The Effect of Child Weight on Academic Performance:

Evidence using Genetic Markers

Stephanie von Hinke Kessler Scholder [1]
Carol Propper [2]
Frank Windmeijer [3]
George Davey Smith [4]
Debbie A. Lawlor [4]

April 2009

Abstract
This paper examines the relationship between children's weight and academic outcomes using genetic markers as instruments to account for the possible endogeneity of body size. We use medically assessed measures of body size which are more appropriate than the generally used BMI measures. OLS results indicate that leaner children perform better in school tests compared to their heavier counterparts, but the IV results, using genetic markers as instruments, show no evidence that fat mass affects academic outcomes. We compare these IV results to those using the instruments generally adopted in this literature. We show that the results are sensitive to the instrument set and argue that several of the commonly used instruments do not meet the exclusion restrictions required of a valid instrument.

Key words: Child weight; Academic Performance; Educational Outcomes; Instrumental Variables; Mendelian Randomization; Genetic Markers; DXA; Body Mass Index; ALSPAC
JEL: I1, I2, J24

Acknowledgements:
We thank Owen O'Donnell, George Leckie and Gerard van den Berg for fruitful discussions. We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council, the Wellcome Trust and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors, who will serve as guarantors for the contents of this paper. Funding from the ESRC is gratefully acknowledged.

[1] *CMPO, University of Bristol*
[2] *CMPO and Department of Economics, University of Bristol; Imperial College Business School, Imperial College London; CEPR*
[3] *CMPO and Department of Economics, University of Bristol; Centre for Microdata, Methods and Practice, Institute of Fiscal Studies*
[4] *MRC Centre for Causal Analysis in Translational Epidemiology (CAiTE) and Department of Social Medicine, University of Bristol*

There is a growing interest in the relationship between individuals' physical traits and their economic success. The literature has focused on three attributes: beauty (Hamermesh and Biddle, 1994; Mobius and Rosenblat, 2006), height (Case and Paxson, 2008) and body size. The rise in body size across a large set of countries (OECD, 2007) makes the latter particularly pertinent. Studies that attempt to estimate the causal effect of obesity on economic outcomes report mixed results. Of those that focus on adults, some find that weight lowers wages, at least for white females (Cawley, 2004), whilst others find no significant effects (Norton and Han, 2008). Studies that focus on child weight and their academic achievement also report conflicting findings. Some find that weight lowers test scores, though only for girls (Ding et al., 2006; Sabia, 2007; Averett and Stifel, 2007), whilst others report no significant differences (Fletcher and Lehrer, 2008; Kaestner and Grossman, 2008).

This paper examines the effect of children's weight on their academic performance. We focus on children as there is growing evidence that socio-economic differences in adulthood are shaped early in life (Case, Dubotsky and Paxson, 2002; Cunha and Heckman, 2007; Johnson and Schoeni, 2007). To overcome the problem that weight may be endogenous to academic attainment, we exploit variation in genetic make-up. Genes are randomly distributed at conception. Under certain assumptions that are discussed in detail below, they are strictly exogenous. This implies that correlations between the genetic variant and the outcome of interest cannot be due to reverse causation, behavioural, or environmental factors, including those that occur *in utero*. We therefore use recently identified genetic markers for weight as instrumental variables to identify the causal effect of child weight on academic achievement.[1] These markers, in contrast with others used in recent studies in the economics literature, have been shown to be associated with weight in large population samples.

In addition to the endogeneity issue, recent literature (Burkhauser and Cawley, 2008)

[1] The use of genetic variation as instruments for endogenous variables is known in the epidemiologic literature as 'Mendelian Randomization' (Davey Smith and Ebrahim, 2003).

has identified the need for a more accurate measure of obesity than the commonly used Body Mass Index (BMI).[2] We use a direct measure of the child's body fat mass, determined by a dual-energy X-ray absorptiometry (DXA) scan and compare the robustness of the results to two other measures that proxy fat mass (BMI and Bioelectrical Impedance Analysis, BIA). Our outcome of interest is an objective, independent and comparable assessment of children's achievement. We use the child's score on the UK's nationally set exam taken by all 14-year olds in the English public school system (known as Key Stage 3 tests). We also use data with rich information on a wide range of family background variables from a large cohort of UK children, allowing us to control for several measures of mother's health and behaviour that may affect both body size and educational outcomes, but which are not observed generally in survey data.

When accounting for the endogeneity of child weight by using carefully selected genetic markers, we find that there is no causal relationship between early obesity and children's academic performance at age 14. In contrast, OLS estimates indicate that heavier children perform worse in school tests compared to their leaner counterparts.

In addition to answering our question of interest, we contribute to the broader literature that has sought to exploit newly discovered genotypes to examine the relationship between physical attributes and economic outcomes. First, we detail the specific conditions that need to be met for genetic markers to be used as instruments. These conditions have not been well defined in the current economic literature, but the increasing availability of biomedical data makes understanding of these conditions crucial to the successful use of genotypes as instruments for phenotypes. Second, we compare the performance of genetic instruments to those non-genetic instruments that have been used in the literature on the effect of child weight on educational performance. We examine an additional two sets of

[2] BMI is weight in kg divided by height in metres squared.

instruments: maternal pre-pregnancy weight (as in Averett and Stifel, 2007; Sabia, 2007) and the child's lagged weight categories (Kaestner and Grossman, 2008). We argue that these do not satisfy the exclusion criteria for a valid instrument and produce misleading results.

I. Motivation

There are three ways in which childhood obesity may be related to educational outcomes. First, obesity could cause lower academic performance. Second, poor results may cause obesity. And third, instead of there being a causal relationship, the association may be driven by other unobserved characteristics relating to both weight and academic outcomes.

In terms of the first, various pathways have been suggested through which a child's excess body weight may affect its educational outcomes. First, obese children have been shown to be significantly more absent from school (Geier et al., 2007), which in turn may affect their educational outcomes. Second, associations between obesity and health during childhood may affect educational achievement. For example, obese children are more likely to have sleep apnoea or other sleeping disorders (Redline et al., 1999), which are negatively related to cognitive functioning. Third, overweight children may be treated differently by teachers, parents and peers, affecting their (learning) environment (Schwartz and Puhl, 2003). Obese children may be bullied, lowering their self-esteem and harming their educational development. Lower popularity may also lead to ostracism; if this means that, rather than engaging in social activities, children spend more time on their studies, this would lead to better school outcomes. Similarly, fewer recreational activities can increase children's weights and simultaneously increase the time that can be spent on studying (Kaestner and Grossman, 2008).

The reverse causal relationship implies that poor school outcomes cause differences in body weights, rather than obesity causing differences in academic performance. Perhaps

some children eat excessively to compensate for doing poorly at school. Or conversely, stress caused by poor achievement may lead to reduced appetites and subsequent weight loss (Sabia, 2007).

In terms of the third, even after controlling for an extensive set of background characteristics, there may still be a host of unobserved family or child factors that are related to both obesity and outcomes. For example, socio-economic position may affect both diet and attitudes to schooling, and with that affect both weight and school performance. Unobserved time discount rates may be positively related to child weight (i.e. children may be overweight because they place less value on the future) and – with the same reasoning – negatively related to the child's human capital investment and educational outcomes. Similarly, rather than engaging in physical activity, overweight children may have an (unobserved) preference (e.g. level of discipline) to devote this time to studying which in turn increases their academic achievement.

The current studies of the relationship between children's weight and educational performance that do not exploit genetic variation reach mixed conclusions. Sabia (2007) examines 14-17 year-olds from the National Longitudinal Study of Adolescent Health (Add Health) and finds a negative relationship between white girls' weights and their educational achievement. These results are robust to an individual fixed effects approach and to a specification using IV, with mother and father's self-reported obesity status as instruments. But the estimates of any effect are small: it takes a weight difference of approximately 150 pounds (68 kg) for there to be a half-letter grade difference in Grade Point Average, all else equal. Averett and Stifel (2007) examine the effects of two types of malnutrition – being underweight or overweight - on educational attainment using data on the children of female cohort members of the National Longitudinal Survey of Youth 1979 (C-NLSY79), from 1986 to 2002. Focusing on elementary school-age children aged 6-13, they show that malnourished

children have lower educational outcomes compared to well-nourished children. This finding remains when using individual fixed effects or IV, with mother's pre-pregnancy BMI and its square as instruments. Kaestner and Grossman (2008) examine 5 to 12 year-olds of the C-NLSY79 between 1986 and 2004. They regress the change in educational attainment over two years on the level of weight and use IV, specifying the child's lagged under- and overweight percentiles as instruments for current weight and find no evidence that children's academic progress is affected by their weight.

II. Use of Genetic Markers as Instruments for Children's Weight

A. Estimation of the Effect of Weight on Academic Performance.

We examine the impact of a child's weight on their educational outcomes at age 14. We model weight as part of a child education production function:

$$S_i = f\left(W_i, X_i, P_i, u_i\right), \tag{1}$$

where S_i, the academic performance of child *i,* is a function of child weight W_i, a set of child and family background characteristics X_i and parental health and behaviour P_i. The variable u_i represents the unobserved component, which includes both unobserved child attributes and unobserved parental/family behaviour. If we begin with a simple OLS model:

$$S_i = \beta_0 + \beta_1 W_i + u_i \tag{2}$$

the parameter of interest, the relationship between child weight and academic achievement, is β_1. To (2) we can include the additional sets of covariates in X_i and P_i, which allows us to explore how the relationship between child weight and academic achievement changes when controlling for various observed inputs in the child education production function.

