DISCUSSION PAPER SERIES

No. 9752

THE ALLOCATION OF TIME IN SLEEP: A SOCIAL NETWORK MODEL WITH SAMPLED DATA

Xiaodong Liu, Eleonora Patacchini and Edoardo Rainone

INTERNATIONAL TRADE AND REGIONAL ECONOMICS, LABOUR ECONOMICS and PUBLIC POLICY

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP9752.php**

ISSN 0265-8003

THE ALLOCATION OF TIME IN SLEEP: A SOCIAL NETWORK MODEL WITH SAMPLED DATA

Xiaodong Liu, University of Colorado, Boulder
Eleonora Patacchini, Syracuse Unviersity, EIEF and CEPR
Edoardo Rainone, Banca d'Italia and Sapienza University of Rome

Discussion Paper No. 9752
November 2013

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **INTERNATIONAL TRADE AND REGIONAL ECONOMICS, LABOUR ECONOMICS and PUBLIC POLICY**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

CEPR Discussion Paper No. 9752

November 2013

ABSTRACT

The Allocation of Time in Sleep: a Social Network Model with Sampled Data*

We analyze peer effects in sleeping behavior using a representative sample of U.S. teenagers from the National Longitudinal Survey of Adolescent Health. The sampling design of the survey causes the conventional 2SLS estimator to be inconsistent. We extend the NLS estimator in Wang and Lee (2013a) to estimate network models with sampled observations on the dependent variable. When accounting for sampling, we find that the sleeping behaviour of the friends is important to shape own sleeping behaviour, besides the impact of individual, family and friend characteristics.

JEL Classification: C13, C21, I15, I19 and J22
Keywords: health, missing data, nonlinear least squares, siblings, social interactions, spatial autoregressive model and time use

Xiaodong Liu
Department of Economics
University of Colorado at Boulder
256 UCB
Boulder, Colorado 80309-0256
USA

Email: xiaodong.liu@colorado.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=173030

Eleonora Patacchini
Syracuse University
426 Eggers Hall
Syracuse, NY 13244-1020
USA

Email: epatacch@maxwell.syr.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=158584

Edoardo Rainone
La Sapienza University of Rome
P.le Aldo Moro 5
00185 Roma
ITALY

Email: edoardo.rainone@uniroma1.it

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=173120

*This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due to Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data les is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis. The views expressed here do not necessarily reflect those of Banca d'Italia.

Submitted 14 November 2013

"Sleep that knits up the ravelled sleave of care, The death of each day's life, sore labour's bath, Balm of hurt minds, great Nature's second course, Chief nourisher in life's feast."

Shakespeare, Macbeth

1 Introduction

Nearly a third of a person's life is spent in slumber. In the U.S. those with insomnia spend about $1 billion a year on prescription sleep aids, and another $1 billion on over-the-counter sleep medications (Yaniv, 2004). The economic costs, both direct (expenditure within the health system) and indirect (absenteeism, low productivity, and work-related injuries) of sleep disorders in the U.S. in 2004 was estimated to be $109 billion (Hillman et al., 2006).

Yet, sleeping behaviour has received relatively little attention in economics. While sleep is primarily a function of the body's internal biological clock (circadian rhythm), individual choice also plays an important role in determining the timing and duration of sleep. Biddle and Hamermesh (1990) posit a simple economic model that accounts for the endogenous nature of sleep choice, but empirical work on the subject has been very limited.

In particular, there is virtually no evidence on the importance of social interactions in shaping sleeping behaviour. In many circumstances, the decision of agents to exert effort in some activity cannot adequately be explained by their personal characteristics and the intrinsic utility derived from the activity. Rather, its rationale may be found in how peers and others value this activity. There is indeed strong evidence that the behaviour of individual agents is affected by that of their peers.[1] The individual utility when allocating time in work or leisure may depend on the same choice made by peers. As a consequence, social interactions might be important for understanding the duration of sleep, which is the residual activity.[2]

In this paper, we exploit the unique information contained in the National Longitudinal Survey of Adolescent Health (AddHealth) to provide evidence on sleeping patterns among

[1] The integration of models of social interactions within economic theory is an active and interesting area of research. See the recent *Handbook of Social Economics*, Benhabib et al. (2011).

[2] Biddle and Hamermesh (1990) study the demand for sleep in this perspective without social incentives.

adolescents in the U.S. Sleeping behaviour during teenage years is of particular interest because of its effect on human capital formation. Research suggests that lack of sleep reduces attendance, increases tardiness, and lowers grades of adolescent students (Eide and Showalter, 2012). Furthermore, lack of sleep in youth is correlated with health and behavioral problems such as moodiness, depression, difficulty controlling behaviour, and increased frustration - all of which make learning in school difficult (National Sleep Foundation; (Mitru et al., 2002)).

The AddHealth data contain unique information on friendship relationships among a representative sample of students from U.S. high school teenagers together with basic information on individual, family, neighborhood and school characteristics (in-school survey). The survey design also includes a questionnaire administered to a random sample of those students collecting information on more sensitive topics (health issues, crime, drug, sexual behaviour, etc.), including time and duration of sleep on week days during the school year (in-home survey). The use of this additional information, however, comes at a cost. The in-home sampling scheme may result in missing observations on the behaviour of friends who were not sampled, and induce measurement error to the endogenous peer effect variable given by the average behaviour of friends. As a result, the existing estimation methods for network models of social interactions (see, e.g., Bramoullé et al., 2009; Lee et al., 2010) are not generally valid.[3]

Recently, social network studies have drawn a great deal of attention. Network models are widely used to represent relational information among interacting units and the implications of these relations. Most inference for social network models assumes that the all possible links are observed and that all the relevant information is available. This is clearly not true in practice, as much network data is collected though sample surveys. In a recent paper, Sojourner (2013) considers a linear-in-means social interaction model with missing observations on covariates. He shows that random assignment of agents to peer groups can

[3]This issue is typically neglected in most empirical papers using the information on friends together with the in-home survey in the AddHealth data set.

help to overcome the missing data problem. On the other hand, Chandrasekhar and Lewis (2011) consider the estimation of network models with sampled observations on network links. They propose a set of analytical corrections for commonly used network statistics and a two-step estimation procedure using graphical reconstruction. Our case is different. We observe all the network links and the covariates for all nodes, but we have sampled observations on the dependant variables.

The social network model considered in this paper has the specification of a spatial autoregressive (SAR) model with group-specific fixed effect. Kelejian and Prucha (2010) consider the estimation of the SAR model with missing observations on the dependent variable and covariates. They suggest two-stage least squares (2SLS) estimators that are based on a subset of the sample so that the dependent variable and covariates are observed, and the spatial lags are either completely observed or partially observed with an asymptotically negligible measurement error. Our set up is similar to the one proposed by Wang and Lee. Wang and Lee consider the estimation of the SAR model with missing observations on the dependent variable for cross-sectional data (Wang and Lee, 2013a) and for random effect panel data (Wang and Lee, 2013b). They propose the generalized method of moments (GMM) estimator, the nonlinear least squares (NLS) estimator, and the 2SLS estimator with imputation. They show that the three estimators are consistent and robust against unknown heteroskedasticity. In this paper, we extend the NLS estimator in Wang and Lee (2013a) to estimate social network models with sampled observations on the dependent variable.

