[FOR ONLINE PUBLICATION]

SUPPLEMENTARY MATERIAL FOR "**Body Mass Index and Social Interactions from Adolescence to Adulthood**"

A The Linear-in-Means Model

The baseline LMM is conceptually very simple. Usually not derived from any predefined individual decision problem, this model allows individual behavior to linearly depend on some individual-specific characteristics as well as on group-specific factors, which include some group observable characteristics and the expected aggregate behavior of the others in the group.[1] This makes it easily interpretable as a regression model, and therefore interesting to the econometrician. However, as pointed out by Manski (1993), the LMM suffers from a special kind of identification problem - the so called reflection problem - due to difficulties in disentangling two different group-effects, namely contextual and endogenous effects. Therefore, in such a framework measuring the impact of social interactions is typically challenging. Consider the simple version of the model, where estimation concerns are not yet addressed. Assume to have G non-overlapping, *a priori* determined groups, each of them made of N^g individuals. Individual choice is assumed to be the result of the following process:

$$y_{ig} = a + y_{ig}^e \beta + \mathbf{x}_g' \gamma + \mathbf{r}_{ig}' \delta + \varepsilon_{ig} \ , \ \text{where} \ \begin{matrix} g = 1, \dots, G \\ i = 1, \dots, N^g \end{matrix} \ . \tag{1}$$

The individual-specific terms are defined by a $r \times 1$ vector of observable characteristics, $\mathbf{r}_{ig}$, and ε_{ig}, a random and unobservable scalar assumed to be independent and identically distributed across individuals. As to group-specific factors, these are divided into a $k \times 1$ vector of predetermined characteristics, $\mathbf{x}_g$, and the expected average choice in the group, y_{ig}^e. These two terms are conceptually different, the former being interpreted as contextual effects and the latter as an endogenous effect, and those exist under the condition that β is non-zero and γ has at least a non-zero element. The key effect is exerted by y_{ig}^e, since it creates reciprocal reactions between individual decisions.

The information set for y_{ig}^e includes values of r_{ig} for other individuals within i's group, as well as the equilibrium expected choice level that occurs for her group. Individuals are assumed to be unable to observe the choices of others, y_{-ig}, or their random payoff terms ε_{ig}. Alternative information assumptions will not affect the qualitative properties of the model. For the LMM, self-consistency amounts to:

$$y_{ig}^e = y_g^e = \frac{a + \mathbf{x}_g' \gamma + \mathbf{r}_g' \delta}{1 - \beta} = \frac{a + \mathbf{x}_g' \gamma}{1 - \beta} + \frac{\mathbf{r}_g' \delta}{1 - \beta}, \tag{2}$$

where $\mathbf{r}_g$ is the average of $\mathbf{r}_{ig}$ within group g.

Notice that such an assumption on the aggregate outcome implies a unique equilibrium: there exists only one expected average choice level that is consistent with the model, given individual and group characteristics. Therefore, equation (2) maps these characteristics into a single y_g^e.

An identification problem in this framework could arise because endogenous and contextual effects may co-move. Indeed, under the self-consistency assumption, the contextual variables determine the endogenous variable, as indicated by condition (2). Given that the identification failure is a consequence of the correlation, by construction, between the endogenous and the contextual effects, Manski (1993) renamed it 'reflection problem', which is not too dissimilar from the basic identification problem in linear regressions with linearly dependent covariates.

[1] LMM can be the result of an optimal decision problem framed around agent's choice, as illustrated in Brock and Durlauf (2001).

Manski's original argument is that every contextual effect might be defined as the average of a corresponding individual characteristic. For example, if one controls for student's maternal education one also introduces average (school) maternal education so that $\mathbf{x}_g = \mathbf{r}_g$. Condition (2) becomes

$$y_g^e = \frac{a + \mathbf{x}_g' (\gamma + \delta)}{1 - \beta}, \tag{3}$$

meaning that the regressor $y_{ig}^e = y_g^e$ in (1) is linearly dependent on the regressors a and $\mathbf{x}_g$ in (1), so the parameters are not identified. Substituting (3) into (1):

$$y_{ig} = \frac{a}{1 - \beta} + \frac{\beta}{1 - \beta} \mathbf{x}_g' (\gamma + \delta) + \mathbf{r}_{ig}' \delta + \varepsilon_{ig}. \tag{4}$$

We can therefore state the following two remarks on the identification of social interaction effects in a LMM:

