

**A Discrete Choice Model with Social Interactions;
an Analysis of High School Teen Behavior**

Peter Kooreman
Adriaan Soetevent *

University of Groningen

December 2001
preliminary

Abstract

We develop an empirical discrete choice model that explicitly allows for endogenous social interactions. We analyze the issues of multiple equilibria, statistical coherency, and estimation of the model by means of simulation methods. In an empirical application, we analyze a data set containing information on the individual behavior of some 8000 high school teenagers from almost 500 different school classes. We estimate the model for five types of teen discrete choice behavior, smoking, truanting, moped ownership, cell phone ownership, and asking parents' permission for purchases. We find strong social interaction effects for behavior closely related to school (truanting), somewhat weaker social interaction effects for behavior partly related to school (smoking, moped and cell phone ownership) and no social interaction effects for behavior far away from school (asking parents' permission for purchases). Intra-gender interactions are much stronger than cross-gender interactions.

Keywords: discrete choice; social interactions; multiple equilibria; teenage behavior

JEL classification: C35, D12

*Corresponding author: Peter Kooreman, Department of Economics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands. Phone: + 31 50 363 4533. Fax: + 31 50 363 7337. E-mail: p.kooreman@eco.rug.nl.

1 Introduction

Early contributions by Veblen (1899), Duesenberry (1949), Leibenstein (1950), Pollak (1976), and others show that economists have recognized the potential importance of social interactions for a long time. Yet, is it only recently that researchers have begun attempts to measure social interactions empirically.

The slow rate of accumulation of empirical analyses on social interactions is related to several difficulties. One important problem is identification. If one observes a correlation between individual behavior and reference group behavior, it is generally difficult to distinguish between genuine, endogenous social effects (two pupils have high grades because they are mutually motivated by the high grades of the other pupil) and spurious social effects (e.g. two pupils in a class have high grades because they have the same teacher); cf. Manski (1993, 2000).

A second problem is that a person's reference group – the group of individuals to which (s)he attaches nonzero weights in making decisions – is usually not easily determined. In most empirical studies, the assumptions made in this respect are largely ad hoc, and determined by data availability. Alessie and Kapteyn (1991), Kapteyn *et al.* (1997), and Aronsson *et al.* (1999) define the reference group of an individual as the group of persons in the population within the same age group and with the same education level. A more attractive alternative is to use subjective information on an individual's reference group, as in Woittiez and Kapteyn (1998). However, their information on the members of the reference group of a sampled individual is limited as these reference group members are not themselves included in the sample. Once a reference group has been defined there is a potential problem related to its endogeneity. For example, people may self-select into reference groups on the basis of similarity of individual characteristics. Failure to control for this may yield biases in estimated endogenous social

interaction effects.

Yet another problem is related to the discrete nature of many variables of interest in research on social interactions, with smoking being a prominent example. In a discrete choice model with endogenous social interactions, the choices of other individuals are explanatory variables in the equation describing the choice behavior of a given individual. For estimation and other purposes, the reduced form (or “social equilibrium” or “solution”) of the model is required. While the reduced form is straightforwardly obtained in a linear model with continuous variables, its derivation is more complicated in the case of discrete variables. As already noted by authors analyzing the simultaneous probit model (see e.g. Heckman, 1978 and Maddala, 1983), such models may not have a solution or may have multiple solutions. This in turn may yield problems regarding the statistical coherency of the model. Existing empirical studies allowing for social interactions usually focus on choices characterized by continuous variables, such as consumption and savings, or have analyzed discrete choices on an aggregated level; see e.g. Glaeser et al. (1996).

The primary aim of this paper is to contribute to a solution of the problems related to the discrete nature of choice variables. In section 2 we present a model based on the assumption that observed choices represent an equilibrium of a static discrete game played by all interacting agents. We analyze the issues of multiple equilibria and statistical coherency. Section 3 discusses estimation of the model by means of simulation methods.

The remainder of the paper is devoted to an empirical application. We analyze a sample of almost 500 school classes with detailed information on the individual behavior of the pupils within each class. We take the class as the natural reference group for each pupil within that class. While teenage behavior is obviously also influenced by persons outside the class, it is generally believed that class mates play a dominant role in shaping teenagers’

preferences and behaviors. Since in principle all pupils in a sampled class are interviewed, the current data set has unusually rich information on the behavior of all members of a sampled individual's reference group.

We estimate the model to analyze five types of teen discrete choice behavior, smoking, truanting, moped ownership, cell phone ownership, and asking parents' permission for purchases. We find strong social interaction effects for behavior closely related to school (truanting), somewhat weaker social interaction effects for behavior partly related to school (smoking, moped and cell phone ownership) and no social interaction effects for behavior far away from school (asking parents' permission for purchases). Intra-gender interactions are much stronger than cross-gender interactions.¹

2 Discrete Choice Interactions and Multiple Equilibria

Preliminaries

Consider a social group consisting of individuals indexed by i ; $i = 1, \dots, N$. Each individual makes a binary choice denoted by y_i ; $y_i \in \{-1, 1\}$. Hence, the total number of possible choice combinations in the group is 2^N . A *choice pattern* is defined as an element (y_1, y_2, \dots, y_N) from the set of all possible choice combinations.

As usual in discrete choice models, we introduce a latent variable y_i^* , which is related to the observed discrete choice variable y_i by the threshold condition $y_i = I(y_i^* > 0)$.² Let x_i be a row vector of observable exogenous

¹While the model presented in the current paper is similar in nature to the discrete choice models of Brock and Durlauf (2001a, 2001b), there are some notable differences. Their models are devised to describe aggregate behavioral outcomes in social groups in which agents observe the choices of other individuals imperfectly. Equilibrium properties are derived under the assumption that the number of observed choices made by others tends to infinity, and that interactions are symmetric. The present model describes the behavior of relatively small groups of a given size in which other individuals' choices can be assumed to be fully observable, and allows for asymmetric interactions between individuals.

² $I(z) = 1$ if z is true, and $I(z) = -1$ otherwise.

variables and β a vector of corresponding coefficients to be estimated. The latent variable is specified as the sum of a linear function of the explanatory variables, $x_i\beta$, and an error term, ϵ_i , representing all unobserved explanatory variables. Initially, we assume the error term ϵ_i to be independent of $\epsilon_j, j \neq i$, and all exogenous variables.