Our results show that the conventional 2SLS is inconsistent without accounting for sampling. In our case, 2SLS fails to detect the presence of peer effects. When sampling is taken into account, we instead find that the sleeping behaviour of the friends is important in shaping own sleeping behaviour, besides the impact of individual and friends characteristics. We use the approach recently proposed by Goldsmith-Pinkham and Imbens (2013) to investigate testable implications of network endogeneity, finding no sign of troubling individual level unobservables that may invalidate our results. Our results are also robust when using

an unique information on siblings to eliminate possible unobserved family factors.

We start our analysis by describing our data in Section 2. Section 3 presents the network model, together with the identification and estimation strategy. We discuss our estimation results in Section 4, whereas Section 5 contains some robustness checks. Section 6 concludes.

2 Data and Descriptive Evidence

Our data source is the AddHealth data that has been designed to study the impact of the social environment (i.e. friends, family, neighborhood and school) on adolescents' behaviour in the United States by collecting data on students in grades 7-12 from a nationally representative sample of roughly 130 private and public schools in years 1994-95. Every student attending the sampled schools on the interview day is asked to compile a questionnaire (in-school survey) containing questions on respondents' demographic and behavioral characteristics, education, family background and friendship. Most notably, students were asked to identify their best friends from a school roster - up to five males and five females. The limit in the number of nominations, however, is not binding (not even by gender),[4] and in the large majority of cases (more than 90%) the nominated best friends are in the same school. Hence, it is possible to reconstruct the entire geometry of the friendship networks within each school. In addition, by matching the identification numbers of the friendship nominations to respondents' identification numbers, one can obtain information on the characteristics of nominated friends. This sample contains information on roughly 90,000 students. These features make these data almost unique. It is extremely rare to have information on the universe of network contacts (here school friends), together with their detailed characteristics.[5] The survey design also includes a longer questionnaire (in-home survey) containing questions related to more sensitive individual and household information which is adminis-

[4] Less than 1 percent of the students in our sample show a list of ten best friends, less than 3 percent a list of five males and roughly 4 percent name five females. On average, they declare to have 4.35 friends with a small dispersion around this mean value (standard deviation equal to 1.41).

[5] The information on social network contacts collected in other existing surveys is about "ego-networks", i.e. the respondent is asked to name few personal contacts and provides (self-reported) information about an extremely limited number of their charactersitics.

tered to a subset of adolescents. We use the *core sample* of in-home survey which provides information on a random and self-weighting subset of adolescents, about 12,000 individuals.[6] The in-home questionnaire contains detailed information about the timing and duration of sleep. The questions has been slightly reformulated over time to measure sleeping patterns more precisely. Indeed, the (in-home survey) students are interviewed again one year later, in 1995–96 (wave II).[7] We derive the information on sleeping patterns by using the wave II question: During the school year, what time do you usually go to bed on week nights?[8,9]

Figure 1 plots the empirical distribution. The graph shows a notable dispersion around the mean "bed time" value (mean equal to 10:37pm and standard deviation equal to 58.7 minutes). About 50% of the students go to bed between 10pm and 11.30pm.

Figure 2 shows the distribution of students by GPA distinguishing between students with different sleeping patterns. It appears that students with sleep deficit (red curve) show a statistically significant lower performance at school.[10] In other words, a student that goes to bed earlier is more likely to have a higher GPA.

Table 1 and Figure 3 collects some further evidence on the relationship between sleeping patterns and other relevant characteristics. We run a principal component analysis (PCA)[11] on body mass index (BMI), GPA, general health, use of alcohol and cigarette smoking. The first principal component explains over one third of the total inertia. Table 1 shows that this variation is associated to differences between two clusters of students, one with high body

[6] The *core sample* contains roughly the 60% of the individuals interviewed in the in-home survey (which are about 20,000 individuals). The difference is due to the fact that in the in-home sampling design some types of individuals are oversampled.

[7] Those subject are also interviewed again in 2001-02 (wave III), and again in 2007-08 (wave IV). For the purposes of this paper, we do not use this longitudinal information. The friendship nominations are only collected when the students were at school (i.e. in waves I and II).

[8] The questions formulated in wave I do not differentiate between the school period and summer time.

[9] We rescaled each hour in 100 units, so for instance half an hour is transformed to a distance of 50. We dropped individuals declaring going to sleep before 5pm and after 6am.

[10] The rejection of the null hypothesis in a Kolmogorov-Smirnov test confirms the difference between these two distributions.

[11] PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables (called principal components). This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for the largest portion of variability in the data).

mass index, poor school performance, poor general health, drinking alcohol and smoking cigarettes (type A students), and the other with the opposite profile (type B students). In other words, splitting the population between type A and type B individuals maximizes the between-group variation and minimizes the within-group variation. Figure 3 shows that type A students tend to sleep for fewer hours than type B students. This is in line with an (expected) relationship between sleeping behaviour and individual socio-economic profile (Eide and Showalter, 2012).

3 Regression Analysis

Our aim is to assess the actual empirical relationship between the individual sleeping behaviour and the sleeping behaviour of the peers using the unique information provided by the AddHealth data. This exercise requires facing the traditional challenges in identifying endogenous social interaction effects, while also overcoming a further (and so far neglected) issue stemming from the sampling design of the AddHealth survey. We present the network model in Section 3.1, whereas the estimation of network models with sampling on the dependant variable is considered in detail in Section 3.2.

3.1 The network model

Consider a population of n individuals partitioned into $\bar{r}$ networks. For the n_r individuals in the rth network, their connections with each other are represented by an $n_r \times n_r$ adjacency matrix $G_r^* = [g_{ij,r}^*]$ where $g_{ij,r}^* = 1$ if individuals i and j are friends and $g_{ij,r}^* = 0$ otherwise.[12] Let $G_r = [g_{ij,r}]$ be the row-normalized G_r^* such that $g_{ij,r} = g_{ij,r}^* / \sum_{k=1}^{n_r} g_{ik,r}^*$.

Given the network adjacency matrix G_r, we assume $y_{i,r}$, the sleeping behaviour of individual i in network r, is given by the following network model

$$y_{i,r} = \phi \sum_{j=1}^{n_r} g_{ij,r} y_{j,r} + \sum_{k=1}^{p} x_{ik,r}\beta_k + \sum_{k=1}^{p} (\sum_{j=1}^{n_r} g_{ij,r} x_{jk,r} \gamma_k) + \eta_r + \epsilon_{i,r}. \quad (1)$$

[12] For ease of presentation, we focus on the case where the connections are undirected and no agent is isolated so that G_r^* is symmetric and $\sum_{j=1}^{n} g_{ij,r}^* \neq 0$ for all i. The result of the paper holds for a directed network with an asymmetric G_r^*.

