A discrete choice model with social interactions; with an application to high school teen behavior

Soetevent, Adriaan R.; Kooreman, Peter

A Discrete Choice Model with Social Interactions; with an Application to High School Teen Behavior

Adriaan R. Soetevent
Peter Kooreman *

University of Groningen

Version January 2004

Abstract

We develop an empirical discrete choice interaction model with a finite number of agents. We characterize its equilibrium properties – in particular the correspondence between the interaction strength, the number of agents, and the set of equilibria – and propose to estimate the model by means of simulation methods.

In an empirical application, we analyze the individual behavior of some 8000 high school teenagers from almost 500 different school classes. We find endogenous social interaction effects to be strong for behavior closely related to school (truancy), somewhat weaker for behavior partly related to school (smoking, cell phone ownership, and moped ownership) and absent for behavior far away from school (asking parents' permission for purchases). Intra-gender interactions are generally much stronger than cross-gender interactions.

Keywords: discrete choice; social interactions; equilibrium properties; teenage behavior
JEL classification: C35, D12

*University of Groningen, Department of Economics, P.O. Box 800, 9700 AV Groningen, The Netherlands. E-mail: a.r.soetevent@eco.rug.nl, p.kooreman@eco.rug.nl.
We thank Rob Alessie, Ulf Bockenholt, Marco Haan, Yannis Ioannides, Joyce Jacobsen, Brian Krauth, Bertrand Melenberg, Bert Schoonbeek, and Michel Wedel for helpful comments and discussions. In addition, we benefitted from comments by seminar participants at ESEM2002, RAND Corporation, Tilburg University, University of Amsterdam, and University of California at Santa Barbara. Soetevent's research was supported by a grant from the MacArthur Research Network on Social Interactions and Economic Inequality, and by the Netherlands Organization for Scientific Research (NWO). Part of this paper was written while Soetevent was a visitor at the University of Wisconsin at Madison, whose hospitality he gratefully acknowledges.

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. The slow progress of empirical research in this area is to a large extent related to a number of methodological problems. As described by Manski (1993, 2000) and others, a major difficulty is to disentangle endogenous social interactions (which imply a social multiplier effect) from other types of social interactions (which do not imply a multiplier effect). Another problem is the endogeneity of reference groups. Recent years have shown an increasing number of empirical studies searching for credible empirical evidence on social interactions, in part by using data that are quasi-experimental in nature; see Sacerdote (2001), Durlauf and Moffitt (2003), and Duflo and Saez (2003) for examples.

The present paper focuses on methodological problems related to a specific but frequently encoutered situation: social interactions in small groups when choice variables are discrete. 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.

In Section 2 we present the model and characterize its equilibrium properties, in particular the correspondence between the interaction strength, the number of agents, and the set of equilibria. We also show that – contrary to

standard binary choice models – these equilibrium properties depend on the choice of support of the dependent variable ($\{0,1\}$ or $\{-1,1\}$). Section 3 proposes to estimate the model by means of simulation moments, assuming that observed choices represent an equilibrium of the static discrete game played by all interacting agents. Section 4 is devoted to an empirical application. We analyze a sample of almost 500 high school classes with detailed information on the individual behavior of the students within each class. As all students in a sampled class are interviewed in principle, the data set has rich information on the behavior of potentially important peers of each respondent. We estimate the model for five types of teen discrete choice behavior: Smoking, truancy, moped ownership, cell phone ownership, and asking parents' permission for purchases. To control for sorting into schools and omitted variables that induce a positive correlation between peers, we also estimate versions that allow for school specific fixed effects and for within-class correlation of error terms. We find strong social interaction effects for behavior closely related to school (truancy), 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 generally much stronger than cross-gender interactions.

A number of recent papers have analyzed social interactions in a discrete choice framework. Brock and Durlauf (2001a and 2003) use a random fields approach to study aggregate behavioral outcomes in an economy in which social interactions are imbedded in individual decisions. Equilibrium properties of this model are derived by imposing a rational expectations condition on the subjective choice probabilities of the agents and by assuming that the number of agents is sufficiently large that each agent ignores the effect of his own choice on the average choice level. In contrast, the present paper describes behavior in relatively small groups of a given size in which

choices of other individuals can be assumed to be fully observable. In this case, it is more appropriate to make an individual's payoff dependent on the actual choice of others in his group. For this reason, the equilbria in the current model can be interpreted as one-shot pure Nash equilibria. In a recent paper Tamer (2003) proposes a semiparametric estimator which allows – under certain conditions – for consistent point estimation of the model in the $N = 2$ case without making assumptions regarding nonunique outcomes. Its extension and empricial implementation to $N >> 2$ have not been fully developed as yet. Gaviria and Raphael (2001) analyze school-based peer effects in the individual discrete choice behavior of tenth-graders. However, their econometric model ignores multiplicity of equilibria.

2 Discrete Choice Interactions and Multiple Equilibria

2.1 Preliminaries

Consider a population of N individuals indexed by i, $i = 1, 2, \ldots, N$. Each player i faces a binary choice and these choices are denoted by an indicator variable y_i which has support $Y_i = \{-1, 1\}$. Y_i is the strategy set of player i and $Y = \times_{i=1}^{N} Y_i$. Elements of Y are called *strategy profiles* or *choice patterns*. A strategy profile is denoted by $\mathbf{y} = (y_i, \mathbf{y_{-i}})$, where $\mathbf{y_{-i}} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N)'$. Note that the number of elements in Y is 2^N. Each individual makes a choice in order to maximize a payoff function $V : Y \to \mathbb{R} \cup \{-\infty\}$. For ease of exposition we will sometimes refer to $y = 1$ as "smoking" and to $y = -1$ as "non-smoking", although we will also consider other types of behavior in the empirical part of the paper.

In the standard economic approach, the payoff function is dependent on individual characteristics. Following the notation in Brock and Durlauf (2001b), we assume that these characteristics can be divided into an observable vector $\mathbf{x_i}$ and a random shock $\epsilon_i(y_i)$ that is unobservable to the

modeller but observable to agent i. Moreover, in interactions-based models explicit attention is given to the influence of the behavior of others on each individual's choice. Each choice is then described as

$$(1) \quad \max_{y_i \in Y_i} V(y_i, \mathbf{x_i}, \mathbf{y_{-i}}, \epsilon_i(y_i)).$$

Similar to Brock and Durlauf (2001b), we assume that the payoff function V can be additively decomposed into three terms:

$$(2) \quad V(y_i, \mathbf{x_i}, \mathbf{y_{-i}}, \epsilon_i(y_i)) = u(y_i, \mathbf{x_i}) + S(y_i, \mathbf{x_i}, \mathbf{y_{-i}}) + \epsilon_i(y_i),$$

where the first term $u(y_i, \mathbf{x_i})$ denotes deterministic private utility, $S(y_i, \mathbf{x_i}, \mathbf{y_{-i}})$ denotes deterministic social utility and ϵ_i denotes random private utility. In this paper we assume the social utility term to have the following form

$$S_i \equiv S(y_i, \mathbf{x_i}, \mathbf{y_{-i}}) = \frac{\gamma}{2(N-1)} y_i \sum_{j \neq i} y_j.$$

Define $\mathbf{y_{-ij}} = \mathbf{y} \backslash \{y_i, y_j\}$ so that $(y_i, \mathbf{x_i}, \mathbf{y_{-i}}) = (y_i, \mathbf{x_i}, y_j, \mathbf{y_{-ij}})$. Note that

$$(3) \quad \begin{array}{l} \{V(1, \mathbf{x_i}, 1, \mathbf{y_{-ij}}, \epsilon_i(y_i)) - V(-1, \mathbf{x_i}, 1, \mathbf{y_{-ij}}, \epsilon_i(y_i))\} - \\ \{V(1, \mathbf{x_i}, -1, \mathbf{y_{-ij}}, \epsilon_i(y_i)) - V(-1, \mathbf{x_i}, -1, \mathbf{y_{-ij}}, \epsilon_i(y_i))\} = \\ \{S(1, \mathbf{x_i}, 1, \mathbf{y_{-ij}}) - S(-1, \mathbf{x_i}, 1, \mathbf{y_{-ij}})\} - \\ \{S(1, \mathbf{x_i}, -1, \mathbf{y_{-ij}}) - S(-1, \mathbf{x_i}, -1, \mathbf{y_{-ij}})\} = \\ \frac{2\gamma}{N-1}. \end{array}$$

Thus, for $\gamma > 0$ the utility of smoking (versus non-smoking) when another person smokes as well is larger than the utility of smoking (versus non-smoking) when another person does not smoke. In this case the parameter γ measures the strategic complementarity between the choice of any pair of individuals; for $\gamma < 0$ it measures the extent to which the choices are strategic substitutes.[1] In fact, for $\gamma > 0$ ($\gamma < 0$), the model falls into the class of supermodular (submodular) games. Supermodular (submodular) games are games in which each player's strategy set is partially ordered and the marginal returns to increasing one's strategy (in this paper moving

[1] When $\gamma = 0$, the model reduces to the standard binary choice formulation without externalities.

from $y = -1$ to $y = 1$) rise (decrease) with increases in the competitors' strategies.[2]

Conditional on the choice by individual i, deterministic private utility is assumed to be a linear function of exogenous characteristics $\mathbf{x_i}$, i.e. $u(1, \mathbf{x_i}) = \beta_1' \mathbf{x_i}$ and $u(-1, \mathbf{x_i}) = \beta_{-1}' \mathbf{x_i}$.