A simple case

In a discrete choice model with endogenous social interactions the choices of other individuals enter as additional explanatory variables in the specification of y_i^* . Consider the specification

$$(1) \quad \begin{cases} y_i^* = x_i\beta + s_i + \epsilon_i \\ y_i = 1 & \text{if } y_i^* > 0 \\ y_i = -1 & \text{if } y_i^* \leq 0, \end{cases}$$

where

$$s_i = \frac{\gamma}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^N y_j$$

for $i = 1, \dots, N$. Note that $\frac{1}{N-1} \sum_{j \neq i} y_j$ is the difference between the number of individuals other than i choosing $y = 1$ and the number of individuals other than i choosing $y = -1$, as a fraction of the total number of other individuals, $N - 1$.³ A positive γ reflects an inclination to conform to the behavior of others, a negative γ an inclination to deviate from the behavior of others. Let M denote the number of individuals choosing $y = 1$, i.e. $M = \sum_{i=1}^N I(y_i = 1)$. In the sequel it appears convenient to use $\sum_{j \neq i} y_j = M \cdot 1 + (N - M) \cdot (-1) - y_i = 2M - N - y_i$.

³Earlier work on simultaneous discrete choice models has used the categorization $y_i \in \{0, 1\}$ rather than $y_i \in \{-1, 1\}$. While the difference is immaterial in a standard discrete choice model, it is not in the present framework. With $y_i \in \{-1, 1\}$ the model is invariant with respect to interchanging the two choices, whereas it is not with $y_i \in \{0, 1\}$. Note, moreover, that the y_j 's rather than the y_j^* 's enter as explanatory variables – we assume that i 's behavior is influenced by j 's actual behavior (y_j) rather than by j 's “intended behavior” (y_j^*). For alternative specifications see, e.g., Maddala (1983).

A choice pattern (y_1, \dots, y_N) is an *equilibrium* if and only if it is consistent with (1) for all i , i.e. if after substitution of these values of y_i in s_i , we have $y_i^* > 0$ for all i with $y_i = 1$, and $y_i^* \leq 0$ for all i with $y_i = -1$. In the model without social interactions (i.e. $\gamma = 0$) a given set of values of x_i 's, ϵ_i 's, β , and γ obviously defines a unique choice pattern. An important feature of the model with social interactions is that, for a given set of values of x_i 's, ϵ_i 's, β , and γ , several choice patterns may be consistent with (1). As an example consider the model for the case $N = 2$:

$$(2) \quad \begin{cases} y_1^* = x_1\beta + \gamma y_2 + \epsilon_1 \\ y_2^* = x_2\beta + \gamma y_1 + \epsilon_2 \\ y_i = 1 & \text{if } y_i^* > 0 \\ y_i = -1 & \text{if } y_i^* \leq 0, \end{cases}$$

for $i = 1, 2$. If $\gamma = 1$ and $x_1\beta + \epsilon_1 = x_2\beta + \epsilon_2 = -\frac{1}{2}$, for example, the choice patterns $(-1, -1)$ and $(1, 1)$ are both consistent with (2).

As noted by Bjorn and Vuong (1984) and Kooreman (1994), the first equation in (1) and the first and second equation in (2) can be interpreted as reaction functions, where y_i^* is the difference between the utility person i derives from choosing $y_i = 1$ and the utility he derives from choosing $y_i = -1$, conditional on the choices y_j made by all other individuals. Equilibria can then be interpreted as (one-shot) pure Nash equilibria of a game played between all group members.

A more complicated model: discrete choices in school classes

Consider a set of school classes indexed by k , $k = 1, \dots, K$. Class k has N_{Gk} girls and N_{Bk} boys. Pupils are indexed by $i = 1, \dots, N_k$, with $N_k = N_{Gk} + N_{Bk}$. Let M_{Gk} denote the total number of girls in class k choosing $y = 1$ and M_{Bk} the total number of boys in class k choosing $y = 1$.⁴ For ease of exposition we will refer to $y = 1$ as “smoking” and

⁴Obviously, one could in principle refine the specification of social groups beyond the boy-girl distinction, for example on the basis of ethnicity, or by allowing the effect of

to $y = -1$ as “non-smoking”, although we will also consider other types of behavior.

We now specify

$$(3) \quad \begin{cases} y_{ik}^* = x_{ik}\beta + s_{ik} + \epsilon_{ik} \\ y_{ik} = 1 \\ y_{ik} = -1 \end{cases} \quad \begin{cases} \text{if } y_{ik}^* > 0 \\ \\ \text{if } y_{ik}^* \leq 0 \end{cases}$$

where

$$(4) \quad s_{ik} = \begin{cases} \gamma_{GG}(2M_{Gk} - N_{Gk} - y_{ik}) + \gamma_{GB}(2M_{Bk} - N_{Bk}) / (N_k - 1) & \text{if } i \text{ is a girl} \\ \gamma_{BB}(2M_{Bk} - N_{Bk} - y_{ik}) + \gamma_{BG}(2M_{Gk} - N_{Gk}) / (N_k - 1) & \text{if } i \text{ is a boy} \end{cases}$$

(Note that if i is a girl, then $2M_{Gk} - N_{Gk} - y_{ik}$ is the number of other girls smoking minus the number of other girls not smoking; $2M_{Bk} - N_{Bk}$ is the number of boys smoking minus the number of boys not smoking; etc..)

Thus, the model distinguishes between interactions among boys, interactions among girls, and interactions between boys and girls. In the first equation in (4) γ_{GG} measures how girls are affected by other girls, and γ_{GB} measures how girls are affected by boys; in the second equation γ_{BB} measures how boys are affected by other boys, and γ_{BG} measures how boys are affected by girls.

