ISER Working Paper Series

Inequality in Pupils' Educational Attainment: How Much Do Family, Sibling Type and Neighbourhood Matter?

Cheti Nicoletti and Birgitta Rabe
Institute for Social and Economic Research
University of Essex

No. 2010-26
22 July 2010

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

www.iser.essex.ac.uk

Non-technical summary

There is a long-standing interest in the relative influence of family and neighbourhood background on socio-economic outcomes such as income and education. Over many years, social scientists have used sibling similarity in socio-economic outcomes to measure the importance of family background, where any sibling resemblance indicates that family background matters. However, the sibling similarity measures both family and community influences, as siblings share both the family and the neighbourhood environment they grow up in. By looking at the similarity between siblings and between unrelated neighbours it is possible to put bounds on the possible magnitude of family and neighbourhood effects.

In this paper we exploit English register data with an exceptionally large sample size to explore the relative importance of family, sibling type and neighbourhood on school outcomes. In particular we look at, and compare, test results at the end of primary school, at age 11, and at the end of compulsory schooling, usually at age 16. We find that the shared neighbourhood can account for at most 10-15% of the differences in educational attainment between pupils in England, whereas the influence of the shared family background is estimated to be between 0.47 and 0.56 for eleven and between 0.48 and 0.62 for sixteen year olds. Living in an urban neighbourhood has a slightly larger influence than living in a rural one, possibly because the contact among neighbouring children is higher in densely populated areas.

Differentiating by sibling type, we find that sibling similarity is higher between same gender siblings than between siblings of differing gender, and it is higher at age eleven than at age sixteen. Identical twins' similarity in educational outcomes is between 0.25 and 0.35 points higher than that of fraternal twins, which presumably is caused at least in part by the identical genes of identical twins. We also find that sibling similarity in test results of primary school pupils declines markedly as the age gap between them increases, possibly because they experience changes in family circumstances at different ages. Finally, the differences between siblings are larger when their school starting age differs, arguably because there is a benefit (penalty) of being among the older (younger) pupils in a school year. This effect is apparent at age 11 but disappears by age 16.

Inequality in pupils' educational attainment:
how much do family, sibling type and neighbourhood matter?

Cheti Nicoletti

Birgitta Rabe

Institute for Social and Economic Research
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
brabe@essex.ac.uk

Abstract

We explore the relative influence of family and neighbourhood on educational attainment and how this varies by sibling type. Using English register data we find sibling correlations in exam scores of 0.563 at the end of primary school and of 0.621 at the end of compulsory schooling. The neighbourhood explains at most 10-15% of the variance in educational attainment; whereas the family explains at least 43%. This percentage is significantly higher for twins and for siblings of the same sex. It is also higher for closely spaced siblings and siblings with a similar school starting age but only at age 11.

Keywords: Siblings, twins, neighbourhood, educational attainment, multilevel model

JEL Code: J13, R23

Acknowledgements: We would like to thank Anna Vignoles and participants at the seminar series at the Institute of Education and at the ESPE 2010 conference where previous versions of this paper were presented. This work was supported by the Economic and Social Research Council through MiSoC.

1. Introduction

There is a long-standing interest in the relative influence of family and neighbourhood background on socio-economic outcomes such as income and education. Over many years, social scientists have used sibling correlations in socio-economic outcomes to measure the importance of family background, where any sibling resemblance indicates that family background matters. However, the sibling correlation measures both family and community influences, as siblings share both the family and the neighbourhood context they grow up in. Solon et al. (2000) use a variance decomposition approach to put bounds on the possible magnitude of family and neighbourhood effects, estimating correlations between siblings and between unrelated neighbours. This can be used to put an upper and a lower bound on the 'pure' family influence, i.e. on the proportion of variation in the outcome explained by shared family factors.

There are, however, a number of family background factors that are not shared by siblings. Examples are factors that differ between the siblings because they are brought up at different times, because parents allocate their resources unequally between the children in the family, or because siblings do not share all genetic traits (unless they are identical twins). These unshared family factors could be the key to explaining why inequalities in socio-economic outcomes are fairly large even within families. They can be captured by comparing sibling types, defined for example by the extent to which siblings share environments (Solon 1999).

In this paper we exploit English register data with an exceptionally large sample size to explore the relative importance of family, sibling type and neighbourhood on educational attainments. In particular we look at, and compare, educational attainments at the end of primary school, at age 11, and at the end of compulsory schooling, usually at age 16. We add to the previous literature in various ways. To the best of our knowledge we are the first to apply the variance decomposition into family and neighbourhood influences (Solon et al. 2000) to the UK, thus contributing to international comparisons of such influences. Moreover, we decompose the variance in exam results which arguably are a finer measure of educational attainment than the years of education that are commonly used in other papers. Furthermore, we are able to differentiate the results by sibling type, where some of our sibling typologies have rarely been used in previous sibling research. Specifically we look at sibling differences in educational outcomes by sibling gender combination, age difference, school starting age difference, and genetic similarity (comparing monozygotic (MZ) twins with dizygotic (DZ) twins and non-twin siblings). We develop a method to identify MZ and DZ twin correlations separately although the twin types are not defined in our data. Insights into the sources and magnitude of sibling differences by sibling type are informative for studies that rely on sibling fixed effects to control for family background. Finally, we use multilevel models to estimate the variance components, and we show that the lower bound on the family influence has a straightforward interpretation; it measures the influence of what we call the *relative family effect*.

We find that the shared neighbourhood can account for at most 10-15% of the variance in educational attainment, whereas the influence of the shared family background is estimated to be between 0.467 and 0.563 for eleven and between 0.478 and 0.621 for sixteen year olds. Living in an urban neighbourhood has a slightly larger influence than living in a rural one, possibly because the contact among neighbouring children is higher in densely populated areas. Differentiating by sibling type, we find that the sibling correlations are higher between same gender siblings than between siblings of differing gender, and they are higher at age eleven than at age sixteen. Monozygotic (identical) twins' correlations in educational attainments are between 0.25 and 0.35 points higher than those of dizygotic (fraternal) twins, which presumably is caused at least in part by the identical genes of monozygotic twins. We also find that sibling correlations in attainments of primary school pupils decline markedly as the age gap between them increases, possibly because they experience changes in family circumstances at different ages. Finally, the differences between siblings are larger when their school starting age differs, arguably because there is a benefit (penalty) of being among the older (younger) pupils in a school year. This effect is apparent at age 11 but disappears by age 16.

2. Background

Sibling correlations in socio-economic outcomes are summary measures of the importance of shared family and community background in explaining the outcome in question. Attempts to disentangle the relative importance of family and neighbourhood are complicated by the fact that they are strongly correlated, as children who grow up in communities with schools, peers and role models that lead to favourable adult outcomes also live in families with favourable characteristics (Björklund and Jäntti 2009). Solon et al. (2000) use a variance decomposition approach to put bounds on the possible magnitude of family and neighbourhood effects, estimating correlations between siblings and between unrelated neighbours. The neighbour correlation captures pure neighbourhood factors, but also family traits that are likely positively correlated within the neighbourhood because of sorting mechanisms. The neighbourhood correlation in children's outcomes is therefore an upper bound on the importance or magnitude of neighbourhood factors in explaining this outcome, i.e. it is an upper bound on the proportion of variation in the children's outcome explained by the neighbourhood effect. Furthermore, the difference between the sibling and neighbour correlation provides a lower bound on the importance of the family influence. The advantage of this approach to estimating neighbourhood and family influence is that it does not require observing any of the family and neighbourhood characteristics that may explain children's outcomes. Therefore it avoids issues of measurement error, omission and arbitrary selection of family and neighbourhood characteristics.

Previous papers have shown that any sibling resemblance in a variety of outcomes arises much more from growing up in the same family than from growing up in the same neighbourhood. Specifically, Solon et al. (2000) estimate the sibling correlation in years of education to be 0.5, of which the shared neighbourhood contributes at most around 0.1, leaving a lower bound of the pure family influence of approximately 0.4. Subsequent papers have shown that sibling and neighbourhood correlations are lower for adult outcomes such as income and earnings, with sibling correlations ranging from around 0.2 (Raaum et al. 2006; Lindahl 2009) to 0.3 (Page and Solon 2003a; 2003b)[1] and neighbourhood correlations ranging from 0.003 (Lindahl 2009, for girls) to 0.06 (Raaum et al. 2006, for boys). After purging the neighbourhood correlation of observed family characteristics, the influence of the neighbourhood is even lower. Lindahl (2009) has also used school and class mate correlations to investigate which part of siblings' correlations can be explained by shared experiences in the school and classroom respectively. The school influences are small at around 0.02 for years of education and about 0.01 for average income, and the class influences are slightly larger.

There are numerous channels through which neighbourhoods could affect the educational outcomes of the pupils living in them. These could include social interactions between members of the community such as schoolmates and friends living in the same neighbourhood ("peer effects") or influences of adults that either live or work in the neighbourhood and serve as role models. They could, on the other hand, include physical characteristics of the neighbourhood such as safety, recreational facilities and the like. Our comparison of the neighbourhood influence of communities with differing population density (urban and rural neighbourhoods) is supportive of the social interaction channel.

There is an extensive literature which spans several disciplines from psychology, sociology and behavioural genetics to economics that looks at which specific within-family background factors are responsible for sibling differences in outcomes. Starting with genetics, some researchers have compared family members with varying degrees of genetic relatedness, such as MZ and DZ twins, full and half siblings to parcel out genetic and environmental influences (e.g. Taubman 1976; Guo and Wang 2002; Bjoerklund et al. 2005, Rabe-Hesketh et al. 2007). The papers generally find that similarity in outcomes increases with genetic relatedness, but disentangling these influences in detail requires quite strong assumptions.

[1] Notice however that Solon et al. (1991) and Mazumder (2008) find estimates for the sibling correlation ranging from 0.4 to 0.5 when considering correlations in the permanent component of income and earnings.