The best response function of individual i given the choices of the other individuals can now be represented as

$$(4) \quad \begin{cases} y_i^* = \beta' \mathbf{x_i} + s_i + \epsilon_i & \\ y_i = 1 & \text{if } y_i^* > 0 \\ y_i = -1 & \text{if } y_i^* \leq 0 \end{cases}$$

where

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

and where y_i^* denotes the difference between the utility individual i derives from choosing $y_i = 1$ and the utility he derives from choosing $y_i = -1$, conditional on $\mathbf{y_{-i}}$, that is,

$$y_i^* = V(1, \mathbf{x_i}, \mathbf{y_{-i}}, \epsilon_i(1)) - V(-1, \mathbf{x_i}, \mathbf{y_{-i}}, \epsilon_i(-1)),$$

with $\beta \equiv \beta_1 - \beta_{-1}; \epsilon_i \equiv \epsilon_i(1) - \epsilon_i(-1)$.

Define $\mathbf{x} \equiv (\mathbf{x_1'}, \mathbf{x_2'}, \ldots, \mathbf{x_N'})'$ and $\epsilon \equiv (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)'$. A strategy profile $\mathbf{y}$ is a pure Nash equilibrium profile if and only if it is consistent with (4) 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 $y_i = -1$. Let $Q(\beta, \gamma, \mathbf{x}, \epsilon, N)$ denote the number of pure Nash equilibria given $\{\beta, \gamma, \mathbf{x}, \epsilon\}$ and the population size N. That is, for $N \geq 2$,

$$(5) \quad Q(\beta, \gamma, \mathbf{x}, \epsilon, N) =$$

$$\sum_{t=1}^{2^N} \left[\prod_{i=1}^{N} I\left(\epsilon_i > -\beta' \mathbf{x_i} - \frac{\gamma}{N-1} \sum_{j \neq i} y_{jt} \right)^{\frac{1+y_{it}}{2}} I\left(\epsilon_i \leq -\beta' \mathbf{x_i} - \frac{\gamma}{N-1} \sum_{j \neq i} y_{jt} \right)^{\frac{1-y_{it}}{2}} \right]$$

[2] Milgrom and Roberts (1990, p. 1255). See also Vives (1990) and the textbook treatments of Topkis (1998) and Vives (1999).

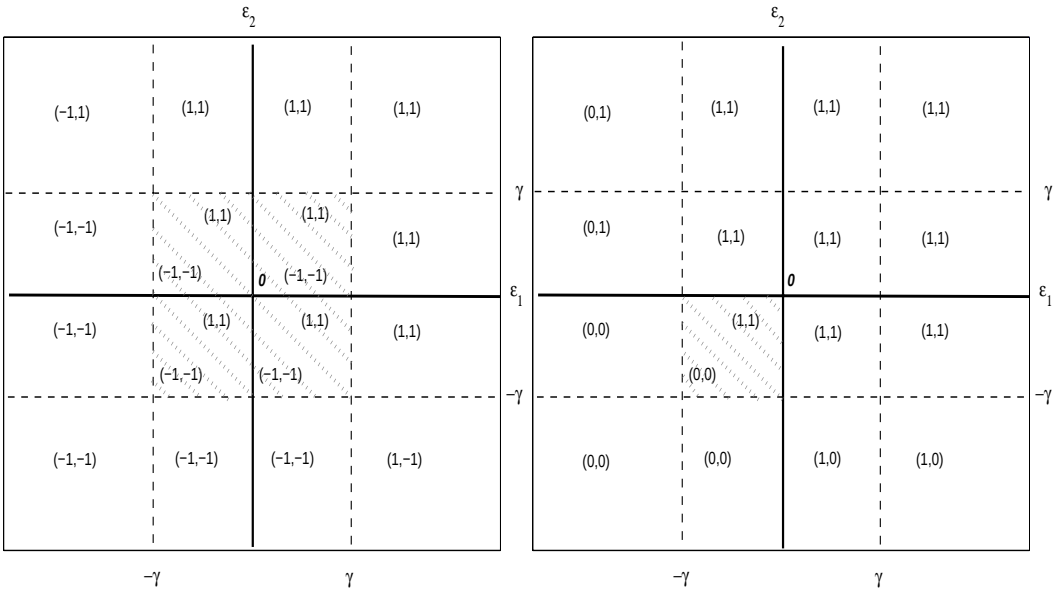

Figure 1: Multiple equilibria in ϵ-space ($N = 2, \gamma > 0, \beta'\mathbf{x_1} = \beta'\mathbf{x_2} = 0$) for support Y_i (left panel) and support $\tilde{Y}_i$ (right panel).

with $I(\cdot)$ an indicator function.[3][4] In the model without social interactions (i.e. $\gamma = 0$) each combination of $\{\beta, \gamma = 0, \mathbf{x}, \epsilon\}$ obviously defines a unique equilibrium, and thus $Q(\beta, 0, \mathbf{x}, \epsilon, N) = 1$.

An important feature of the model with social interactions is that, for a given combination of $\{\beta, \gamma \neq 0, \mathbf{x}, \epsilon\}$, several strategy profiles may be consistent with (4). For example, if $N = 2, \gamma = 1$, and $\beta'\mathbf{x_1} + \epsilon_1 = \beta'\mathbf{x_2} + \epsilon_2 = -\frac{1}{2}$, profiles $\mathbf{y} = (1, 1)'$ and $\mathbf{y} = (-1, -1)'$ are both consistent with (4). In the left panel of figure 1, equilibrium profiles for this two-person game are drawn in ϵ-space. The shaded area is the area with multiple equilibria.

[3] We follow the convention $0^0 = 1$.

[4] If the disturbances are i.i.d. with cumulative distribution function $F(\cdot)$, the expected number of equilibria can be expressed as

$$E[Q(\beta, \gamma, \mathbf{x}, N)] = \int Q(\beta, \gamma, \mathbf{x}, \epsilon, N) dF(\epsilon) =$$
$$\sum_{t=1}^{2^N} \left[\prod_{i=1}^{N} \left(1 - F\left(-\beta'\mathbf{x_i} - \frac{\gamma}{N-1} \sum_{j \neq i} y_{jt}\right)\right)^{\frac{1+y_{it}}{2}} F\left(-\beta'\mathbf{x_i} - \frac{\gamma}{N-1} \sum_{j \neq i} y_{jt}\right)^{\frac{1-y_{it}}{2}} \right].$$

See Soetevent (2004) for some properties of $\partial E[Q(\beta, \gamma, \mathbf{x}, N)]/\partial\gamma$.

2.2 Choice of support $\{-1, 1\}$ versus $\{0, 1\}$

It is of some importance to discuss the choice of support $Y_i = \{-1, 1\}$ instead of the alternative $\tilde{Y}_i = \{0, 1\}$. The latter is the common choice in standard binary choice models where the difference is just a matter of scaling and therefore immaterial. In this section, we will show that the specific choice of support *does* affect the equilibrium properties of binary choice interaction models. This fact has hitherto not been explicitly recognized in the literature. Krauth (2001) for example, taciturnly switches to $\tilde{Y}_i$ as support in his development of the small sample analog of the Brock-Durlauf model, whereas these authors themselves employ Y_i. Key idea is that in using support Y_i, the model is symmetric and therefore invariant with respect to interchanging the two choices. This is not the case with $\tilde{Y}_i$.

This difference between the two models becomes clear when one compares the equilibria for the two-person game in ϵ-space under the assumption that exogenous variables are irrelevant ($\beta' \mathbf{x_1} = \beta' \mathbf{x_2} = 0$). The left panel of figure 1 uses support Y_i and is symmetric with respect to the line $\epsilon_1 + \epsilon_2 = 0$. The right panel, which uses support $\tilde{Y}_i$, is not.

Compared to the left panel of figure 1, one observes that in the right panel the shaded area with multiple equilibria is reduced and restricted to the points where the private utility difference of smoking for both players is negative ($\beta' \mathbf{x_i} + \epsilon_i = \epsilon_i < 0, i = 1, 2$). When using $\tilde{Y}_i$, one implicitly assumes that only positive choices have a social effect. A justification for this choice of support might be given from an evolutionary point of view, for example by arguing that everybody starts as a non-smoker. In that case only the teenagers who start smoking give a signal while the number of non-smokers is irrelevant. Note, however, that the decision not to smoke can convey just as strong a signal to others, especially in environments with many smokers.[5]

[5] To give an example, suppose that in a class with 9 teenagers, 3 of them would smoke were social interactions absent ($\gamma = 0$), that is, $y_i^* = \beta' \mathbf{x_i} + \epsilon_i > 0$ for three of them and $y_i^* \leq 0$ for the others. How would one interpret in this instance the observation

In other contexts however, $\tilde{Y}_i$ may be the preferred support. Consider for example the context in which firms have to make a decision to enter a certain market (Tamer, 2002). It is plausible that this decision is only dependent on how many other firms decide to enter the market and that the number of firms that decide not to enter is irrelevant. All results in the sequel are derived while working with support Y_i.