Alternatively, the social interactions term might be specified as

$$(5) \quad s_{ik} = \begin{cases} \gamma_{GG}(2M_{Gk} - N_{Gk} - y_{ik}) / (N_{Gk} - 1) + \gamma_{GB}(2M_{Bk} - N_{Bk}) / N_{Bk} & \text{if } i \text{ is a girl} \\ \gamma_{BB}(2M_{Bk} - N_{Bk} - y_{ik}) / (N_{Bk} - 1) + \gamma_{BG}(2M_{Gk} - N_{Gk}) / N_{Gk} & \text{if } i \text{ is a boy} \end{cases}$$

To appreciate the difference consider the case $\gamma_{GG} = \gamma_{GB} = \gamma_{BB} = \gamma_{BG}$. Then in specification (5) the groups of boys and girls have the same impact on i , irrespective of their relative sizes. According to specification (4) the impact of a gender group increases with its relative size, which we consider more plausible.

younger and of older class mates to be different. Such a refinement is beyond the scope of the present paper.

3 Estimation by simulation

In order to calculate the probability that a particular choice pattern will emerge as an equilibrium, we first reconsider the model specified in (2). For this model we have the following conditions for the four potential equilibria:

$$\begin{aligned}
(1, -1) &\Leftrightarrow x_1\beta - \gamma + \epsilon_1 > 0; & x_2\beta + \gamma + \epsilon_2 < 0, \\
(-1, 1) &\Leftrightarrow x_1\beta + \gamma + \epsilon_1 < 0; & x_2\beta - \gamma + \epsilon_2 > 0, \\
(1, 1) &\Leftrightarrow x_1\beta + \gamma + \epsilon_1 > 0; & x_2\beta + \gamma + \epsilon_2 > 0, \\
(-1, -1) &\Leftrightarrow x_1\beta - \gamma + \epsilon_1 < 0; & x_2\beta - \gamma + \epsilon_2 < 0.
\end{aligned}$$

It is easily verified that the four corresponding regions in the (ϵ_1, ϵ_2) -space partly overlap; see figure 1 – one of the subregions supports both $(-1, -1)$ and $(1, 1)$. Following Bjorn and Vuong (1984) and Kooreman (1994), assume that in case of multiple equilibria one of them will be observed with probability equal to one over the number of equilibria. From this assumption and the equilibrium conditions given above it then follows that

$$\begin{aligned}
(6) \quad P(1, -1) &= P(\epsilon_1 > -x_1\beta + \gamma; \epsilon_2 < -x_2\beta - \gamma), \\
P(-1, 1) &= P(\epsilon_1 < -x_1\beta - \gamma; \epsilon_2 > -x_2\beta + \gamma), \\
P(1, 1) &= P(\epsilon_1 > -x_1\beta - \gamma; \epsilon_2 > -x_2\beta - \gamma) - \frac{1}{2}A, \\
P(-1, -1) &= P(\epsilon_1 < -x_1\beta + \gamma; \epsilon_2 < -x_2\beta + \gamma) - \frac{1}{2}A,
\end{aligned}$$

where

$$A = P(-x_1\beta - \gamma < \epsilon_1 < -x_1\beta + \gamma; -x_2\beta - \gamma < \epsilon_2 < -x_2\beta + \gamma)$$

(the probability mass corresponding to the shaded area in figure 1). Note that without subtracting $\frac{1}{2}A$ in $P(1, 1)$ and $P(-1, -1)$, we would have $P(1, 1) + P(1, -1) + P(-1, 1) + P(-1, -1) = 1 + A > 1$ for $\gamma > 0$. Subtracting $\frac{1}{2}A$ equally divides A between choice patterns $(1, 1)$ and $(-1, -1)$. This treatment of multiple equilibria ensures that the four probabilities add up to unity, and thus that the model is statistically coherent.

We now turn to the more general case specified in (3) and (4). Suppose that, for a class k , we observe a choice pattern $(y_1, y_2, \dots, y_N) \equiv \mathbf{y}$ (we suppress subscript k). Then maximum likelihood estimation requires

to calculate the probability $P(\mathbf{y})$ that we observe \mathbf{y} , for any given set of parameter values.

The support in ϵ -space for choice pattern \mathbf{y} is

$$(7) \quad \begin{cases} \epsilon_i > -x_i\beta - s_i(\mathbf{y}) & \text{if } y_i = 1 \\ \epsilon_i < -x_i\beta - s_i(\mathbf{y}) & \text{if } y_i = -1 \end{cases}$$

Denote the region in ϵ -space defined in (7) by $W(\mathbf{y}, \theta)$, with θ being the parameters to be estimated. Given the independence of the ϵ_i 's, the probability that (7) is satisfied, $P(\epsilon \in W(\mathbf{y}, \theta))$, can be calculated straightforwardly. Since $W(\mathbf{y}, \theta)$ may also support equilibria other than \mathbf{y} , we have $P(\epsilon \in W(\mathbf{y}, \theta)) \geq P(\mathbf{y})$. In case of social groups with the size of a school class, the procedure for determining the number of equilibria in the various subregions of the $(\epsilon_1, \dots, \epsilon_N)$ -space is more complicated. First, the number of subregions to be distinguished increases exponentially, and, second, in each subregion we have to check in principle whether each of the 2^N choice patterns can be an equilibrium. We therefore use a simulation based method.

We assume $(\epsilon_1, \dots, \epsilon_N)$ to follow a normal distribution with zero mean and identity covariance matrix. Consider R random draws (indexed by r , $r = 1, \dots, R$) from the joint distribution of $(\epsilon_1, \dots, \epsilon_N)$ on $W(\mathbf{y}, \theta)$. For each draw, we calculate the number of equilibria. Note that \mathbf{y} is either the single equilibrium or one of the multiple equilibria. Let Ω_r be the set of equilibria corresponding to draw r and let E_r denote the number of elements in Ω_r (i.e. E_r is the number of equilibria at draw r). Then the probability $P(\mathbf{y})$ that choice pattern \mathbf{y} will be observed is consistently estimated by

$$(8) \quad P(\mathbf{y}) = P(\epsilon \in W) \cdot \frac{1}{R} \sum_{r=1}^R \frac{1}{E_r}$$

We have found that $R = 1000$ generates estimated probabilities that are sufficiently precise as inputs in the maximum likelihood procedure. Note that since $E_r \geq 1$ we have $\frac{1}{R} \sum_{r=1}^R \frac{1}{E_r} \leq 1$. We also found that the probability of a single equilibrium is usually larger than 80 percent, thus we usually

have $\frac{1}{R} \sum_{r=1}^R \frac{1}{E_r} > 0.8$. As a consequence, our empirical results are relatively insensitive with respect to the assumption regarding the treatment of multiple equilibria.