To explain differences between siblings, researchers have also looked at the effects on socio-economic outcomes of birth order in the context of constraints on parental investments in their children. Birth order may affect the allocation of parental resources such as time inputs between siblings, with the first-born usually receiving the largest share and subsequent children receiving less and less. While there is a growing body of literature showing that outcomes such as educational attainments decline with family size and birth order (e.g. Black et al. 2005), the effects of sibling age differences are less often researched (see e.g. Conley et al. 2007, Conley and Glauber 2008 for exceptions). We expect the negative effects of birth order on younger siblings to be smaller if the children are widely spaced (Powell and Steelman 1995), as older siblings may require less parental input once they have passed their formative years and may in turn spend time with the later born siblings. Another hypothesis in favour of greater similarity of widely spread siblings compared to closely spread ones is that siblings who are close in age will be in direct competition with each other and will seek to magnify small differences in their abilities or seek out different niches within the family (Conley et al. 2007). On the other hand, there may be more similarity between closer spaced siblings than between those spaced further apart because they experience the same family environment, including critical transitions such as family income shocks or a parents' divorce, at similar ages. Moreover, parents may try to attenuate differences between siblings, particularly if they are spaced closely together (Grilliches 1979).

While our data do not allow us to investigate birth order effects, we are able to study differences in sibling correlations between MZ and DZ twins and non-twin siblings, and to distinguish in detail the age gap effects between siblings, up to a maximum of approximately two years age difference.[2] Moreover, we look at another source of sibling inequality that to the best of our knowledge has not been analysed to date. This is the difference in school outcomes caused by differences in the school starting age. The school starting age is determined in the UK, like in many other countries, by birth date. Children born on 31 August start school and take exams up to one year earlier than children born on 1 September. It is well documented that the older children in each school year outperform the younger members of the cohort, and these differences persist well into secondary school (see e.g. Angrist and Krueger 1992; Bedard and Dhuey 2006, Crawford et al. 2007). Therefore differences in siblings' educational attainment would be expected to reflect differences in the season of birth.

[2] Birth spacing can be the outcome of parental choice (e.g. Rosenzweig 1986) and can therefore be endogenous to any child outcomes. For example, lower educated families may space their children closer together than families with higher education. However, this would only be a problem for our analysis if the parental choice of birth spacing were related to sibling differences, e.g. through the parental propensity to compensate for such differences. We assume that this is not the case.

The literature on intergenerational mobility has also focused on how family effects on children's outcomes change across age. It is generally found that the effect of parent's socio-economic status increases with age. In England, for example, Goodman and Gregg (2010) find that the gap in cognitive abilities and educational attainments between poor and rich children widens constantly during pre-school and primary school. This widening of the gap seems to slow down during secondary school but by the age of 16 the socio-economic gradient is still large (see Goodman and Gregg 2010; Ermisch and Del Bono 2010). We add to this literature by assessing how both family and neighbourhood influences change when looking at educational attainments at the end of primary school and of compulsory schooling.

3. Model and Econometric Methods

Variance decomposition approach

In this section we describe the variance decomposition approach used to bound the effects of family and neighbourhood on children's outcomes (see Solon et al. 2000). Estimation issues are discussed in the next section.

Let y_{cfs} denote our outcome of interest, educational attainment, for sibling *s* in family *f* in neighbourhood *c*, and let us assume the following model:

$$y_{cfs} = \alpha' X_{cf} + \beta' Z_c + u_{cfs} , \qquad (1)$$

where X_{cf} and Z_c are vectors of all family and neighbourhood characteristics relevant to explain y_{cfs}, α and β are the corresponding vectors of coefficients, and u_{cfs} is an error term independent of family and neighbourhood characteristics and identically and independently distributed (i.i.d.) with mean zero and variance σ_u^2.

In the ideal situation where there are no omitted or mis-measured family or neighbourhood characteristics, we can estimate the family and neighbourhood effects, $(\alpha' X_{cf})$ and $(\beta' Z_c)$, by simply regressing the educational attainment on the set of observed explanatory variables. In less ideal situations the estimation of these effects will be biased. However, even in absence of any observed explanatory variable, it is possible to estimate the correlations in educational attainment between siblings and between neighbouring children, and these provide upper bounds on the proportion of variance of y_{cfs} explained by family background and neighbourhood characteristics, $V(\alpha' X_{cf})/V(y_{cfs})$ and $V(\beta' Z_c)/V(y_{cfs})$ which we call family and neighbourhood *influence* (see Solon et al. 2000).

The correlation between siblings is:

$$Corr_S=Cov(y_{cfs},y_{cfs'})/V(y_{cfs})=[V(\beta'Z_c)+V(\alpha'X_{cf})+2\ \mathrm{Cov}(\alpha'X_{cf},\beta'Z_c)]/V(y_{cfs}),$$

The sibling correlation is thus equal to the sum of three positive addends: the variance explained by the neighbourhood influence, $V(\beta'Z_c)/V(y_{cfs})$, the variance explained by the 'pure' family influence, $V(\alpha'X_{cf})/V(y_{cfs})$, and twice the ratio $\mathrm{Cov}(\alpha'X_{cf},\beta'Z_c)]/V(y_{cfs})$. The latter term represents sorting of families into neighbourhoods and is presumably positive because we assume that families with advantaged characteristics sort into less deprived neighbourhoods. Hence the sibling correlation is higher than the proportion of variation in y_{cfs} explained by family background and thus puts an upper bound on the family influence.

Similarly, the correlation between pupils from the same neighbourhood,

$$Corr_N=Cov(y_{cfs},y_{cf's'})/V(y_{cfs})=[V(\beta'Z_c)+\mathrm{Cov}(\alpha'X_{cf},\alpha'X_{cf'})+2\ \mathrm{Cov}(\alpha'X_{cf},\beta'Z_c)]/V(y_{cfs}),$$

provides an upper bound for the explanatory power of the neighbourhood influence, $V(\beta'Z_c)/V(y_{cfs})$. Notice that we assume that $\mathrm{Cov}(\alpha'X_{cf},\alpha'X_{cf'})$ is positive because families with similar characteristics sort into the same neighbourhoods.

Finally, by subtracting the correlation between children in the same neighbourhood from the sibling correlation, we can compute a lower bound on the variance explained by the 'pure' family influence, $V(\alpha'X_{cf})/V(y_{cfs})$. This is a lower bound because

$$Corr_S - Corr_N = [V(\alpha'X_{cf}) - \mathrm{Cov}(\alpha'X_{cf},\alpha'X_{cf'})]/V(y_{cfs}),$$

and as before, we assume that $\mathrm{Cov}(\alpha'X_{cf},\alpha'X_{cf'})>0$ because we expect families with similar characteristics to sort into the same neighbourhoods.

To produce tighter bounds on the neighbourhood influence Altonji (1988) proposes a two-step procedure (see Solon et al. 2000). Let us assume that we can observe a subset of all the relevant family characteristics and let us partition X_{cf} into two sub-vectors of observed and unobserved characteristics $X_{1,cf}$ and $X_{2,cf}$, $X_{cf}=[X_{1,cf}\ , X_{2,cf}\ ]$. The first step consists in regressing y_{cfs} on the subset of observed family characteristics $X_{1,cf}$ and dummy variables for each neighbourhood. The second step uses the family effect predicted from the first step to estimate the covariance between pupils living in the same neighbourhood, $Cov(\widehat{\alpha}_1'X_{cf},\ \widehat{\alpha}_1'X_{cf'})$. This is a lower bound on $Cov(\alpha'X_{cf},\alpha'X_{cf'})$ because only a subset of the predictors are used for its estimation. Therefore $[Corr_N$- $Cov(\widehat{\alpha}_1'X_{cf},\ \widehat{\alpha}_1'X_{cf'})/\,Var(\mathrm{y}_{\mathrm{cfs}})]$ provides a tighter upper bound for the variance explained by the neighbourhood influence.

This new upper bound on the neighbourhood influence can be subtracted from the upper bound of the family influence to produce the following tighter lower bound on the family influence

$$Corr_S\text{-}\left[Corr_N\text{-}Cov\left(\hat{\alpha}_1^{'}X_{cf}, \hat{\alpha}_1^{'}X_{cf'}\right)\right]$$

$$=V(\alpha'X_{cf})/V(y_{cfs})\text{-}[Cov(\alpha'X_{cf},\alpha'X_{cf'})\text{-}\ Cov\left(\widehat{\alpha}_1^{'}X_{cf}, \widehat{\alpha}_1^{'}X_{cf'}\right)]/V(y_{cfs}).$$

Another method that has been used to produce a tighter upper bound on the neighbourhood influence is to estimate the correlation between pupils living in the same neighbourhood net of observed family characteristics, $Corr_{N,NET}$.[3] However, this procedure does not necessarily produce a lower bound on the family influence. For this reason, in this application we use the procedure suggested by Altonji (1988) to produce tighter upper bounds on the neighbourhood influence, which we call adjusted upper bounds. Adjusted upper bounds are used to provide tighter lower bounds on the family influence.

Estimation Method

Following Guo and Wang (2002), Mazumder (2008) and Lindahl (2009) we estimate the intraclass correlations $Corr_S$ and $Corr_N$ by using mixed models (multilevel models). To estimate the sibling correlation, Lindahl (2009) adopts the following type of mixed model:

$$y_{cfs} = \gamma_0 + u_{cf} + u_{cfs} \qquad (2)$$

where γ_0 is the overall mean of y_{cfs}, u_{cf} is a family random component i.i.d. as normal with mean zero and variance σ_f, u_{cfs} is a child specific error term normally i.i.d. with mean zero and variance σ_u^2, and u_{cf} and u_{cfs} are mutually independent.