The possible endogeneity of child weight is characterised by the fact that the unobservable confounders u_i determine educational outcomes S_i, but may also determine

weight W_i, leading to biased OLS estimates. The bias can be expressed as the final term in:

$$\hat{\beta}_1^{OLS} = \beta_1 + \frac{Cov(W_i u_i)}{Var(W_i)} \tag{3}$$

The existing literature generally tries to deal with the endogeneity problem by either estimating child fixed effects models when children are observed multiple times, or by using IV. The fixed effects specification deals with the endogeneity problem only if the unobserved factors that jointly affect child weight and educational outcomes are constant over time. Any time-varying unobservables such as changes in children's peer groups, and changes in family or household circumstances that affect both school performance and children's weight (gain) may therefore still bias the estimates. Additionally, the fixed effects model does not deal with any reverse causality running from school outcomes to obesity.

The IV method estimates the causal effect β_1 by introducing instrumental variables Z_i that are associated with W_i, but only associated with S_i indirectly through its association with W_i. Formally, the assumptions are:

Assumption 1: $E(Z_i W_i) \neq 0$

Assumption 2: $E(Z_i u_i) = 0$

Figure A1, Appendix A, summarises the model assumptions for IV estimation in a directed acyclic graph (DAG) representing conditional independence relationships implied by the model for S, W and u, with the core conditions that must be satisfied by an instrumental variable Z. Each node represents a variable, with square nodes being observed and circular nodes being unobserved. Directed edges between the variables indicate causal direction.

Two instrument sets commonly used in studies of child weight on educational attainment are the child's or parental/maternal weights in previous periods.[3] Both are likely

[3] Studies that examine the effects of body size on adult labour market outcomes have also used sibling's weights as instrumental variables (e.g. Cawley, 2004; Norton and Han, 2008). In stark contrast and similar to this paper,

to be related to the child's current weight, satisfying the first assumption. However, whether the instruments satisfy the second assumption is debatable.

The use of a child's weight at earlier ages is often justified by arguing that this deals with problems of reversed causation, as current outcomes cannot affect previous weight. The correlation between children's past and current weights, however, can be as high as 0.95. Such substantial correlation suggests that a child's earlier weight is more or less a perfect predictor of its current weight and raises doubts about its use as an IV. For previous weight to be a valid instrument, the component of child weight that is uncorrelated to its previous weight needs to contain all of the correlation with the unobserved characteristics of school performance u_i. Put another way: all factors contributing to the high correlation between weight and lagged weight must be unrelated with these unobserved components. With the (unobserved) family environment being an important determinant of both weight and educational outcomes, there is no reason to believe that this assumption holds. For example, high unobserved time discount rates can decrease children's school performance and increase their weights. If this affects weight at all ages, both lagged and contemporaneous weight will be endogenous. This situation is depicted in Figure A2, Appendix A, showing an undirected edge between the instrument (the child's lagged weight, *L*) and the unobserved confounder u, indicating a violation of assumption 2.

The IV estimate of the weight coefficient from equation (2), using the child's lagged weight L_i to instrument for current weight W_i, is:

$$\hat{\beta}_1^{IV} = \beta_1 + \frac{Cov(L_i u_i)}{Cov(L_i W_i)}, \tag{4}$$

where $\frac{Cov(L_i u_i)}{Cov(L_i W_i)}$ is the IV bias. With the child's lagged weight being more or less a perfect

---

others argue that familial weights reflect unobserved family characteristics and include them to proxy for these confounders (Gronniger, 2005). We do not observe siblings in the data used here.

predictor of current weight, $Cov(L_i W_i)$ is large, satisfying assumption 1. However, this also implies that the covariance between lagged weight and the unobservables, $Cov(L_i u_i)$, is very similar to the covariance between current weight and the unobservables $Cov(W_i u_i)$ in equation (3), violating assumption 2. Hence, our prior is for the IV estimates that use the child's lagged weight as the instruments to be very similar to the OLS estimates.

The evidence of a genetic component in weight is often used to justify the choice of maternal weights as 'quasi-genetic' instruments (Averett and Stifel, 2007; Sabia, 2007). Maternal (lagged) weight however, is likely to be correlated with family resources, unmeasured preferences or choices, and educational inputs. For example, discrimination against obese females in the labour market (Cawley, 2004) can affect the family's financial resources that are available for inputs into the child education production function. Likewise, there is a large literature arguing that the family environment plays the main role in shaping children's food preferences, and that genetics play a minimal part (see e.g. the review in Birch, 1999). With the family environment also being an important determinant in the child education production function, there is no reason to believe that the use of maternal weight as instruments solves the endogeneity problem. The situation is similar to that depicted in Figure A2, replacing L_i with M_i: maternal (lagged) weight. Formally, the IV estimator using maternal (lagged) weight M_i as the instrument can be written as:

$$\beta_1^{IV} = \beta_1 + \frac{Cov(M_i u_i)}{Cov(M_i W_i)}. \qquad (5)$$

If the relationship between weight and outcomes is driven by, for example, unobserved family environments or socio-economic position u_i, the covariance between mother's weight and the unobservables, $Cov(M_i u_i)$, is likely to have the same sign as the covariance between child weight and the unobservables, $Cov(W_i u_i)$ in equation (3). In fact,

as a longer exposure to the (unobserved) environment for the mother may lead to stronger correlations with weight compared to that for the child, the former may actually be larger than the latter. In addition, as maternal weight is correlated with child weight, $Cov(M_i W_i)$ is likely to be large, satisfying assumption 1, though smaller than $Cov(L_i W_i)$ of equation (3). Hence, dividing a relatively larger negative covariance $Cov(M_i u_i)$ by a relatively smaller denominator, our prior is for these IV estimates to be more negative compared to the OLS estimates and compared to the findings that use the child's lagged weight as the instruments. More generally, the question is whether maternal weight can be used as a 'quasi-genetic' instrument, or whether it should act as a proxy for the unobserved family environment. In the latter case, it can be argued that maternal weight is part of, or a proxy for, parental inputs *P*, clearly invalidating the requirement for being a valid instrument. It also implies that mother's weight should be included as a covariate in the regressions, as depicted in the DAG shown in Figure A3, Appendix A.

B. Mendelian Randomization

Mendelian randomization – using genetic variation as instruments for endogenous variables – is closely linked to other study designs used in the epidemiologic and economic literature. First, it is closely related to Randomised Controlled Trials, where the allocation of treatment is randomised over all eligible individuals, as there is an equal probability that either allele is transmitted to offspring. Individuals who carry the genetic variant should have similar observable and unobservable characteristics compared to those without the variant. Second, Mendelian randomization is linked to the treatment effect literature, which often examines the effect of treatment on a subgroup of individuals who are induced by the instrument to change the variable of interest (i.e. the endogenous regressor). For example, Angrist and Krueger (1993) estimate the returns to schooling for those who are induced to remain in

school due to being born at different times of the year. The treatment effect for these '*compliers*' is also known as the Local Average Treatment Effect (LATE). Estimates obtained from Mendelian randomization experiments however, can be interpreted as the population Average Treatment Effect (ATE), as all individuals are *compliers*.[4]

We estimate equation (1) and use the child's genotype as an instrument for its weight. Although 2SLS is widely used in economics, the use of genetic markers in this field is new. Epidemiological studies emphasize the importance of carefully examining several situations and (biological) processes that may violate the IV assumptions in Mendelian randomization experiments (e.g. Davey Smith and Ebrahim, 2003). The existing studies in economics, however, have mainly failed to do so. We therefore begin with a discussion of each of the conditions that need to be met to obtain causal estimates of the effect of the phenotype on the outcome of interest.[5] We discuss concepts defined in the epidemiologic literature and relate them to the IV assumptions as used in the economics literature. In this discussion, we focus on our research question i.e. the effect of child weight (the phenotype) on academic achievement (the outcome of interest) and examine the issues that arise in this context.[6]

The first condition is the robustness of the genetic instrument in explaining the observed physical attribute. Mendelian randomization can only be used with genetic markers that have been robustly shown to affect the phenotype. This means it relies on *prior* knowledge about the association between genotype and phenotype, as shown in a large number of independent studies. This latter point is especially important, as many genetic association studies fail to replicate (Colhoun, McKeigue and Davey Smith, 2003). Without a robust and consistent population association, even if a significant *sample* correlation between genotype and phenotype exists, IV assumption 1 is violated. Any correlation may simply be

[4] Only in the case of gene-environment interactions, i.e. if a gene is only expressed in certain environments and not in others, is the estimated effect a LATE rather than an ATE.
[5] For a brief description of some key concepts in genetic epidemiology see Appendix B.
[6] For a more general discussion of the use of Mendelian randomization, see Lawlor et al. (2008).

due to factors such as measurement error, chance, or poor quality DNA. But even if a suitable and robust genetic instrument is available, it may explain little of the variation in observed phenotype, leading to low explanatory power in the first stage regression. Moreover, if the alleles shift the weight distribution by a very small amount, the effect of weight on educational attainment will be identified by this difference in mean weight. Hence, as IV only relates the amount of phenotypic variation explained by the genetic variant to the outcome of interest, this emphasizes the need for very large sample sizes. This, of course, is not a problem specific to Mendelian randomization, but refers to a more general problem of weak instruments, often encountered in IV studies (Angrist and Krueger, 1991).