Alternatively, $P(\mathbf{y})$ could be estimated directly using

$$(9) \quad P(\mathbf{y}) = \frac{1}{R} \sum_{r=1}^R \frac{I(\mathbf{y} \in \Omega_{\mathbf{r}})}{E_r}$$

with R the number of draws from the joint distribution of $(\epsilon_1, \dots, \epsilon_N)$ on \mathfrak{R}^N . However, this would require the number of draws R to be of a much larger magnitude to achieve the same precision as achieved when using (8).

We now provide a lemma that helps to reduce the number of potential equilibria that have to be checked. Note first that $\sum_{i=1}^N y_i = k$ implies that the number of agents with $y = 1$ is $M = \frac{1}{2}(N + k)$.

Lemma 1: *Suppose model (1) has an equilibrium with $\sum_{i=1}^N y_i = k$. Let $\gamma \geq 0$. Then $\max_{\{i|y_i=-1\}}(z_i) < \min_{\{i|y_i=1\}}(z_i) - \frac{2\gamma}{N-1}$, where $z_i \equiv x_i\beta + \epsilon_i$.*

Proof: Consider an agent i with $y_i = 1$ and an agent j with $y_j = -1$. Suppose $z_j > z_i - \frac{2\gamma}{N-1}$. Then $y_j^* = z_j + \left(\frac{k+1}{N-1}\right)\gamma > z_i + \left(\frac{k-1}{N-1}\right)\gamma = y_i^*$. But since $y_i = 1$ and $y_j = -1$ implies $y_i^* > 0 > y_j^*$, we have a contradiction.

From Lemma 1 it follows that, with $\gamma \geq 0$, the M agents with $y_i = 1$ are those with the M largest values of z_i . To determine whether there exists an equilibrium with $\sum_{i=1}^N y_i = k$, we therefore first rank observations on the basis of the values of z_i . Denote the ordered values as $z_{(1)} < z_{(2)} < \dots < z_{(N)}$. Then we have an equilibrium with $\sum_{i=1}^N y_i = k$, if and only if the inequalities

$$(10) \quad \begin{aligned} z_{(1)} + \frac{k+1}{N-1}\gamma &< \dots < z_{(N-M)} + \frac{k+1}{N-1}\gamma < 0 < \\ z_{(N-M+1)} + \frac{k-1}{N-1}\gamma &< \dots < z_{(N)} + \frac{k-1}{N-1}\gamma, \end{aligned}$$

with $1 \leq M = \frac{1}{2}(N + k) \leq N$, are satisfied. An equilibrium with $M = 0$ occurs if and only if $z_i < 0$ for all i ; an equilibrium with $M = N$ occurs if

and only if $z_i + \gamma > 0$ for all i . As a result, we only have to check $N + 1$ out of the 2^N choice patterns as possible equilibria ($M = 0, 1, \dots, N$).

Suppose that model (3)-(4), with all γ 's positive, has an equilibrium with M_{Gk} smoking girls and M_{Bk} smoking boys. It is straightforward to show that Lemma 1 implies that the smoking girls are those with the largest values of z_i in the subset of girls, and that the smoking boys are those with the largest values of z_i in the subset of boys. As a result, we only have to check $(N_{Gk} + 1)(N_{Bk} + 1)$ out of the 2^{N_k} choice patterns as potential equilibria.

If one or several γ 's are negative, it is possible to have $z_j > z_i$ combined with $y_j = -1$ and $y_i = 1$. This prevents a reduction of potential equilibria similar to the procedure described above. Therefore, with negative γ 's, estimation of the model requires – for each evaluation of the likelihood function, for each simulation within a likelihood evaluation – to check all 2^{N_k} possible equilibria for class k . This is computationally demanding but not infeasible given the social group sizes in the current application.

Having calculated for each class the probability that the observed choice pattern occurs using (8), we estimate the model by maximum likelihood.

4 The data: the Dutch National School Youth Survey

We will estimate the model outlined in the previous sections using data from the Dutch National School Youth Survey (NSYS) from the year 2000.⁵

Although in principle all pupils in a sampled class participate in the survey, some pupils are excluded from the data. In some cases this is because a pupil was absent when the questionnaires were filled out, in other cases

⁵Previous surveys were conducted in 1984, 1990, 1992, 1994, and 1996. The NSYS is a joint effort of the Social and Cultural Planning Office of The Netherlands (SCP) and the Netherlands Institute for Family Finance Information (NIBUD). In each survey year a random sample of high schools in The Netherlands is drawn. A participating school is compensated by means of a report summarizing the survey results for that school. The series of surveys is not a panel, although some schools have participated more than once.

because information on some of the variables is missing. The data set used in estimation contains information on 7534 pupils in 487 classes in 66 schools. It contains information on the teenagers' individual characteristics, time use, income and expenditures, subjective information on norms and values, and information on various behaviors and durable goods ownership. There is only limited information on the parents (including education and working hours) and no information on siblings. Tables 1, 2 and 3 provide sample information at the individual level, the class level, and the school level, respectively.

All information is self-reported. Thus, strictly speaking our analysis measures social interactions in how teenagers report on their behavior. The results for "asking parents' permission for purchases" may provide some insight in potential differences between social interactions in reported behavior and in actual behavior. Asking parents for permission before making a purchase is an aspect of out-of-class behavior. Since this primarily concerns the relationship between a pupil and his or her parents, we expect very weak or no social interaction effects in this type of actual behavior. However, if pupils copy each others' responses to the survey questions when filling out the questionnaire, spurious social interaction effects might be found.