2.3 Equilibrium properties

This section provides three propositions on the equilibrium properties of model (4). Proposition 1 guarantees equilibrium existence. It turns out that the situation with strategic complements ($\gamma > 0$) is characterized by fundamentally different equilibrium behavior than the one with strategic substitutes ($\gamma < 0$). Moreover, in the latter case it makes a difference whether the population has an even or an odd number of members. Propositions 2 and 3 provide strict upper bounds on the number of equilibria, for the case with strategic complements and for the case with strategic substitutes, respectively.

Define $z_i \equiv \beta' \mathbf{x_i} + \epsilon_i$ and $k \equiv \sum_{i=1}^{N} y_i$, that is, k is the net number of agents choosing $y = 1$.[6] Rank observations on basis of the values of z_i. Denote the ordered values as $z_{[1]} \geq z_{[2]} \geq \ldots \geq z_{[N]}$. Denote the corresponding values of y for the agent with $z_{[j]}$ as $y_{[j]}$. Note that the latter are not ordered, such that it is not precluded that e.g. $y_{[j]} < y_{[j+1]}$.

Proposition 1 Equilibrium existence

For every combination $\{\beta, \gamma, \mathbf{x}, \epsilon\}$ *there exists at least one vector* $\mathbf{y} \equiv (y_1, y_2,$

of zero smokers in this class? A natural explanation is that due to a social effect, the six non-smokers keep the potential smokers from smoking. Support Y_i allows for this explanation, since the difference in social utility of smoking when nobody else smokes equals $\frac{\gamma}{N-1}\sum_{j\neq i} y_j = \gamma\frac{-8}{8} < 0$ for $\gamma > 0$. On the contrary, with $\tilde{Y}_i$ as underlying support, $\frac{\gamma}{N-1}\sum_{j\neq i} y_j = 0$ irrespective of γ, such that social interactions cannot offer an explanation.

[6] Note that given N, only those values of k for which $N + k$ is an even number are possible. This follows from the observation that $k = a \cdot 1 - (N - a), a \in \{0, 1, \ldots, N\}$ can be rewritten as $N + k = 2a$.

$\ldots, y_N)'$ *for which (4) holds.*

Proof: See the Appendix.

Proposition 2 Maximum number of equilibria (strategic complements)

For every combination $\{\beta, \gamma > 0, \mathbf{x}, \epsilon\}$*, the discrete interaction model (4) with* N *agents can have at most* $d(N)$ *distinct equilibria, with*

$$(6) \quad d(N) = \lfloor \frac{N}{2} + 1 \rfloor.$$

Moreover, for every number N*, there exists a combination of* $\{\beta, \gamma > 0, \mathbf{x}, \epsilon\}$ *for which* $Q(\beta, \gamma, \mathbf{x}, \epsilon, N) = d(N)$*.*

Proof: See the Appendix.

The first part of proposition 2 states that in case of strategic complements the maximal number of equilibria grows linearly in N. The second part ensures that the upper bound on the number of equilibria is strict.

Proposition 3 Maximum number of equilibria (strategic substitutes)

For every combination $\{\beta, \gamma < 0, \mathbf{x}, \epsilon\}$*, the discrete interaction model (4) with* N *agents can have at most* $d(N)$ *distinct equilibria, with*

$$\begin{aligned} d(N) = d^e(N) &= \frac{N!}{(N/2)!(N/2)!} && \text{if } N \text{ is even, and} \\ d(N) = d^o(N) &= \frac{N!}{\{(N+1)/2\}!\{(N-1)/2\}!} && \text{if } N \text{ is odd.} \end{aligned}$$

Moreover, for every even (odd) number N*, there exists a combination of* $\{\beta, \gamma < 0, \mathbf{x}, \epsilon\}$ *for which* $Q(\beta, \gamma, \mathbf{x}, \epsilon, N) = d^e(N)$ *(*$Q(\beta, \gamma, \mathbf{x}, \epsilon, N) = d^o(N)$*).*

Proof: See the Appendix.

Proposition 3 states that for the situation with strategic substitutes, the maximal number of equilibria grows exponentially in N. As in the case with strategic complements, the upper bound on the number of equilibria is strict. Note that $\frac{d^e(N)}{d^o(N-1)} = 2$ for all even N and $\lim_{N\to\infty} \frac{d^o(N)}{d^e(N-1)} \uparrow 2$ for N odd. That is, in the limit adding one agent to the population doubles the upper bound on the number of equilibria.
It is also worth mentioning that with strategic substitutes $|k| = |\sum_{i=1}^{N} y_i|$ decreases monotonically to 0 (1) as $\gamma \to -\infty$ for N even (N odd). In fact, this result holds more generally: in equilibrium, the difference between the number of agents choosing $y = 1$ and the number of agents choosing $y = -1$, is smaller when γ is more negative, other things equal.

2.4 Extension to more general interactions

The model considered so far only allows for identical interactions between all individuals in the group. One can think more general interactions, where the degree of interaction between two given individuals depends on e.g. their socio-economic characteristics. In this section, we briefly discuss the consequences of one particular extension of the model given by (4) in which the degree of interaction is made gender-dependent. This leads to four different interaction parameters: γ_{GB} measures the effect of boys on girls; γ_{BG} from girls on boys, and γ_{GG} and γ_{BB} the intra-gender effects between girls and boys, respectively. Specify

$$(7) \quad \begin{cases} y_i^* = \beta' \mathbf{x_i} + S_i + \epsilon_i & \\ y_i = 1 & \text{if } y_i^* > 0, \\ y_i = -1 & \text{if } y_i^* \leq 0. \end{cases}$$

where

$$S_i = \begin{cases} (\gamma_{GG} \sum_{j=1, j\neq i}^{N} y_j^G & +\gamma_{GB} \sum_{j=1}^{N} y_j^B &)/(N-1) & \text{if } i \text{ is a girl,} \\ (\gamma_{BG} \sum_{j=1}^{N} y_j^G & +\gamma_{BB} \sum_{j=1, j\neq i}^{N} y_j^B &)/(N-1) & \text{if } i \text{ is a boy,} \end{cases}$$

with $y_j^G \equiv y_j \cdot I(j \text{ is a girl})$ and $y_j^B \equiv y_j \cdot I(j \text{ is a boy})$, $\forall j$.

Corollary 1 *For every combination* $\{\beta, \gamma_{BB} \geq 0, \gamma_{GG} \geq 0, \gamma_{GB}, \gamma_{BG}, \mathbf{x}, \epsilon\}$ *there exists at least one vector* $\mathbf{y} \equiv (y_1, y_2, \ldots, y_N)'$ *for which (7) holds.*

Proof: See the Appendix.

The equivalent of proposition 2 for the extended model follows automatically:

Corollary 2 *For every combination* $\{\beta, \gamma_{BB} > 0, \gamma_{GG} > 0, \gamma_{GB}, \gamma_{BG}, \mathbf{x}, \epsilon\}$, *the discrete interaction model given by (7) with* N_G *girls and* N_B *boys can have at most* $d^*(N_B, N_G)$ *distinct equilibria, where*

$$d^*(N_B, N_G) \quad = \quad \lfloor \frac{N_B}{2} + 1 \rfloor \cdot \lfloor \frac{N_G}{2} + 1 \rfloor.$$

Moreover, for all N_G *and* N_B, *there exists a combination of* $\{\beta, \gamma_{BB} \geq 0, \gamma_{GG} \geq 0, \gamma_{GB}, \gamma_{BG}, \mathbf{x}, \epsilon\}$ *for which the maximum number of equilibria is obtained.*

It is noteworthy that the values of the cross-gender interaction parameters γ_{GB} and γ_{BG} do not play a role in determining the maximum number of equilibria.

3 Estimation by simulation

To estimate the model we require the probability $P(\mathbf{y})$ that we observe $\mathbf{y}$, for any given set of parameter values.

A choice pattern $\mathbf{y}$ observed for a particular group is either a single equilibrium or one of multiple equilibria. The support in ϵ-space for choice pattern $\mathbf{y}$ is

$$(8) \quad \begin{cases} \epsilon_i > -\beta' \mathbf{x_i} - s(\mathbf{y_{-i}}) & \text{if } y_i = 1 \\ \epsilon_i \leq -\beta' \mathbf{x_i} - s(\mathbf{y_{-i}}) & \text{if } y_i = -1 \end{cases}$$

where

$$s(\mathbf{y}_{-\mathbf{i}}) = \frac{\gamma}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} y_j,$$

for all $i, i = 1, \ldots, N$. Denote the region in ϵ-space defined in (8) by $W(\mathbf{y}, \theta)$, with θ being the parameters to be estimated. Since $W(\mathbf{y}, \theta)$ may also support equilibria other than $\mathbf{y}$, we have $P(\epsilon \in W(\mathbf{y}, \theta)) \geq P(\mathbf{y})$.

Following Bjorn and Vuong (1983) and Kooreman (1994) we make a randomization assumption in case of multiple equilibria: whenever the model generates multiple equilibria we assume that one of them will occur with probability equal to one over the number of equilibria. To determine the number of equilibria in the various subregions of $W(\mathbf{y}, \theta)$ we use a simulation based method. Consider R random draws (indexed by r, $r = 1, \ldots, R$) from the joint distribution of $(\epsilon_1, \ldots, \epsilon_N)$ on $W(\mathbf{y}, \theta)$. For each draw, we calculate the number of equilibria. Recall that by construction of $W(\mathbf{y}, \theta)$, $\mathbf{y}$ 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 the frequency simulator

$$(9) \quad P_1(\mathbf{y}) = P(\epsilon \in W(\mathbf{y}, \theta)) \cdot \frac{1}{R_1} \sum_{r=1}^{R_1} \frac{1}{E_r}$$

This procedure guarantees the statistical coherency of the model, i.e. $\sum_{t=1}^{2^N} P_1(\mathbf{y} = \mathbf{y}_t) = 1$, where $\mathbf{y}_t, t = 1, \ldots, N$ is the enumeration of all elements in Y.