The error terms u_{cfs} in models (1) and (2) are identical. Moreover, since siblings living in the same family share the same neighbourhood the random family component u_{cf} in model (2) captures both neighbourhood and family effects and is identical to ($\beta'Z_c+\alpha'X_{cf}$) in model (1). The sibling correlation $Corr_S=V(\beta'Z_c+\alpha'X_{cf})/V(y_{cfs})$ is equal to $\sigma_f^2/(\sigma_f^2+\sigma_u^2)$ and can be estimated consistently by restricted maximum likelihood of model (2) as suggested by Mazumder (2008).

Similarly, Lindahl (2009) estimates the correlation between pupils living in the same neighbourhood by considering the following mixed model:

$$y_{cfs} = \beta_0 + \varepsilon_c + \varepsilon_{cfs} \qquad (3)$$

[3] Lindahl (2009) for example estimates $Corr_{NET,S}$ by adding observed family characteristics into a mixed model (see model 3 below).

where β_0 is the overall mean of y_{cfs}, ε_c is a neighbourhood random component i.i.d. as normal with mean zero and variance σ_c^2, ε_{cfs} is a pupil specific error term normally i.i.d. with mean zero and variance σ_ε^2, and ε_c and ε_{cfs} are mutually independent.

Model (3) obviously differs from model (1). In the following we clarify how the random components ε_c and ε_{cfs} in model (3) relate to the components of model (1). By formalizing the link between these two models we are also able to show that the lower bound for the family influence is equal to the influence of what we call the *relative family effect*.

The neighbourhood random component ε_c in model (3) differs from the neighbourhood influence ($\beta'Z_c$) in model (1) except when there is no sorting of families into neighbourhoods i.e. if $Cov(\alpha'X_{cf},\alpha'X_{cf'})=Cov(\alpha'X_{cf},\beta'Z_c)=0$. More in general, in the presence of sorting, ε_c captures ($\beta'Z_c$) as well as the variation of ($\alpha'X_{cf}$) across neighbourhoods. If we assume that this variation across neighbourhoods is equal to the variation of ($\alpha'\,\overline{X}_c$), where $\overline{X}_c$ is the average of the family characteristics in neighbourhood c, then

$\varepsilon_c=\beta'Z_c+\alpha'\,\overline{X}_c$,

$\varepsilon_{cfs}= \varepsilon_{cf}+u_{cfs}$ and

$\varepsilon_{cf}= \alpha'\,(X_{cf}- \overline{X}_c)$,

where $\beta'Z_c$, $\alpha'X_{cf}$ and u_{cfs} are the neighbourhood and family effects and error term as in model (1).

Given assumptions imposed by model (1), it is easy to prove that the independence conditions between ε_c and ε_{cfs} and between ε_{cfs} and $\varepsilon_{cf's'}$ imposed by model (3) are satisfied if the following two assumptions hold,

A1. independence between $\varepsilon_{cf}=\alpha'\,(X_{cf}- \overline{X}_c)$ and ($\beta'Z_c$),

A2. independence between ε_{cf} and $\varepsilon_{cf'}=\alpha'\,(X_{cf'}- \overline{X}_c)$.

Both assumptions are quite credible. Assumption A1 is satisfied if we assume that the deviation of family characteristics from the neighbourhood mean ($X_{cf}- \overline{X}_c$) are independent of the neighbourhood characteristics Z_c, whereas assumption A2 holds if there is independence between ($X_{cf}- \overline{X}_c$) for two unrelated children living in the same neighbourhood.

The random family component $\varepsilon_{cf}=\alpha'(X_{cf}- \overline{X}_c)$ captures the effect of deviations of family characteristics of family f in neighbourhood c from the average family characteristics in that neighbourhood. So it measures that part of the family influence arising from a family having characteristics which differ from the average of others living in the same neighbourhood. We call this effect the *relative family effect*.

Under assumption A1 and A2 the correlation between unrelated children living in the same neighbourhood becomes

$$Corr_N= [V(\beta'\mathrm{Z}_c)+\mathrm{Var}(\alpha'\,\overline{X}_c)+2\,\mathrm{Cov}(\alpha'\,\overline{X}_c\,,\beta'\mathrm{Z}_c)]/V(y_{cfs})=V(\beta'\mathrm{Z}_c+\alpha'\,\overline{X}_c)/V(y_{cfs}), \quad (4)$$

and it is equal to the ratio $\sigma_c^2/(\sigma_c^2+\sigma_\varepsilon^2)$, which can be estimated by restricted maximum likelihood of model (3).[4]

Furthermore, the difference between $Corr_S$ and $Corr_N$

$$Corr_S\text{-}Corr_N= \mathrm{Var}[\alpha'(X_{cf}\text{-}\,\overline{X}_c)]/V(y_{cfs}),$$

not only provides a lower bound for the influence of family characteristics, but can also be interpreted as that part of the variance $V(y_{cfs})$ explained by the relative family effect, which we call the influence of the *relative family effect*.

Heterogeneity of the neighbourhood and family variance components

In models (2) and (3) we have assumed that the random components have the same constant variance for all pupils. In our empirical application we relax this restrictive assumption. We extend model (3) and allow both the neighbourhood component ε_c and the residual error ε_{cfs} to have variance which changes between pupils living in urban and rural neighbourhoods. In other words model (3) becomes:

$$y_{cfs} =\beta_0+\varepsilon_{R,c}\mathrm{d}_{R,cfs}+\varepsilon_{U,c}\mathrm{d}_{U,cfs} +\varepsilon_{cfs} \qquad (4)$$

[4] Notice that restricted maximum likelihood estimation of σ_c^2 represents the covariance between all pairs of children living in a same neighbourhood, including pairs of siblings. Therefore σ_c^2 could in part capture the family effect. This implies that $\sigma_c^2/(\sigma_c^2+\sigma_v^2)$ is probably an upper bound on $V(\beta'\mathrm{Z}_c)\,/V(y_{cfs})$ which is less tight than $Cov(y_{cfs},y_{cf's'})/V(y_{cfs})$ computed using only pairs of unrelated children living in the same neighbourhood. Solon et al. (2000) compute the neighbourhood correlation by excluding pairs of siblings. More precisely, they compute the covariance between y_{cfs} for all possible pairs of unrelated children within each neighbourhood and then combine these within covariances using different weighting methods to take account of the unbalanced structure of the data. When the data are not extremely unbalanced these different weighting methods produce similar results (see Solon et al. 2000 and Rauum et al. 2006). However, this result does not hold in general (for example data used to estimate school mates' correlation typically has a lot of variation in the number of children across schools) and it is difficult to choose among different weighting methods which are, after all, arbitrary. For this reason we prefer to adopt mixed models for the estimation of our correlations.

where β_0 is the overall mean, $d_{R,cfs}$ and $d_{U,cfs}$ are dummy variables taking value *1* if the neighbourhood *c* is rural and urban respectively, $\varepsilon_{R,c}$ and $\varepsilon_{U,c}$ are family components mutually independent and i.i.d. as normal with mean zero and variances σ_U^2 and σ_R^2, ε_{cfs} is independent of the neighborhood random component and independently normally distributed with mean zero and variance $\sigma_{\varepsilon U}^2$ for urban neighbourhoods and $\sigma_{\varepsilon R}^2$ for rural neighbourhoods. Given this new model the correlation between two pupils living in the same rural neighbourhood is $Corr_{RN}=\sigma_R^2/(\sigma_R^2+\ \sigma_{\varepsilon R}^2)$, whereas the correlation between two pupils living in the same urban neighbourhood is $Corr_{UN}=\sigma_U^2/(\sigma_U^2+\ \sigma_{\varepsilon U}^2)$.

Similarly we extend model (2) to allow the variance of the family component ε_{cf} and the residual error term u_{cfs} to vary for pairs of mixed sex siblings, pairs of brothers and pairs sisters, i.e. we consider the following model:

$$y_{cfs} = \gamma_0+\varepsilon_{FF,cf}d_{FF,cfs}+\varepsilon_{MM,cf}d_{MM,cfs} +\varepsilon_{FM,\, cf}d_{FM,cfs} +u_{cfs} \tag{5}$$

where γ_0 is the overall mean of y_{cfs}, $d_{k,cfs}$ are dummy variables for different typologies of siblings ($d_{FF,cfs}$ for a pair of sisters, $d_{MM,cfs}$ for a pair of brothers and $d_{FM,cfs}$ for a pair of mixed sex siblings), $\varepsilon_{FF,cf}$, $\varepsilon_{MM,cf}$ and $\varepsilon_{FM,cf}$ are random family components i.i.d. as normal with mean zero and variances σ_{FF}^2, σ_{MM}^2 and σ_{FM}^2 respectively, u_{cfs} is independent of the family neighborhood component and independently normally distributed with mean zero and variances σ_{uFF}^2, σ_{uMM}^2 and σ_{uFM}^2. Then the correlation between sisters is $Corr_S^{FF}=\sigma_{FF}^2/(\sigma_{FF}^2+\sigma_u^2)$, between brothers it is $Corr_S^{MM}=\sigma_{MM}^2/(\sigma_{MM}^2+\sigma_u^2)$, and between mixed gender sibling it is $Corr_S^{FM}=\sigma_{FM}^2/(\sigma_{FM}^2+\sigma_u^2)$.

We also estimate an extension of model (5) to allow the variance of the family component ε_{cf} to vary between twins and non-twins siblings, i.e. we consider the following model

$$y_{cfs} = \gamma_0+\textstyle\sum_{k=1}^{K} \varepsilon_{k,cf}\, d_{k,cf}+\ u_{cfs} \tag{6}$$

where γ_0 is the overall mean of y_{cfs}, $d_{k,cfs}$ are *K* dummy variables for different typologies of siblings, $\varepsilon_{k,cf}$ is a family component i.i.d. as normal with mean zero and variance σ_k^2, u_{cfs} is independent of the family neighborhood components and independently normally distributed with mean zero and variance that changes by typology of sibling. To be more specific we use 2 sets of dummies to allow for different correlations between

1. twins of different gender, twin brothers, twin sisters ($d_{T,FM,cf}$, $d_{T,MM,cf}$ and $d_{T,FF,cf}$),
2. non-twin siblings of different gender, non-twin brothers, non-twin sisters ($d_{NT,FM,cf}$, $d_{NT,MM,cf}$ and $d_{NT,FF,cf}$).