The second situation is that behaviours may be affected by the genotype. As children inherit their genes from their parents, in a study of children's outcomes it is important to consider whether parents' behaviours or preferences are affected by their genotype. For example, mothers who carry 'fat' alleles may be discriminated against in the labour market because of their higher weights (Cawley, 2004).[7] This may affect her behaviour or preferences for her child's education, violating assumption 2.

The third situation relates to the mechanisms through which genetic variants affect fat mass. These are often unknown.[8] If the mechanism involves changes in behaviour or preferences that also affect the outcome of interest, assumption 2 will be violated.

The fourth situation relates to the allocation of genes across populations. Although

[7] This example merely gives an idea of a possible violation of the IV assumptions. In reality, this situation is very unlikely, as genotypes have small phenotypic effects. Being overweight or obese will therefore mainly be due to other factors, including non-genetic ones, rather than due to being homozygotic for the alleles used here.

[8] A decade ago, researchers mainly used the 'candidate gene approach' to examine associations between SNPs and phenotypes. This approach consists of testing a specific hypothesis: based on biological knowledge, researchers examined the association between one particular variant and the phenotype. However, we now know that these studies produced many false-positive findings. Currently, genome wide association studies (GWAS) are used. This approach genotypes 500,000 to 1,000,000 SNPs in one go and relates all SNPs to the phenotype of interest in a hypothesis-free way. Stringent criteria are used for GWAS p-values to take account of this hypothesis-free approach. Studies are either two-stage studies, where GWAS is performed on one sample of individuals, after which the small number of SNPs that reach GWAS levels of statistical significance are typed in other independent samples to examine the extent of robustness. Or studies consist of a number of independent GWAS, where only those SNPs that have consistent associations across all studies are interpreted as robust.

allocation of genes is random *within* families, any assortative mating may affect this randomness *between* families, leading to population stratification. Population stratification refers to a situation in which there exists a systematic relationship between the allele frequency and the outcome of interest in different sub-populations. This can lead to an association between the two at the population level without an actual causal relationship. For example, allele frequencies can vary across ethnic groups. Any systematic differences in educational outcomes across these subpopulations that are not due to a genetic make-up may therefore lead to biased estimates of the effect of weight by violating assumption 2. As a result, in studies which use genes as instruments it is necessary to test whether certain population subgroups are more likely to carry the genetic variant than others. The second, third and fourth situations can all be depicted by the DAG shown in Figure A2, Appendix A, substituting the instrument L with the genetic marker, which is related to unobserved population subgroups, parental preferences or behaviour u_i that affect both weight and outcomes.

Finally, there is the situation of a relationship between the genetic instrument and other genetic variants that may affect the outcome of interest. Mendel's second law states that the inheritance of one trait is independent of the inheritance of another. However, it has been shown that this does not always hold and that some variants are likely to be co-inherited, especially if they are physically close to each other on the chromosome. Depending on the effects of the co-inherited variant, this so-called 'Linkage Disequilibrium' (LD) can bias the estimates. This is shown in Appendix A, Figures A4-a and A4-b, where G_1 denotes our instrumental variable. If G_1 is in LD with another polymorphic locus G_2 that affects the phenotype W, the IV estimates remain unbiased (Figure A4-a). However, if G_1 is in LD with a polymorphic locus G_2 that affects the outcome S, assumption 2 is violated and the estimate will be biased (Figure A4-b). Relatedly, there is the situation of pleiotropy, where one genetic

marker has multiple phenotypic effects. The case is similar to that of LD, and will invalidate the IV approach if the pleiotropic effect influences children's educational outcomes *S*, but not if it affects only the phenotype *W*.[9]

Whether assumption 2 is violated by the conditions discussed here can be tested using tests of over-identification, but as in all IV studies this remains an assumption and the 'truth' can never be stated with certainty. Various studies have examined the relationship between genetic variants and individual or family characteristics. These studies provide insight into whether genetic markers are likely to be related to background characteristics or preferences. Davey Smith et al. (2008), for example, estimate pairwise correlations between non-genetic variables and genetic markers and compare the number of correlations that are statistically significant with the number expected by chance if all variables were uncorrelated. This sheds light on the degree of confounding that Mendelian randomization studies may be subjected to.[10] They show significant correlations between behavioural, socioeconomic and physiological factors, with 45% of the 4,560 pairwise correlations being significant at the 1% level. In contrast, genetic variants show no greater association with each other, or with the behavioural, socioeconomic and physiological factors than what would be expected by chance. In an attempt to shed some light on whether genotypes are related to individual's preferences and behaviours, Bhatti et al. (2005) explore differences in polymorphism frequencies by willingness to participate in epidemiologic studies. They examine three studies with different recruitment designs and different participation incentives. Conditional on having provided blood or saliva samples, they investigate whether genotype frequencies

[9] Another biological process that may affect estimates in Mendelian randomization studies is canalisation. This refers to a process by which potentially disruptive influences on the outcome of interest are buffered by foetal or post-natal developmental processes. Canalisation therefore alters the association between genotype and outcome, without any change in the relationship between genotype and phenotype. Canalisation is not likely to affect our estimates, as the pathways through which weight may affect educational attainment are generally not biological. Rather, they are socially or culturally constructed and therefore unlikely to be influenced by developmental processes that result in canalisation.

[10] They use a wide range of non-genetic indicators, such as body size, pulse, vitamin levels, type and frequency of the consumption of various foods, weekly hours of exercise, social class, education, housing tenure, smoking, birth weight, number of siblings, nurse estimation of life expectancy, etc.

differ by the timing of non-response to questionnaires (early, late and never responders). They find no evidence of correlations between genotypes and response characteristics.

C. Empirical Evidence on the Impact of Child Weight using Genetic Markers

Ding et al. (2006) examine the effects of several health conditions, one of which is weight, on adolescent's academic achievement. Their IV results show large and significant negative effects on female's Grade Point Average (GPA), but not for males. GPAs for obese girls are on average 0.8 points lower than those for non-obese girls. They use four markers: the dopamine transporter (*DAT1*), the dopamine D2 receptor (*DRD2*), tryptophan hydroxylase (*TPH*) and cytochrome P4502B6 (*CYP2B6*). Fletcher and Lehrer (2008) take a similar approach to Ding et al. (2006), but use a different dataset (the Add Health data) to exploit within-family genetic inheritance, slightly different instruments, and find no effects of obesity on academic achievement. In addition to *DAT1* and *DRD2*, their instruments include the dopamine D4 receptor (*DRD4*), the serotonin transporter (*5HTT*), monoamine oxidase (*MAOA*) and cytochrome P4502A6 (*CYP2A6*). Finally, Norton and Han (2008) examine the effects of excess weight on labour market outcomes using *DAT1* and *DRD4* to instrument for BMI and find no relationship.

The discussion in section IIB above highlights the importance of the choice of genetic variants in Mendelian randomization experiments. The first condition is that consistent and robust associations should have been shown between the genotype and phenotype in a large number of independent studies. The three economic studies cited above do not appear to have taken this approach. Rather than basing their selection on associations that are robustly shown in the literature, their choice of instruments seems rather ad hoc: using either forward stepwise estimation (Ding et al., 2006) or selecting those SNPs that have statistically significant *sample* correlations in the first stage (Fletcher and Lehrer, 2008). In fact, both

these studies acknowledge there is "weak and inconsistent evidence in the medical literature of the association between [their] genetic markers and health status or behaviours". Fletcher and Lehrer (2008) also state that "the scientific literature has not identified a unique (…) obesity gene". Norton and Han (2008) base their selection of SNPs on a study by Guo, North, and Choi (2006), who argue that there is a negative association between *DRD4* and obesity. This relationship, however, has not been replicated in other independent studies. For example, Hinney et al. (1999) and Schneider (2001) find no evidence of any relationship between *DRD4* and individuals' weights, and Fletcher and Lehrer (2008) find an insignificant but *positive* association.

In addition, these studies are unable to replicate various associations they note are reported in the literature. For example, Ding et al. (2006) find no association between *DAT1* and obesity, whilst they note the literature reports a positive relationship, and Norton and Han (2008) find a negative correlation. Fletcher and Lehrer (2008) fail to show any correlation between the *A1A1* variant of *DRD2* and obesity. However, given that the evidence of a robust association for these variants is lacking, this is not surprising (Lawlor, Windmeijer and Davey Smith, 2008). Furthermore, Norton and Han (2008) argue that the effects of the genetic markers differ by gender, while Patsopoulos et al. (2007) note that most claims of gender differences are spurious. Finally, Norton and Han (2008) use several markers as additional controls rather than instruments, as they fail the over-identification tests (*SLC6A4*, *MAOA*, *DRD2* and *CYP2A6*). Fletcher and Lehrer (2008) and Ding et al. (2006) use several of these as their instruments.