We have found that $R_1 = 1000$ generates estimated probabilities that are sufficiently precise as inputs in a maximum likelihood procedure. Note that since $E_r \geq 1$ we have $\frac{1}{R_1} \sum_{r=1}^{R_1} \frac{1}{E_r} \leq 1$. If the disturbances are i.i.d., $P(\epsilon \in W(\mathbf{y}, \theta))$ can be straightforwardly evaluated as the product of N univariate probabilities; for a relaxation of the i.i.d. assumption, see subsection 4.6.

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

$$P_2(\mathbf{y}) = \frac{1}{R_2}\sum_{r=1}^{R_2}\frac{I(\mathbf{y}\in\mathbf{\Omega_r})}{E_r} \tag{10}$$

with R_2 the number of draws from the joint distribution of $(\epsilon_1,\ldots,\epsilon_N)$ on $\Re^N$. However, this would require the number of draws to be of a much larger magnitude to achieve the same precision as achieved when using (9).

The characterization of the equilibria in propositions 1, 2, and 3 and their proofs turns out to be extremely helpful in developing an algorithm for estimation. Let $M = \sum_{i=1}^{N} I(y_i = 1)$, i.e. M denotes the number of individuals choosing $y = 1$. Then $\sum_{i=1}^{N} y_i = k$ implies $M = \frac{1}{2}(N+k)$. From the proof of proposition 2 it follows that, with $\gamma > 0$, 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, for a given draw of $(\epsilon_1,\ldots,\epsilon_N)$. An equilibrium with $\sum_{i=1}^{N} y_i = k$ exists if and only if the inequalities

$$\begin{array}{llll} z_{[N]} + \frac{k+1}{N-1}\gamma & \leq \ldots & \leq z_{(M+1)} + \frac{k+1}{N-1}\gamma \leq 0 < \\ z_{[M]} + \frac{k-1}{N-1}\gamma & \leq \ldots & \leq z_{[1]} + \frac{k-1}{N-1}\gamma, \end{array} \tag{11}$$

with $1 \leq M = \frac{1}{2}(N+k) \leq N-1$, are satisfied. An equilibrium with $M = 0$ occurs if and only if $z_i - \gamma \leq 0$ for all i; an equilibrium with $M = N$ occurs if and only if $z_i + \gamma > 0$ for all i. The proof of proposition 2 also shows that two vectors $\mathbf{y}$ and $\tilde{\mathbf{y}}$ that differ in only one element cannot both be equilibria. As a result, we only have to check $d(N) = \lfloor \frac{N}{2} + 1 \rfloor$ out of the 2^N choice patterns as possible equilibria.

Suppose that model (7), with all γ's positive, has an equilibrium with M_G smoking girls and M_B smoking boys. It is straightforward to show 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 $d^*(N_B, N_G) = \lfloor \frac{N_B}{2} + 1 \rfloor \cdot \lfloor \frac{N_G}{2} + 1 \rfloor$ out of the 2^N choice patterns as potential equilibria.

Proposition 3 implies that with one or more negative γ's estimation is computationally more demanding. However, negative γ's were encountered in only a limited number of cases; see also Kooreman (2003).

Having calculated for each group the probability that the observed choice pattern occurs using, the model can be estimated by maximum likelihood.

From an empirical perspective it is important to note that in the estimated models the probability of a single equilibrium usually turns out to be larger than 80 percent, i.e. we usually have $\frac{1}{R_1}\sum_{r=1}^{R_1}\frac{1}{E_r} > 0.8$. The estimation results in this paper's application also appear to be largely insensitive with respect to the assumptions regarding the treatment of multiple equilibria. For example, maximizing a quasi-loglikelihood based on $P(\epsilon \in W(\mathbf{y},\theta))$ yields estimates very similar to those based on $P_1(\mathbf{y})$.

4 Empirical application

4.1 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.[7]

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.

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

[7]Previous 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 Institue 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.

a pupil was absent when the questionnaires were filled out, in other cases because information on some of the variables is missing.

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 endogenous 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.[8]

4.2 Specification of the empirical model

Given the cross-section nature of the data we will not be able to fully account for the identification problems that characterize the empirical analysis of social interactions. In order to provide a proper perspective for the interpretation of the empirical results to be presented, we briefly discuss the identification issues in relation to the present data set: *i)* the definition of the reference group, *ii)* non-random selection into reference groups, and *iii)*

[8] A US data set which is comparable to the present one is the National Education and Longitudinal Study (NELS), see *e.g.* Gaviria and Raphael (2001). Both the Dutch NSYS and the NELS focus on non-cognitive outcomes within schools. The NELS is a biannual survey, first held in 1988, and samples students within roughly 1000 schools. An important difference with the Dutch NSYS is that the NELS surveys only a relatively small group of students within each school. For example, in the 1990 sample used by Gaviria and Raphael, the mean sample size per school was 13.3 students. While the NELS contains information on school averages, these are not available per class, grade, or gender. This limits the possibilities for an analysis of interactions within schools (for example, it is impossible to allow for a school specific fixed effect) and it precludes any analysis of social interactions within classes. Two other US data sets on teenagers with peer group information are the Teenage Attitudes and Practices (TAPS) and the National Longitudinal Survey of Youths (NLSY). However, the TAPS only contains subjective information on a respondent's four best same-sex friends, whereas the NLSY only has subjective peer information based on questions of the type *"What percentage of kids in your grade...?"*.

simultaneity of mutual endogenous interaction effects.

The definition of the reference group

As in any empirical analysis on social interaction we require an assumption regarding the definition of the reference group – Who interacts with whom? A number of empirical papers have defined the reference group of an individual as the group of all persons in the population within the same age group and with the same education level, using the sample analogues as an approximation; see *e.g.* Kapteyn *et al.* (1997) and Aronsson *et al.* (1999). This is a crude definition, largely motivated by data limitations. A more attractive alternative is to use subjective information on an individual's reference group, as in Woittiez and Kapteyn (1998). However, the information on the reference group of a sampled individual is often limited as these reference group members are not themselves included in the sample. The data in the current analysis can be viewed as a *reference group based sample* as all students within a sampled class are interviewed in principle. While teenage behavior is obviously also influenced by persons outside the class, classmates are likely to play a dominant role in shaping teenagers' preferences and behavior. On a weekday, the average student in the sample spends about six hours in his or her school class. The total time spent on school related activities (including homework and commuting) is about eight hours per weekday, more than fifty percent of the daily waking time. Teenagers within the same school or class therefore form social groups that are more clearly defined and delineated than in many other situations in which social interactions are likely to play a role. Obviously, the definition of the reference group could be extended to allow for interactions with students outside the class. Also, one could in principle refine the specification of social groups within the class beyond the boy-girl distinction, for example on the basis of ethnicity, or by allowing the effect of younger and of older

classmates to be different. These extensions are left for future research.

Non-random selection into reference groups

With respect to the selection issue one can make a distinction between *selectors* and *actors*. An actor is the one whose behavior is being analyzed. A selector is the one who decides to which reference group the actor belongs. Contrary to most other studies on social interactions (see Duflo and Saez (2002, 2003) for a recent exmaple) selectors and actors are not identical in the present analysis: Selection into classes and schools is to a large extent determined by parents and school authorities. More importantly, selection into classes is usually based on cognitive abilities whereas the present analysis focuses on non-cognitive behaviors. In fact, we will find that for some behaviors within-class correlation is absent, suggesting that the selection issue is less poignant here than in other studies on social interactions. To control for the selection issue to some extent we will also estimate a version of the model including school specific fixed effects and allowing for within-class correlation of error terms.

Endogenous versus contextual effects

Gaviria and Raphael (2001) argue that students are less exposed to the family background of their school peers than they are exposed to the family background of peers residing in the same neighborhood. They conjecture that in an analysis of interactions through schools contextual effects are less important than in an analysis of interactions through neighborhoods. In their empirical analysis they assume that contextual effects are absent. Kawaguchi (2004) invokes subjective information about the perception of peer behaviors to achieve full identification.[9] He finds that the absence of contextual effects cannot be rejected. The empirical results presented

[9]Identification is based on the problematic assumption that perceived behavior is not determined by actual behavior.

below are based on the assumption that there are no contextual effects. The estimates on the endogenous social interaction effects should therefore be interpreted as upper bounds on the true effects.

The vector $\mathbf{x}$ 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' eduation 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).[10] Tables 1 provides sample statistics for both the endogenous and exogenous variables in the model.

4.3 Estimation results

Table 2 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'

[10] A 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.

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 other parameters.

4.4 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 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.[11]

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 **??**. 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 inter-

[11] Clearly, a more flexible specification would be obtained by allowing for class specific fixed effects. With the current data, the estimation of class specific effects is infeasible. However, below we will estimate a version with class specific random effects.

action 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 truancy, moped ownership, cell phone ownership, and asking parents' permission for purchases.[12] Tables 3 and 4 report the results without and with school specific fixed effects, respectively. (For ease of comparison the first column in table 3 repeats the second column from table 2 and the first column in table 4 repeats the fourth column from table 2.)

