking	0.867	0.011	***	0.846	0.011	***	0.245	0.130	*	0.266	0.052	***
Conditional mean of cigarettes smoked	8.77	0.749	***	8.66	0.570	***	4.32	1.009	***	4.60	0.660	***
Austria												
Probability of Smoking	0.830	0.009	***	0.814	0.009	***	0.138	0.027	***	0.074	0.019	***
Conditional mean of cigarettes smoked	4.65	0.601	***	4.98	0.516	***	4.04	0.745	***	3.18	0.566	***
Italy												
Probability of Smoking	0.790	0.006	***	0.775	0.006	***	0.544	0.021	***	0.182	0.018	***
Conditional mean of cigarettes smoked	4.47	0.285	***	5.04	0.234	***	2.24	0.367	***	2.55	0.286	***
Spain												
Probability of Smoking	0.748	0.007	***	0.709	0.008	***	0.103	0.017	***	0.072	0.010	***
Conditional mean of cigarettes smoked	5.78	0.404	***	5.07	0.296	***	1.81	0.453	***	3.23	0.350	***
Portugal												
Probability of Smoking	0.851	0.006	***	0.844	0.006	***	0.329	0.026	***	0.300	0.021	***
Conditional mean of cigarettes smoked	1.51	0.362	***	4.16	0.303	***	0.22	0.402		2.39	0.298	***
Greece												
Probability of Smoking	0.740	0.007	***	0.732	0.008	***	0.192	0.019	***	0.282	0.019	***
Conditional mean of cigarettes smoked	0.26	0.393		2.23	0.284	***	0.02	0.477		0.29	0.335	

Note: The first row in each country denotes the APE of the lagged decision to smoke or not on the probability of smoking in the current period. The second row denotes, for the ordered probit model, the APE of smoking between 17.5 and 22.5 cigarettes (compared to not smoking at all) during the previous period on the current conditional mean number of cigarettes smoked. For the linear model, the APE measures the effect of the change from not smoking at all to smoking 20 cigarettes during the previous period. ***, **, * denote statistical significance at 1%, 5% and 10%, respectively.

Table 4B. Average Partial Effects of Lagged Smoking Decisions, Females

Country	(1)	(2)		(3)	(4)		(5)	(6)		(7)	(8)	
	Pooled Ordered Probit Model			Pooled Linear Model			Semi-parametric Panel Ordered Probit Model			Semi-parametric Panel Linear Model		
	M. Eff.	Std. Error		M. Eff.	Std. Error		M. Eff.	Std. Error		M. Eff.	Std. Error	
Finland												
Probability of Smoking	0.863	0.012	***	0.829	0.012	***	0.172	0.064	***	0.025	0.011	**
Conditional mean of cigarettes smoked	10.17	0.541	***	10.39	0.396	***	4.74	0.774	***	5.58	0.633	***
Denmark												
Probability of Smoking	0.880	0.010	***	0.868	0.010	***	0.262	0.054	***	0.328	0.043	***
Conditional mean of cigarettes smoked	9.43	0.638	***	9.00	0.541	***	3.24	0.664	***	4.25	0.619	***
United Kingdom												
Probability of Smoking	0.876	0.007	***	0.856	0.007	***	0.203	0.038	***	0.308	0.038	***
Conditional mean of cigarettes smoked	9.76	0.432	***	10.52	0.313	***	3.27	0.602	***	3.33	0.449	***
Ireland												
Probability of Smoking	0.773	0.013	***	0.757	0.012	***	0.043	0.018	**	0.091	0.016	***
Conditional mean of cigarettes smoked	5.99	0.552	***	6.46	0.425	***	2.22	0.761	***	4.46	0.471	***
Belgium												
Probability of Smoking	0.899	0.010	***	0.871	0.012	***	0.219	0.037	***	0.050	0.0183	***
Conditional mean of cigarettes smoked	7.73	0.784	***	8.45	0.606	***	3.09	0.881	***	5.586	0.6545	***
Austria												
Probability of Smoking	0.938	0.005	***	0.796	0.013	***	0.242	0.047	***	0.077	0.025	***
Conditional mean of cigarettes smoked	7.29	1.222	***	7.88	0.478	***	3.27	1.145	***	3.78	0.745	***
Italy												
Probability of Smoking	0.813	0.009	***	0.770	0.009	***	0.521	0.028	***	0.436	0.024	***
Conditional mean of cigarettes smoked	7.54	0.396	***	7.72	0.302	***	0.76	0.546		3.09	0.377	***
Spain												
Probability of Smoking	0.746	0.010	***	0.683	0.010	***	0.073	0.010	***	0.067	0.011	***
Conditional mean of cigarettes smoked	7.37	0.438	***	6.90	0.294	***	1.78	0.539	***	3.653	0.382	***
Portugal												
Probability of Smoking	0.825	0.017	***	0.798	0.017	***	0.149	0.043	***	0.025	0.009	***
Conditional mean of cigarettes smoked	5.89	0.719	***	7.02	0.552	***	-1.21	0.778		2.31	0.789	***
Greece												
Probability of Smoking	0.748	0.011	***	0.726	0.011	***	0.139	0.019	***	0.101	0.012	***
Conditional mean of cigarettes smoked	4.66	0.462	***	4.71	0.349	***	3.91	0.542	***	1.46	0.367	***