In this model, $\sum_{j=1}^{n_r} g_{ij,r} y_{j,r}$ is the average sleeping behaviour of i's direct friends with its coefficient ϕ representing *the endogenous effect*, wherein an individual's choice/outcome may depend on those of his/her friends about the same activity. $x_{ik,r}$, for $k = 1, \cdots, p$, are exogenous control variables. For $k = 1, \cdots, p$, $\sum_{j=1}^{n_r} g_{ij,r} x_{jk,r}$ is the average value of the k-th control variable taking over i's direct friends with its coefficient γ_k representing *the contextual effect*, wherein an individual's choice/outcome may depend on the exogenous characteristics of his/her friends. η_r is a network-specific parameter representing *the correlated effect*, wherein individuals in the same group tend to behave similarly because they face a common environment. $\epsilon_{i,r}$ is an i.i.d. error term with zero mean and finite variance σ^2.

Let $x_{i,r} = (x_{i1,r}, \cdots, x_{ip,r})'$, $\beta = (\beta_1, \cdots, \beta_p)'$ and $\gamma = (\gamma_1, \cdots, \gamma_p)'$. In matrix form, (1) can be rewritten as

$$Y_r = \phi G_r Y_r + X_r \beta + G_r X_r \gamma + \eta_r l_{n_r} + \epsilon_r, \tag{2}$$

where $Y_r = (y_{1,r}, \cdots, y_{n_r,r})'$, $X_r = (x_{1,r}, \cdots, x_{n_r,r})'$, $\epsilon_r = (\epsilon_{1,r}, \cdots, \epsilon_{n_r,r})'$, and l_{n_r} is an $n_r \times 1$ vector of ones.

Let $\text{diag}\{A_j\}_{j=1}^m$ denote a generalized diagonal block matrix with the diagonal blocks being A_j's, where A_j may or may not be a square matrix. Then, for all $\bar{r}$ networks, we can stack the data such that (3) becomes

$$Y = \phi G Y + X\beta + GX\gamma + L\eta + \epsilon, \tag{3}$$

where $Y = (Y_1', \cdots, Y_{\bar{r}}')'$, $G = \text{diag}\{G_r\}_{r=1}^{\bar{r}}$, $X = (X_1', \cdots, X_{\bar{r}}')'$, $L = \text{diag}\{l_{n_r}\}_{r=1}^{\bar{r}}$, $\eta = (\eta_1, \cdots, \eta_{\bar{r}})'$, and $\epsilon = (\epsilon_1', \cdots, \epsilon_{\bar{r}}')'$.

The identification and estimation of endogenous, contextual, and correlated effects have been the main interests of social network models. The conventional identification and estimation strategy in the literature (see, e.g., Lee, 2007; Bramoullé et al., 2009; Lee et al., 2010)

relies on the assumption that $\mathrm{E}(\epsilon_r|G_r, X_r, \eta_r) = 0$.[13] Based on this assumption, Bramoullé et al. (2009) show that if intransitivities exist in networks so that I_n, G, G^2, G^3, are linearly independent, then model (2) is identified. For estimation, we first eliminate the incidental parameters η using a within-transformation projector $J = \mathrm{diag}\{J_r\}_{r=1}^{\bar{r}}$, where $J_r = I_{n_r} - \frac{1}{n_r} l_{n_r} l'_{n_r}$. As $JL = 0$, premultiplying (3) by J, we have

$$JY = \phi JGY + JX\beta + JGX\gamma + J\epsilon.$$

Let $Z = (GY, X, GX)$ and $\theta = (\phi, \beta', \gamma')'$. For the instrumental variable (IV) matrix $Q = (X, GX, G^2X)$, the two-stage least squares estimator is given by

$$\hat{\theta}_{2sls} = (\hat{Z}'JZ)^{-1}\hat{Z}'JY, \tag{4}$$

where $J\hat{Z} = JQ(Q'JQ)^{-1}Q'JZ$ is the predicted JZ from the first-stage regression.

In the following section, we focus on the sampling issue of the network model that has been largely ignored by the literature.

3.2 Estimation of peer effects with sampling

In our and many other studies, the analysis of the network model (1) has been made possible by the use of a unique database on friendship networks from the AddHealth data.[14] As we explain in Section 2, students are asked to identify their best friends from the school roster in the in-school survey. Thus, we can observe all friendship links in the networks. However, as some more sensitive individual information - (i.e. sleeping behaviour) - is in the in-home survey, we only have this information for the sampled students.[15]

Without loss of generality, suppose the first m_r ($m_r > 1$) individuals in network r are sampled. Suppose we can observe network connections $G_r = [g_{ij,r}]$ and controls $x_{i,r}$ for all individuals in network r, but we can only observe $y_{i,r}$'s of sampled individuals. For the

[13] We will investigate the validity of this assumption for this empirical study in Section 5.

[14] See, e.g. Lin (2010), Patacchini and Zenou (2008) and the references herein.

[15] The use of the *core sample* is crucial because otherwise the sampled students are not random.

sampled individuals, $i = 1, \cdots, m_r$, (1) becomes

$$y_{i,r} = \phi \sum_{j=1}^{m_r} g_{ij,r} y_{j,r} + x'_{i,r}\beta + \sum_{j=1}^{n_r} g_{ij,r} x'_{j,r}\gamma + \eta_r + \epsilon^*_{i,r}. \tag{5}$$

By comparing (1) and (5), we have $\epsilon^*_{i,r} = \phi \sum_{j=m_r+1}^{n_r} g_{ij,r} y_{j,r} + \epsilon_{i,r}$. Therefore, the error term of model (5) contains two types of errors - the error due to unobserved individual heterogeneity $\epsilon_{i,r}$ and the measurement error due to the sampling design $\phi \sum_{j=m_r+1}^{n_r} g_{ij,r} y_{j,r}$. The measurement error could be correlated with the control variables and, as a result, the 2SLS given by (4) may not be consistent.

To further illustrate this point, we rewrite (5) in matrix form. Let

$$G_r = \begin{bmatrix} G_r^S \\ G_r^N \end{bmatrix} = \begin{bmatrix} G_r^{SS} & G_r^{SN} \\ G_r^{NS} & G_r^{NN} \end{bmatrix},$$

where G_r^S is an $m_r \times n_r$ matrix of the first m_r rows of G_r and G_r^{SS} is an $m_r \times m_r$ matrix of the first m_r columns of G_r^S. Then, for the sampled individuals, we have

$$Y_r^S = \phi G_r^{SS} Y_r^S + X_r^S \beta + G_r^S X_r \gamma + \eta_r l_{m_r} + \epsilon_r^*, \tag{6}$$

where $Y_r^S = (y_{1,r}, \cdots, y_{m_r,r})'$ denotes the $m_r \times 1$ vector of observations on the dependent variable of the sampled individuals, $X_r^S = (x_{1,r}, \cdots, x_{m_r,r})'$ denotes the $m_r \times p$ matrix of observations on the control variables of the sampled individuals, and $\epsilon_r^* = \epsilon_r^S + \phi G_r^{SN} Y_r^N$ with $\epsilon_r^S = (\epsilon_{1,r}, \cdots, \epsilon_{m_r,r})'$ and $Y_r^N = (y_{m_r+1,r}, \cdots, y_{n_r,r})'$. As $\mathrm{E}(\epsilon_r | G_r, X_r, \eta_r) = 0$, we have

$$\mathrm{E}(\epsilon_r^* | G_r, X_r, \eta_r) = \mathrm{E}(\epsilon_r^S + \phi G_r^{SN} Y_r^N | G_r, X_r, \eta_r) = \phi G_r^{SN} \mathrm{E}(Y_r^N | G_r, X_r, \eta_r).$$