The significance of the fixed effects varies across the five types of behavior. For truancy, smoking, and moped ownership the fixed effects are not significant (see bottom row of table 2), while for cell phone ownerhip and asking parents' permission they are significant. The discussion of estimation results below is therefore based on table **??** for smoking, truancy, and moped ownership, and on table 2 for the other two choice behaviors.

For truancy, the intra-gender effects are stronger than for smoking. Moreover, we now also have significant cross-gender interactions. The probability of truancy 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 truancy.

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

[12]The variable 'truancy' in the empirical analysis is based on the question *"How often have you been playing truant during the last (school)month?"*. As truanters have a larger probability of being absent when the questionnaire is being filled out, there is a potential selection bias. The effect on the estimated social interaction coefficients, however, is likely to be small. The absence of a group of truanters with strong mutual interactions might bias the estimated γ's towards zero, but the presence of a group on non-truanters with strong mutual interactions will have the opposite effect. Moreover, tentative calculations indicate that the probability of a student truanting on a random schoolday is in the order of one percent.

than for girls. It strongly increases in age (the legal minimum age for riding 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 ownerhip 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 suggests that pupils do not copy each other's responses when filling out the questionnaire. It also indicates that the effects found for the other four types of choice behavior represent genuine endogenous social interaction effects rather than unobserved social group effects.

4.5 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 hypothetical 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 **??**, we find that in equilibrium the expected number of truanters is 3.14 (the probability of truancy is 0.191

for girls and 0.201 for boys).[13]

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 truancy 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 16 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 truancy probability of her daughter increases from 0.191 to 0.210. However, it also changes the equilibrium truancy probabilities of the other girls (from 0.1909 to 0.1915) and boys (from 0.2002 to 0.2012). As a result, the expected of number of truanters in class increases not only by 0.019 (0.210-0.191), but by 0.031.

4.6 Correlated within-class error terms

As an additional check on the robustness of the empirical results we also estimated the model for smoking with a slightly more general correlation pattern of the error terms within a class (but without school specific fixed effects). We assume the covariance matrix Σ of $(\epsilon_1, \ldots, \epsilon_N)$ to be a 'one-factor' matrix such that $\Sigma = \{\rho_{ij}\}$ with $\rho_{ij} = \rho$ if $i \neq j$ and $\rho_{ij} = 1$ if $i = j$.

[13] All numbers are based on simulations with R=100000.

To calculate the probabilities $P(\epsilon \in W(\mathbf{y}, \theta))$ we use a decomposition simulator which effectively depends on only a one-dimensional random variable; cf. Stern (1992).[14]

We first estimated this version without social interaction effects. We then found the estimated ρ to be highly significant ($\hat{\rho}$=0.098, t-value 8.6, loglikelihood -2146.8) with the other parameters largely unaffected. When estimating the model with social interaction effects, the estimated ρ is virtually equal to zero and highly insignificant, with the other parameters being identical to those in the second column of table 2. These results are another indication that the γ's are measures of genuine endogenous social interactions effects rather than a reflection of unmeasured class specific effects.

5 Conclusion

We derived a number of equilibrium properties for the binary choice interaction model with a finite number of agents. Both for the case with strategic complements and strategic substitutes, equilibrium existence was proved and tight upper bounds were derived for the size of the set of equilibria, given the number of agents and the degree of interaction between them. We also briefly discussed the consequences for the set of equilibria when the model is extended to allow for gender-dependent interactions. The main finding here is that the cross-gender parameters are irrelevant in the derivation of the upper bounds.

In our application to teenagers' discrete choices, we found strong social interaction effects for behavior closely related to school (truancy), somewhat

[14] Let the random variables $u_1, \ldots, u_N$, and v be independently normally distributed with zero means; $var(u_i) = 1 - \rho$, $i = 1, \ldots, N$ and $var(v) = \rho$. (We require $\rho > 0$; the procedure for $\rho < 0$ is slightly different. Note, however, that the positive definiteness of Σ implies $-\frac{1}{N-1} < \rho < 1$.) Let $\epsilon_i = u_i + v$, $i = 1, \ldots, N$. Then $Cov(\epsilon) = \Sigma$, with Σ defined in the main text. Now $P(\epsilon_1 < z_1, \ldots, \epsilon_N < z_N) = P(u_1 < z_1 - v, \ldots, u_N < z_N - v) = \int \Phi\left(\frac{z_1 - v}{\sqrt{1-\rho}}\right) \ldots \Phi\left(\frac{z_N - v}{\sqrt{1-\rho}}\right) . f(v) dv$, with $\Phi(.)$ the standard normal cumulative distribution function and $f(v)$ a $N(0, \rho)$ density function. The integral is simulated by drawing v from $f(.)$ and then evaluating $\Phi\left(\frac{z_1 - v}{\sqrt{1-\rho}}\right) \ldots \Phi\left(\frac{z_N - v}{\sqrt{1-\rho}}\right)$ conditional on v.

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.

The work presented in this paper indicates various possible extensions for future research. An example is to allow for more general interaction structures, for example by making interaction parameters dependent on socio-economic characteristics. Another, more general issue – typically neglected in the empirical social interactions literature to date – is the question which type of equilibrium concept is appropriate. The fact that classmates interact daily, usually for many years, and often become friends suggests that non-cooperative Nash equilibria may not always be plausible.

While the present data set has a number of important advantages in terms of information on reference group members, the empirical results are subject to the usual qualifications regarding inferences about social interactions based on cross-section data. Future steps toward increasing our understanding of social interactions will require more informative data and models characterized by a tight link between game theory and econometrics.

Appendix: Proofs

Proof of Proposition 1: Equilibrium existence

The case for $\gamma = 0$ is obvious. We prove proposition 1 for the game with strategic complements ($\gamma > 0$) and the game with strategic substitutes ($\gamma < 0$) separately. For the first case, existence can be readily proved by showing that the game belongs to the class of supermodular games. Existence then immediately follows from using Theorem 5 in Milgrom and Roberts (1990, p. 1265). In this appendix however, we will follow for both cases the alternative route of proving equilibrium existence through finding an explicit equilibrium for all combinations of $\{\beta, \gamma, \mathbf{x}, \epsilon\}$. This procedure may give more insight into some of the peculiarities of the model.

Every possible combination of $\{\beta, \gamma > 0, \mathbf{x}, \epsilon\}$ clearly falls into one of the three following categories

(i) $z_{[1]} \leq 0$;

(ii) $z_{[N]} > 0$;

(iii) $z_{[1]} > 0, z_{[N]} \leq 0$;

We show that for each $\mathbf{z}$ in every category there is an associated $\mathbf{y}$ for which (4) holds, for all values $\gamma > 0$.

(i) $z_{[1]} \leq 0$:
$y_i = -1, i = 1, 2, \ldots, N$ $(k = -N)$ is an equilibrium solution, since $y^*_{[1]} = z_{[1]} - \gamma\frac{N-1}{N-1} \leq 0$. This implies that $y^*_{[i]} \leq 0$ $\forall i$ since $\gamma\frac{N-1}{N-1}$ is a constant and $z_{[i]}$ weakly decreases with i.

(ii) $z_{[N]} > 0$:
$y_i = 1, i = 1, 2, \ldots, N$ $(k = N)$ is an equilibrium solution, since $y^*_{[i]} = z_{[i]} + \gamma\frac{N-1}{N-1} > 0, \forall i$.

(iii) $z_{[1]} > 0, z_{[N]} \leq 0$:

Define $M \equiv 0$ if $z_{[j]} \leq -\gamma\frac{(2j-N-1)}{N-1}, \forall j,\ j \in \{1,2,\ldots,N\}$ and $M \equiv \arg\max_i \left(z_{[j]} > \gamma\frac{(2i-N-1)}{N-1}; \forall j \leq i\right)$ otherwise. Five examples of sequences of $z_{[i]}$ with $N = 6$ and $\gamma = 1$ are plotted in figure 2 together with the corresponding values of M. The solid line represents the equation $z_{[i]} = -\gamma\frac{(2i-N-1)}{N-1}$.

If $M = 0$, $y_{[i]} = -1$, $i = 1,2,\ldots,N$ is an equilibrium solution, since $y^*_{[i]} = z_{[i]} - \gamma \leq z_{[1]} - \gamma \leq 0$, $\forall i$. (See the +-sequence in figure 2.)

If $M > 0$, $y_{[i]} = 1$ for $i = 1,2\ldots,M$ and $y_{[i]} = -1$ for $i = M+1, M+2,\ldots,N$ ($k = M - [N-M] = 2M - N$) is an equilibrium solution, since $y^*_{[i]} = z_{[i]} + \gamma\frac{2M-N-1}{N-1} > 0$ for $i = 1,2,\ldots,M$ and $y^*_{[j]} \leq y^*_{[M+1]} = z_{[M+1]} + \gamma\frac{2M-N+1}{N-1} = z_{[M+1]} + \gamma\frac{2(M+1)-N-1}{N-1} \leq 0$ for all $j = M+1, M+2,\ldots,N$.