Note: The first row in each country denotes the APE of the lagged decision to smoke or not on the probability of smoking in the current period. The second row denotes, for the ordered probit model the APE of smoking between 17.5 and 22.5 cigarettes (compared to not smoking at all) during the previous period on the current conditional mean number of cigarettes smoked. For the linear model, the APE measures the effect of the change from not smoking at all to smoking 20 cigarettes during the previous period. ***, **, * denote statistical significance at 1%, 5% and 10%, respectively.

Table 5. Correlation of Attitudes and Restrictions on Smoking with Persistence Indicators from the Data and the Semi-parametric Panel Ordered Probit Model

Item	(1) (2) Correlation with Smoking Inflows from Data	(3) (4) Correlation with Smoking Outflows from Data	(5) (6) Correlation with Smoking Inflows from Model	(7) (8) Correlation with Smoking Outflows from Model
Smoke can cause health problems to nonsmokers, EB 2005	-0.218	-0.077	-0.096	0.098
Smoke can cause health problems to nonsmokers, EB 1992	0.304	0.346	0.131	0.092
Encouraged to quit if got scientific proof that smoking can cause serious illnesses, EB 1994	0.001	0.226	-0.096	0.131
Heard about passive smoking, EB 1992	0.108	-0.152	-0.094	0.156
Smoking causes cancer and death, EB 1994	0.201	0.486 **	0.203	-0.141
In favour of smoking ban in any indoor public space, EB 1994	0.202	0.499 **	-0.075	0.126
Advertisements for cigarettes should not be regulated in anyway, EB 1994	-0.216	-0.490 **	0.007	-0.069
Smoking in government facilities is banned, WHO	0.322	0.208	-0.284	0.397 *
Smoking in private sector working facilities is restricted, WHO	0.216	0.099	-0.062	0.200
Smoking in restaurants is restricted, WHO	-0.274	-0.341	-0.143	0.141

Note: ***, **, * denote statistical significance at 1%, 5% and 10%, respectively.
Source: Information on attitudes about smoking is taken from the Eurobarometer (EB) Survey, various years. Data on smoking restrictions are from the WHO Tobacco Control Database (see http://data.euro.who.int/tobacco/), various years. Smoking restrictions are classified by the WHO Tobacco Control Database as voluntary agreements, partial restrictions and bans. For indoor workplaces and offices and restaurants, our indicators take the value 1 if there is no regulations, 2 if there is a voluntary agreement, 3 if there is a partial restriction and 4 if smoking is prohibited in that setting. For government facilities, our indicator takes the value 1 if smoking is prohibitted and the value 0 otherwise.