To obtain $\mathrm{E}(Y_r^N | G_r, X_r, \eta_r)$, we need to inspect the reduced form equation of the model. If

$(I_{n_r} - \phi G_r)$ is nonsingular, the reduced form equation of (2) is given by

$$Y_r = (I_{n_r} - \phi G_r)^{-1}(X_r\beta + G_r X_r\gamma) + \frac{\eta_r}{1-\phi} l_{n_r} + (I_{n_r} - \phi G_r)^{-1}\epsilon_r. \tag{7}$$

Let $D_r^N = [0_{(n_r-m_r)\times m_r}, I_{n_r-m_r}]$ denote an $(n_r - m_r) \times n_r$ matrix of the last $(n_r - m_r)$ rows of an identity matrix. Then, it follows from (7) that

$$\mathrm{E}(Y_r^N|G_r, X_r, \eta_r) = D_r^N \mathrm{E}(Y_r|G_r, X_r, \eta_r) = D_r^N (I_{n_r} - \phi G_r)^{-1}(X_r\beta + G_r X_r\gamma) + \frac{\eta_r}{1-\phi} l_{n_r-m_r}.$$

Therefore,

$$\mathrm{E}(\epsilon_r^*|G_r, X_r, \eta_r) = \phi G_r^{SN} \mathrm{E}(Y_r^N|G_r, X_r, \eta_r) = \phi G_r^{SN} D_r^N (I_{n_r} - \phi G_r)^{-1}(X_r\beta + G_r X_r\gamma) + \frac{\phi\eta_r}{1-\phi} G_r^{SN} l_{n_r-m_r}.$$

As $\mathrm{E}(\epsilon_r^*|G_r, X_r, \eta_r)$ is not zero in general, the 2SLS estimator given by (4) may not be consistent for (6).

To avoid the measurement error due to sampling, we consider the NLS approach suggested by Wang and Lee (2013a) based on the reduced form equation (7). Let $D_r^S = [I_{m_r}, 0_{m_r\times(n_r-m_r)}]$ be an $m_r \times n_r$ matrix of the first m_r rows of an identity matrix. Then,

$$Y_r^S = D_r^S Y_r = D_r^S (I_{n_r} - \phi G_r)^{-1}(X_r\beta + G_r X_r\gamma) + \frac{\eta_r}{1-\phi} l_{m_r} + u_r, \tag{8}$$

where $u_r = D_r^S (I_{n_r} - \phi G_r)^{-1}\epsilon_r$. As $\mathrm{E}(u_r|G_r, X_r, \eta_r) = 0$, a regression estimator based on (8) would be consistent.

First, to eliminate the incidental parameters η_r, we apply a within transformation using the projector $J_r^S = I_{m_r} - \frac{1}{m_r} l_{m_r} l'_{m_r}$ so that (8) becomes

$$J_r^S Y_r^S = J_r^S h_r(\theta) + J_r^S u_r,$$

where $h_r(\theta) = D_r^S (I_{n_r} - \phi G_r)^{-1}(X_r\beta + G_r X_r\gamma)$ with $\theta = (\phi, \beta', \gamma')'$. The NLS estimator of

θ is given by

$$\hat{\theta}_{nls} = \arg\min_{\theta} \sum_{r=1}^{\bar{r}} [Y_r^S - h_r(\theta)]' J_r^S [Y_r^S - h_r(\theta)]. \tag{9}$$

Let $J^S = \text{diag}\{J_r^S\}_{r=1}^{\bar{r}}$ and $D^S = \text{diag}\{D_r^S\}_{r=1}^{\bar{r}}$. Following a similar argument in Wang and Lee (2013a), the NLS estimator $\hat{\theta}_{nls}$ is consistent with an asymptotic distribution

$$\sqrt{n}(\hat{\theta}_{nls} - \theta) \xrightarrow{d} N(0, \Sigma_{nls}),$$

where $\Sigma_{nls} = \lim_{n\to\infty} n(C'B'BC)^{-1}C'B'\Omega BC(C'B'BC)^{-1}$, with $B = J^S D^S (I - \phi G)^{-1}$, $C = [G(I - \phi G)^{-1}(X\beta + GX\gamma), X, GX]$ and $\Omega = \sigma^2 BB'$.[16]

3.3 A simulation experiment

We conduct a Monte Carlo simulation in which we compare the 2SLS estimator which is commonly used for the estimation of peer effects and the NLS estimator given in (9). The setup of our simulations is as follows. The population numerosity is 500 nodes and the number of separated networks is 50, resulting in subnetworks of 10 nodes. Each node is allowed to have three connections as a maximum and zero as a minimum with a uniform distribution within the subnetwork to which it belongs. Links are formed randomly. We consider sampling rates of 40 percent, 60 percent, 80 percent, 100 percent . For each rate and for each estimator, we estimate 5,000 times model (1) using one variable x. The control variable x and the network fixed effect η are randomly generated by a normal distribution $N(0, 1)$. The innovation ϵ is generated by a normal distribution $N(0, \sigma^2)$. We set $\lambda = 0.3$, $\beta = 1.0$, $\gamma = 1.0$, and $\sigma^2 = 2$ in the data generating process.[17] Table 2 reports the results of our Monte Carlo study. The NLS estimates roughly coincide with the true parameter values. The 2SLS estimates are downwards biased, with the magnitude of the bias increasing as the sampling rate decreases. The NLS and 2SLS have similar performance when all individuals

[16] As in Wang and Lee (2013a), we assume the number of sampled individuals is proportional to n so that the convergence rate of the estimator can be written in terms of n.

[17] Conclusions of our simulation study are not sensitive to the parameters values. For the sake of brevity we do not show the output of all simulations.

are sampled (i.e. the sampling rate is zero). We have also repeated our simulations when varying the maximum number of connections (i.e. the network density) and using various distributions (other from uniform). The results are stable across the different specifications.[18]

4 Estimation Results

Having in mind the simulation results, we move to the empirics and follow the same comparative approach among different methods.

Our main estimation results are reported in Table 3. The dependent variable is the time students go to bed. During the school days, this variable captures the time allocated to sleep - the later a student goes to bed, the lower is her/his sleep duration. The different columns show the results with an increasing set of controls. In the first specification, we include individual demographic characteristics, family background characteristics, contextual effects (the average of peers' characteristics) and network fixed effects. We introduce scores in mathematics and history/social science in the second specification, and finally we include a risky behaviour factor in the third specification.[19] The results can be summarized as follows.