This model allows us to estimate the correlations between monozygotic and dizygotic sibling as explained below.

Finally, focusing on non-twin siblings we also estimate two new models to evaluate whether sibling correlations decrease when the age gap between them widens, and whether it shrinks for siblings with similar schooling starting age. These models are identical to model (6) but adopt the following two different definitions of sibling typology:

1. non-twin siblings with an age gap of between 10 and 14, between 15 and 19, between 20 and 24 and above 24 months, and
2. non-twins siblings with a starting school age gap of 0, 1, 2 and 3 quarters.

Identification of correlations for dizygotic and monozygotic twins

Model (6) allows us to estimate the variance of the random family component and of the error term (hence the sibling correlation) separately for dizygotic (DZ) and monozygotic (MZ) twins. In the following we show how this is possible even in situations where we cannot distinguish MZ and DZ twins in the data set as is the case in our application.

The identification of the variance of the family component for DZ twins of different gender, $\sigma^2_{DZ,FM}$, is straightforward because there are no MZ twins of different sex. Therefore the correlation for mixed gender DZ twins is equal to the correlation for twins of different gender, which can be estimated using model (6). We compute the corresponding variance for DZ twin brothers, $\sigma^2_{DZ,MM}$, (sisters $\sigma^2_{DZ,FF}$) as the sum of the variance of mixed sex DZ twins and the gap in the variance between non-twin brothers (sisters) and non-twin siblings of different sex, which presumably is a good approximation of the corresponding difference between DZ twin brothers (sisters) and mixed sex twin siblings.

The computation of the variance of the family component for MZ twin brothers, $\sigma^2_{MZ,MM}$, and MZ twin sisters, $\sigma^2_{MZ,FF}$, is slightly more complicated because we are able to identify twin sisters and twin brothers but we cannot distinguish between MZ and DZ twins. To compute $\sigma^2_{MZ,MM}$ and $\sigma^2_{MZ,FF}$ we exploit the fact that

$$\sigma^2_{T,MM} = \sigma^2_{MZ,MM}\, p_{MZ,MM} + \sigma^2_{DZ,MM}\, p_{DZ,MM}, \qquad (7)$$

where $\sigma^2_{T,MM}$ is the variance of the family component for all twin brothers (including MZ and DZ twins), and $p_{MZ,MM}$ and $p_{DZ,MM}$ are the proportions of twin brothers who are MZ and DZ twins ($p_{MZ,MM}$+$p_{DZ,MM}$=1). The unknown term $\sigma^2_{MZ,MM}$ can be computed as a function of $\sigma^2_{T,MM}$, $\sigma^2_{DZ,MM}$, $p_{MZ,MM}$ and $p_{DZ,MM}$. We have already shown how to estimate $\sigma^2_{DZ,MM}$ and $\sigma^2_{T,MM}$ directly using model (6). $p_{MZ,MM}$ and $p_{DZ,MM}$ can be estimated making use of the fact that 50% of DZ twins are different gender twins, 25% are DZ twins sisters and the remaining 25% are DZ twin brothers. Therefore the total number of DZ twin brothers and sisters can each be estimated to be half the number of different gender DZ twins, $N_{DZ,FM}$, and we can compute

$$p_{DZ,MM} = 0.5\ N_{DZ,FM}/N_{T,MM} \text{ and } p_{MZ,MM} = 1 - p_{DZ,MM}$$

where $N_{T,MM}$ is the total number of twin brothers. Replacing the computed values for $\sigma^2_{DZ,MM}, \sigma^2_{T,MM}$, $p_{MZ,MM}$ and $p_{DZ,MM}$ in equation (7), ultimately we can derive the variance of the random family components for MZ brothers, $\sigma^2_{MZ,MM}$. Similarly we can estimate the corresponding variance for MZ sisters, $\sigma^2_{MZ,FF}$.

Following the same line of reasoning we can derive the variance of the error component separately for MZ and DZ twins. Finally we can compute the correlation separately for MZ and DZ twins as the ratio between the family component variance and the sum of the family component and error variances.

4. Data

The analysis is based on the National Pupil Database (NPD). This is a longitudinal register dataset for all children in state schools in England which combines pupil level attainment data with pupil characteristics as they progress through primary and secondary school. It also holds individual pupil level attainment data for pupils in independent schools who partake in the tests/exams. Pupil characteristics are collected in annual school censuses and include, for example, age, gender, ethnicity, the pupil's language group, a low-income marker and information on any Special Education Needs. Pupil level attainment data during compulsory schooling includes Foundation Stage Profiles as assessed by teachers at age 5 as well as National Curriculum assessments typically taken at ages 7, 11, 14 and 16 that comprise a mixture of teacher-led and test-based assessment depending on the age of the pupils.

The advantage of using the NPD for our analysis is that it is a census and as such contains the population of all pupils in state schools. It allows us to identify the whole set of siblings and neighbouring children of the relevant age groups in state schools and thus has a very large sample size. This makes it possible to study sibling and neighbour correlations for various sub-groups, and to assess how the relative importance of family and neighbourhood on educational attainment differ over time.

Sibling and twin definition

The NPD includes address data, released under special conditions, which allows us to match siblings in the data set. The first year that full address details were collected in the NPD across all pupil cohorts was 2007. Siblings are therefore defined as pupils in state schools aged 4-16 and living together at the same address in January 2007. Siblings that are not school-age, those in independent schools and those living at different addresses in January 2007 are excluded from our sibling definition. Step and half siblings are included if they live at the same address, and we are not able to distinguish them from biological siblings.

We define as living at the same address those pupils with identical postcodes and house number/house name, as well as flat and block number where applicable. Extensive data cleaning was necessary to extract information on house number or name, flat and block number, as data on these items was not always entered in the dedicated fields, and occasionally one field contains information relating to two items, e.g. 'Flat 2, Merton House'. Special attention was given to the cleaning and extraction of flat and block information as we assume that a higher proportion of disadvantaged pupils live in flats than houses. The matching of siblings was carried out using 1) postcode and house number/name for addresses with no flat or block number; 2) postcode, house number/name and flat number for addresses without block number; 3) postcode, house number/name, flat and block number; 4) postcode, flat and block number where house number/name was missing. Of the 7.246 million pupil files with address information contained in the 2007 school census, only 4,158 cases had insufficient address information to produce a match using these criteria, and 1,212 cases were dropped where more than ten siblings were identified at an address, and it is possible that they were falsely identified as siblings (false positives).[5]

We define as twins any pair of siblings – living at the same address - that have the same month and year of birth. There is the possibility that this twin definition includes unrelated same-aged children living at the same address, but we do not expect this to occur often enough to influence our results. The twins defined in this way include both MZ and DZ twins which we are not able to identify separately. In the previous section we described how to derive MZ and DZ twin correlations. For Key Stage 2 (Key Stage 4) we have 6,319 (6,010) different gender (DZ) twin pairs out of a total of 24,018 (21,716) twin pairs in our sample. Because of the assumption that half of DZ twin pairs will be different gender and half will be same gender, we infer that we have a total of 12,638 (12,020) DZ

[5] Even after extensive data cleaning there will be false positives in the data. These are, for example, pupils living at the same house number within a postcode but different streets; pupils living in different flats/blocks at same house number where this could not be identified; pupils living in boarding houses. Likewise, we expect to have false negatives, i.e. cases where siblings live at the same address but we have not been able to identify this. This may occur through data input errors (typos), omissions or entering more than one item of address information in a field where the correct information could not be extracted. However, in the vast majority of cases the address information and hence the matching of siblings was unambiguous

twin pairs in the sample. This means that there are 11,380 (9,696) MZ twin pairs which is 47% (45%) of the twin sample.

Neighbourhood definition

We define a pupil's neighbourhood in terms of where he or she lived at the time of the 2007 school census. In many cases, however, the family may have lived elsewhere before or indeed after 2007. We therefore assume that the 2007 neighbourhood is a good proxy for longer-run neighbourhood environment. Previous research has shown that even when families move, the neighbourhoods they move to are usually similar to the ones they move from (Kunz et al. 2003). Rabe and Taylor (2010) show that this is particularly true of school-age children in Britain. We define neighbourhoods at the level of lower layer super output areas (LSOAs). There are 32,482 LSOAs in England which were constructed using measures of proximity (to give a reasonably compact shape) and social homogeneity (type of dwelling and type of tenure, to encourage areas of similar social background). Each LSOA has constant boundaries and a mean population of 1,500 and a minimum of 1,000 individuals. Although LSOAs are primarily a statistical geography and thus far from being a perfect definition of a neighbourhood, they do allow more meaningful fine-grained area analysis at the local level than the more heterogeneous Census tracts or wards.

Outcome and observed background

The outcomes of interest are test results at two different stages of a pupil's school career, at the end of primary school (Key Stage 2) and at the end of compulsory schooling (Key Stage 4). In year 6, usually at age 11, pupils take National Curriculum tests in the three core subjects of English, Mathematics and Science. These provide records of attainment in the subjects, including separate levels for reading and writing as part of the overall English grade. At the end of compulsory schooling, usually at age 16, pupils enter General Certificate of Secondary Education (GCSE) or equivalent vocational or occupational exams. GCSEs are not compulsory, but are by far the most common qualification. Pupils decide which GCSE courses to take, and because English, Mathematics and Science are compulsory study subjects, virtually all students take GCSE examinations in these topics, plus others of their choice, with a total of ten different subjects normally taken. In addition to GCSE examinations, a pupil's final grade may also incorporate coursework elements. In this paper we focus on the GCSE results in the core subjects English, Mathematics and Science which makes the outcome directly comparable to the Key Stage 2 results, and we would argue that they are closer to measuring the ability of a pupil than the all-subject score (the sum of the points obtained in each

GCSE or equivalent exam). The inclusion of all subjects chosen by a pupil is likely to inflate the outcome by high attainment in subjects which are 'easier', such as information technology for example. We also run our models on the all-subject score for comparison.