Table 6. Simulation of the Smoking Rate with State Dependence Set to Zero

Country	(1)	(2)		(3)	(4)		(5)	(6)	
	Difference using the Semi-Parametric Panel Ordered Probit Model			Difference using the Pooled Ordered Probit Model			Difference in Differences (1) – (3)		
	Est.	Std. Error		Est.	Std. Error		Est.	Std. Error	
				Panel A. Males					
Finland	0.088	0.017	***	0.235	0.004	***	-0.147	0.016	***
Denmark	0.099	0.014	***	0.262	0.006	***	-0.163	0.015	***
United Kingdom	0.069	0.010	***	0.226	0.004	***	-0.157	0.011	***
Ireland	0.027	0.007	***	0.201	0.006	***	-0.173	0.007	***
Belgium	0.134	0.040	***	0.240	0.005	***	-0.107	0.041	***
Austria	0.047	0.010	***	0.287	0.005	***	-0.240	0.009	***
Italy	0.198	0.007	***	0.258	0.003	***	-0.060	0.006	***
Spain	0.059	0.009	***	0.267	0.004	***	-0.208	0.008	***
Portugal	0.143	0.008	***	0.270	0.004	***	-0.126	0.008	***
Greece	0.085	0.009	***	0.363	0.005	***	-0.278	0.008	***
				Panel B. Females					
Finland	0.032	0.010	***	0.156	0.004	***	-0.124	0.010	***
Denmark	0.173	0.015	***	0.279	0.006	***	-0.106	0.013	***
United Kingdom	0.114	0.010	***	0.230	0.003	***	-0.116	0.010	***
Ireland	0.021	0.009	***	0.214	0.006	***	-0.193	0.008	***
Belgium	0.065	0.013	***	0.183	0.004	***	-0.117	0.012	***
Austria	0.080	0.014	***	0.185	0.004	***	-0.105	0.015	***
Italy	0.095	0.003	***	0.117	0.002	***	-0.022	0.003	***
Spain	0.044	0.005	***	0.157	0.004	***	-0.113	0.005	***
Portugal	0.033	0.005	***	0.068	0.002	***	-0.035	0.005	***
Greece	0.037	0.006	***	0.165	0.004	***	-0.128	0.005	***

Note: Figures represent the average over the three years 1999-2001 of the difference between smoking rates obtained using the estimated coefficients of state dependence and counterfactual rates obtained when setting those coefficients equal to zero. ***, **, * denote statistical significance at 1%, 5% and 10%, respectively.

Table 7. Simulation of the Smoking Rate using Ireland's Coefficients of State Dependence

Country	(1)	(2)		(3)	(4)		(5)	(6)	
	Difference in Smoking Rate			Difference in Rates of Transitions into Smoking			Difference in Rates of Transitions out of Smoking		
	Est.	Std. Error		Est.	Std. Error		Est.	Std. Error	
				Panel A. Males					
Finland	0.045	0.006	***	-0.020	0.002	***	-0.150	0.019	***
Denmark	0.073	0.007	***	-0.028	0.002	***	-0.191	0.014	***
United Kingdom	0.049	0.005	***	-0.017	0.002	***	-0.162	0.014	***
Belgium	0.093	0.026	***	-0.024	0.008	***	-0.281	0.038	***
Austria	0.016	0.002	***	-0.008	0.001	***	-0.038	0.004	***
Italy	0.145	0.007	***	-0.035	0.004	***	-0.378	0.029	***
Spain	0.010	0.004	***	-0.007	0.003	***	-0.026	0.012	**
Portugal	0.091	0.005	***	-0.033	0.004	***	-0.270	0.025	***
Greece	0.038	0.004	***	-0.014	0.003	***	-0.071	0.009	***
				Panel B. Females					
Finland	0.022	0.008	***	-0.006	0.002	***	-0.082	0.025	***
Denmark	0.136	0.007	***	-0.036	0.005	***	-0.394	0.024	***
United Kingdom	0.078	0.006	***	-0.029	0.003	***	-0.267	0.024	***
Belgium	0.043	0.007	***	-0.011	0.001	***	-0.120	0.018	***
Austria	0.056	0.008	***	-0.018	0.002	***	-0.225	0.045	***
Italy	0.082	0.002	***	-0.010	0.001	***	-0.518	0.028	***
Spain	0.020	0.003	***	-0.010	0.002	***	-0.101	0.018	***
Portugal	0.023	0.002	***	-0.006	0.001	***	-0.275	0.027	***
Greece	0.015	0.003	***	-0.004	0.001	***	-0.049	0.009	***

Note: Rates of transitions in and out of smoking are defined as in Table 1. Figures represent the average over the three years 1999-2001 of the difference between outcomes obtained using each country's own coefficients of state dependence and counterfactual outcomes obtained using Ireland's coefficients of state dependence. The model used for all calculations is the semi-parametric dynamic panel ordered probit. ***, **, * denote statistical significance at 1%, 5% and 10%, respectively.