First, with the exception of peer effects, point estimates and standard errors are stable across specifications and estimators. The results are in line with the expectations. Biddle and Hamermesh (1990) model the demand for sleep as a function of wage and leisure. In their model, the higher the value of an additional worked hour (i.e. the higher the wage), the lower is the time allocated to sleep. Although we deal with students rather than workers, the general mechanisms still apply. If one interprets the return of school performance as wage, then we expect a negative correlation between student grade and sleep duration because incentives to spend hours in studying increase over the school years. Similarly, if time spent in risky behaviour is seen as leisure time, then an increase in risky activities should negatively impact the amount of time allocated to sleep.

Second, the peer effect estimated coefficient is significantly different from zero for all

[18] We do not report these further results for brevity. They remain available upon request.

[19] The Risky Behavior Factor is the score of a factor analysis run on use of alchool, cigarette smoking and general health. The results are robust to alternativfe sets of controls.

specifications when estimated using the NLS estimator, while it is never significantly different from zero when using 2SLS. In addition, our estimator shows both point estimates and standard errors which are stable across specifications. In terms of magnitude, in the average group of four people, an additional hour of sleep of each of the friends translates to about 45 minutes in the individual sleeping duration.

Note that this empirical evidence is in line with the simulation results, since the downwards bias here leads the 2SLS to suggest that no peer effect is at work, unlike with NLS.

5 Robustness Checks

5.1 Endogenous network formation

An important feature of our identification strategy is the use of network fixed effects. In most cases individuals sort into groups non-randomly. For example, kids whose parents are low educated or worse than average in unmeasured ways would be more likely to sort with low human capital peers. If the variables that drive this process of selection are not fully observable, potential correlations between (unobserved) group-specific factors and the target regressors are major sources of bias. It is thus difficult to disentangle the endogenous peer effects from the correlated effects, i.e. from effects arising from the fact that individuals in the same group tend to behave similarly because they face a common environment. Network fixed effect are a remedy for the selection bias that originates from the possible sorting of individuals with similar unobserved characteristics into a network. The underlying assumption is that such unobserved characteristics are common to the individuals within each network. This is reasonable in our case study where the networks are quite small (see Section 2). However, if there are student-level unobservables that drive both network formation and outcome choice, then this strategy fails.

Recently, Goldsmith-Pinkham and Imbens (2013) highlight the fact that endogeneity of this sort can be tested. Signals of individual-level correlated unobservables would motivate the use of parametric modeling assumptions and Bayesian inferential methods to integrate a

network formation with the study of behaviour over the formed networks. We present below the results which are obtained by applying the approach proposed by Goldsmith-Pinkham and Imbens (2013) in our case.

Model (6) can be written as follows:

$$Y_r = \phi G_r Y_r + X_r\beta + G_r X_r \gamma + \eta_r l_{n_r} + \underbrace{\zeta v_r + e_r}_{\epsilon_r}, \tag{10}$$

where $v_r = (v_{1,r}, \cdots, v_{n_r,r})'$ denotes a vector of unobserved characteristics at the individual level and $e_r = (e_{1,r}, \cdots, e_{n_r,r})'$ is a vector of random disturbances.

Let us consider a network formation model where the variables that explain $g_{ij,r}$ are distances in terms of observed and unobserved characteristics between students i and j:

$$g_{ij,r} = \alpha + \sum_{m=1}^{M} \delta_m |x_{i,r}^m - x_{j,r}^m| + \theta |v_{i,r} - v_{j,r}| + \eta_r + u_{ij,r}. \tag{11}$$

Homophily in the unobserved characteristics implies that $\theta^l < 0$, i.e. that the closer two individuals are in terms of unobservables, the higher is the probability that they are friends. If ζ is different from zero, then these unobservables have a direct effect on outcome as well.

A testable implication of the presence of this problem would be to find in the data a positive and statistical significant correlation between the predicted probability to observe a link between i and j, $q_{ij} = \hat{g}_{ij}$, and the difference between residuals of i and j in the outcome equation (6), $|\hat{\epsilon}_{i,r} - \hat{\epsilon}_{j,r}|$, when $g_{ij} = 1$.

The intuition is as follows. If we observe in the data that two students are friends, i.e. $g_{ij} = 1$, and a low value of q_{ij}, then it means that we are not explaining network formation with the observed characteristics. As a result, we should find low values of q_{ij} associated with low values of $|\hat{\epsilon}_{i,r} - \hat{\epsilon}_{j,r}|$, i.e. friendship between i and j is explained by similarity in unobserved rather than observed characteristics. A similar argument can be applied for nonfriend pairs, $g_{ij} = 0$.

Table 4 contains our evidence, which is obtained when performing a logit estimation of model (11). The upper panel reports the results when $g_{ij} = 1$. We use the empirical distributions of the predicted probabilities q_{ij}^l to measure *low values* of q_{ij}. We choose three different thresholds. Specifically, we define low values of q_{ij}^l those below the 25% or 35% or 45% percentile. The results can be summarized as follows.

(i) First, we fail to predict the existence of a link in less than 4% of the cases.

(ii) Second, in those cases, we find no sign of correlation of the sort discussed above.

(iii) Those results are robust when moving through the different thresholds.

In order to get more confidence in our exercise, we perform the following experiment. We deliberately leave out one individual characteristic, which will then act as unobserved factor (to the econometrician). We exclude grade, which is relevant both in the link formation process and in determining "bed time". If our exercise detects this problem, then we should obtain a correlation between q_{ij} and $|\hat{\epsilon}_{i,r} - \hat{\epsilon}_{j,r}|$ positive and significantly different from zero. The last columns of Table 4 report the results. One can see that the correlation is now constantly different from zero, irrespective of the threshold used. The lower panel of Table 4 shows the results when $g_{ij} = 0$. The evidence is similar.

As a result, conditional on the (unusually) large set of individual characteristics provided by the AddHealth, peer characteristics and network effects, we find no evidence of network endogeneity.

5.2 Siblings

Let us conclude our analysis with a further robustness check.

The restricted-use version of the AddHealth dataset contains sibling pairs data. For each respondent, we know who is the sibling, her/his characteristics, the nominated friends and her/her friends' characteristics. We exploit this unique source of information to test whether peer effects are still significantly different from zero if we introduce sibling fixed effects. If our peer effect estimate is simply picking up unobserved individual characteristics, then we should find no effect when washing away the influence of factors that are common for siblings

who grew up in the same family and consequently have been educated by the same persons, lived in the same neighborhood and more generally faced a wide number of common shocks.

Almost all our sample of siblings (about 97%) are in the same social network, i.e. are indirectly connected through a chain of friends. However, they have different direct friends. So this is the source of variation which is exploited in our sibling fixed effect strategy.

Table 5 shows the estimation results. The coefficient estimates are reduced in magnitude and the parameters are less precisely estimated due to the reduced sample size. However, the substance of the results remain unchanged: the peer effect estimate remains significantly different from zero when using the NLS estimator in all specifications.