We focus on GCSEs because they mark the first major branching point in a young person's educational career. Poor GCSE attainment is a considerable obstacle which precludes young people from pursuing more advanced educational courses. Young people with low levels of GCSE attainment are usually more likely to leave education at the minimum school leaving age and their qualification level frequently disadvantages them in the labour market. Lower levels of GCSE attainment are also likely to have a longer term impact on experiences in the adult labour market (McIntosh 2002). Key Stage 2 National Curriculum tests form a good point of comparison at a younger age because they are taken at the end of primary school and as such equally mark a turning point in the pupil's school career. Moreover, schools give most attention to these tests rather than those taken at other times because schools are likely to be judged by parents on the outcomes and they play a prominent role in setting up school league tables. Finally, Key Stage 2 and 4 exams are marked externally and contain fewer teacher assessments and therefore arguably contain less measurement error than Key Stage 1 and 3 exams.

In the Key Stage 2 exams, pupils can usually attain a maximum of 33 points in each subject, but teachers will provide opportunities for very bright pupils to test to higher levels. The points are then transformed into levels of achievement which are reported back to pupils and parents. We use as an outcome measure the average points achieved across the three core subjects English, Mathematics and Science. In Key Stage 4 pupils receive a grade for each GCSE course, where pass grades include A*, A, B, C, D, E, F, G. Those who fail the course receive an U (unclassified). To derive a continuous outcome variable from these GCSE grades we use a scoring system developed by the Qualifications and Curriculum Authority which assigns points to each of the achievable grades. A pass grade G receives 16 points, and 6 points are added for each unit improvement from grade G. The total point score is the sum of the points obtained in English, Mathematics and Science. Students who do not pass any GCSE receive a score of zero. We refer to this point score as the Key Stage 4 score.

In the UK, like in many other countries, girls consistently outperform boys in their educational attainment (e.g. Burgess et al. 2004). To make test scores comparable between boys and girls, we purge them of the gender differences. We regress test scores on gender and use the residuals from these regressions as outcome variables.

The NPD annual school census allows identification of a number of family background variables which we use to tighten the upper bound on the neighbourhood effect. These include binary variables coding whether or not a pupil is of white British ethnicity and whether or not the first language spoken at home is English. Moreover, we can identify whether or not a pupil is eligible for free school meals (FSM). FSM eligibility is linked to parents' receipt of means-tested benefits such as income support and income-based jobseeker's allowance and has been used in many studies as a low-income marker. Finally, we use as family background variable the number of siblings in 2007 and its square. This is an approximation to the true number of siblings as it is derived from our matching of pupils at the same address in 2007 and only includes school-age siblings in state schools at that point in time.

We are also able to merge geographically coded data into the data set, using LSOA identifiers. We restrict this to an indicator of whether a neighbourhood lies in a rural or urban area, where urban is defined as settlements with a population over 10,000.

Estimation sample

For our analysis we select two samples from the National Pupil Database. The first is used to estimate the neighbour correlations and therefore includes all pupils (singletons and siblings) that took Key Stage 4 exams in 2007 or in one of the two following currently available years (2008, 2009), totaling 1.725 million English pupils. For these pupils we also have their exam results at Key Stage 2. We exclude Key Stage 4 years before 2007 because we would not be able to trace and match pupils leaving school after their GCSE exams, pre 2007, in the 2007 address data. The sample we select from the NPD thus includes only neighbouring children (and siblings) that are closely spaced, i.e. one and two years apart in the school year. We remove all pupils with missing data on any of the background variables from the dataset which leads to a reduction of 3.2%. There are also some missing and zero-value cases for Key Stage 2 and 4 scores, 4.4% and 9.8% respectively, which we drop from our analyses.[6] But, to avoid unnecessarily reducing the sample size, we retain pupils with missing or zero information on Key Stage 2 when analyzing Key Stage 4 and vice versa.

The second sample is used to estimate sibling correlations and therefore concentrates on siblings only, now excluding singleton pupils. The resulting sample includes 347,793 siblings. The estimation of the correlation between all possible distinct pairs of siblings within each household has as unit of analysis the sibling-pair. Therefore we expand the dataset to include all sibling pair combinations within each household producing a number of children-pair observations of 373,270. In the vast majority of cases there are only two siblings living in the same household and taking GCSE

[6] The larger number of missing outcomes in Key Stage 4 is partly a result of concentrating on core GCSE subjects as an outcome which excludes pupils choosing to take vocational or occupational exams.

exams in the 2007-2009 period, and the percentage of two-sibling households with respect to the total households is 96%. In some parts of the analysis we further partition this sample into a sample of twins and a sample of non-twin siblings.

Table 1: Sample description

	All pupils sample		Siblings sample	
	mean	Std. deviation	mean	Std. deviation
Key stage 2 score, girls	27.28	4.13	26.97	4.34
Key stage 2 score, boys	26.91	4.38	26.67	4.54
Key stage 4 score, girls	119.01	28.11	118.14	28.90
Key stage 4 score, boys	115.60	28.55	115.08	29.15
male	0.51		0.51	
twins	0.03		0.15	
number of school-age siblings	1.90	0.95	2.68	0.97
white British	0.83		0.80	
first language English	0.91		0.88	
free school meal eligible	0.13		0.16	
urban	0.82		0.81	
pupils per neighbourhood, KS2	57.30	17.78	13.64	6.56
Pupils per neighbourhood, KS4	53.74	16.67	12.72	6.18
Number of observations	1,724,851		347,793	

Notes: National Pupil Database, 2007-2009: Pupils taking GCSE or equivalent exams in 2007-2009. Non-missing cases of Key Stage 2 score 1,653,957/328,683; non-missing cases of Key Stage 4 score 1,560,600/308,401 (all pupils/siblings sample).

Table 1 describes main characteristics of the full and siblings estimation samples. In both samples girls achieve higher mean exam scores at Key Stages 2 and 4 than boys do. For the sample containing all pupils, the table shows that half of the pupils are male and 3% of pupils are twins. On average there are 1.9 school-age children in every household with at least one pupil taking Key Stage 4 exams over the time-period 2007-2009. 83% of the pupils in the sample are of white British ethnicity, and 91% speak English as their first language. 13% of pupils in the full sample are eligible for free school meals and 82% of them live in a neighbourhood which is located in an urban area. For each neighbourhood we observe on average about 54 pupils taking Key Stage 4 exams and 57 taking Key Stage 2 exams. This is a considerably larger sample size than those used in previous studies. The characteristics of the pupils contained in the sibling sample differ from the full sample in that, as expected, the number of school-age children per household (2.7) as well as the proportion of twins (15%) is higher. Moreover, the proportion of pupils of white British origin and those speaking English as their first language is slightly lower in the sibling sample, whereas the proportion of children that are eligible for free school meals is higher at 16%. There is also a slight variation in the mean educational attainment between the two samples, with pupils in the sibling sample attaining lower scores on average than those in the full sample.

5. Results

Estimates of upper and lower bounds on the neighbourhood and family influence

Table 2 presents neighbour and sibling correlations in Key Stage 2 and 4 attainment. The upper panel shows the correlations between neighbouring pupils and was estimated using the full sample.[7] The neighbour correlation can be interpreted as the proportion of the variation in educational attainment explained by the neighbourhood influence, and this will be an upper bound because it is inflated by neighbours' similarity in family background. The overall neighbourhood influence (model 3 in section 2) is estimated to be at most 0.102 for pupils at the end of primary school and 0.145 at the end of compulsory schooling. Adjusting the neighbour correlations using observed family characteristics as in the procedure suggested by Altonji (1988) tightens this upper bound by very little, probably because the set of observed covariates is quite limited (see notes to table 2). The neighbourhood correlations are modest and comparable in size to those obtained by Solon et al. (2000) and by Raaum et al. (2006) for years of schooling.

[7] In contrast to some previous papers we do not estimate the neighbourhood influence separately for boys and girls because we allow for the possibility of cross-gender neighbourhood influences, i.e. we assume that a pupil's peer group can consists of both boys and girls in the neighbourhood.