D. The Genetic Markers used in the Present Study

We use two SNPs that have been shown consistently to relate to child weight. Using a total of 38,759 individuals aged between 7 and 80 from 13 different cohorts of European

ancestry, Frayling et al. (2007) explore the association between *FTO* and BMI, the risk of being overweight and the risk of being obese. They find a positive association with all measures of weight for individuals in all cohorts, in all countries, of all ages and of both sexes, with no difference between males and females. They show that *FTO* is specifically associated with fat mass, and much weaker with lean mass. In addition, there is no association between *FTO* and birth weight, or *FTO* and height, indicating that the relationship with BMI is entirely driven by individuals' weights. They find that each copy of the risk allele is associated with an increase in BMI of 0.2 units at age 7, to 0.4 units at age 11. For the average age-specific height, this refers to a weight increase of 0.3 and 0.9 kg respectively. Hence, 11-year-olds who are homozygous for the rare allele are on average 1.8 kg heavier compared to those carrying no rare alleles.

Several different SNPs near *MC4R* also have been associated with obesity. We use that identified by Loos et al. (2008) here. They find a positive relationship with BMI in genome-wide association data from 16,876 individuals and confirm the relationship in an additional 60,352 adults and 5,988 children. They find no differences by gender, and no effects on birth weight or height, again suggesting the association is mediated solely through an effect on weight.

Our choice of genetic markers can be related to the conditions for suitable use of genetic markers as instruments discussed in IIB above. First, the prior findings of robust associations between the genetic markers and child weight justifies their use as instruments. In addition, the effect of *FTO* (*MC4R*) on child weight is relatively large; each allele leads to an increase of 0.9 (0.3) kg. Our instruments are therefore not weak (as we will show below). Second, we are able to examine whether maternal characteristics and behaviours are significantly related to the child's genetic markers as our data contains a wide range of variables on maternal behaviour. Third, the possible mechanisms through which SNPs affect

weight are increasingly studied in the medical literature. Although this work is ongoing, the current evidence suggests that the SNPs used here are associated with diminished satiety (Wardle et al., 2008) and are expressed in the hypothalamus (Willer et al., 2008). The main functions of the hypothalamus are to control food intakes and metabolic processes. Fourth, as we observe genotypes for both the mother and the child, we can examine the allele frequencies in both generations. If frequencies are orthogonal across generations, i.e. if there is no assortative mating, they are in 'Hardy-Weinberg Equilibrium' (HWE). We test for this below. Note that although *FTO*-allele frequencies are known to vary by ethnic group (Frayling et al., 2007), population stratification is not likely to be a problem in our data, as our cohort is recruited from a specific geographically defined region with a predominantly white population.[11] In addition, we observe and control for whether the child is non-white. We also examine whether genotypes are related to child and family background characteristics by testing whether there are specific patterns in observable characteristics between non-carriers, heterozygotes and homozygotes. In this exercise, we examine all covariates used in our analysis as well as an additional random set of background variables that we do not include in our specifications. The latter will provide further evidence of our instruments satisfying IV assumption 2.

Fifth, Linkage Disequilibrium (LD) would bias the IV estimates if the linkage is with another variant that affects children's educational outcomes such as IQ. Although there is some evidence of IQ heritability, specific variants or chromosomes have not been identified. In addition, as LD is not likely to occur for genetic markers on different (non-homologous) chromosomes, and the degree of linkage is a function of the distance between the loci, the 'ability marker' would have to sit on the same chromosome and be physically close to our genetic instruments. Although we cannot rule this out, it does not seem plausible. However,

[11] Only 5.1% of children in the region have a non-European, non-Caucasian parent (Golding et al., 2001).

even if our genetic variant is in LD with an 'ability marker', the literature (e.g. Plomin et al., 1994) suggests that 'ability markers' have very small phenotypic effects, leading to little bias in the estimates.

Finally, pleiotropy would invalidate the IV approach if the pleiotropic effect influences children's educational outcomes. Frayling et al. (2007) explored the relationship between *FTO* and type II diabetes and found a strong association between the two. However, this relationship disappeared once they controlled for BMI, suggesting that the association with diabetes is mediated through BMI. Similar positive associations were found between *FTO* and insulin, glucose and triglycerides (Freathy et al., 2008) but again become insignificant after adjusting for BMI. Hence, even if type II diabetes is related to children's educational attainment, the effect of the instrument goes via BMI. Therefore the IV assumptions are not violated.

III. Data

Our data are from a cohort of children born in one geographic area (Avon) of England. Women eligible for enrolment in the population-based Avon Longitudinal Study of Parents and Children (ALSPAC) had an expected delivery date between 1 April 1991 and 31 December 1992. Approximately 85% of these mothers enrolled, leading to about 14,000 pregnancies. The Avon area has approximately 1 million inhabitants and is broadly representative of the UK as a whole, although slightly more affluent than the general population (Golding et al., 2001).[12] Detailed information on the children and their families has been collected using a variety of sources, including self-completed questionnaires, data extraction from medical and educational records, in-depth interviews, and biological samples. Hence, the variables in ALSPAC relate to a wide range of child health and development,

[12] See www.alspac.bris.ac.uk for more a detailed description of the representativeness of the sample, its enrolment, and response rates.

family background, family inputs and school measures.

We observe a total of 12,620 children who survived past the age of 1 and returned at least one questionnaire. Our sample selection process is as follows. First, we select those children for whom we observe their genotypes, leaving us with approximately 7,700 children. Second, we drop children with missing phenotypes. Children were invited to attend specially designed clinics, where their anthropometric measures were recoded. As not all children attended these clinics, our sample sizes reduce to just over 4,500, equally distributed between boys and girls. Third, as we aim to explore whether maternal BMI can be used as quasi-genetic instruments or whether it should be included as a covariate, we drop observations where these are missing. Finally, we restrict the sample to those children for whom we observe their educational outcomes, leading to a final sample size of just over 3,500 children. We deal with missing values on other covariates by using mean substitution and including an indicator for the value being missing.[13]

A. Measures of Academic Achievement

Our main outcome measure is the child's Key Stage 3 (KS3) score. The KS3 exam is a nationally set exam, taken by all 14-year-olds in English state (public) schools[14]. This measure of children's performance is therefore objective and comparable across all children. Children's scores for three subjects (English, maths and science) are obtained from the National Pupil Database, a census of all pupils in England within the state school system, which is matched into ALSPAC. We use an average score for the three subjects, standardised on the full sample of children for whom data is available, with mean 100 and standard deviation 10.

[13] In robustness checks below, we present results after imputing the missing values for all variables apart from the genotypes, resulting in a sample size of 7,706 children.
[14] 93 percent of English children attend state schools.

B. Measures of Child Weight, Fat Mass and the Genetic Markers

Our main measure for child weight is the child's body fat mass (adjusted for age in months, height and height squared), as measured by a dual-energy X-ray absorptiometry scan (DXA). This method scans the whole body, dividing it into body fat, lean tissue mass, and bone density. The measurements are taken at age 11 and 9. Our primary focus is on DXA measures at 11: when we examine lagged fat mass we use the DXA score at age 9. For ease of interpretation and for comparison with later measures, DXA scores are standardised on the full sample of children for whom data is available, with mean 100 and standard deviation 10. For the genetic markers, we use two SNPs that have been consistently found to relate to weight: *FTO* (rs9939609) and *MC4R* (rs17782313).[15]

C. Contextual Variables and Mother's Health and Behaviour

The child's initial health status is measured by its birth weight, together with ordered indicators for the intensity of mother's breastfeeding as proxies for early nutrition (never, <1 month, 1-3 months and 3+ months). As children's educational outcomes are known to differ with within year-age, the analyses include dummies for children's age (in months), as well as an indicator for whether the child is non-white. We also include variables indicating the number of older and younger siblings under 18 in the household.

We include several controls for socio-economic status: log equivalised family income and its square, four binary indicators for mother's educational level, the mother's parents' educational level, an indicator for whether the child is raised by a lone parent, dummy variables for the family's social class, maternal age at birth, and parents' employment status when the child is 21 months. We also include a measure of small (local) area deprivation: the

[15] The rs-number (reference SNP, or RefSNP) is an identification tag that uniquely positions the polymorphism in the genome. All genotyping was performed by KBioscience (http://www.kbioscience.co.uk). SNPs were genotyped using the KASPar chemistry, which is a competitive allele-specific PCR SNP genotyping system using FRET quencher cassette oligos (http://www.kbioscience.co.uk/genotyping/genotyping-chemistry.htm).

Index of Multiple Deprivation (IMD).[16]

In addition to these generally observed controls, our data allows us to also account for further measures of mother's health and behaviour, which may be correlated to both child weight and educational attainment. In addition, we use these to test whether mother's behaviour differs significantly for the different genotypes. We include two binary variables which measure whether the mother smoked or drank alcohol in the first three months of pregnancy to control for the *in utero* environment. We include measures of the mother's 'locus of control', a psychological concept that describes whether individuals attribute successes and failures to internal or external causes. Those with an internal (low) locus of control see themselves as responsible for the outcomes of their actions. Those with an external (high) locus of control believe that successes and failures are chance-determined. We also include two measures of maternal mental health during pregnancy.[17] We include several measures of parental involvement or interest in the child's development.[18] Finally, we use a continuous variable that measures the extent to which the parents engage in active (outdoor) activities with their children.[19]

D. Descriptive Statistics

[16] Family income is an average of two observations (when the child is aged 3 and 4) and is in 1995 prices). The educational dummies are having poorer qualifications than those taken at school leaving age (less than O-level), O-level, having qualifications that permit higher educational study (A-level) and having a university degree. We use the standard UK classification of social class based on occupation (professional (I), managerial and technical (II), non-manual skilled (IIInm), manual skilled (IIIm), semi-skilled (IV) and unskilled (V)). IMD is based on six deprivation domains, including health deprivation and disability; employment; income; education, skills and training; housing; and geographical barriers to services. Increasing scores on this measure indicate greater deprivation. The IMD measures relates to areas containing around 8000 persons.