6 Conclusions

There is remarkably little evidence on the determinant of individual differences in sleep duration. By implementing sound econometric techniques, our study is able to provide novel evidence in this respect. We have two contributions to the literature. One, we extend the NLS estimator in Wang and Lee (2013a) to estimate social network models with sampled observations on the dependent variable. Two, we analyze peer effects in sleeping behaviour using a representative sample of U.S. teenagers, finding not-negligible endogenous effects. That is, besides the impact of individual and friend characteristics, we show that the sleeping behaviour of the friends is important in shaping own sleeping behaviour. Unique information on siblings and their friends allows us to check the robustness of our results to unobserved family factors.

Adolescent sleep patterns deserve particular attention because of their potential to affect school performance. Side effects associated with sleep deprivation - inattention, irritability, hyperactivity, and impulse control problems - are likely to show up in school. It is important for educators to screen for sleep problems when concerns exist about a student's attention or behavior problems. Our analysis suggests that an effective intervention should not only be measured by the possible sleep disorder reduction it implies but also by the group interactions

it engenders.

References

Benhabib, J., Bisin, A. and Jackson, M. O. (eds) (2011). *Handbook of Social Economics*, Vol. 1A, North-Holland.

Biddle, J. E. and Hamermesh, D. S. (1990). Sleep and the allocation of time, *Journal of Political Economy* **98**: 922–943.

Bramoullé, Y., Djebbari, H. and Fortin, B. (2009). Identification of peer effects through social networks, *Journal of Econometrics* **150**: 41–55.

Chandrasekhar, A. and Lewis, R. (2011). Econometrics of sampled networks. Working paper, MIT.

Eide, E. R. and Showalter, M. H. (2012). Sleep and student achievement, *Eastern Economic Journal* **38**: 512–524.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects, *Journal of Business and Economic Statistics* **31**: 253–264.

Hillman, D., Murphy, A., Antic, R. and Pezzullo, L. (2006). Economic cost of sleep disorders, *Sleep* **29**: 299–305.

Kelejian, H. H. and Prucha, I. (2010). Spatial models with spatially lagged dependent variables and incomplete data, *Journal of Geographical Systems* **12**: 241–257.

Lee, L. F. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects, *Journal of Econometrics* **140**: 333–374.

Lee, L. F., Liu, X. and Lin, X. (2010). Specification and estimation of social interaction models with network structures, *The Econometrics Journal* **13**: 145–176.

Lin, X. (2010). Identifying peer effects in student academic achievement by a spatial autoregressive model with group unobservables, *Journal of Labor Economics* **28**: 825–860.

Mitru, G., Millrood, D. L. and Mateika, J. H. (2002). The impact of sleep on learning and behavior in adolescents, *Teachers College Record* **104**: 704–726.

Patacchini, E. and Zenou, Y. (2008). The strength of weak ties in crime, *European Economic Review* **52**: 209–236.

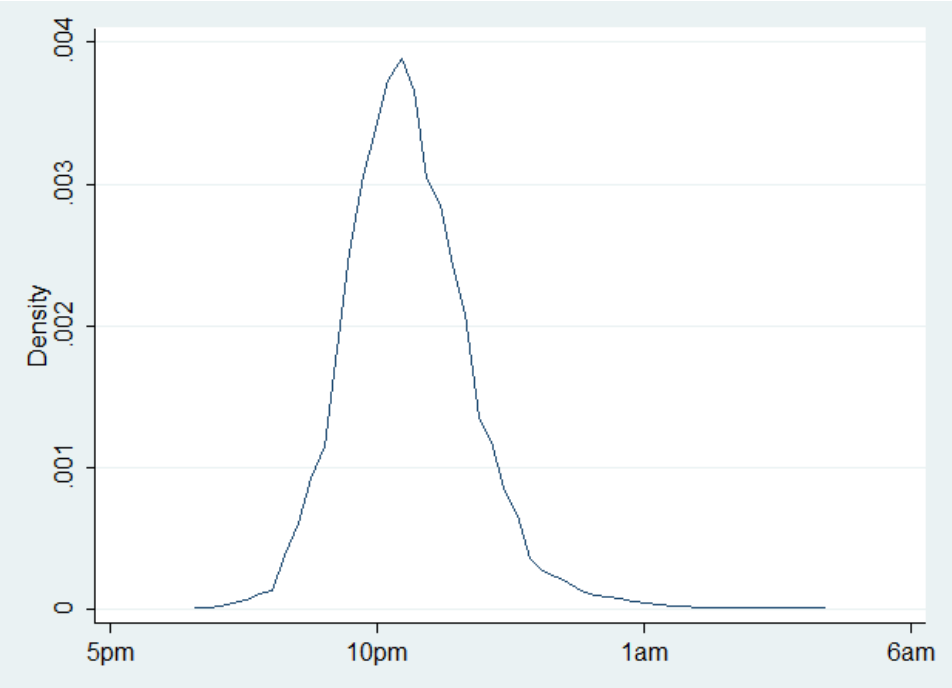
Sojourner, A. (2013). Identification of peer effects with missing peer data: Evidence from project star, *The Economic Journal* **123**: 574–605.

Wang, W. and Lee, L. F. (2013a). Estimation of spatial autoregressive models with randomly missing data in the dependent variable, *Econometrics Journal* **16**: 73–102.

Wang, W. and Lee, L. F. (2013b). Estimation of spatial panel data models with randomly missing data in the dependent variable, *Regional Science and Urban Economics* **43**: 521–538.

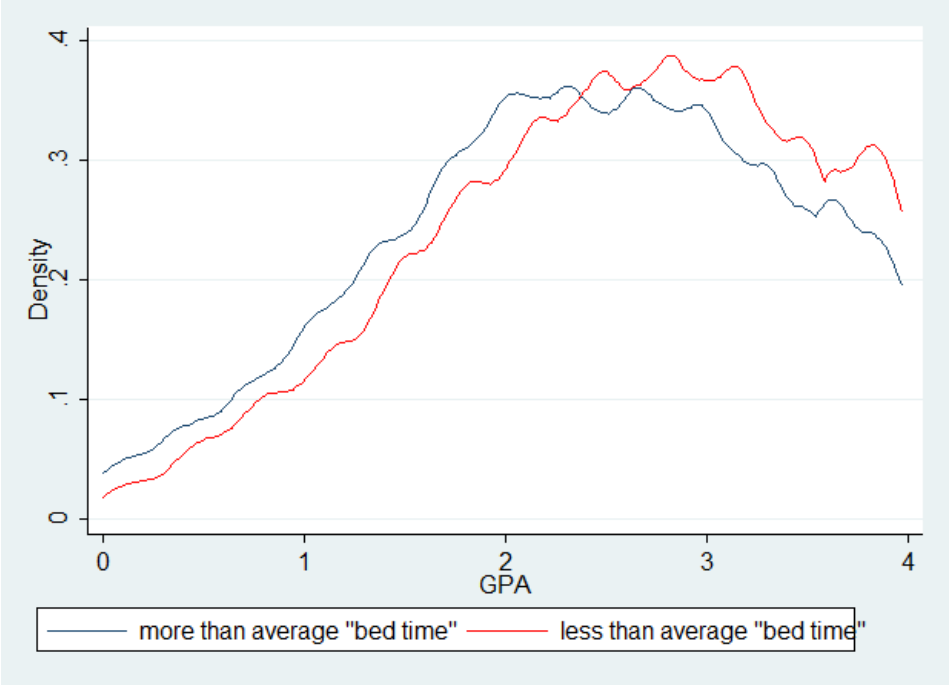
Yaniv, G. (2004). Insomnia, biological clock, and the bedtime decision: an economic perspective, *Health Economics* **13**: 1–8.