The vector x includes age, and dummy variables for gender, for being non-Dutch (based on the question "Do you consider yourself to be Dutch?"), for the type of education (MAVO (lower level), HAVO (intermediate level), and VWO (higher level), with 'vocational' as reference category), for catholic, for protestant, and for living in a single parent family (based on the question "Do you live in a family with father and mother?"). Unfortunately, a large proportion of teenagers do not know their parents' education level (41 and 36 percent for father's and mother's education level, respectively). We therefore choose not to include parents' education levels as explanatory variables. However, we do include the father's working time

and the mother’s working time (for a pupil with a single parent the working time of the missing parent is set equal to the sample average).⁶

5 Empirical results

Table 4 presents four versions of the estimated model for smoking. The first column contains estimation results for the model without social interactions (i.e. with $\gamma_{GG} = \gamma_{GB} = \gamma_{BB} = \gamma_{BG} = 0$). The probability of smoking strongly increases in age. The effect of gender is insignificant. The higher the level of the type of education, the smaller the probability that a pupil smokes. We also find that pupils from single parent households and pupils whose mother has a paid job have a significantly larger probability to smoke. The variables non-Dutch, catholic, and protestant negatively affect pupils’ smoking behavior. The effects are largely consonant with earlier empirical studies on smoking behavior; see for example, Gruber and Zinman (2001) and Gruber (2001).

Column two presents results for the model with social interactions. All social interaction coefficients are positive and highly significant. The largest one is γ_{BB} , measuring the boy-boy interaction, followed in size by γ_{GG} , measuring the interaction between girls. The coefficients γ_{GB} and γ_{BG} , measuring the cross gender interactions are also significant, though smaller in size. Note that the inclusion of the social interaction coefficients hardly affects the parameters.

Fixed effects

Smoking behavior in all classes of a given school is likely to be affected by a number of unobserved school specific factors, like smoking behavior of teachers, the school’s policy regarding smoking, and proximity of tobacco

⁶A number of studies have reported indicators for self-esteem to be important explanatory variables in the analysis of teenage behavior; see e.g. Smetters and Gravelle (2001). We choose not to include such a variable because of its potential endogeneity.

outlets. Unobserved school specific factors may also be related to a non-random assignment of pupils to schools. For example, parents who smoke themselves may be less likely to send their children to a school in which smoking is strictly prohibited. Significant social interaction coefficients may then merely reflect the failure to control for these unobserved effects. We therefore also estimate a version with school specific fixed effects.⁷

The inclusion of school specific fixed effects amounts to estimating 64 additional parameters (one school is reference category, another school is deleted because it has non-smokers only). The results are reported in the third and fourth column of table 4. While, in column four, the cross-gender interaction effects are not significant for this specification, the within gender interactions are still sizeable and significant, with again the boy-boy interaction being stronger than the girl-girl interaction. The other coefficients now have somewhat larger standard errors, but this has a negligible effect on the significance of explanatory variables. More importantly, a χ^2 -test shows that the fixed effects are jointly insignificant ($p = 0.201$).

We have also estimated the model for truancing, moped ownership, cell phone ownership, and asking parents' permission for purchases, without school specific fixed effects (table 5) and with school specific fixed effects (table 6). (For ease of comparison the first column in table 5 repeats the second column from table 4 and the first column in table 6 repeats the fourth column from table 4).

The significance of the fixed effects varies across the five types of behav-

⁷Clearly, a more flexible specification would be obtained by allowing for class specific fixed effects. An example in which a fixed effect at the class level would be appropriate is when the detrimental health effects of smoking are discussed in one class, but not in other classes in the same school. Given that the classes are observed only at a single point in time, the estimation of class specific fixed effects is infeasible. We would like to argue, however, that fixed effects mainly operate at the school level. Most high school teachers are not tied to one particular class (as in elementary schools), but teach in various classes across the school. Proximity of tobacco outlets and school policy regarding smoking obviously also operate at the school level. Finally, conditional on school choice, grade and educational level (for which we control by either school specific fixed effects or by including appropriate regressors), the assignment of pupils to classes is largely random.

ior. For truanting, smoking, and moped ownership the fixed effects are not significant (see bottom row of table 6), while for cell phone ownership and asking parents' permission they are significant. The discussion of estimation results below is therefore based on table 5 for smoking, truanting, and moped ownership, and on table 6 for the other two choice behaviors.

For truanting, the intra-gender effects are stronger than for smoking. Moreover, we now also have significant cross-gender interactions. The probability of truanting sharply increases in age, is larger for non-Dutch pupils, and decreases in the level of education. The mother's working time also has a significant positive effect on truanting.

Moped ownership is the only type of behavior where we find a large gender effect: The probability of moped ownership is much larger for boys than for girls. It strongly increases in age (the legal minimum age for driving a moped in The Netherlands is 16) and decreases in the level of education. It is also the only type of behavior where we have a clear asymmetry in social interactions between genders. For a boy, the probability of moped ownership is strongly affected by moped ownership of other boys and of girls. Moped ownership for girls, on the other hand, is not affected by social interactions.

For cell phone ownership we again find an increasing effect of age and a decreasing effect of education. Teenagers from a single parent family have a much larger probability of owning a cell phone. Only the girl-girl social interaction effect is significant.

The probability of asking parents' permission before purchasing something strongly decreases in age, and is smaller for non-Dutch pupils and for pupils in a single parent household. It also significantly decreases in mother's working time. The four social interaction coefficients are (jointly) insignificant. This also indicates that pupils do not copy each other's responses when filling out the questionnaire.

The magnitude of the social interaction effects

In order to gain some insight in the magnitude of the social interaction effects implied by the estimated γ 's consider a reference class (largely based on median values of exogenous variables). This is a MAVO class composed of 8 girls and 8 boys; all of them are aged 14, Dutch, non-protestant, non-catholic, and come from a two-parent household with a father working 36 hours per week and a mother working 16 hours per week. Using the estimated parameters from table 5, we find that in equilibrium the expected number of truanters is 3.14 (the probability of truanting is 0.191 for girls and 0.201 for boys).⁸

Now suppose that a surely truanting girl is added to this class (i.e. we add a girl with characteristics such that her probability of truanting is virtually equal to 1, irrespective of the behavior of others). Without social interaction effects, the expected fraction of truanters would rise from 0.196 (3.14/16) to 0.244 (4.14/17), a 24 percent increase. Taking social interaction effects into account, the new equilibrium fraction of truanters rises to 0.278 (4.73/17), an increase of 41 percent compared to the original level. If a surely non-truanting girl is added to this class, the expected fraction decreases from 0.196 (3.14/16) to 0.185 (3.14/17) without social interaction effects (a 6 percent decrease), and to 0.169 (2.88/17) with social interaction effects (a 15.8 percent decrease).