Table A.1. Number of Observations for the Unbalanced and Balanced Panels, by Country and Gender

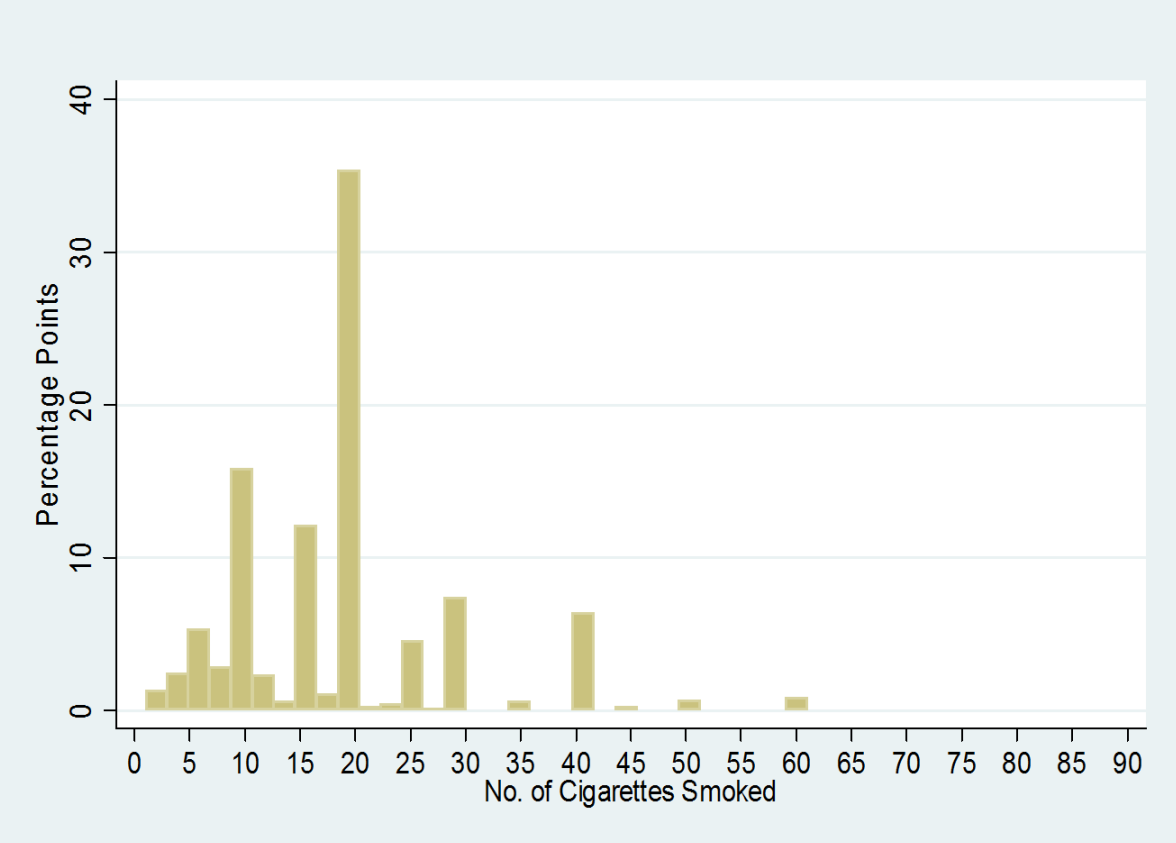
Country	(1)	(2)	(3)	(4)
	Males		Females	
	Unbalanced Panel	Balanced Panel	Unbalanced Panel	Balanced Panel
Finland	7,214	5,547	7,398	5,772
Denmark	4,289	3,606	4,438	3,738
United Kingdom	9,045	7,866	10,489	9,306
Ireland	5,274	3,975	5,495	4,251
Belgium	4,858	3,828	5,492	4,389
Austria	6,382	5,415	6,659	5,694
Italy	16,629	13,497	17,001	13,944
Spain	13,388	10,965	13,938	11,712
Portugal	11,885	10,380	12,590	11,082
Greece	9,702	8,181	10,316	8,973