Note that for sequences of $z_{[i]}$'s for which $M = N$ (like the sequence of circles and x-es in figure 2), $y_{[i]} = -1$, $i = 1,2,\ldots,N$ is another equilibrium solution iff. $z_{[1]} \leq \gamma$. In figure 2, this condition holds for the sequence of x-es but not for the sequence of circles. □

Strategic substitutes ($\gamma < 0$)

In this case, we distinguish between the case where the number of subjects N is even and the case where this number is odd.

N even Let $\gamma < 0$. Define $m \equiv \arg\max_i \left(z_{[i]} > 0\right)$. Suppose that $m > N/2$, that is, the majority of the subjects have a value of z greater than zero. Define the non-overlapping non-empty intervals $I_0 \equiv \left[0, \frac{z_{[m]}(N-1)}{2m-N-1}\right)$; $I_{m-N/2} \equiv \left[\frac{z_{[N/2+1]}(N-1)}{2(N/2+1)-N-1}, \infty\right) = [z_{[N/2+1]}(N-1), \infty)$ and, if $m > N/2+1$, $I_r \equiv \left[\frac{z_{[m-r+1]}(N-1)}{2(m-r+1)-N-1}, \frac{z_{[m-r]}(N-1)}{2(m-r)-N-1}\right)$, for $r = 1,2,\ldots,m-N/2-1$.

First consider the case $m > N/2+1$. Since the intervals are non-

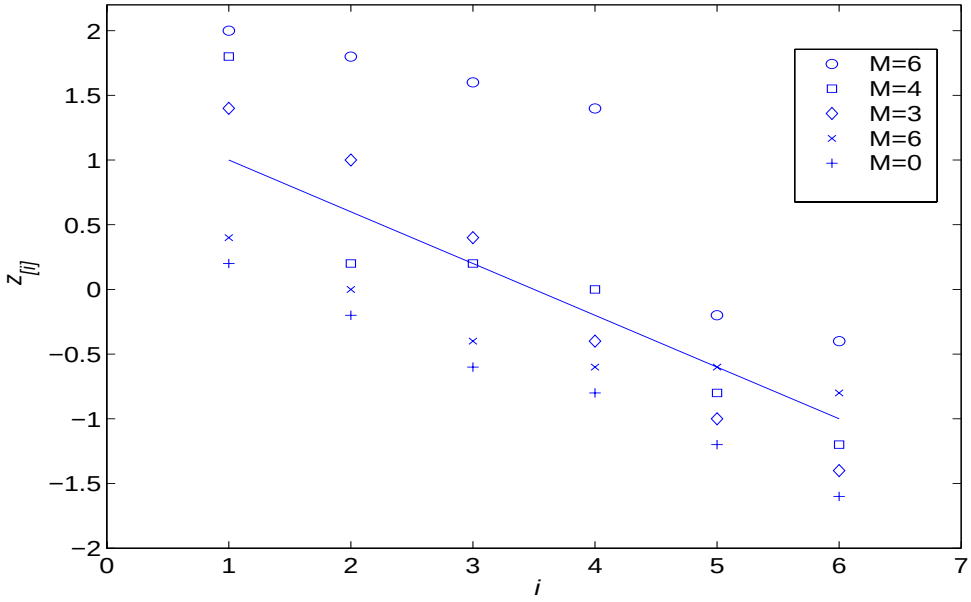

Figure 2: Five examples of $z_{[i]}$-sequences and the corresponding solutions for $M \equiv \arg\max_i \left(z_{[j]} > \gamma \frac{(2i-N-1)}{N-1}; \forall j \leq i \right)$ for the case with $N = 6$ and $\gamma = 1$.

overlapping and since $I_0 \cup I_1 \cup \ldots \cup I_{m-N/2} = [0, \infty)$, $-\gamma$ is in one and only one of these intervals. If $-\gamma \in I_0$, $\mathbf{y} = (1, 1, \ldots, 1_m, -1, \ldots, -1)'$, $(k = 2m - N)$ is an equilibrium, since for this solution $y^*_{[1]} \geq \ldots \geq y^*_{[m]} = z_{[m]} + \gamma \frac{2m-N-1}{N-1} > 0$ and $y^*_{[N]} \leq \ldots \leq y^*_{[m+1]} = z_{[m+1]} + \gamma \frac{2m-N+1}{N-1} \leq 0$. If $-\gamma \in I_r$, for $r = 1, 2, \ldots, m-N/2-1$, $\mathbf{y} = (1, 1, \ldots, 1_{m-r}, -1, \ldots, -1)'$ $(k = 2(m-r)-N)$ is an equilibrium, since for this solution $y^*_{[1]} \geq \ldots \geq y^*_{m-r} = z_{[m-r]} + \gamma \frac{2(m-r)-N-1}{N-1} > 0$ and $y^*_{[N]} \leq \ldots \leq y^*_{[m-r+1]} = z_{[m-r+1]} + \gamma \frac{2(m-r)-N+1}{N-1} \leq 0$. If $-\gamma \in I_{m-N/2}$, $\mathbf{y} = (1, 1, \ldots, 1_{N/2}, -1, \ldots, -1)'$ $(k = 0)$ is an equilibrium, since for this solution $y^*_{[1]} \geq \ldots \geq y^*_{[N/2]} = z_{[N/2]} + \gamma \frac{-1}{N-1} > 0$ and $y^*_{[N]} \leq \ldots \leq y^*_{[N/2+1]} = z_{[N/2+1]} + \gamma \frac{1}{N-1} \leq 0$.

If $m = N/2 + 1$, then $I_0 \cup I_{m-N/2} = I_0 \cup I_1 = [0, \infty)$. Applying similar reasoning, one can verify that $\mathbf{y} = (1, 1, \ldots, 1_{N/2+1}, -1, \ldots, -1)'$, $(k = 2)$ is an equilibrium when $-\gamma \in I_0$ and that $\mathbf{y} = (1, 1, \ldots, 1_{N/2}, -1, \ldots, -1)'$ $(k = 0)$ is an equilibrium when $-\gamma \in I_1$.

If $m = N/2$, then $\mathbf{y} = (1, 1, \ldots, 1_{N/2}, -1, \ldots, -1)'$ is an equilibrium for all $-\gamma \in (0, \infty)$, since $y^*_{[1]} \geq \ldots \geq y^*_{[N/2]} = z_{[N/2]} + \gamma \frac{-1}{N-1} > z_{[N/2]} > 0$ and $y^*_{[N]} \leq \ldots \leq y^*_{[N/2+1]} = z_{[N/2+1]} + \gamma \frac{1}{N-1} < z_{[N/2+1]} \leq 0$.

Due to symmetry, the above argument can be applied for $m < N/2$ with m replaced by $\tilde{m} \equiv N - m \geq N/2$ and the roles of the outcomes +1 and -1 interchanged.

N odd The above argument can also be applied for odd N. Suppose that $m > (N+1)/2$ and define $I_0 \equiv \left[0, \frac{z_{[m]}(N-1)}{2m-N-1}\right)$, $I_{m-(N+1)/2} \equiv \left[\frac{z_{[(N+1)/2+1]}(N-1)}{2\left(\frac{N+1}{2}+1\right)-N-1}, \infty\right) = \left[\frac{z_{[(N+1)/2+1]}(N-1)}{2}, \infty\right)$ and, if $m > (N+1)/2+1$, $I_r \equiv \left[\frac{z_{[m-r+1]}(N-1)}{2(m-r+1)-N-1}, \frac{z_{[m-r]}(N-1)}{2(m-r)-N-1}\right)$, for $r = 1, 2, \ldots, m - (N+1)/2 - 1$.

Taking the case that $m > (N+1)/2 + 1$, it follows that for $-\gamma \in I_0$, $\mathbf{y} = (1, 1, \ldots, 1_m, -1, \ldots, -1)'$ $(k = 2m - N)$ is an equilibrium; for $-\gamma \in I_r$, $r = 1, 2, \ldots, m - (N+1)/2 - 1$, $\mathbf{y} = (1, 1, \ldots, 1_{m-r}, -1, \ldots, -1)'$ $(k = 2(m-r) - N)$ is an equilibrium; and for $-\gamma \in I_{m-(N+1)/2}$, $\mathbf{y} = (1, 1, \ldots, 1_{(N+1)/2}, -1, \ldots, -1)'$ $(k = 1)$ is an equilibrium.

If $m = (N+1)/2 + 1$, then $I_0 \cup I_{m-(N+1)/2} = I_0 \cup I_1 = [0, \infty)$. Applying similar reasoning, one can verify that $\mathbf{y} = (1, 1, \ldots, 1_{(N+1)/2+1}, -1, \ldots, -1)'$ $(k = 3)$ is an equilibrium when $-\gamma \in I_0$ and that $\mathbf{y} = (1, 1, \ldots, 1_{(N+1)/2}, -1, \ldots, -1)'$ $(k = 1)$ is an equilibrium when $-\gamma \in I_1$.

If $m = (N+1)/2$, then $\mathbf{y} = (1, 1, \ldots, 1_{(N+1)/2}, -1, \ldots, -1)'$ is an equilibrium for all $-\gamma \in (0, \infty)$, since $y^*_{[1]} \geq \ldots \geq y^*_{[(N+1)/2]} = z_{[(N+1)/2]} + \gamma \cdot 0 > 0$ and $y^*_{[N]} \leq \ldots \leq y^*_{[(N+1)/2+1]} = z_{[(N+1)/2+1]} + \gamma \frac{2}{N-1} < z_{[N/2+1]} \leq 0$. Again, the case with $m < (N+1)/2$ follows from symmetry. □

Proof of Proposition 2: Maximum number of equilibria (strategic complements)

The proof for strategic complements uses the following lemma:

Lemma 1 *Let $\gamma > 0$. Suppose model (4) has an equilibrium $\mathbf{y}$. Then*

$$\min_{\{i|y_i=1\}} z_i - \max_{\{i|y_i=-1\}} z_i > \frac{2\gamma}{N-1},$$

where $z_i \equiv \beta' \mathbf{x_i} + \epsilon_i$.