The model also implies that a change in the value of an exogenous variable of only one of the pupils in principle affects the behavior of all pupils in class. Suppose, for example, that the mother of one of the girls in the reference class increases her working hours to 46 per week. Then the equilibrium truanting probability of her daughter increases from 0.191 to 0.210. However, it also changes the equilibrium truanting probabilities of the other girls (from 0.1909 to 0.1915) and boys (from 0.2012 to 0.2002). As a result,

⁸All numbers are based on simulations with R=100000.

the expected of number of truanters in class increases not only by 0.019 (0.210-0.191), but by 0.031.

6 Conclusion

The model presented and estimated in this paper represents, in our view, the simplest and most natural approach to incorporate social interactions in empirical discrete choice models. In our application to teenagers' discrete choices, we found strong social interaction effects for behavior closely related to school (truanting), somewhat weaker social interaction effects for behavior partly related to school (smoking, moped and cell phone ownership) and no social interaction effects for behavior far away from school (asking parents' permission for purchases). The latter result suggests that the effects found for the other four types of choice behavior represent genuine endogenous social interaction effects rather than unobserved social group effects.

While the present data set has a number of important advantages in terms of reference group definition and information on reference group members, the empirical results are subject to the usual qualifications regarding inference on the basis of cross section data alone. The analysis of data collected at several points in time on the same teenagers, preferably with exogenous reassignment of pupils to other classes within the same school, would be another step towards increasing our understanding of social interactions.

References

- Alessie, R.J.M. and A. Kapteyn (1991), “Habit formation, interdependent preferences and demographic effects in the almost ideal demand system”, *Economic Journal*, **101**, 404–419.
- Aronsson, T., Blomquist S. and H. Sacklén (1999), “Identifying interdependent behaviour in an empirical model of labor supply”, *Journal of Applied Econometrics*, **14**, 607–626.
- Bjorn, P. and Q. Vuong (1984), *Simultaneous Models for Dummy Endogenous Variables: A Game Theoretic Formulation with an Application to Household Labor Force Participation*, Working Paper, California Institute of Technology.
- Duesenberry, J.S. (1949), *Income, Saving and the Theory of Consumer Behavior*, Harvard University Press, Cambridge, MA.
- Glaeser, E. L., B. Sacerdote and J. A. Scheinkman (1996), “Crime and social interactions”, *Quarterly Journal of Economics*, 507–548.
- Gruber, J. (2001), “Youth smoking in the 1990’s: Why did it rise and what are the long-run implications?”, *American Economic Review*, **91**(2), 85–90.
- Gruber, J. and J. Zinman (2001), *Youth Smoking in the U.S.: Evidence and Implications*, University of Chicago Press, Chicago, 69–120.
- Heckman, J. J. (1978), “Dummy endogenous variables in a simultaneous equation system”, *Econometrica*, **46**, 931–960.
- Kapteyn, A., S. Van de Geer H. Van de Stadt and T. Wansbeek (1997), “Interdependent preferences: An econometric analysis”, *Journal of Applied Econometrics*, **12**, 665–686.
- Kooreman, P. (1994), “Estimation of econometric models of some discrete games”, *Journal of Applied Econometrics*, **9**, 255–268.

- Leibenstein, H. (1950), “Bandwagon, snob, and Veblen effects in the theory of consumer’s demand”, *Quarterly Journal of Economics*, **64**, 183–207.
- Maddala, G.S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Econometric Society Monographs No. 3.
- Manski, C. F. (2000), “Economic analysis of social interactions”, *Journal of Economic Perspectives*, **14**, 115–136.
- Manski, C.F. (1993), “Identification of endogenous social effects: The reflection problem”, *The review of Economic Studies*, **60**, 531–542.
- Pollak, R. A. (1976), “Interdependent preferences”, *American Economic Review*, **66**(3), 309–321.
- Smetters, K. and J. Gravelle (2001), “The exchange theory of teenage smoking and the counterproductiveness of moderate regulation”, *NBER Working Paper No. 8262*.
- Veblen, T. (1899), *The Theory of the Leisure Class*, MacMillan, New York.
- Woittiez, I. and A. Kapteyn (1998), “Social interactions and habit formation in a model of female labour supply”, *Journal of Public Economics*, **70**, 185–205.

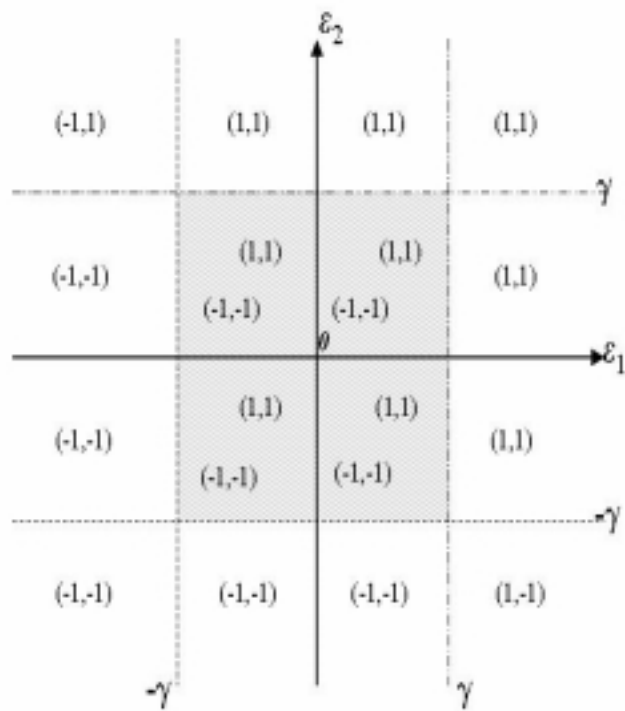


Figure 1: Multiple equilibria ($\gamma > 0, x_1\beta = x_2\beta = 0$)