[17] These are the Edinburgh Post-natal Depression Score (EPDS) and Crown-Crisp Experimental Index (CCEI) at 18 weeks gestation. EPDS indicates the extent of post-natal depression; CCEI captures a broader definition of mental health, measuring general anxiety, depression and somaticism. Higher scores mean the mother is more affected.

[18] A continuous variable ranging from 0-10 is included measuring the mother's 'teaching score'. This is constructed from questions that measure whether the mother is involved in teaching her child (depending on the child's age) certain songs, the alphabet, being polite, etc. We use an average score from three measures at ages 18, 30 and 42 months to capture longer-term involvement. Likewise, a variable is included indicating whether the mother reads or sings to the child, allows the child to build towers or other creations, and so on.

[19] This includes several recreational pursuits (including going to the park or playground and going swimming).

We begin by testing the assumption that genotype frequencies are orthogonal across generations using the HWE test on the child's and mother's *FTO* and *MC4R*. The test and its results are detailed in Appendix C. The results support the null of random distribution of alleles and of no differences in genotype frequencies between the mother and child. We next examine whether the instruments are correlated with fat mass by plotting the empirical DXA distribution functions for non-carriers, heterozygotes and homozygotes for the *FTO* risk allele (as in Angrist, Graddy and Imbens 2000). Figure 1a shows large differences between the distributions, with homozygotes having most fat mass. Figure 1b presents the differences between these distribution functions and plots the unnormalized weight functions for heterozygotes and non-carriers (solid line) and for homozygotes and non-carriers (dashed line). These are defined as

$$\frac{1}{N_1}\sum_{i|heterozygote} 1\left[DXA_i \geq j\right] - \frac{1}{N_0}\sum_{i|non-carrier} 1\left[DXA_i \geq j\right] \qquad \text{and}$$

$$\frac{1}{N_1}\sum_{i|homozygote} 1\left[DXA_i \geq j\right] - \frac{1}{N_0}\sum_{i|non-carrier} 1\left[DXA_i \geq j\right] \qquad \text{for } DXA_{\min} \leq j \leq DXA_{\max}$$

respectively, where $1[\cdot]$ is an indicator function equalling one when the expression between brackets holds. The figure shows that the two *FTO* genotypes affect a similar and large area of the DXA distribution (between the values of 85 and 125). Only the very top and bottom of the distribution is not captured by the instruments. Heterozygotes' and homozygotes' weights are similar in the lower part of the DXA distribution (85 – 95), but the middle and upper part are weighed more heavily for homozygotes. Figure 2 shows a similar picture for *MC4R*, with both genotypes associated with the bulk of the DXA distribution, though with slightly smaller weights.

Table 1 presents summary statistics for non-carriers, heterozygotes and homozygotes of the *FTO* and *MC4R* risk alleles. The table shows that children with one risk-allele (heterozygotes) of either marker have significantly higher DXA scores than those without.

Similarly, those with two risk alleles (homozygotes) have higher DXA scores than heterozygotes. For the relatively small sample sizes for *MC4R*-homozygotes, the difference is not significant. There is also a positive relationship between the number of risk alleles and the probability of being overweight (defined as being in the top 15% of the DXA distribution). The table shows that these differences are driven by children's weights rather than their heights, which do not show significant differences for both variants. The average DXA score at age 9 shows similar trends to that at age 11. Mother's pre-pregnancy BMI is higher for children with one or two risk-alleles compared to none, but the difference is not significant.

Finally, the contextual variables and the indicators for mother's health and behaviour do not show any clear patterns of significant differences across the child's genotypes.[20] For example, *MC4R*-heterozygotes have a lower mean birth weight, and their mothers have lower teaching scores. *FTO*-homozygotes are more likely to be in social class II (managerial or technical occupations), and their mothers have a higher average locus of control. However, there is no obvious structure in the magnitude or significance of these means, supporting the assumption that the genotypes are distributed randomly in the population and are unrelated to other child or family characteristics. Also note that the two genetic markers are uncorrelated (*r=0.00*), and that both markers lead to a weight increase. If this rise in weight affects (e.g.) mother's behaviour, we would expect to see similar patterns or differences in mean characteristics for both markers. Only for mother's locus of control do we find increasing scores for both variants; none of the other behaviours or preferences show similar patterns, supporting the assumption that they are not affected by the genotype (condition two in section IIB above).[21]

[20] There are also no differences in mother's health and behaviour by mother's genotype (results available from the authors).

[21] Appendix D presents the relationship between our SNPs and an additional set of background variables that are not included in our analyses. Apart from strong positive associations with waist and hip circumference, it shows no clear patterns of significant differences, providing further evidence that our instruments satisfy assumption 2.

IV. Results

We begin by examining the non-parametric relationship between educational attainment (KS3 scores at age 14) and the child's fat mass (DXA score at age 11). Figure 3 shows a clear negative relationship, which is linear over the full range of the DXA distribution. Table 2 presents the OLS results from the regression of KS3 on DXA. Each column subsequently adds more control variables. The raw correlation between educational attainment and fat mass is negative, with a one standard deviation increase in fat mass associated with an almost 0.1 standard deviation decrease in test scores (column 1). Column 2 controls for mother's pre-pregnancy BMI and BMI squared, leading to a slight decrease in the effect of fat mass on KS3, although the coefficient remains highly significant. Accounting for the contextual variables (column 3) and mother's health and behaviour (column 4) brings the estimate closer to zero, but it remains negative and significant.[22] Accounting for the contextual variables also renders the effect of the linear and quadratic terms in mother's BMI jointly insignificantly different from zero, suggesting that maternal BMI is picking up part of the effect of contextual variables. The estimated coefficients of the other covariates (not shown) are in line with priors and other analyses of educational attainment on these UK tests.[23] In summary, these OLS findings suggest that there is a correlation, albeit a small one, between child fat mass and educational attainment.

We now turn to the 2SLS estimates. We examine three sets of instruments. First, following Averett and Stifel (2007) and Sabia (2007), we instrument weight by mother's pre-pregnancy BMI and BMI squared.[24] Second, as Kaestner and Grossman (2008), we use the

[22] Non-linearities in DXA were explored, but show no clear patterns (available upon request).
[23] Girls perform better, and the child's age is positively related to performance. Test scores are lower for those brought up by a stepfather rather than the natural father. Mother's education and father's social class are positively related to the child's test score. Mother's employment status negatively affects the child's performance, with larger coefficients for full-time employment compared to part-time. Test scores are lower for those living in more deprived areas and for those whose mothers have an external locus of control. Mother's teaching and child related activity scores show positive associations with school performance.
[24] Sabia (2007) uses mother's and father's obesity status. However, as father's height and weight is only known for a much smaller sample, we take Averett and Stifel's (2007) approach and use mother's BMI and its square.

child's lagged under- and overweight categories (more specifically, the percentiles 0-5, 6-15, 16-84, 85-94 and 95-100 of the distribution) as instruments for later fat mass.[25] Finally, we present our preferred instruments, instrumenting the child's phenotype with its genotype.

Table 3 presents the specification tests for the first-stage regressions using the three sets of instruments and controlling for all covariates. All instruments have a significant association with DXA, as shown by the F-statistic of IV strength. The smallest F-statistic, using the genetic markers, still exceeds the Stock and Yogo (2002) critical values with a value of 17.6. If we use only the stronger FTO SNP, the F-statistic rises to almost 30, confirming the strength of our genetic markers.[26] The LM test for under-identification has large values in all three columns, indicating that all instruments perform well on this criterion. The Hansen J-test, although not formally rejected, is close to being significant when using either the mother's or child's lagged weight, suggesting the instruments may be related to the error term of the structural equation. The specification in column 3 is not rejected, supporting the assumption that the genetic instruments are valid.

The only difference between columns 1 and 2 and between columns 1 and 3 is the addition of the instrumental variables, as all specifications include a linear and quadratic in mother's BMI.[27] Hence, these instruments cause the decrease in the F-statistic of joint significance of mother's BMI and BMI squared (final row in Table 3). Compared to column 1 however, the F-statistic in column 3 is not significantly smaller, implying that mother's BMI, part of which is genetic, is not strongly related to these two variants. Similarly, even after controlling for the child's lagged weight (column 2), the F-statistic for mother's BMI remains

[25] Kaestner and Grossman (2008) regress the *change* in educational attainment on the level of child weight, using the child's lagged weight categories as the instruments. To make the analysis more comparable across our alternative instrument sets, we use the *level* of educational performance as the dependent variable.