Table 2: Sibling and neighbour correlations at Key Stages 2 and 4

	Key Stage2 Correlation(SE)		**Key Stage 4** Correlation (SE)	
Neighbour correlation (upper bound on the neighbourhood influence)				
Neighbours, model (3)	0.102	(0.001)	0.145	(0.001)
Neighbours, adjusted, model (3)	0.096	(0.001)	0.143	(0.001)
Neighbours, urban, model (4)	0.107	(0.001)	0.151	(0.001)
Neighbours, urban adjusted, model (4)	0.099	(0.001)	0.148	(0.001)
Neighbours, rural, model (4)	0.078	(0.002)	0.119	(0.002)
Neighbours, rural adjusted, model (4)	0.077	(0.002)	0.118	(0.002)
N (neighbours)	1,653,957		1,560,600	
Sibling correlation (upper bound on the family influence)				
Siblings, model (2)	0.563	(0.002)	0.621	(0.002)
Brothers, model (5)	0.589	(0.003)	0.633	(0.003)
Sisters, model (5)	0.604	(0.003)	0.659	(0.003)
Mixed gender siblings, model (5)	0.525	(0.003)	0.588	(0.002)
N (sibling pairs)	350,968		329,103	
Difference between sibling and neighbour correlation (lower bound on the family influence) *Relative family effect*				
Siblings, models (2) and (3)	0.467	(0.002)	0.478	(0.002)
Brothers, models (5) and (3)	0.493	(0.003)	0.490	(0.003)
Sisters, models (5) and (3)	0.508	(0.003)	0.516	(0.003)
Mixed gender siblings, models (5) and (3)	0.429	(0.003)	0.445	(0.002)
Non-linear Wald tests				
	Coefficient	p-value	Coefficient	p-value
$H_{o\ \mathrm{urbanicity}}$: $Corr_{UN}$=$Corr_{RN}$, model (3)	0.037	0.000	0.034	0.000
$H_{o\ \mathrm{gender}}$: $Corr_S^{FF}$=$Corr_S^{MM}$, model (5)	0.015	0.000	0.025	0.000
$H_{o\ \mathrm{gender}}$: $Corr_S^{FF}$=$Corr_S^{FM}$, model (5)	-0.079	0.000	-0.071	0.000
$H_{o\ \mathrm{gender}}$: $Corr_S^{MM}$=$Corr_S^{FM}$, model (5)	-0.064	0.000	-0.046	0.000

Notes: National Pupil Database, 2007-2009: Pupils taking GCSE exams in 2007-2009. Neighbour estimates adjusted for: white British ethnicity, first language English, low income group, number of school-age siblings in state schools and its square. Relative family effect derived using adjusted neighbour correlation. Standard errors and Wald tests calculated using the delta method.

We let the importance of neighbourhood factors in explaining pupils' outcomes vary by whether or not the neighbourhood is located in a rural or urban area (see model 4 in section 2) . This may affect outcomes, say, because the physical proximity of neighbours in urban areas compared to rural ones may intensify the influence of the neighbourhood factors, perhaps by increasing interaction within the neighbourhood peer group. The results of these estimates are displayed in the first panel of table 2. Indeed, the upper bound on the proportion of variation in the pupil's outcome explained by rural neighbourhood characteristics is approximately 20-30% lower than the one observed for urban neighbourhoods, and this is true both for exams taken in primary and secondary school and for both adjusted and unadjusted correlations. We conduct non-linear Wald tests of the equality of neighbourhood correlations for urban and rural neighbourhoods ($Corr_{UN}$=$Corr_{RN}$) which we calculate using the delta method. The results are displayed in the bottom panel of the table and show that equality is rejected at standard levels of significance.

Secondary school pupils in England often commute outside of their neighbourhood boundaries to attend different schools, whereas this is less often the case for primary school students who usually attend the local school within the so-called catchment area of residence. This would suggest that neighbouring pupil's attainment should be more correlated at the younger than at the older age because of the difference in shared school background. We find that the opposite is the case. The adjusted neighbourhood influence rises by 49% between ages eleven and sixteen, albeit from a low initial level. A possible explanation is that at age sixteen pupils are often allowed to interact independently with neighbouring teenagers and to engage in what is known as 'hanging out' on the streets, whereas this is less common for eleven year olds. This would suggest that peer group interaction drives the neighbourhood influence on educational attainment.

The second panel of table 2 shows the upper bound of the family influence on educational attainment, i.e. the sibling correlation, estimated using the sibling sample (see model 2 in section 2). The correlation between siblings in exam results is 0.563 at Key Stage 2 and 0.621 at Key Stage 4. These results are higher than that obtained by Solon et al. (2000) for years of education using US data, with a correlation of 0.513. The sibling correlation found for Norway by Raaum et al. (2006) for years of schooling is lower at 0.42, indicating possibly higher intergenerational mobility in the Nordic welfare state than in the US and UK which are liberal welfare states.

The third panel of table 2 shows the lower bound on the family influence, which is given by the difference between the upper bound and the adjusted correlation between pupils living in the same neighbourhood. The range of possible values for the family influence on Key Stage 2 scores is (0.467, 0.563) and on Key Stage 4 scores it is (0.478, 0.621). The two intervals overlap in part so it is not possible to draw strong conclusions on whether family influence increases or decreases from age 11 to age 16. If anything, the family influence seems to increase, which is in line with previous research (Goodman and Gregg 2010).

The second and third panels of table 2 also display the upper and lower bounds on the family influence separately for different gender combinations. The results show higher correlations for same gender than for mixed gender siblings at the end of primary school and at the end of compulsory schooling. There are also differences between sister-pair and brother-pair correlations. Wald tests displayed in the bottom panel show that equality of the correlations between sisters and between brothers ($Corr_S^{\mathrm{FF}}$=$Corr_S^{\mathrm{MM}}$), between sisters and mixed sex siblings ($Corr_S^{\mathrm{FF}}$=$Corr_S^{\mathrm{FM}}$), and between brothers and mixed sex siblings ($Corr_S^{\mathrm{MM}}$=$Corr_S^{\mathrm{FM}}$) can be rejected at standard levels of significance. Note that the outcome measures we use in the analysis are purged of the mean gender differences in school outcomes between girls and boys, so that the lower mixed gender correlations are not a reflection of such differences.

While we can provide only bound estimates for the family influence, we are able to provide point estimates for the *relative family influence* i.e. for the share of the total variance in educational attainment which is explained by deviations of family's characteristics from the average in their neighbourhood. We have shown in section 2 that under some minor assumptions this relative family influence is given by the lower bound of the family influence. How much the characteristics of a family differ from other families living in the same neighborhood appears to be a very important factor in explaining pupil's attainments, indeed it does explain about 47% of the variance in the Key Stage 2 and about 48% of the variance in the Key Stage 4 exam score. These results seem to suggest a substantial contribution of family background and an extraordinarily high importance of differences in family characteristics within a neighborhood in explaining pupil's educational attainments.

Estimates of upper bounds on the family influence by sibling type

Up to now we have focused on the sibling correlation as a measure of the influence of shared family and neighborhood background on educational attainment. In addition to the family and neighbourhood factors shared by siblings there will be non-shared family and neighbourhood influences that affect siblings differently. Non-twin siblings and dizygotic twins, for example, have only half of their genes in common. Moreover, even if they share the same family events (such as family disruption and income shocks), they experience these events at different points in their life. Furthermore, parents can treat their children differently. To assess the magnitude of such differentiating influences, we compare the correlation in educational attainments of MZ twins who have identical genes with those of DZ twins and non-twin siblings who share only half of their genes. Differences between DZ twin and non-twin sibling correlations can be interpreted as the effect of growing up at the same or at a different time and hence being exposed to an environment which is more or less similar. Finally, we investigate non-twin sibling correlations by differences in the sibling age gap and the school starting age.

We begin by estimating correlations in educational attainments between twins and non-twin siblings using model (6) introduced in section 2, which allows the family random component to vary across different typologies of siblings. We do not report the lower bounds on the family influence which can be easily computed by subtracting from each sibling correlation the adjusted neighbourhood influence (0.96 and 0.143 for Key Stages 2 and 4 respectively). The first panel in table 3 shows twins correlations for all twins in the sample, not distinguishing between MZ and DZ twins at this point. The mixed gender twins are DZ twins by definition and have a considerably lower correlation in attainments than same gender twins. Our twin brother correlations can be compared to the recent results of Bjoerklund et al. (2010) who, based on Swedish data, look at twin brother correlations in IQ. Their correlations in IQ are around 0.65 while ours are higher at roughly 0.8. This seems to indicate that family background matters more in the UK than in the Nordic welfare state.

The second and third panels in table 3 show correlations for MZ and DZ twins at the end of primary school and at the end of compulsory schooling. As expected, the correlation between MZ brothers and sisters are considerably higher than they are for DZ twins. We estimate the correlation of Key Stage 2 scores to be 0.905 (0.895) for MZ brothers (sisters), and the correlation in Key Stage 4 scores to be 0.889 (0.908) for MZ brothers and sisters respectively. This is a remarkably high correlation. When estimated on the alternative outcome at Key Stage 4 which includes all subjects chosen by pupils instead of concentrating only on the core subjects English, Math and Science, the correlation is significantly lower for MZ twins. This could suggest that our outcome variable comes quite close to measuring innate ability, whereas the all-subject score is a reflection of choices taken by MZ twins. The exposition to external influences over time and personality development as well as deliberate differential treatment by parents may lead to greater between-twin differentiation in choices than indicated by measures of innate ability.

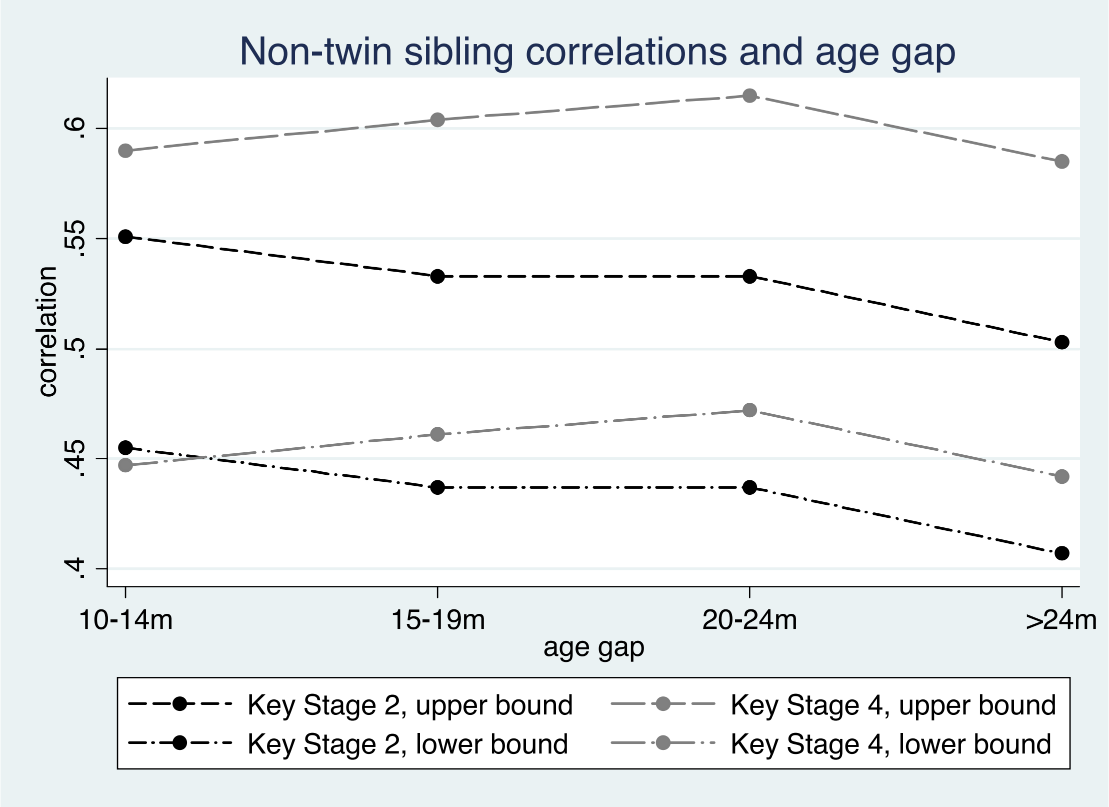
Table 3: Model (6) estimates of the sibling correlation in educational attainment for twins and non-twin siblings at Key Stages 2 and 4, by gender.