Table 1: Sample statistics at the individual level (7,534 observations)

	mean	median	st. dev.	min.	max.
girl	0.5167	1.0000	0.4998	0.0000	1.0000
age	14.2520	14.0000	1.4437	11.0000	21.0000
non-Dutch	0.0881	0.0000	0.2835	0.0000	1.0000
single parent hh.	0.0832	0.0000	0.2762	0.0000	1.0000
MAVO	0.3211	0.0000	0.4669	0.0000	1.0000
HAVO	0.1968	0.0000	0.3976	0.0000	1.0000
VWO	0.1724	0.0000	0.3778	0.0000	1.0000
working time father	36.0284	36.0000	12.6600	0.0000	46.0000
working time mother	15.4080	16.0000	15.1320	0.0000	46.0000
catholic	0.2360	0.0000	0.4246	0.0000	1.0000
protestant	0.1856	0.0000	0.3888	0.0000	1.0000
smoking	0.0897	0.0000	0.2858	0.0000	1.0000
truanting	0.1886	0.0000	0.3912	0.0000	1.0000
asking for permission	0.8600	1.0000	0.3470	0.0000	1.0000
moped	0.0657	0.0000	0.2478	0.0000	1.0000
cell phone	0.2104	0.0000	0.4076	0.0000	1.0000
girls (3,893 observations)					
smoking	0.0917	0.0000	0.2886	0.0000	1.0000
truanting	0.1811	0.0000	0.3851	0.0000	1.0000
asking for permission	0.8513	1.0000	0.3559	0.0000	1.0000
moped	0.0301	0.0000	0.1708	0.0000	1.0000
cell phone	0.2009	0.0000	0.4007	0.0000	1.0000
boys (3,641 observations)					
smoking	0.0876	0.0000	0.2828	0.0000	1.0000
truanting	0.1966	0.0000	0.3975	0.0000	1.0000
asking for permission	0.8693	1.0000	0.3372	0.0000	1.0000
moped	0.1038	0.0000	0.3051	0.0000	1.0000
cell phone	0.2205	0.0000	0.4147	0.0000	1.0000

Table 2: Sample statistics at the class level (487 observations)

	mean	median	st. dev.	min.	max.
class size	15.4702	15.0000	4.6244	8.0000	30.0000
fraction of girls	0.5193	0.5238	0.1486	0.1111	0.8947
MAVO	0.3294	0.0000	0.4683	0.0000	1.0000
HAVO	0.1771	0.0000	0.3767	0.0000	1.0000
VWO	0.1643	0.0000	0.3659	0.0000	1.0000
smoking					
<i>fraction y = 1</i>					
class	0.0894	0.0714	0.0946	0.0000	0.4348
fraction of girls	0.0896	0.0000	0.1249	0.0000	0.6667
boys	0.0908	0.0000	0.1328	0.0000	0.6667
truanting					
<i>fraction y = 1</i>					
class	0.1884	0.1429	0.1691	0.0000	0.8000
girls	0.1799	0.1250	0.2075	0.0000	1.0000
boys	0.2033	0.1667	0.2198	0.0000	1.0000
asking for permission					
<i>fraction y = 1</i>					
class	0.8597	0.8750	0.1187	0.3847	1.0000
girls	0.8523	0.8750	0.1627	0.0000	1.0000
boys	0.8651	0.9091	0.1600	0.0000	1.0000
moped					
<i>fraction y = 1</i>					
class	0.0662	0.0476	0.0832	0.0000	0.4167
girls	0.0287	0.0000	0.0697	0.0000	0.5000
boys	0.1068	0.0000	0.1455	0.0000	0.7500
cell phone					
<i>fraction y = 1</i>					
class	0.2113	0.1818	0.1573	0.0000	0.9091
boys	0.2019	0.1667	0.1990	0.0000	1.0000
girls	0.2234	0.2000	0.2007	0.0000	1.0000

Table 3: Sample statistics at the school level (66 observations)

	mean	median	st. dev.	min.	max.
# classes	7.3030	6.0000	6.4664	2.0000	48.0000
# pupils	113.2273	88.0000	98.8511	17.0000	698.0000
fraction of girls	0.5201	0.5175	0.0710	0.3200	0.7647
smoking					
<i>fraction y = 1</i>					
class	0.1018	0.0935	0.0527	0.0000	0.2400
girls	0.0997	0.09377	0.0650	0.0000	0.3200
boys	0.1041	0.0923	0.0712	0.0000	0.3333
truanting					
<i>fraction y = 1</i>					
class	0.2021	0.1740	0.1291	0.0000	0.6552
girls	0.2000	0.1905	0.1392	0.0000	0.7857
boys	0.2065	0.1786	0.1467	0.0000	0.7500
asking for permission					
<i>fraction y = 1</i>					
class	0.8479	0.8537	0.0790	0.5862	1.0000
girls	0.8353	0.8539	0.0927	0.5000	1.0000
boys	0.8585	0.8714	0.0908	0.6250	1.0000
moped					
<i>fraction y = 1</i>					
class	0.0705	0.0691	0.0468	0.0000	0.2414
girls	0.0334	0.0306	0.0360	0.0000	0.1786
boys	0.1119	0.1052	0.0876	0.0000	0.5000
cell phone					
<i>fraction y = 1</i>					
class	0.0705	0.0691	0.0468	0.0000	0.2414
boys	0.0334	0.0307	0.0360	0.0000	0.1786
girls	0.1119	0.1052	0.0876	0.0000	0.5000

Table 4: Estimation results; smoking (t-values in parentheses)

	with fixed effects			
	no SI	with SI	no SI	with SI
constant	-4.18 (-19.1)	-3.16 (-10.2)	-3.84 (-11.6)	-3.41 (-8.5)
girl	0.039 (0.9)	0.004 (0.0)	-0.005 (0.1)	-0.034 (-0.1)
age	0.189 (12.3)	0.156 (8.3)	0.169 (7.4)	0.158 (6.5)
non-Dutch	-0.274 (-3.3)	-0.248 (-2.8)	-0.214 (-2.0)	-0.215 (-2.0)
single parent family	0.188 (2.8)	0.183 (2.7)	0.170 (2.2)	0.176 (2.3)
MAVO	0.173 (3.6)	0.148 (2.3)	0.269 (3.1)	0.233 (2.4)
HAVO	-0.042 (-0.7)	-0.034 (-0.5)	-0.110 (-1.2)	-0.087 (-0.8)
VWO	-0.238 (-3.8)	-0.194 (-2.4)	-0.308 (-2.9)	-0.268 (-2.3)
father's working time	0.002 (1.0)	-0.000 (1.0)	0.001 (0.7)	0.002 (0.8)
mother's working time	0.004 (3.3)	0.005 (3.2)	0.005 (3.3)	0.005 (3.2)
catholic	-0.197 (-4.1)	-0.174 (-3.3)	-0.160 (-2.3)	-0.162 (-2.3)
protestant	-0.136 (-2.4)	-0.126 (-1.9)	-0.167 (-1.8)	-0.158 (-1.7)
γ_{BB}	—	0.880 (4.7)	—	0.491 (2.3)
γ_{BG}	—	0.533 (2.1)	—	0.223 (0.8)
γ_{GB}	—	0.569 (2.6)	—	0.188 (0.8)
γ_{GG}	—	0.765 (4.6)	—	0.386 (1.9)
log-likelihood function	-2153.9	-2107.2	-2133.8	2097.2