Proof of Lemma 1

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

The lemma's effect is that it restricts the maximum number of potential equilibria to $N+1$. The following observation is an immediate consequence of lemma 1:

1 *In any equilibrium the agents with $y_i = 1$ are those with the largest values for z_i.*

Now consider two vectors $\mathbf{y}$ and $\tilde{\mathbf{y}}$ that differ in one element only. Without loss of generality, assume that $y_i = 1$ and $\tilde{y}_i = -1$ for some i. Define $\mathbf{y_{-i}} \equiv (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N)'$ and $\tilde{\mathbf{y}}_{\mathbf{-i}} \equiv (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{i-1}, \tilde{y}_{i+1}, \ldots, \tilde{y}_N)'$. Since $\mathbf{y_{-i}} = \tilde{\mathbf{y}}_{\mathbf{-i}}$, it follows that $y_i^* = z_i + \frac{\gamma}{N-1}\sum_{j\neq i} y_j = z_i + \frac{\gamma}{N-1}\sum_{j\neq i} \tilde{y}_j = \tilde{y}_i^*$ given a combination of $\{\beta, \gamma, \mathbf{x}, \epsilon\}$. This implies that $y_i = \tilde{y}_i$ and we arrive at a contradiction. Note that this result holds irrespective of γ being positive or negative. The following observation is thus obtained:

2 *Two vectors $\mathbf{y}$ and $\tilde{\mathbf{y}}$ that differ in only one element cannot both belong to the set of equilibria.*

From the observations 1 and 2 it follows that the number of equilibria for a given combination of $\{\beta, \gamma > 0, \mathbf{x}, \epsilon\}$ can be at most $d = \lfloor \frac{N}{2} + 1 \rfloor$, where $\lfloor w \rfloor$ denotes the largest integer not larger than w. To give an example: When the number of agents $N = 8$, the maximum number of equilibria can be at most $\lfloor \frac{N}{2} + 1 \rfloor = 5$. Due to statements 1 and 2, the strategy profiles of

these equilibria must be strictly ordered and differ in at least two elements. This leaves the following five strategy profiles as the only candidates:

$(1,1,1,1,1,1,1,1)'$; $(1,1,1,1,1,1,-1,-1)'$; $(1,1,1,1,-1,-1,-1,-1)'$; $(1,1,-1,-1,-1,-1,-1,-1)'$; $(-1,-1,-1,-1,-1,-1,-1,-1)'$.

This proves the first part of proposition 2. The proof of the second part – the upper bound on the number of equilibria is strict – runs as follows. Denote the d equilibria that are to be sustained as[15]

$$\begin{aligned}
\mathbf{y^1} &= (1,1,\ldots,1)' \\
\mathbf{y^2} &= \begin{cases} (1,\ldots,1,-1,-1)' & \text{if } N \text{ is even,} \\ (1,\ldots,1,-1,-1,-1)' & \text{if } N \text{ is odd.} \end{cases} \\
&\vdots \\
\mathbf{y^j} &= \begin{cases} (1,\ldots,1_{N-2(j-1)},-1_{N-2(j-1)+1},\ldots,-1)' & \text{if } N \text{ is even,} \\ (1,\ldots,1_{N-2(j-1)-1},-1_{N-2(j-1)},\ldots,-1)' & \text{if } N \text{ is odd,} \end{cases} \\
&\quad \text{with } j = 3,\ldots,d-1. \\
\mathbf{y^d} &= (-1,-1,\ldots,-1)'.
\end{aligned}$$

First note that $\mathbf{y^1}$ can be sustained as an equilibrium outcome if and only if $z_{[N]} > -\gamma$ and that $\mathbf{y^d}$ can be sustained as an equilibrium outcome if and only if $z_{[1]} \leq \gamma$. Further note that $\mathbf{y^{d-i}}$, $i = 1,\ldots,d-2$ can be sustained as equilibria if and only if $z_{[2i]} > \gamma\frac{N-4i+1}{N-1}$ and $z_{[2i+1]} \leq \gamma\frac{N-4i-1}{N-1}$. The fact that these necessary and sufficient conditions on the values of z can be satisfied simultaneously completes the proof.

Proof of Proposition 3: Maximum number of equilibria (strategic substitutes)

In order to prove proposition 3, we will use:

Lemma 2 *For a given combination* $\{\beta, \gamma < 0, \mathbf{x}, \epsilon\}$, $\mathbf{y}$ *and* $\tilde{\mathbf{y}}$ *are both equilibria of (4), only if* $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \tilde{y}_i$.

[15] When N is odd, there has to be one equilibrium that differs in at least three elements when compared to any of the other equilibria. Without loss of generality we assume the last three elements of $\mathbf{y}$ to be the three elements that move together.

The proof of lemma 2 uses:

Lemma 3 *If for a given combination* $\{\beta, \gamma < 0, \mathbf{x}, \epsilon\}$ *there exists an equilibrium* $\mathbf{y}$ *with* $y_{[j]} = -1$ *and* $y_{[j+1]} = 1$*, then there also exists an equilibrium* $\tilde{\mathbf{y}}$ *with* $\tilde{y}_{[j]} = 1$ *and* $\tilde{y}_{[j+1]} = -1$ *and* $\tilde{y}_{[i]} = y_{[i]}$ *for* $i \neq j, j+1$.

Proof of Lemma 3

From the fact that $\mathbf{y}$ is an equilibrium with $y_{[j]} = -1$ and $y_{[j+1]} = 1$, it follows that

$$\begin{aligned} y^*_{[j]} &= z_{[j]} + \gamma \frac{k+1}{N-1} \leq 0 \\ y^*_{[j+1]} &= z_{[j+1]} + \gamma \frac{k-1}{N-1} > 0. \end{aligned}$$

However, since $\gamma < 0$, we have

$$\begin{aligned} \tilde{y}^*_{[j]} &= z_{[j]} + \gamma \frac{k-1}{N-1} \geq z_{[j+1]} + \gamma \frac{k-1}{N-1} > 0 \\ \tilde{y}^*_{[j+1]} &= z_{[j+1]} + \gamma \frac{k+1}{N-1} \leq z_{[j]} + \gamma \frac{k+1}{N-1} \leq 0. \end{aligned}$$

It then follows that $\tilde{\mathbf{y}}$ with $\tilde{y}_{[i]} = y_{[i]}$ for $i \neq j, j+1$ and $\tilde{y}_{[j]} = 1$ and $\tilde{y}_{[j+1]} = -1$ is also an equilibrium. □

Having proved lemma 3 we can now prove lemma 2.

Proof of Lemma 2

Suppose that $\mathbf{y}$ with $\sum_{i=1}^N y_i = k$ and $\tilde{\mathbf{y}}$ with $\sum_{i=1}^N \tilde{y}_i = \tilde{k}$ and $\tilde{k} \neq k$ are both equilibria of (4), given a combination $\{\beta, \gamma < 0, \mathbf{x}, \epsilon\}$. From lemma 3 it follows that this is true only if $\mathbf{y^k} = (1_1, 1_2, \ldots, 1_{\frac{N+k}{2}}, -1_{\frac{N+k+2}{2}}, \ldots, -1_N)$ and $\mathbf{y^{\tilde{k}}} = (1_1, 1_2, \ldots, 1_{\frac{N+\tilde{k}}{2}}, -1_{\frac{N+\tilde{k}+2}{2}}, \ldots, -1_N)$ are both equilibria given $\{\beta, \gamma < 0, \mathbf{x}, \epsilon\}$. Assume without loss of generality that $\tilde{k} > k$, that is: $\tilde{k} - k \geq 2$. Let ν be the first subject whose choice is -1 in equilibrium $\mathbf{y^k}$ and $+1$ in equilibrium $\mathbf{y^{\tilde{k}}}$. Then, for this subject

$$z_{[\nu]} + \gamma \frac{k+1}{N-1} \leq 0 \text{ and } z_{[\nu]} + \gamma \frac{\tilde{k}-1}{N-1} > 0.$$

But also

$$z_{[\nu]}+\gamma\frac{\tilde{k}-1}{N-1}=z_{[\nu]}+\gamma\frac{\tilde{k}-k+k+1-2}{N-1}=\underbrace{z_{[\nu]}+\gamma\frac{k+1}{N-1}}_{\leq 0}+\underbrace{\gamma\frac{(\tilde{k}-k)-2}{N-1}}_{\leq 0}\leq 0,$$

and the contradiction follows. □

The message of lemma 2 is that for a given value of $\gamma < 0$, two different equilibria $\mathbf{y}$ and $\tilde{\mathbf{y}}$ can co-exist only if $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \tilde{y}_i$. That is, both equilibria must have the same number of subjects with outcome +1 and with outcome -1.