	Key Stage 2 score Correlation (SE)		**Key Stage 4 score** Correlation (SE)	
All twins				
Twin brothers	0.801	(0.004)	0.799	(0.004)
Twin sisters	0.794	(0.004)	0.821	(0.004)
Mixed gender twins	0.561	(0.009)	0.619	(0.008)
Monozygotic twins				
MZ brothers	0.905	(0.006)	0.889	(0.007)
MZ sisters	0.895	(0.007)	0.908	(0.006)
Dizygotic twins				
DZ brothers	0.558	(0.010)	0.636	(0.009)
DZ sisters	0.584	(0.010)	0.661	(0.009)
DZ mixed gender twins	0.561	(0.009)	0.619	(0.008)
Non-twin siblings				
Brothers	0.512	(0.004)	0.600	(0.003)
Sisters	0.536	(0.004)	0.623	(0.003)
Mixed gender	0.515	(0.003)	0.583	(0.003)
Non-linear Wald tests				
	Coefficient	p-value	Coefficient	p-value
H$_o$: $Corr_{MZ}^{FF}=Corr_{MZ}^{MM}$	0.057	0.000	0.057	0.000
H$_o$: $Corr_{NT}^{FF}=Corr_{NT}^{MM}$	0.024	0.000	0.023	0.000
H$_o$: $Corr_{NT}^{FF}=Corr_{NT}^{FM}$	0.022	0.000	0.040	0.000
H$_o$: $Corr_{NT}^{MM}=Corr_{NT}^{FM}$	0.003	0.568	-0.017	0.000

Notes: National Pupil Database, 2007-2009: Pupils taking GCSE or equivalent exams in 2007-2009. Standard errors and Wald tests calculated using the delta method. Wald tests for equality of correlations between DZ twins of different gender combinations not shown as the differences are by assumption identical to those of non-twin siblings, see section 2.

Comparing DZ twins with non-twin siblings (panels 3 and 4 of table 3), we find that DZ twin correlations are 9% higher than those of their non-twin siblings at Key Stage 2 and 6% at Key Stage 4. As they share the same proportion of genes (50%), these differences can only be explained by differing family and neighbourhood environmental factors. In contrast to DZ twins, siblings born at different times are exposed to different family and community environments to the extent that these environments change over time.

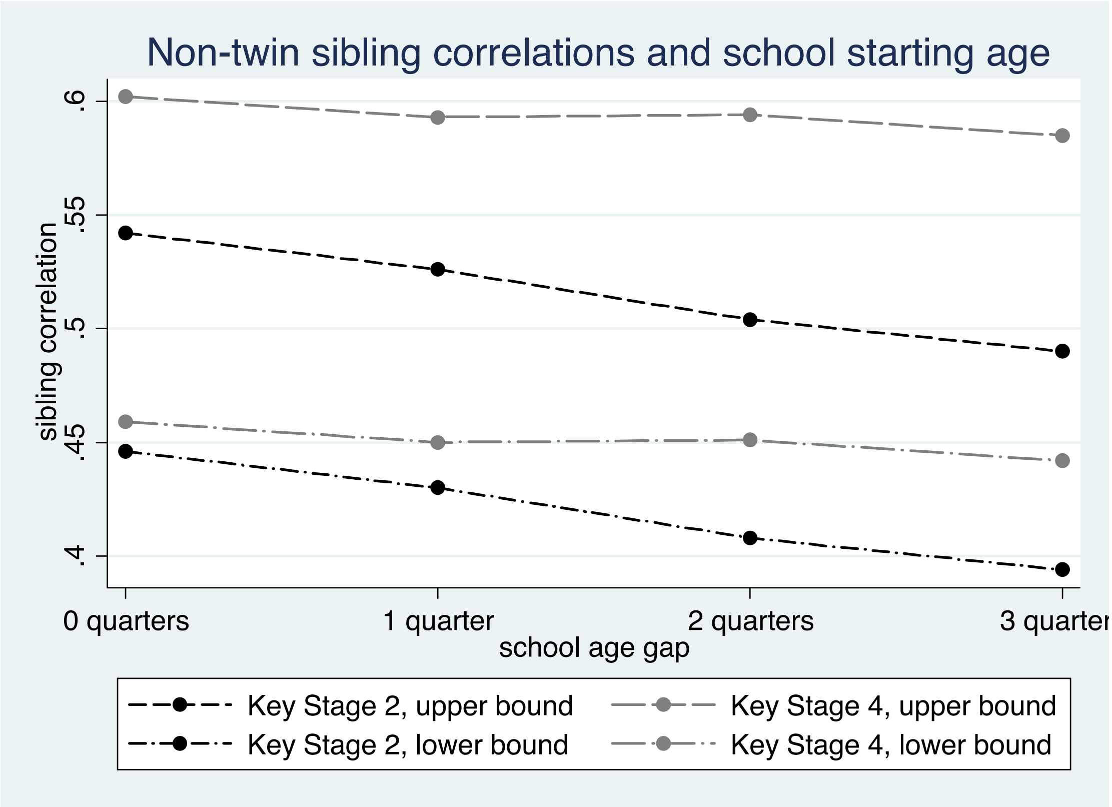
Figure 1:



Notes: National Pupil Database, 2007-2009. Model (6) estimates of upper bound of family influence. Lower bound derived using adjusted neighbor correlation.

To investigate how sibling correlations evolve with an increasing sibling age gap we estimate sibling correlations at Key Stages 2 and 4 for siblings that were born within 10-14 months of each other, for those with an age gap of 15-19 months and 20-24 months respectively, and for those born more than two years apart. We plot these correlations as well as the corresponding lower bounds on the family influence in figure 1. At Key Stage 2, shown in black, there is a negative trend in age difference with siblings spaced closer together having higher correlations in attainment than those born further apart. Non-linear Wald tests (not displayed) show that the differences between the correlations at different age gap intervals are statistically significant at the 1% level at Key Stage 2. This is consistent with the expectation that siblings spaced further apart experience a different family environment and are thus less similar. The sibling correlation declines by roughly 0.05 points in the course of a year. At Key Stage 4, shown in grey in figure 1, the relationship between age gap and sibling correlation in attainment is less clear, as the correlation appears to increase with the age gap and then decrease again for siblings spaced more than two years apart. Non-linear Wald tests (not displayed) show that some of the differences in the estimates at different age gaps are not statistically different from zero.[8] A possible explanation for this finding is that there are two conflicting forces at work that pan out differently at different ages. The first one is the decreasing similarity in family environment for widely spaced siblings that makes them more dissimilar. The second one is the fact that closely spaced siblings are in immediate competition with each other and this may lead them to diverge in their outcomes, magnifying small differences and seeking out niches within the family (Conley et al. 2007).

Figure 2:



Notes: National Pupil Database, 2007-2009. Model (6) estimates of upper bound of family influence. Lower bound derived using adjusted neighbor correlation.

Next we look at how sibling correlations evolve depending on differences in their school starting age. We have created 4 sibling groups to distinguish these effects. The first group contains all siblings whose birth month is 0-2 months apart and therefore places them in the same quarter at

[8] Specifically, the difference between age gap 10/14 and 14/19 and age gap 14/19 and 20/24 are not statistically different from zero.

school. The remaining categories group siblings who are one, two or three quarters apart in the school year because of their birth months. Figure 2 shows upper and lower bounds of the sibling correlation at Key Stages 2 and 4 by the gap in the school starting age. At Key Stage 2 there is a negative trend with siblings that have a larger school starting age gap being more dissimilar than those with a smaller gap. The differences in sibling correlation by school age gap are statistically significant and are associated with a decline of approximately 0.02 points for each quarter the siblings are further apart in the school year. By the time the pupils finish compulsory education, however, the season of birth effects have diminished. The correlations in attainment are now less than 0.01 points apart for each additional quarter of school age difference, but not all of these differences are statistically significant. This is in line with previous research showing that school starting age effects are large at the beginning of school but the differences disappear towards the end of the school career (Crawford et al. 2007).

6. Sensitivity analysis

Comparing simple sibling-pair correlations with correlations based on multilevel models. There are two main advantages of using multilevel models to estimate sibling correlations. The first is that we can consistently estimate correlations for unbalanced data (i.e. clustered data with different size clusters). This is particularly relevant for the neighbourhood correlation and less relevant in this application for the sibling correlation because in our sample we observe mainly two-sibling families. The second is that we can produce formal tests for the equality of correlations between different typologies of siblings. This comes at the cost of imposing a normality assumption. Simple sibling-pair correlations do not impose this normality assumption and produce consistent estimates of the sibling correlation at least when considering only two-sibling families. For this reason we compare the correlations for different typologies of siblings estimated by using simple sibling-pair correlations and the multilevel model (6) and considering only two-sibling families. We obtain estimated correlations which are basically identical and with confidence intervals almost completely overlapping.