Table 5: Estimation results (t-values in parentheses)

	smoking	truanting	moped	cell phone	permission
constant	-3.16 (-10.2)	-2.74 (-9.4)	-4.52 (-12.8)	-2.52 (-11.2)	4.07 (16.3)
girl	0.004 (0.0)	-0.024 (-0.3)	-0.870 (-3.0)	0.036 (0.4)	-0.090 (-0.6)
age	0.156 (8.3)	0.156 (8.1)	0.255 (14.0)	0.145 (9.6)	-0.197 (-13.4)
non-Dutch	-0.248 (-2.8)	0.127 (1.9)	-0.178 (-1.9)	0.142 (2.3)	-0.159 (-2.5)
single parent family	0.183 (2.7)	0.037 (0.6)	-0.034 (-0.4)	0.277 (5.0)	-0.246 (-4.2)
MAVO	0.148 (2.3)	0.094 (1.5)	-0.131 (-1.9)	0.039 (0.7)	-0.107 (-2.0)
HAVO	-0.034 (-0.5)	0.131 (1.7)	-0.215 (-2.9)	-0.072 (-1.2)	-0.140 (-2.4)
VWO	-0.194 (-2.4)	0.048 (0.6)	-0.408 (-4.5)	-0.254 (-3.5)	-0.042 (-0.6)
father's working time	0.002 (1.0)	-0.000 (-0.2)	0.002 (1.2)	-0.002 (-1.1)	-0.004 (-2.6)
mother's working time	0.005 (3.2)	0.003 (2.1)	0.003 (1.6)	0.002 (1.8)	-0.004 (-2.7)
catholic	-0.174 (-3.3)	-0.126 (-2.6)	0.0103 (0.2)	-0.019 (-0.4)	0.233 (5.0)
protestant	-0.126 (-1.9)	-0.117 (-2.2)	-0.083 (-1.1)	-0.280 (-5.0)	0.273 (5.1)
γ_{BB}	0.880 (4.7)	0.829 (6.8)	0.486 (2.4)	0.562 (5.1)	0.303 (2.1)
γ_{BG}	0.533 (2.1)	0.535 (3.5)	0.497 (2.0)	0.434 (2.8)	0.082 (0.5)
γ_{GB}	0.569 (2.6)	0.465 (2.9)	0.346 (1.1)	0.467 (2.7)	0.128 (0.8)
γ_{GG}	0.765 (4.6)	1.171 (10.3)	0.153 (0.6)	0.830 (8.2)	0.220 (2.0)
log-likelihood function	-2133.8	3254.6	-1586.9	-3599.9	-2832.7

Table 6: Estimation results; with school specific fixed effects (t-values in parentheses)

	smoking	truanting	moped	cell phone	permission
constant	-3.41 (-8.5)	-2.92 (-8.1)	-5.40 (-12.8)	-3.34 (-11.7)	4.39 (13.7)
girl	-0.034 (-0.1)	-0.024 (-0.3)	-0.824 (-2.8)	0.028 (0.3)	-0.094 (-0.6)
age	0.158 (6.5)	0.158 (7.1)	0.282 (11.4)	0.189 (11.0)	-0.207 (-11.5)
non-Dutch	-0.215 (-2.0)	0.125 (1.7)	-0.175 (-1.4)	0.051 (0.7)	-0.183 (-2.5)
single parent family	0.176 (2.3)	0.036 (0.5)	-0.036 (-0.4)	0.249 (4.0)	-0.227 (-3.4)
MAVO	0.233 (2.4)	0.198 (2.2)	-0.136 (-1.2)	0.018 (0.3)	-0.196 (-2.4)
HAVO	-0.087 (-0.8)	0.118 (1.2)	-0.161 (-1.4)	-0.184 (-2.5)	-0.161 (-1.9)
VWO	-0.268 (-2.3)	0.002 (0.0)	-0.394 (-3.2)	-0.463 (-5.6)	-0.021 (-0.2)
father's working time	0.002 (0.8)	-0.000 (-0.2)	0.003 (1.2)	-0.001 (-0.5)	-0.004 (-2.6)
mother's working time	0.005 (3.2)	0.003 (2.0)	0.003 (1.8)	0.002 (1.8)	-0.004 (-2.6)
catholic	-0.162 (-2.3)	-0.106 (-1.8)	-0.030 (-0.4)	-0.056 (-1.1)	0.2000 (3.4)
protestant	-0.158 (-1.7)	-0.159 (-2.4)	-0.210 (-1.8)	-0.228 (-3.1)	0.255 (3.3)
γ_{BB}	0.491 (2.3)	0.829 (6.8)	0.197 (0.9)	-0.099 (-0.8)	-0.156 (-1.0)
γ_{BG}	0.223 (0.8)	0.359 (2.2)	0.101 (0.4)	-0.148 (-1.0)	-0.317 (-1.6)
γ_{GB}	0.188 (0.8)	0.277 (1.6)	0.044 (0.1)	-0.191 (-1.2)	-0.298 (-1.6)
γ_{GG}	0.386 (1.9)	1.023 (8.0)	-0.140 (-0.4)	0.244 (2.2)	-0.205 (-1.4)
log-likelihood function	-2097.2	3220.2	-1563.8	-3500.9	-2782.04
Significance fixed effects (p-values)	0.201	0.286	0.945	0.000	0.002