[26] Distinguishing between heterozygosity and homozygosity for the genetic markers leads to slightly smaller F-statistics. However, the coefficient for being homozygous is twice that of being heterozygous, suggesting the relationship is linear. We therefore use one indicator for *FTO* and one for *MC4R*, although the results are robust to other genetic instrument specifications.

[27] They are specified as the instruments in column 1, while columns 2 and 3 include them as covariates.

just under half the size compared to that in column 1. Child (lagged) weight is a result of both environmental and genetic influences. Over and above this, the effect of maternal BMI is likely to proxy additional family and environmental characteristics. This therefore suggests that maternal BMI should not be used as a 'quasi-genetic' instrument, but should be added as a covariate in the regressions to proxy for these unobservables.

Table 4 presents the second stage regression results. Panels A, B and C refer to the different instrument sets. Column 1 replicates the OLS results from Table 2, whilst columns 2 – 5 show the findings after instrumenting for fat mass. Controlling for all covariates, the OLS results show that fat mass negatively affects school performance. Using mother's pre-pregnancy BMI and BMI squared as instruments (panel A), this relationship remains negative and highly significant. The estimates suggest that one standard deviation increase in fat mass relates to 0.1 standard deviation decrease in KS3 (column 5). The IV estimate is more than 2.5 times larger than the OLS, suggesting that the OLS underestimates the true effect. Panel B, which uses the child's lagged fat mass as instruments, also shows negative effects of fat mass on KS3, although the inclusion of contextual variables renders these insignificant. In contrast, when genetic markers are used as instruments in panel C, the estimates show *positive* effects of fat mass on educational performance. With relatively large standard errors however, we cannot reject the null that there is no effect on academic outcomes.

The findings in panel A confirm Averett and Stifel (2007) and Sabia (2007), who use similar instruments. However, they are in stark contrast with the results in panel C, and also somewhat with those in panel B. The only driver behind these differences is the choice of instrument, since the model specification is identical in all other aspects. As both maternal and child weight are likely to be correlated with family resources, preferences and educational inputs, they seem a poor choice of instrument.

The size and patterns of our estimates in panels A and B correspond to our priors as

discussed in section IIA: larger negative coefficients when using maternal BMI as the instruments (panel A), and similar coefficients to those found in OLS when using the child's lagged fat mass as the instruments (panel B). These findings add to the argument that both maternal and child lagged weight should not be used as instruments for current weight. In addition, comparing the estimates of panels A and B as we move across the columns, including more covariates, shows that the impact of DXA decreases with each addition, with the largest drop from column 3 to 4. Changes in the DXA estimate only occur if the predicted DXA from the first stage regression is correlated with the variables added to the model. However, as each subsequent column adds covariates to both stages of the 2SLS regression, it is unclear how to interpret a change in the second stage DXA estimate. Examining the first stage coefficients though, shows they are very similar for panels A and C (results available upon request).[28] This implies that the difference in predicted DXA between panels A and C is solely driven by the instruments chosen. With the effect of DXA decreasing monotonically as we include more controls in panel A, this suggests that predicted DXA is picking up the effects of the child and family background characteristics. As discussed above, these are also related to children's test scores. In contrast, panel C does not show any clear patterns in the magnitude of DXA, suggesting that predicted DXA is not systematically related to background characteristics that are also related to academic outcomes. This therefore again casts doubt on the appropriateness of using maternal or child lagged weight as instruments for child current weight.

V. Robustness Checks

A. Measurement of Child Fat Mass

Our results suggest that once the endogeneity of child weight is accounted for using

[28] The first stage coefficients differ for Panel B, as the estimates can then no longer be interpreted as affecting children's *weight*, but as affecting children's *weight gain*.

children's genotypes as IVs, there are no differences in academic performance between children of different weights. Burkhauser and Cawley (2008) show that the parameter estimates of obesity on employment are sensitive to different definitions of obesity, as defined by BMI or percentage body fat from a Bioelectrical Impedance Analysis (BIA). Although DXA is a direct measure of body fat, we also observe several other indicators and can use these to examine the robustness to different proxies for fat mass. First, we observe the child's DXA score at age 9 as well as at age 11, allowing us to examine whether the effect of fat mass is the same regardless of when it is measured in the mid-childhood years. Second, we use the child's BMI at age 9 and at age 11 as a proxy for fat mass. Third, we compute a measure indicating the percentage body fat, calculated as fat mass as a percentage of total body mass, based on the DXA scan. Fourth, we use the percentage body fat as measured by BIA, as Burkhauser and Cawley (2008).[29] Finally, we check for non-linearities, using the child's overweight status, defined as being in the top 15th percentile of the DXA distribution. We focus on the regressions that use genetic markers as instruments and control for all covariates, including mother's BMI, as this is our preferred specification. To allow for comparisons, DXA and BMI are standardised with mean 100 and standard deviation 10. The percentage body fat is entered as a percentage.

Table 5 presents the results. In contrast to the OLS results, all estimates are positive, though they are small and not significant. The estimated DXA coefficients at ages 9 and 11 suggest that precisely when DXA is measured in this two-year-span does not matter. Like DXA, the BMI coefficient is larger at age 9 than age 11, though again not significantly different from zero. One standard deviation increase in fat mass, as measured by BMI or DXA, is related to an increase in KS3 scores of between 0.05 and 0.11 standard deviations.

[29] To measure children's fat mass with BIA, examiners attach electrodes to the child's heel and toe and pass very small electrical currents through the body, measuring weight and impedance. These measurements can be used to calculate fat and fat-free mass as the resistance to an electric current is inversely related to the amount of fat-free mass in the body; the water in muscles and lean tissue conducts the electricity while fat does not.

The estimates from the two measures of percentage body fat are also slightly different in magnitude: the effect of a one percent increase in body fat measured by BIA is almost three times the size of that measured by DXA (though both are small and neither is significant). The estimate of being overweight is positive, but not significant. So while the parameter estimates do appear sensitive to the definition of fat mass used, for all measures we cannot reject the null that fat mass has no effect on children's academic performance.[30]

B. Multiple Imputation to Increase the Sample Size

Although the IV regressions show no evidence of weight affecting outcomes, the standard errors are relatively large. Indeed, the OLS and IV estimates are not significantly different due to the relative imprecision of the latter. One way of dealing with this is to use a larger sample. For all variables apart from the genetic markers, we therefore impute their missing values using Multiple Imputation by Chained Equations (MICE; Royston, 2004).[31] This leads to a doubling of our sample to 7,706 observations.[32] The (second stage) IV results, using maternal pre-pregnancy BMI, the child's lagged weight categories, and the child's genetic markers, are presented Table 6, panels A, B and C respectively. The sign and magnitude of the coefficients are similar to those estimated on the smaller sample (Table 4).[33] Adding more controls to the regression also leads to similar (decreasing) patterns in the estimated coefficients of panels A and B, but again not in panel C. The standard errors are

[30] As growth patterns and academic performance differ by gender, we also test whether our findings are similar for boys and girls and examine earlier outcomes (tests at age 11). We find no significant differences between the genders and the direction of the findings again suggests that, if any relationship exists, it is that fat mass *in*creases rather than *de*creases test scores.

[31] As the genetic markers do not show any systematic correlation to the other covariates in the model apart from child weight or fat mass, we cannot impute its values; the markers are distributed randomly.

[32] Due to the strong within individual correlation in DXA and BMI over time, and as we observe the child's DXA score, its BMI and its weight at various ages, we have strong predictors if child DXA is missing at age 11. In fact, only 36 out of the 7,706 children have missing observations for all measures of weight. Similarly, a child's performance on the Key Stage tests (taken at 7 and 11 as well as 14) is highly correlated over time. As we observe the child's scores on the entry assessment test, as well as their KS exams at age 7 and 14, this will help in imputing any missing KS3 scores. For 787 children, we do not observe any of these outcome measures.

[33] The OLS and first stage IV results are available upon request; they are also similar to those estimated on the smaller sample used above.

slightly smaller than those derived using the smaller sample. However, they remain large and the imputation does not allow us to statistically distinguish between OLS and IV.

VI. Conclusions

We examine the relationship between child weight and educational outcomes, using data from a rich cohort of UK children. To account for the endogeneity of child weight, we examine genetic markers. We compare these to two instrument sets recently used in the literature to date: maternal pre-pregnancy weight and the child's lagged weight.

The OLS results show that leaner children perform better in school tests compared to their heavier counterparts, but our preferred genetic IV contradicts these findings. We find no evidence that children's fat mass affects their academic performance. The observed association therefore appears to be driven by unobserved characteristics. We also find that using genetic IVs produces different results from the other two instrument sets. Using the child's lagged weight as instruments for current weight leads to (patterns of) estimates that are very similar to those obtained by OLS, though with slightly larger standard errors. The estimates that use maternal BMI as the instruments confirm the OLS results: fat mass negatively affects educational outcomes. We argue that these results arise from incorrect specification of the instruments: both are correlated with unobservables that also affect academic performance.

The increasing availability of biomedical data, in combination with a growing medical literature on the effects of carrying specific genetic markers, opens up the use of new instruments for examination of physical traits on economic outcomes. However, even with strictly exogenous instruments that comply with all IV assumptions, the limitations of genetic markers need to be understood. If the markers account for only a small amount of phenotypic variation, this implies the need for one of three things. These are (1) larger sample sizes than

those used in many current studies to increase the precision of the estimates; (2) observing more genetic variants to increase the explained variation in phenotype; or (3) waiting for genetic markers to be discovered with large phenotypic effects, though as larger effects tend to be discovered before smaller ones, the last route seems unlikely to be one that will be useful.