Remark 1 *In the empirical model* (1) *the set of regressors* $(1, y_g^e, \mathbf{x}_g, \mathbf{r}_{ig})$ *requires the estimation of* $2 + k + r$ *parameters.*

Remark 2 *Assuming reflection* $\mathbf{r}_g = \mathbf{x}_g$ *in the reduced form* (4) *the set of regressors* $(1, \mathbf{x}_g, \mathbf{r}_{ig})$ *allows us to identify* $1 + k + r$ *parameters. Hence, the endogenous effect parameter,* β*, remains unidentified.*

It is then clear why in the LMM framework identification of parameters is a major challenge. In Section two of the paper we show how to achieve identification of the endogenous effect parameter, β.

B GMM and Correlated Effects in Dynamic Linear Panel Data Models

Consider system (5) with correlated effects both at the individual level, f_i, and group level, α_g.

$$\begin{aligned} y_{t,ig} &= y_{t-1,ig}\varphi + y_{t,ig}^e \beta + \mathbf{x}_{t,g}' \gamma + \mathbf{r}_{t,ig}' \delta + e_{t,ig}, \ |\varphi| < 1 \\ e_{t,ig} &= \alpha_g + u_{t,ig}, \\ u_{t,ig} &= f_i + \varepsilon_{t,ig} \end{aligned} \tag{5}$$

By recursion we can write (for $t = 1, ..., T$) :

$$\begin{aligned} y_{t,ig} = {} & \left(1 + \varphi + ... + \varphi^{t-1}\right) f_i + \left(1 + \varphi + ... + \varphi^{t-1}\right) \alpha_g + \varphi^t y_{0,ig} + \\ & + [y_{t,ig}^e \beta + y_{t-1,ig}^e \beta\varphi + ... + y_{1,ig}^e \beta\varphi^{t-1}] \\ & + [\mathbf{x}_{t,g}'\gamma + \mathbf{x}_{t-1,g}'\gamma\varphi + ... + \mathbf{x}_{1,g}'\gamma\varphi^{t-1}] + \\ & + [\mathbf{r}_{t,ig}'\delta + \mathbf{r}_{t-1,ig}'\delta\varphi + ... + \mathbf{r}_{1,ig}'\delta\varphi^{t-1}] + \\ & + [\varepsilon_{t,ig} + \varphi\varepsilon_{t-1,ig} + ... + \varphi^{t-1}\varepsilon_{1,ig}] \end{aligned} \tag{6}$$

We can write system (6) in compact form as:

$$E[y_{t,ig} \mid \mathbf{W}_i] = \mathbf{W}_i^{'} \, \mathbf{\Pi} + \eta(f_i \; + \alpha_g)$$

where $\mathbf{W}_i = \left[y_{0,ig}, \; y^e_{t,ig}, ..., y^e_{1,ig}, \; \mathbf{x}_{t,g}, ..., \mathbf{x}_{1,g}, \; \mathbf{r}_{t,ig}, ..., \mathbf{r}_{1,ig} \right]$. The $\mathbf{\Pi}$ matrix is defined in terms of the coefficients of the linear predictors of the dependent variable at each period given all explanatory variables at all periods. Hence, for the individual effect, f_i, and the group effect, α_g, we therefore have:

$$E[f_i, \alpha_g \mid \mathbf{W}_i] = 0$$

B.1 A Montecarlo Exercise

We perform a simple simulation exercise in order to detect the finite sample properties of the system GMM estimation of the empirical model (5). The values of the two main coefficients of interest φ and β are consistent with the values we obtain from our general empirical analysis. Table B.1 summaries all the results for different T values. A GMM estimation generally produces relatively unbiased estimates of φ and β and small RMEs. Increasing T generally ameliorates these results by reducing the RMSE.

Table B1: Montecarlo simulation for different parameters values

	T=4			T=10			T=20		
Parameters:	Mean	Bias	RMSE	Mean	Bias	RMSE	Mean	Bias	RMSE
$\varphi = 0.2$	0.201	0.001	0.009	0.201	0.001	0.006	0.199	-0.001	0.004
$\beta = 0.1$	0.102	0.002	0.091	0.098	-0.002	0.056	0.097	-0.002	0.032
$\varphi = 0.4$	0.402	0.002	0.009	0.401	0.001	0.005	0.399	-0.001	0.004
$\beta = 0.2$	0.202	0.002	0.090	0.198	-0.002	0.056	0.197	-0.003	0.032
$\varphi = 0.6$	0.602	0.002	0.009	0.601	0.001	0.006	0.599	-0.002	0.004
$\beta = 0.4$	0.403	0.003	0.091	0.398	-0.002	0.056	0.599	-0.001	0.004
$\varphi = 0.8$	0.802	0.002	0.009	0.801	0.001	0.006	0.799	-0.001	0.004
$\beta = 0.6$	0.602	0.002	0.091	0.598	-0.002	0.056	0.597	-0.003	0.032