Repeated application of lemma 3 shows that a strategy profile $\mathbf{y}$ with $\sum_{i=1}^{N} y_i = k$ can only be an equilibrium if the ordered (with respect to the z_i's) strategy profile $\mathbf{y} = (1_1, 1_2, \ldots, 1_k, -1_{k+1}, \ldots, -1_N)'$ is an equilibrium. This result will prove to be useful later on in deriving upper bounds for the number of equilibria that may be sustained for a given value of γ.

To complete the proof of Proposition 3, note that the first part of lemma 2 implies that the maximum number of possible equilibria subject to the condition $\sum_{i=1}^{N} y_i = k$ is obtained when k is chosen to equal 0 (+1 or -1) when N is even (odd). In that case, there are $N/2$ ($(N+1)/2$ or $(N-1)/2$) agents choosing $+1$ and the others choosing -1, giving the upper bounds on the number of possible equilibria as given by $d(N)$ in proposition 3.

What is left to show is that there exists a combination of $\{\beta, \gamma < 0, \mathbf{x}, \epsilon\}$ for which the maximum number of equilibria is obtained. From lemma 2 we know that, given a combination of $\{\beta, \gamma < 0, \mathbf{x}, \epsilon\}$, every element in the equilibrium set must have the same number of agents choosing $y = 1$. For N is even, the set can thus only have $d^e(N)$ elements when the set contains *all* strategy profiles for which the number of agents choosing $y = 1$ equals the number of agents choosing $y = -1$. For each of these profiles to be an equilibrium, it must be optimal for *each* agent i to choose $y_i = 1$ given that $\sum_{j \neq i} y_j = -1$ and to choose $y_i = -1$ given that $\sum_{j \neq i} y_j = 1$. In particular,

it must hold that for each element the number of agents splits

$$\begin{aligned} z_{[1]} + \gamma \frac{1}{N-1} &\leq 0 \text{ and;} \\ z_{[N]} + \gamma \frac{-1}{N-1} &> 0. \end{aligned}$$

For γ negative enough, this condition is satisfied irrespective of the values of $z_{[1]}, \ldots z_{[N]}$.

For N is odd, the equilibrium set can only contain $d^o(N)$ elements when the set contains all strategy profiles for which $\sum_{i=1}^{N} y_i = 1$ or all strategy profiles for which $\sum_{i=1}^{N} y_i = -1$. The necessary and sufficient conditions for each of the profiles for which $\sum_{i=1}^{N} y_i = 1$ to be an equilibrium, are

$$\text{(A.1)} \quad z_{[1]} + \gamma \frac{2}{N-1} \leq 0 \qquad \text{and} \qquad z_{[N]} > 0,$$

and the corresponding conditions for the strategy profiles with $\sum_{i=1}^{N} y_i = -1$ are

$$\text{(A.2)} \quad z_{[1]} \leq 0 \qquad \text{and} \quad z_{[N]} + \gamma \frac{-2}{N-1} > 0.$$

From these conditions it follows that the equilibrium set with $d^o(N)$ elements for which $\sum_{i=1}^{N} y_i = 1$ ($\sum_{i=1}^{N} y_i = -1$) is only obtainable when all z values are positive (non-positive). Together this proves proposition 3. □

Lemma 2 and the observation that for the equilibria in the proof of proposition 1 $|k| = |\sum_{i=1}^{N} y_i|$ monotonically decreases as $\gamma \to -\infty$, together lead to the following corollary[16] that for all equilibria, $|k|$ decreases monotonically to 0 (1) as $\gamma \to -\infty$, given N even (odd). This result is consonant with intuition: variation in behavior increases when the utility derived from being different increases.

[16]The corresponding result for positive interactions is that $|\sum_{i=1}^{N} y_i| \nearrow N$ as $\gamma \to \infty$. That is, in the limit all agents conform to $y = 1$ or to $y = -1$ regardless their private utility such that variation in behavior is minimized.

Corollary 3 *For the equilibria* $\mathbf{y}$ *of the discrete choice interaction model given by (4),*

$$|\sum_{i=1}^{N} y_i| \searrow 0 \quad \textit{as } \gamma \to -\infty \textit{ and } N \textit{ is even,}$$

$$|\sum_{i=1}^{N} y_i| \searrow 1 \quad \textit{as } \gamma \to -\infty \textit{ and } N \textit{ is odd.}$$

Proof of Corollary 1

Define $\forall i,\ z_i^G \equiv \beta' \mathbf{x_i} + \frac{\gamma_{GB} \sum_{j=1}^{N} y_j^B}{N-1} + \epsilon_i$ if i is a girl and $z_i^B \equiv \beta' \mathbf{x_i} + \frac{\gamma_{BG} \sum_{j=1}^{N} y_j^G}{N-1} + \epsilon_i$ if i is a boy. Denote the ordered values of z_i^G (z_i^B) as $z_{[i]}^G$ ($z_{[i]}^B$) such that $z_{[1]}^G \geq z_{[2]}^G \geq \ldots \geq z_{[N_G]}$ ($z_{[1]}^B \geq z_{[2]}^B \geq \ldots \geq z_{[N_B]}$), with N_G (N_B) denoting the total number of girls (boys) in the sample.

The line of reasoning used in the proof of proposition 1 now can be applied to the subset of girls (boys), with $z_{[i]}$ replaced by $z_{[i]}^G$ ($z_{[i]}^B$) and γ replaced by γ_{GG} (γ_{BB}). □

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.

Arcidiano, Peter and Sean Nicholson (2002), *Peer Effects in Medical School*, NBER Working Paper #9025

Aronsson, T., Blomquist, S., and H. Sacklen (1999), "Identifying Interdependent Behavior 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 Technlogy.

Brock, W.A. and S.N. Durlauf (2001a), "Discrete Choice with Social Interactions", *Review of Economic Studies*, 68, 235-260.

Brock, W.A. and S.N. Durlauf (2001b), "Interactions-Based Models", in J.J. Heckman and E. Leamer (eds.): *Handbook of Econometrics*, vol. 5, North-Holland.

Brock, W.A. and S.N. Durlauf (2003), *Multinomial Choice with Social Interactions*, SSRI Working Paper # 2003-1.

Duflo, Esther and Emmanuel Saez (2002), "Participation and investment decisions in a retirement plan: the influence of colleagues' choices", *Journal of Public Economics*, 85, 121-148.

Duflo, Esther and Emmanuel Saez (2003), "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment", *Quarterly Journal of Economics*, 118, 815-842.

Duesenberry, J.S. (1949), *Income, Saving, and the Theory of Consumer Behavior*, Harvard University Press.

Durlauf, S.N. and R.A. Moffitt (2003), "Introduction to the Special Issue: Empirical Analysis of Social Interactions", *Journal of Applied Econometrics*, vol. 18, 499. 815-842.

Gaviria, A. and S. Raphael (2001), "School-Based Peer Effects and Juvenile Behavior", *Review of Economics and Statistics*, 83, 257-268.

Glaeser, E.L., B. Sacerdote, and J.A. Scheinkman (1996), "Crime and Social Interactions", *Quarterly Journal of Economics*, 111, 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, 85-90.

Gruber, J. and J. Zinman (2001), "Youth Smoking in the US: Evidence and Implications", in J. Gruber (ed.), *Risky Behavior among Youth: an Economic Analysis*, University of Chicago Press, 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.

Kawaguchi, D. (2004), "Peer Effects on Substance Use among American Teenagers", *Journal of Population Economics*, forthcoming.

Kooreman, P. (1994), "Estimation of Econometric Models of Some Discrete Games", *Journal of Applied Econometrics*, 9, 255-268.

Kooreman, P. (2003), *Time, Money, Peers, and Parents; Some Data and Theories on Teenage Behavior*, Working Paper, University of Groningen.

Krauth, B. (2001), *Social Interactions in Small Groups*, Working Paper Simon Fraser University.

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*, Cambridge University Press.

Manski, C.F. (1993), "Identification of Endogeneous Social Effects: The Reflection Problem", *The Review of Economic Studies*, 60, 531-542.

Manski, C.F. (2000), "Economic Analysis of Social Interactions", *Journal of Economic Perspectives*, 14, 115-136.

Milgrom, P. and J. Roberts (1990), "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities", *Econometrica*, 58, 1255-1277.

Pollak, R.A. (1976), "Interdependent Preferences", *American Economic Review*, 66, 309-321.

Sacerdote, B. (2001), "Peer Effects with Random Assignment: Result for Dartmouth Roommates", *Quarterly Journal of Economics*, 116, pp. 681-703.Smetters, K. and J. Gravelle (2001),

Soetevent, A.R. (2004), *Social Interactions and Economic Outcomes*, Ph.D. thesis, University of Groningen.

Stern, S. (1992), "A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models", *Econometrica*, 60, 943-952.

Tamer, E. (2002), *Empirical Strategies for Estimating Discrete Games with Multiple Equilibria*, Working Paper, Princeton University.

Tamer, E. (2003), "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria", *Review of Economic Studies*, vol. 70, 147-167.

Topkis, D.M. (1998), *Supermodularity and Complementarity*, Frontiers of Economic Research, Princeton University Press, New Jersey.

Veblen, T. (1899), *The Theory of the Leisure Class*, MacMillan, New York.

Vives, X. (1999), *Oligopoly Pricing; Old Ideas and New Tools*, MIT Press Cambridge, Massachusetts.

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.

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

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

Table 3: Estimation results (t-values in parentheses)

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

Table 4: Estimation results; with school specific fixed effects (t-values in parentheses)

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