References

Angrist, J. D., K. Graddy, and G. W. Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equation Models with an Application to the Demand for Fish." *Review of Economic Studies* 67: 499-527.

Angrist, J. D., and A. B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106(4):979-1014.

Averett, S. L., and D. C. Stifel. 2007. "Food for Thought: The Cognitive Effects of Childhood Malnutrition in the United States." http://aysps.gsu.edu/events/2007/averett_stifel_obesity_june.pdf**.**

Bhatti, P. et al. (2005). "Genetic Variation and Willingness to Participate in Epidemiologic Research: Data from Three Studies." *Cancer Epidemiology Biomarkers and Prevention* 14:2449-53.

Birch, L. 1999. "Development of Food Preferences." *Annual Review of Nutrition*, 19: 41-62

Burkhauser, R., and J. Cawley. 2008. "Beyond BMI: The Value of More Accurate Measures of Fatness and Obesity in Social Science Research." *Journal of Health Economics*, 27: 519-29.

Case, A., D. Dubotsky, and C. Paxson. 2002. "Economic Status and Health in Childhood: The Origins of the Gradient." *American Economic Review*, 92(5): 1308-34.

Case, A., and C. Paxson. 2008. "Stature and status: Height, Ability and Labor Market Outcomes" *Journal of Political Economy*, 116(3): 499-532.

Cawley, J. 2004. "The Impact of Obesity on Wages." *Journal of Human Resources*, 39(2): 451-74.

Colhoun, H., P. McKeigue, and G. Davey Smith. 2003. "Problems of Reporting Genetic Associations with Complex Outcomes." *The Lancet*, 361: 865-72.

Cunha, F., and J. Heckman. 2007. "The Technology of Skill Formation." *American Economic Review*, 97(2): 31-47.

Davey Smith, G., and S. Ebrahim. 2003. "'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease?" *International Journal of Epidemiology*, 32: 1-22.

Davey Smith G. et al. 2008. "Clustered Environments and Randomized Genes: A Fundamental Distinction between Conventional and Genetic Epidemiology." *PLoS Medicine* 4: 1985-92.

Ding, W., et al. 2006. "The Impact of Poor Health on Education: New Evidence using Genetic Markers." NBER Working paper 12304.

Fletcher, J., and S. Lehrer. 2008. "Using Genetic Lotteries within Families to Examine the Causal Impact of Poor Health on Academic Achievement." http://www.econ.sfu.ca/Seminars/SeminarPapers/Lehrer.pdf

Frayling, T., et al. 2007. “A Common Variant in the FTO Gene is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity.” *Science,* 316: 889-94.
Freathy, R. et al. 2008. “Common Variation in the FTO Gene Alters Diabetes-Related Metabolic Traits to the Extent Expected Given Its Effect on BMI” *Diabetes* 57:1419-26.
Geier, A., et al. 2007. “The Relationship between Relative Weight and School Attendance among Elementary Schoolchildren.” *Obesity,* 15(8): 2157-61.
Golding, J. et al. 2001. “ALSPAC - The Avon Longitudinal Study of Parents and Children: I. Study Methodology.” *Pediatric and Perinatal Epidemiology,* 15: 74-87.
Gronniger, T. 2005. “Familial Obesity as a Proxy for Omitted Variables in the Obesity-Mortality Relationship.” *Demography,* 42(4): 719-35.
Guo, G., K. North, and S. Choi. 2006. “DRD4 Gene Variant associated with Body Mass: The National Longitudinal Study of Adolescent Health.” *Human Mutation,* 27(3): 236-41.
Hamermesh, D., and J. Biddle. 1994. “Beauty and the Labor Market.” *American Economic Review,* 84(5): 1174-94.
Hinney, A. et al. 1999. “No Evidence for Involvement of Polymorphisms of the Dopamine D4 Receptor Gene in Anorexia Nervosa, Underweight, and Obesity.” *American Journal of Medical Genetics,* 88(6): 594-7.
Johnson, R., and R. Schoeni. 2007. “The Influence of Early-Life Events on Human Capital, Health Status and Labor Market Outcomes over the Life Course.” Population Studies Center Research Report 07-616.
Kaestner, R., and M. Grossman. 2008. “Effects of Weight on Children's Educational Achievement.” NBER Working paper 13764.
Kivimäki, M. et al. Forthcoming. “Lifetime Body Mass Index and Later Atherosclerosis Risk in Young Adults: Examining Causal Links using Mendelian Randomization in the Cardiovascular Risk in Young Finns Study” *European Heart Journal*
Lawlor, D., F. Windmeijer, and G. Davey Smith. 2008. “Is Mendelian Randomization “Lost in Translation”?” *Statistics in Medicine,* 27:2750-55.
Lawlor, D. et al. 2008. “Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology.” *Statistics in Medicine,* 27: 1133-63.
Loos, R., et al. 2008. “Common Variants Near *MC4R* are Associated with Fat Mass, Weight and Risk of Obesity.” *Nature Genetics,* 40(6): 768-75.
Lewontin, R. and C. Cockerham. 1959. “The Goodness-of-fit Test for Detecting Natural Selection in Random Mating Populations” *Evolution* 13(4): 561-4.
Mobius, M. and T. Rosenblat. 2006. “Why Beauty Matters.” *American Economic Review* 96(1): 222-35.
Norton, E., and E. Han. 2008. “Genetic Information, Obesity and Labor Market Outcomes.” *Health Economics, 17(9): 1089-1104.*
OECD. 2007. *Health at a Glance 2007: OECD Indicators.* OECD Publishing.
Patsopoulos, N., A. Tatsioni, and J. Ioannidis. 2007. “Claims of Sex Differences: an Empirical Assessment in Genetic Associations.” *JAMA,* 298: 880-93.
Plomin, R. 1986. *Development, Genetics and Psychology*. Mahwah, Lawrence Erlbaum Associates Inc.
Plomin, R. et al. 1994. “DNA Markers Associated with High Versus Low IQ: The IQ Quantitative Trait Loci (QTL) Project.” *Behavior Genetics* 24(2):107-18.
Redline, S. et al. 1999. “Risk Factors for Sleep-Disordered Breathing in Children: Association with Obesity, Race and Respiratory Problems.” *American Journal of Respiratory and Critical Care Medicine,* 159(5): 1527-32.
Royston, P. 2004. “Multiple Imputation of Missing Values.” *The Stata Journal* 4(3):227-41.
Sabia, J. 2007. “The Effect of Body Weight on Adolescent Academic Performance.” *Southern Economic Journal,* 73(4): 871-900.

Schneider, J. 2001. “Allelverteilungen der Polymorphismen des Dopamin-D4-Rezeptorgens bei Probanden unterschiedlicher Gewichtsklassen und Patientinnen mit Anorexia nervosa.” PhD Diss. University of Marburg.

Schwartz, M., and R. Puhl (2003). “Childhood Obesity: A Societal Problem to Solve.” *Obesity Reviews,* 4(1): 57-71.

Stock, J., and M. Yogo. 2002. ”Testing for Weak Instruments in Linear IV Regression.” NBER Technical Working Paper 284.

Wardle, J. et al. 2008. “Obesity Associated Genetic Variation in *FTO* is associated with Diminished Satiety” *The Journal of Clinical Endocrinology and Metabolism*, 93(9):3640-3.

Willer, C. et al. 2008. “Six New Loci Associated with Body Mass Index Highlight a Neuronal Influence on Body Weight Regulation.” *Nature Genetics,* 41: 25-34.

Table 1: A selection of descriptive statistics by *FTO* and *MC4R* genotype. Mean (standard deviation)