Notes:

1. *The sample size (N)=14.000 obs. and groups (G)=125*
2. *1000 Montecarlo Replications*
3. *The values of φ and β are consistent with the Add Health model Estimation*

C Design Weighting

The Add Health Study is a US representative, probability-based survey of adolescents in grades 7 through 12 conducted between 1994 and 1995, and extended to 2008 with three in-home interviews. The sample design used to collect the data embeds a certain degree of complexity which should be accounted for. Indeed, failing at considering such complexity may result in biased parameter estimates and incorrect variance estimates. Hence, we corrected for design effects and unequal probability of selection, according to what is suggested in the Add Health user guides.[2] We exploit the longitudinal feature of the dataset, keeping the strength of its innovative design. With the longitudinal data from adolescence, the third and four in-home interviews allow "researchers to map early trajectories out of adolescence in health, achievement, social

[2] http://www.cpc.unc.edu/projects/addhealth/data/guides

relationships, and economic status and to document how adolescent experiences and behaviors are related to decisions, behavior, and health outcomes in the transition to adulthood. The fundamental purpose of this [...] follow-up was to understand how what happens in adolescence is linked to what happens in the transition to adulthood when adolescents begin to negotiate the social world on their own and develop their expectations and goals for their future adult roles." (Harris, 2011). Data have been appropriately weighted to correct their longitudinal format. For details on the Add Health weighting scheme, the reader is cross-referred to Tourangeau and Shin (1999).

D Falsification and Robustness tests

D.1 Falsification Test

We provide a simple falsification test (see Cohen and Fletcher (2008)) where instead of using BMI as our main dependent variable, we include an alternative continuous time-varying variable, i.e the height of all the respondents, information available in Add Health up to wave 4.
For body height we expect past behavior to exhert the strongest effect and the peer effect to be negligible or absent. We re-estimate our model (5) by including past height as well as average height by grade-school groups. Table D.1, shows that the past accounts for most of the current height in the full sample. The peer effect, even though significant,[3] is very small compared to the past.[4]

Table D1: Falsification Test on Height: Estimates using full sample

Dependent Variable: $ln(height)_t$	Model 1		Model 2	
Variables	Coeff.	SE	Coeff.	SE
$ln(Height)_{t-1}$	0.619***	0.014	0.618***	0.014
Average $ln(Height)_t$	0.090***	0.013	0.089***	0.013
Individual Controls	✓		✓	
Contextual Average Effects	✓		✓	
Average Cohorts Height	✓		✓	
Wave Fixed Effects	✓		✓	
Observations	13,882		13,882	
Number of individuals	5,035		5,035	
Number of Instruments	114		108	

*** p<0.01, ** p<0.05, * p<0.1
*Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1*
Notes:
1) *Average levels are calculated at grade level within school*
2) *Model 1: Full Observations considering Health as Exogenous*
3) *Model 2: Full Observations excluding Health*
Community Fixed Effects included as additional Instruments

D.2 Robustness Test: BMI adjusted for Gender and Age

We provide additional evidence on BMI adjusted for age and gender (zBMI). The zBMI growth curve for boys and girls are different and the cut-off levels for both genders are non identical. We estimate our empirical model (5) for both males and females for the three categories (normal, over-weight and obese). The reference levels for zBMI are obtained as the z-score per-

[3]We also conducted separate estimations by dividing the sample by gender in three height categories (short, medium and tall) and found that the peer effect for the female sample is not significant or weakly significant. Whereas for the male sample the peer effect is significant but very small. The results are available upon request.