Table A.2. Estimated Correlations Coefficients of the Two Time-varying Errors

Country	(1)	(2)		(3)	(4)	
	Semi-parametric Mundlak Ordered Probit			Semi-parametric Mundlak Linear		
	Coefficient	Std. Error		Coefficient	Std. Error	
			Panel A. Males			
Finland	-0.456	0.122	***	0.261	0.094	***
Denmark	-0.198	0.180		0.238	0.116	**
United Kingdom	-0.554	0.099	***	0.101	0.124	
Ireland	0.161	0.216		0.852	0.050	***
Belgium	-0.062	0.192		0.315	0.094	***
Austria	-0.254	0.156		0.791	0.052	***
Italy	0.084	0.137		0.784	0.031	***
Spain	0.630	0.109	***	0.913	0.013	***
Portugal	-0.265	0.250		0.756	0.038	***
Greece	0.107	0.158		0.834	0.026	***
			Panel B. Females			
Finland	0.676	0.174	***	0.051	0.136	
Denmark	-0.840	0.068	***	0.490	0.072	***
United Kingdom	-0.171	0.221		0.283	0.063	***
Ireland	-0.460	0.201	**	0.962	0.019	***
Belgium	-0.023	0.238		0.200	0.130	
Austria	0.059	0.304		0.661	0.095	***
Italy	-0.423	0.149	***	0.440	0.038	***
Spain	0.888	0.040	***	0.933	0.013	***
Portugal	0.685	0.164	***	0.935	0.039	***
Greece	-0.118	0.181		0.876	0.020	***

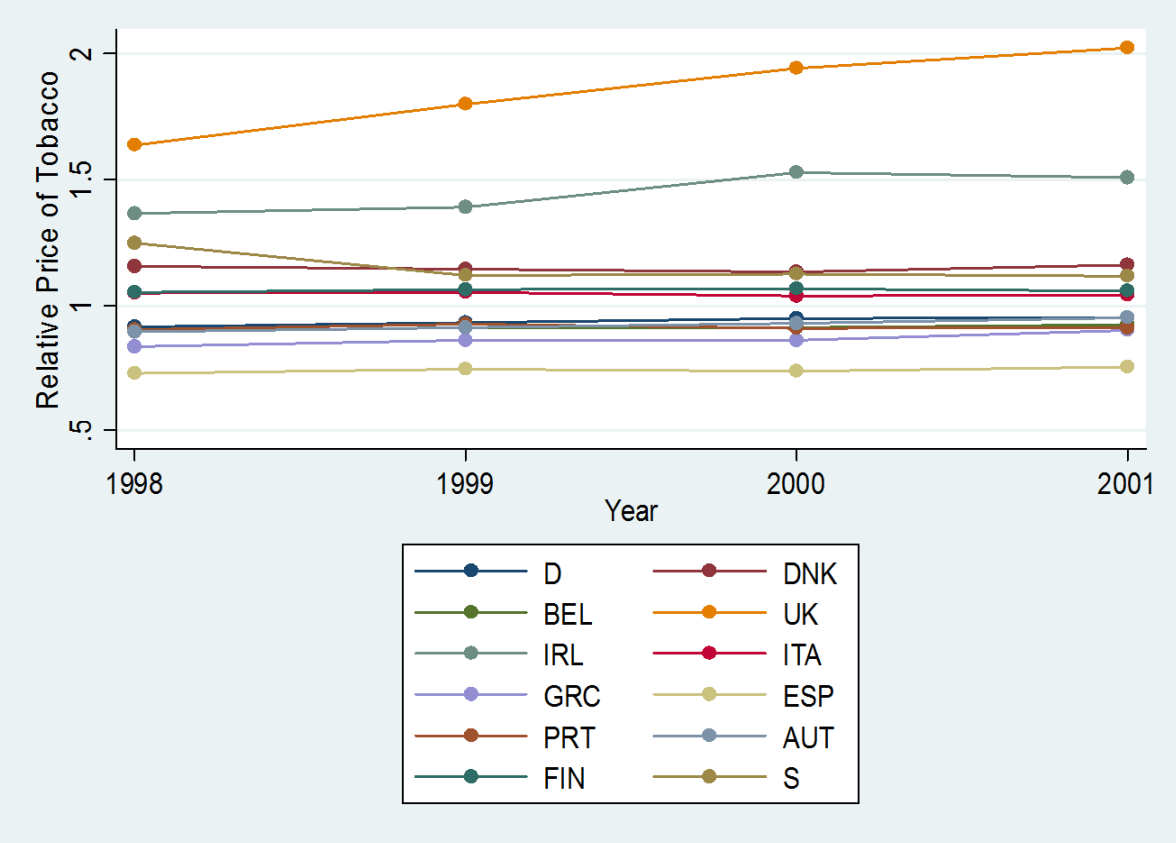
Note: ***, **, * denote statistical significance at 1%, 5% and 10%, respectively.

Fig. 1. Evidence on Heaping



Note: Prevalence of reported values of cigarettes smoked among smokers, in percentage points.
Source: ECHP

Fig. 2 Relative Price of Tobacco



Note: Relative price of tobacco, comparable across countries and time.
Source: Eurostat (for further details see Appendix A.1)