Figure 1: Kernel Density Estimate of “Bed Time”



Notes. Kernel = Epanechnikov, bandwidth = 40.429. We report the distribution of student by the time they go to sleep.

Figure 2: “Bed Time”and School Performance



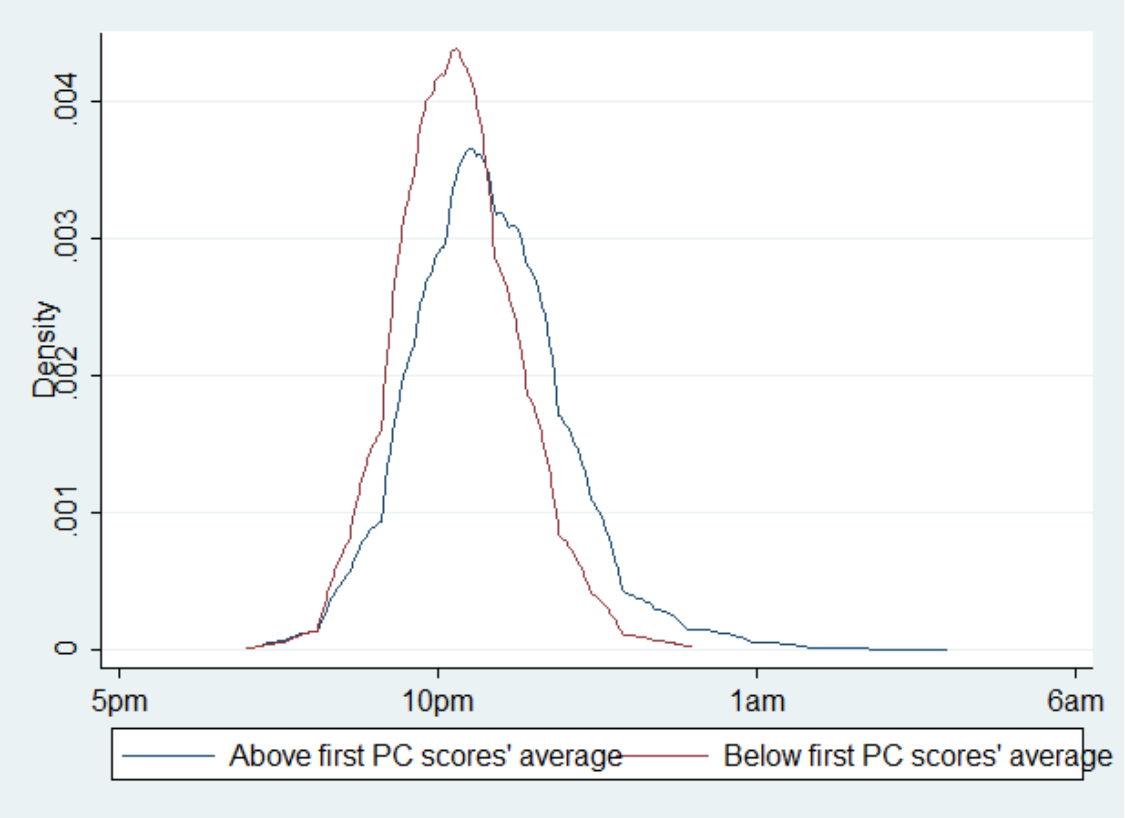
Notes. Kernel = Epanechnikov, bandwidth = 17.793. We report the distributions of students by school performance as measured by GPA, distinguishing between students that sleep more and less than average. GPA is the composite score of a factor analysis run on Mathematics score, English score, History/Social Science score and Science score.

Table 1: Student Characteristics - PCA results -

Variable	Correlation with the first PC
Body mass Index	0.19
GPA	-0.32
General Health	-0.39
Use of Alchool	0.56
Cigarette smoking	0.57

Notes. The first PC explain 32% of the total variance. Body mass index is the ratio between weight (in kilos) and the height squared (in meters). GPA is the composite score of a factor analysis run on Mathematics score, English score, History/Social Science score and Science score. General health is derived from the question: "In general how is your health?", coded as 1= excellent, 2= very good, 3= good, 4= fair, 5= poor. The use of alchool is measured using the question: "During the past twelve months, how often did you: drink beer, wine, or liquor?", coded as 0 = never, 1= once or twice, 2= once a month or less, 3= 2 or 3 days a month, 4= once or twice a week, 5= 3 to 5 days a week, 6= nearly every day. Cigarette smoking uses responses to the question: "During the past twelve months, how often did you: smoke cigarettes?", coded as 0 = never, 1= once or twice, 2= once a month or less, 3= 2 or 3 days a month, 4= once or twice a week, 5= 3 to 5 days a week, 6= nearly every day.

Figure 3: "Bed Time"and First PC



Notes. Kernel = Epanechnikov, bandwidth = 40.429. We report the distributions of Type A students (blu line) and Type B students (red line). Type A students have high body mass index, poor school performance, poor general health, drink alcohol and smoke cigarettes, whereas Type B students have the opposite profile.

Table 2: Simulation Results

Sampling rate	Method	Parameter	Point estimation	Standard error	MSE
40%					
		λ	0.293	0.152	0.152
	NLS	β	1.002	0.093	0.093
		γ	1.013	0.179	0.180
		λ	0.255	0.546	0.548
	2SLS	β	0.995	0.120	0.120
		γ	1.007	0.535	0.535
60%					
		λ	0.294	0.117	0.117
	NLS	β	1.001	0.072	0.072
		γ	1.011	0.138	0.139
		λ	0.252	0.216	0.221
	2SLS	β	0.997	0.081	0.081
		γ	1.013	0.222	0.222
80%					
		λ	0.294	0.100	0.101
	NLS	β	1.001	0.062	0.062
		γ	1.009	0.116	0.116
		λ	0.272	0.132	0.135
	2SLS	β	0.999	0.066	0.066
		γ	1.007	0.141	0.141
100%					
		λ	0.295	0.090	0.093
	NLS	β	1.001	0.050	0.054
		γ	1.008	0.100	0.103
		λ	0.300	0.091	0.091
	2SLS	β	1.000	0.054	0.054
		γ	1.002	0.101	0.101

Notes. Number of replications =5000. Sample size = 500. Number of groups = 50. Number of nodes per group = 10. Maximum number of connections for a node = 3. Distribution of nodes' connections: uniform. Model: $y = \lambda Gy + \beta x + \gamma Gx + \varepsilon$. $\lambda = 0.3$, $\beta = 1.0$, $\gamma = 1.0$, $\sigma^2 = 2$. MSE= $\sqrt{(\theta - \hat{\theta})^2 + \text{var}(\hat{\theta})}; \theta = \lambda, \beta, \gamma$.