Joint modeling of neighbourhood and family effects. Throughout all our analysis we estimate separate models to indentify the neighborhood and family effects. To check whether this could potentially bias our results, we also estimate a two-level model where pupils are clustered within families and families are clustered within neighbourhoods. By using this model to jointly estimate sibling and neighbourhood correlations, we find correlations which are very similar to the ones estimated using the two separate single-level models (2) and (3).

Alternative measures of educational attainments at the Key Stage 4. We use alternative measures of educational attainment at Key Stage 4, (1) the sum of the scores obtained in any the GCSE subjects taken and (2) the sum of the eight best GCSE exam scores. Both neighbourhood and sibling correlations decrease significantly when using these two alternative measures rather than the sum of the scores obtained in English, Mathematics and Science. Apart from English, Mathematics and Science the GCSEs subjects taken by a pupil reflect their personal choice. Therefore the alternative outcomes are less comparable across pupils.

Using the subsample of the two oldest siblings for each family. Our sample of all possible sibling pairs within each family could lead to an over-representation of big size families. For this reason we check whether our results change when focusing only to the two oldest siblings in each family and the results hardly change.

7. Conclusions

Our study confirms earlier research showing that growing up in the same family is much more important for explaining sibling similarity in educational attainment than growing up in the same neighbourhood. Specifically, the sibling correlation, measuring the importance both of the shared family and neighbourhood background, is computed to be 0.563 at age eleven and 0.621 at age sixteen. The proportion of variation in pupil's educational attainment explained by neighbourhood factors is 0.096 and 0.143 for pupils aged eleven and sixteen respectively. In addition to an upper bound on the family effect we are able to derive a lower bound for the family effect which has a straightforward interpretation as relative family effect. This relative family effect measures that part of the family effect arising from a family having characteristics which differ from those of other living in the same neighbourhood. The estimates show that deviations of family's characteristics from observed neighbourhood mean family characteristics account for 47% of the deviation in pupils' Key Stage 2 and 48% in pupils' Key Stage 4 attainments.

We go beyond previous research by exploring differences in sibling correlations by sibling type. Looking at sibling gender combinations we generally find that different gender siblings are less similar than sister and brother pairs respectively. By imposing a few minor assumptions we are able to distinguish MZ and DZ twin correlations although the twin types are not identified in our data. Not surprisingly, the correlations in educational outcomes are highest for MZ twins who have identical genes. They are about 0.9 for both MZ twin brothers and sisters at both Key Stages. DZ twins have a correlation in educational attainment around 0.6 both at the end of primary school and at the end of compulsory schooling. Differences between MZ and DZ twin correlations should be mainly caused by differences in the proportion of genes shared by MZ twins (100%) and DZ twins (50%). Differences in the correlation between DZ twins and non-twin siblings growing up at different times (about 0.05 points at Key Stage 2 and less at Key Stage 4) should be mainly caused by differences in the family and neighbourhood environment. The effects of growing up at different times in the same family also emerge when comparing sibling correlations by age difference. Moreover, even for closely spaced siblings there can be differences in the school starting age which are associated with lower sibling correlations at age eleven but disappear by age sixteen. Finally we explore differences in neighbourhood effects depending on whether the neighbourhood is located in an urban or a rural area. We show that the upper bound on the proportion of variation in the pupil's outcome explained by neighbourhood effects is higher for urban areas than for rural areas, suggesting that peer interaction which is presumably higher in urban areas is responsible for the effect.

References

Altonji, J. G., (1988). "The Effects of Family Background and School Characteristics on Education and Labor Market Outcomes", mimeograph, Northwestern University.

Angrist J. D. and A. B. Krueger (1992). "The Effect of Age at School Entry on Educational Attainment: An Application of InstrumentalVariables with Moments from Two Samples". *Journal of the American Statistical Association*, 418, 328-336.

Bedard, K., E. Dhuey (2006). "The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age-Effects", *Quarterly Journal of Economics*, 121: 1437-1472.

Björklund A., Hederos K. H. and M.. Jäntti (2010). "IQ and Family Background: Are Associations Strong or Weak?", *The B.E. Journal of Economic Analysis and Policy* (*Contributions*), 10(1).

Björklund, A. and M. Jäntti (2009). "Intergenerational Income Mobility and the Role of Family Background", mimeograph, Swedish Institute for Social Research, Stockholm University. In Wiemer Salverda, Brian Nolan, and Timothy M Smeeding, editors, *Oxford Handbook of Economic Inequality*, chapter 20, Oxford University Press, Oxford.

Björklund, A., M. Jäntti, G. Solon (2005). "Influences of Nature and Nurture on Earning Variation: A Report on a Study of Various Sibling Types in Sweden". In S. Bowles, H. Gintis, M. Osborne Groves (eds.), *Unequal chances: Family background and economic success*, New York: Russell Sage Foundation.

Black, S., P.J. Devereux, K.G. Salvanes (2005). "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education", *Quarterly Journal of Economics*, 120(2): 669-700.

Burgess, S., B. McConnell, C. Propper and D. Wilson (2004). "Girls Rock, Boys Roll: An Analysis of the Age 14-16 Gender Gap in English Schools", *Scottish Journal of Political Economy* 51(2): 209-229

Conley, D., R. Glauber (2008), "All in the Family? Family Composition, Resources, and Sibling Similarity in Socioeconomic status", *Research in Social Stratification and Mobility*, 26, 297–306.

Conley, D., K.M. Pfeiffer, M. Velez (2007). "Explaining Sibling Differences in Achievement and Behavioural Outcomes: The Importance of Within- and Between-Family Factors", *Social Science Research*, 26: 1087-1104.

Crawford, C., L. Dearden, C. Meghir (2007). "When You Are Born Matters: The Impact of Date of Birth on Child Cognitive Outcomes in England." Institute for Fiscal Studies, London.

Ermisch J. and E. Del Bono (2010). "Inequality in Achievements during Adolescence", ISER, University of Essex.

Goodman A. and P. Gregg (Eds) (2010). "Poorer Children's Educational Attainment: How Important Are Attitudes and Behaviour?" Report Joseph Rowntree Foundation Report www.jrf.org.uk/publications/educational-attainment-poor-children

Grilliches, Z. (1979). "Sibling Models and Data in Economics: Beginnings of a Survey", *Journal of Political Economy*, 87(5): S37-S64.

Guo, G. and J. Wang (2002). "The Mixed or Multilevel Model for Behavior Genetic Analysis", *Behavior Genetics*, 32(1): 37-49.

Kunz, J., M.E. Page and G. Solon (2003). "Are Point-in-Time Measures of Neighborhood Characteristics Useful Proxies for Children's Long-Run Neighborhood Environment?", *Economics Letters* 79 (May): 231–37.

Lindahl, Lena (2009). "A Comparison of Family and Neighbourhood Effects on Grades, Test Scores, Educational Attainment and Income – Evidence from Sweden", SOFI, Stockholm.

Mazumder, B. (2008). “Sibling Similarities and Economic Inequality in the U.S.”, *Journal of Population Economics* 21: 685-701.

McIntosh (2002). “Further Analysis of the Returns to Academic and Vocational Qualifications”, Department for Education and Skills Research Report RR370.

Page, M. E. and G. Solon (2003a). “Correlations between Brothers and Neighboring Boys in Their Adult Earnings: The Importance of Being Urban.”, *Journal of Labor Economics* 21(4): 831–55.

Page, M.E., and G. Solon (2003b). “Correlations between Sisters and Neighboring Girls in Their Subsequent Income as Adults”, *Journal of Applied Econometrics* 18: 545-562.

Powell, B., L.C. Steelman (1995). “Feeling the Pinch: Child Spacing and Constraints on Parental Economic Investments in Children”, *Social Forces* 73(4): 1465-1486.

Raaum, O., K.-G. Salvanes and E. Sørensen (2003). “The Impact of a Primary School Reform on Educational Stratification: A Norwegian Study of Neighbour and School Mate Correlations”, *Swedish Economic Policy Review* 10: 143-169

Raaum, O., K.-G. Salvanes and E. Sørensen (2006).”The Neighborhood Is Not What It Used to Be”, *Economic Journal* 116(1): 200-22.

Rabe-Hesketh S., Skrondal a., Gjessing H.K. (2008). “Biometrical Modeling of Twin and Family Data Using Standard Mixed Model Software”, *Biometrics*, 64, 280-288.

Rabe, B, and M. Taylor (2010). “Residential Mobility, Quality of Neighbourhood and Life Course Events”, *Journal of the Royal Statistical Society, Series A (Statistics in Society),* 173(3): 531-555.

Rosenzweig, M.R. (1986). “Birth Spacing and Sibling Inequality: Asymmetric Information within the Family”, *International Economic Review*, 27(1): 55-76.

Solon, G., M.Corcoran, R. Gordon and D. Laren (1991). “A Longitudinal Analysis of Sibling Correlations in Economic Status”, *Journal of Human Resources*, 26: 509-534.

Solon, G. (1999). “*Intergenerational Mobility in the Labor Market*”, in O. Ashenfelter and D. Card (eds.) Handbook of Labor Economics, vol. 3: 1761-1800.

Solon, G., M.E. Page and G.J. Duncan (2000). “Correlations between Neighboring Children in Their Subsequent Educational Attainment”, *Review of Economics and Statistics* 82 (August): 383–92.

Taubman, P. (1976). “The Determinants of Earnings: Genetics, Family, and Other Environments: A Study of White Male Twins”, *American Economic Review*, 66(5): 858-870.