[4]Other available variables like having acne/asthma problems are binary variables (having or not skin/asthma problems) or are not available in all waves (such as ADHD - Attention Deficit Hyperactivity Disorder) so they are not suitable to be used in a falsification test via GMM.

centiles from the table classification produced by the WHO.[5] The results are slightly different as compared to our baseline estimations. For the sample of males the results in Table D.2 show that current BMI of normal-weight and obese boys is affected by past BMI and peer effects with a similar magnitude. The opposite effect occurs for the sample of over-weight boys where past BMI is not significant anymore, whereas the peer effect explains most of the current BMI variation. The estimation for the female sample are presented in Table D.3: both past and peer effects are relevant in explaining current BMI, with the past effects having a stronger impact on current BMI in all the three weight categories.

Table D2: Estimates using full sample of Males for Adjusted BMI by age

Dependent Variable: $ln(BMI)_t$	Model 1		Model 2	
Variables	Coeff.	SE	Coeff.	SE
Boys normal-weight during adolescence				
$ln(BMI)_{t-1}$	0.569***	0.028	0.571***	0.027
Average $ln(BMI)_t$	0.349***	0.042	0.341***	0.042
Observations	3,718		3,718	
Number of individuals	1,896		1,896	
Number of Instruments	106		104	
Boys over-weight during adolescence				
$ln(BMI)_{t-1}$	−0.151	0.114	0.151	0.113
Average $ln(BMI)_t$	0.548***	0.113	0.548***	0.113
Observations	312		312	
Number of individuals	165		165	
Number of Instruments	82		82	
Boys obese during adolescence				
$ln(BMI)_{t-1}$	0.558***	0.126	0.621***	0.123
Average $ln(BMI)_t$	0.505***	0.369	0.560***	0.188
Observations	169		169	
Number of individuals	85		85	
Number of Instruments	65		63	
Individual Controls	✓		✓	
Contextual Average Effects	✓		✓	
Average Cohorts BMI	✓		✓	
Wave Fixed Effects	✓		✓	

*** p<0.01, ** p<0.05, * p<0.1

Notes:

1) *Average levels are calculated at Grade level within Schools*

2) *Model 1: Full Observations considering Health as Exogenous*

3) *Model 2: Full Observations excluding Health*

4) *Community Fixed Effects included as Additional Instruments*

[5]The z-scores tables and percentiles are retrieved from https://www.who.int/childgrowth/standards/bmi for age/en/

Table D3: Estimates using full sample of Females for Adjusted BMI by age

Dependent Variable: $ln(BMI)_t$	Model 1		Model 2	
Variables	Coeff.	SE	Coeff.	SE
Girls normal-weight during adolescence				
$ln(BMI)_{t-1}$	0.633***	0.029	0.636***	0.029
Average $ln(BMI)_t$	0.378***	0.041	0.382***	0.041
Observations	4,425		4,425	
Number of individuals	2,244		2,244	
Number of Instruments	105		103	
Girls over-weight during adolescence				
$ln(BMI)_{t-1}$	0.356***	0.168	0.362***	0.169
Average $ln(BMI)_t$	0.346***	0.128	0.343***	0.124
Observations	258		258	
Number of individuals	134		134	
Number of Instruments	82		80	
Girls obese during adolescence				
$ln(BMI)_{t-1}$	0.529***	0.120	0.519***	0.140
Average $ln(BMI)_t$	0.322***	0.099	0.309***	0.095
Observations	197		197	
Number of individuals	104		104	
Number of Instruments	74		72	
Individual Controls	✓		✓	
Contextual Average Effects	✓		✓	
Average Cohorts BMI	✓		✓	
Wave Fixed Effects	✓		✓	

*Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1*

Notes:

1) *Average levels are calculated at grade level within school*

2) *Model 1: Full Observations considering Health as Exogenous*

3) *Model 2: Full Observations excluding Health*

4) *Community Fixed Effects included as Additional Instruments*

References

[1] Brock, W. A., and Durlauf, S. N. (2001). "Discrete choice with social interactions." *Review of Economic Studies*, 68(2), 235 – 260.

[2] Cohen-Cole, E., and Fletcher, J. (2008). "Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis." *British Medical Journal*, 338(7685), 28 – 31.

[3] Harris, K. M. (2011). "Design Features of Add Health" Carolina Population Center, website: www.cpc.unc.edu/projects/addhealth/data/guides/.

[4] Manski, C. F. (1993). "Identification of endogenous social effects: The reflection problem." *Review of Economic Studies*, 60(3): 531 – 542.

[5] Tourangeau, R. and Shin, H. (1999). "National Longitudinal Study of Adolescent health: Grand Sample Weights." National Opinion Research Center and Carolina Population Center. website: http://www.cpc.unc.edu/projects/addhealth/.