Table 3: Peer effect Estimation – Different method comparison- Increasing set of controls

Variable	NLS			2SLS		
Peer effect	0.723**	0.753**	0.726**	-0.213	-0.435	-0.309
	(0.328)	(0.336)	(0.367)	(0.186)	(0.300)	(0.196)
Female	5.579	4.628	5.217	-1.632	-1.350	-1.684
	(5.470)	(5.520)	(5.566)	(5.235)	(5.247)	(5.152)
Grade	27.000***	24.501***	26.812***	24.173***	23.970***	22.780***
	(5.694)	(5.668)	(5.652)	(2.907)	(2.787)	(2.739)
Black	14.624	17.895	22.084	16.208	19.995	22.707
	(16.188)	(16.121)	(16.029)	(12.509)	(12.595)	(15.402)
Asian	20.565	20.006	25.703	24.249	25.255	26.762
	(16.125)	(15.976)	(15.943)	(15.151)	(15.197)	(14.933)
Mathematics score		7.290**	8.491**		9.417***	10.657***
		(3.331)	(3.377)		(3.376)	(3.338)
History/Social Science score		-10.675***	-9.669***		-10.742***	-8.622***
		(3.434)	(3.490)		(3.417)	(3.358)
Risky Behavior Factor			9.557***			10.373***
			(2.518)			(2.343)
Family Characteristics	yes	yes	yes	yes	yes	yes
Contextual effects	yes	yes	yes	yes	yes	yes
Network fixed effects	yes	yes	yes	yes	yes	yes
1,127 Sampled individuals over 3,700 Individuals in 77 Networks						

Notes: Robust standard errors in parentheses, *** p<0,01, ** p<0,05, * p<0,1. Family characteristics include occupation and education of the parents, household size as measured by the number of people living in the household, and a dummy taking value one if the respondent lives in a household with two parents (both biological and non biological) that are married. Parental education is the schooling level of the (biological or non-biological) parent who is living with the child, distinguishing between "never went to school", "not graduate from high school", "high school graduate", "graduated from college or a university", "professional training beyond a four-year college", coded as 0 to 4. We consider only the education of the father if both parents are in the household. Mather and father occupation dummies include the following categories: manager, professional/technical, officer or sales worker, military or security, farm or fishery, other. "None" is the reference group The Risky Behavior Factor is the score of a factor analysis run on use of alcohol cigarette smoking and general health (see the notes to Table 1 for the definition of these variables).

Table 4: Endogeneous network formation -Testable implications

	Dep var. $\left\|\widehat{\varepsilon_i} - \widehat{\varepsilon_j}\right\|$					
$g_{ij} = 1$	**Full set of controls**			**Grade unobserved**		
Threshold (percentile %)	T = 25%	T = 35%	T = 45%	T = 25%	T = 35%	T = 45%
$q_{ij} = \widehat{g_{ij}}$	Nc	168,158.2270	129,781.5472	Nc	61,490.1352**	44,023.8723*
	Nc	(116,612.652)	(85,513.336)	Nc	(28,200.874)	(23,788.400)
Constant	Nc	45.0100	-23.5495	Nc	-127.8422	-12.2902
	Nc	(85.440)	(111.309)	Nc	(218.685)	(209.029)
Network fixed effects	Nc	Yes	Yes	Nc	Yes	Yes
$P(q_{ij} < t \mid g_{ij} = 1)$	**Nc**	**3%**	**4%**	**Nc**	**11%**	**16%**
$P(q_{ij} > t \mid g_{ij} = 1)$	**Nc**	**97%**	**96%**	**Nc**	**89%**	**84%**
$g_{ij} = 0$	**Full set of controls**			**Grade unobserved**		
Threshold (percentile %)	T = 95%	T = 85%	T = 75%	T = 95%	T = 85%	T = 75%
$q_{ij} = \widehat{g_{ij}}$	-21.7473	-1.2709	-54.2087	-1,234.0554*	-1,382.8678**	-1,417.9023**
	(140.834)	(65.105)	(40.550)	(615.024)	(671.879)	(684.427)
Constant	141.2311***	124.2897***	120.6849***	719.5613***	702.9793***	635.2061***
	(33.725)	(8.810)	(4.012)	(77.366)	(67.153)	(50.571)
Network fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
$P(q_{ij} > t \mid g_{ij} = 0)$	**4%**	**12%**	**21%**	**8%**	**13%**	**22%**
$P(q_{ij} < t \mid g_{ij} = 0)$	**96%**	**88%**	**79%**	**92%**	**87%**	**78%**

Notes: nc = not computed, number of observation < 30. Threshold based on percentiles of the empirical distributions of qij. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.$q_{ij}$is estimated with a logit model. Full set of controls as listed in Table 3 are included.

Table 5: Robustness Check. Peer Effect Estimation with Sibling Fixed Effects

Variable	NLS			2SLS		
Peer effect	0,614*	0,547*	0,667*	0,653	0,439	0,128
	(0,319)	(0,298)	(0,395)	(1,063)	(0,895)	(0,943)
Sex	18,243	10,507	-1,433	14,369	12,908	12,321
	(16,608)	(17,126)	(16,901)	(14,351)	(14,361)	(14,272)
Grade	28,115***	26,989***	30,348***	30,817***	28,976***	27,498***
	(6,045)	(8,180)	(7,950)	(5,080)	(4,729)	(4,736)
Black	57,389	29,551	115,329	-10,817	-7,852	-8,849
	(101,692)	(105,599)	(106,494)	(20,428)	(20,016)	(19,927)
Asian	153,274*	90,641	132,325	76,334*	71,505*	70,623*
	(78,011)	(79,331)	(79,610)	(39,991)	(39,423)	(39,131)
Mathematics score		4,332	10,200		13,557*	18,430**
		(9,796)	(10,216)		(8,151)	(8,585)
History/Social Science score		-15,418	-5,331		-5,764	-5,479
		(11,049)	(10,787)		(8,467)	(8,431)
Risky Behavior Factor			12,334			10,526
			(7,743)			(6,825)
Family Characteristics	yes	yes	yes	yes	yes	yes
Contextual effects	yes	yes	yes	yes	yes	yes
Sibling fixed effects	yes	yes	yes	yes	yes	yes
171 Sampled individuals over 3,700 Individuals in 77 Networks						

Notes: Robust standard errors in parentheses, *** p<0,01, ** p<0,05, * p<0,1. Family characteristics include occupation and education of the parents, household size as measured by the number of people living in the household, and a dummy taking value one if the respondent lives in a household with two parents (both biological and non biological) that are married. Parental education is the schooling level of the (biological or non-biological) parent who is living with the child, distinguishing between “never went to school”, “not graduate from high school”, “high school graduate”, “graduated from college or a university”, “professional training beyond a four-year college”, coded as 0 to 4. We consider only the education of the father if both parents are in the household. Mather and father occupation dummies include the following categories: manager, professional/technical, officer or sales worker, military or security, farm or fishery, other. “None” is the reference group The Risky Behavior Factor is the score of a factor analysis run on use of alcohol cigarette smoking and general health (see the notes to Table 1 for the definition of these variables).