	FTO (rs9939609)			*MC4R* (rs17782313)		
	Non-carrier	Heterozygous	Homozygous	Non-carrier	Heterozygous	Homozygous
Weight						
DXA, age 11	98.6	100.1 ***	101.2 ***	99.4	100.1 **	100.5
	(9.26)	(10.03)	(10.47)	(9.71)	(10.02)	(10.43)
DXA overweight	0.107	0.157 ***	0.188 ***	0.135	0.155	0.160
	(0.31)	(0.36)	(0.39)	(0.34)	(0.36)	(0.37)
BMI, age 11	98.6	100.0 ***	101.2 ***	99.5	100.1 *	99.3
	(9.05)	(9.98)	(9.89)	(9.62)	(9.87)	(8.85)
Weight (kg), age 11	42.6	43.5 ***	44.9 ***	43.3	43.5	43.8
	(9.55)	(9.99)	(10.07)	(9.66)	(10.03)	(11.02)
Height (cm), age 11	150.6	150.5	151.2	150.8	150.4	150.3
	(7.14)	(7.21)	(7.13)	(7.07)	(7.31)	(7.36)
DXA, age 9	98.8	100.2 ***	101.1 ***	99.6	100.1	99.6
	(9.13)	(9.89)	(10.04)	(9.67)	(9.81)	(8.68)
BMI mother,	99.8	100.4	100.2	100.1	100.1	100.7
pre-pregnancy	(9.47)	(9.76)	(9.28)	(9.74)	(9.35)	(9.50)
Contextual variables						
Birth weight (g)	3416	3410	3424	3433	3382 ***	3439
	(555)	(524)	(546)	(538)	(540)	(527)
Ln(income)	5.32	5.33	5.34	5.34	5.32	5.30
	(0.44)	(0.41)	(0.42)	(0.42)	(0.43)	(0.38)
Mother's education	0.12	0.15 *	0.12	0.13	0.14	0.15
< O level	(0.33)	(0.36)	(0.32)	(0.34)	(0.34)	(0.36)
Mother's education	0.50	0.44 ***	0.50	0.48	0.45	0.51
O level	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Mother's education	0.25	0.27	0.25	0.25	0.28 **	0.21
A level	(0.43)	(0.44)	(0.43)	(0.43)	(0.45)	(0.41)
Mother's education	0.12	0.14	0.13	0.14	0.13	0.13
Degree	(0.33)	(0.35)	(0.34)	(0.34)	(0.33)	(0.34)
Social class I	0.09	0.10	0.11	0.10	0.10	0.09
	(0.29)	(0.30)	(0.31)	(0.30)	(0.29)	(0.28)
Social class II	0.31	0.33	0.37 **	0.32	0.33	0.30
	(0.46)	(0.47)	(0.48)	(0.47)	(0.47)	(0.46)
Social class IIInm	0.14	0.13	0.11	0.12	0.13	0.15
	(0.34)	(0.33)	(0.31)	(0.33)	(0.33)	(0.36)
Social class IIIm	0.30	0.28	0.28	0.28	0.29	0.27
	(0.46)	(0.45)	(0.45)	(0.45)	(0.45)	(0.45)
Social class IV	0.09	0.09	0.07	0.09	0.08	0.09
	(0.28)	(0.29)	(0.26)	(0.29)	(0.27)	(0.29)
Social class V	0.02	0.03	0.01	0.02	0.02	0.03
	(0.15)	(0.16)	(0.12)	(0.15)	(0.15)	(0.18)
Mother's health and behaviour						
Smoke	0.18	0.18	0.15	0.17	0.18	0.17
	(0.38)	(0.39)	(0.36)	(0.38)	(0.38)	(0.38)
Alcohol	0.56	0.57	0.56	0.57	0.55	0.56
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Mother's locus	98.3	98.9	99.3 **	98.6	98.8	99.4
of control	(9.33)	(9.64)	(9.69)	(9.61)	(9.43)	(9.46)
EPDS	6.3	6.5	6.3	6.4	6.4	6.5
	(4.63)	(4.57)	(4.43)	(4.69)	(4.37)	(4.54)
CCEI	12.3	12.6	12.5	12.3	12.7	12.7
	(7.33)	(7.20)	(7.06)	(7.28)	(7.19)	(6.91)
Teaching score	7.0	7.0	7.0	7.1	7.0 **	7.0
	(0.91)	(0.91)	(0.86)	(0.88)	(0.94)	(0.92)
Activities (indoor)	0.69	0.68	0.69	0.68	0.69	0.70
	(0.19)	(0.20)	(0.20)	(0.20)	(0.19)	(0.16)
Activities (outdoor)	27.9	27.9	28.1	27.9	28.0	27.6
	(4.41)	(4.45)	(4.05)	(4.34)	(4.37)	(4.71)
N	1295	1666	552	2031	1295	187

Notes: * p<0.10; ** p<0.05; ***p<0.01 refers to t-tests of mean(hetero/homozygous) equals mean(non-carriers).

Table 2: OLS regressions of KS3 (age 14) on child DXA (fat mass, age 11)

	(1) KS3	(2) KS3	(3) KS3	(4) KS3
DXA	-0.095***	-0.074***	-0.044**	-0.036**
	(0.015)	(0.016)	(0.014)	(0.014)
Mother's pre-pregnancy BMI		0.029	0.139	0.149
		(0.176)	(0.152)	(0.146)
Mother's BMI squared		-0.000	-0.001	-0.001
		(0.001)	(0.001)	(0.001)
F-statistic of joint significance of mother's BMI and BMI^2		11.29	1.62	2.23
R-squared	0.01	0.02	0.26	0.29
Number of observations	3513	3513	3513	3513
Maternal pre-pregnancy BMI		Yes	Yes	Yes
Contextual variables			Yes	Yes
Mother's health and behaviour				Yes

Notes: * p<0.10; ** p<0.05; ***p<0.01, robust standard errors in parentheses.

Table 3: First stage specification tests of the Instrumental Variable regressions of child weight

	(1) IV: Maternal pre-pregnancy BMI, BMI^2	(2) IV: Lagged child weight (at age 9)	(3) IV: Genetic Markers
IV strength, F-statistic	132.5	1203.8	17.58
Under identification LM test [a]	223.5	729.3	34.57
p-value, Hansen J test	0.169	0.131	0.867
F-stat: joint sign of mother's BMI and BMI^2	132.5 [b]	58.25	130.3

Notes: [a] The Kleibergen-Paap LM statistic for under-identification; [b] The F-statistic of joint significance of mother's BMI and BMI^2 is identical to the F-statistic of IV strength, since mother's BMI and BMI^2 are the instruments in column (1); All controls included.

Table 4: Second stage IV results: KS3 on child DXA

	OLS	2SLS			
	(1)	(2)	(3)	(4)	(5)
	KS3	KS3	KS3	KS3	KS3
A. Maternal BMI as IV					
DXA, age 11	-0.042***		-0.309***	-0.105**	-0.108**
	(0.014)		(0.0523)	(0.0461)	(0.0453)
B. Child's lagged weight as IV					
DXA, age 11	-0.036**	-0.0846***	-0.0623***	-0.0312	-0.0247
	(0.014)	(0.0197)	(0.0211)	(0.0195)	(0.0193)
C. Genetic markers as IV					
DXA, age 11	-0.036**	0.0590	0.0830	0.000	0.0506
	(0.014)	(0.143)	(0.153)	(0.133)	(0.132)
Maternal pre-pregnancy BMI	Yes		Yes	Yes	Yes
Contextual variables	Yes			Yes	Yes
Mother's health and behaviour	Yes				Yes

Notes: Panel A only includes mother's BMI and BMI^2 as instruments; i.e. they are not included in the model, hence the OLS estimate in column 1 differs from that of panel B and C and from Table 2. * p<0.10; ** p<0.05; ***p<0.01, robust standard errors in parentheses.

Table 5: Robustness checks, IV regressions of KS3 on different indicators of child weight

	Measured at	IV estimate	Robust SE	N
DXA	age 11	0.051	(0.132)	3513
DXA	age 9	0.101	(0.134)	3643
BMI	age 11	0.047	(0.138)	3568
BMI	age 9	0.113	(0.127)	3819
% Body fat from DXA	age 11	0.057	(0.139)	3516
% Body fat from BIA	age 11	0.157	(0.267)	2887
Overweight	age 11	1.667	(4.219)	3513

Notes: * p<0.10; ** p<0.05; ***p<0.01, robust standard errors in parentheses, all controls included.

Table 6: Second stage IV results: KS3 on child DXA, imputed sample

	OLS	2SLS			
	(1)	(2)	(3)	(4)	(5)
	KS3	KS3	KS3	KS3	KS3
A. Maternal BMI as IV					
DXA, age 11	-0.043***		-0.431***	-0.154***	-0.165***
	(0.010)		(0.0584)	(0.0493)	(0.0484)
B. Child's lagged weight as IV					
DXA, age 11	-0.037***	-0.0746***	-0.0549***	-0.0108	-0.0048
	(0.011)	(0.0157)	(0.0164)	(0.0145)	(0.0142)
C. Genetic markers as IV					
DXA, age 11	-0.037***	0.0286	0.0477	-0.022	0.0351
	(0.011)	(0.135)	(0.144)	(0.117)	(0.119)
Maternal pre-pregnancy BMI	Yes		Yes	Yes	Yes
Contextual variables	Yes			Yes	Yes
Mother's health and behaviour	Yes				Yes

Notes: Panel A only includes mother's BMI and BMI^2 as instruments; i.e. they are not included in the model, hence the OLS estimate in column 1 differs from that of panel B and C and from Table 2. * p<0.10; ** p<0.05; ***p<0.01, robust standard errors in parentheses.

Figure 1: Distribution and weight functions of DXA for FTO genotypes

Figure 1a

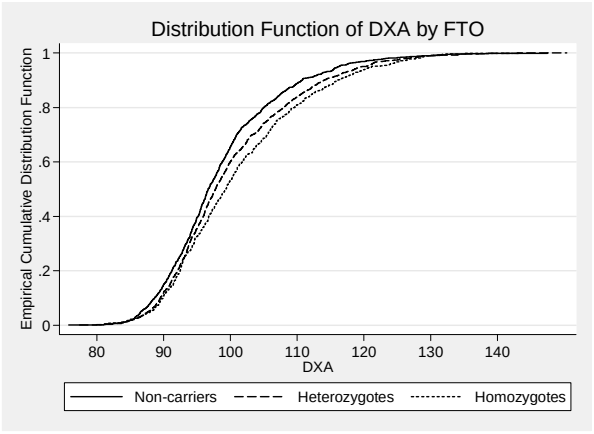


Figure 1b

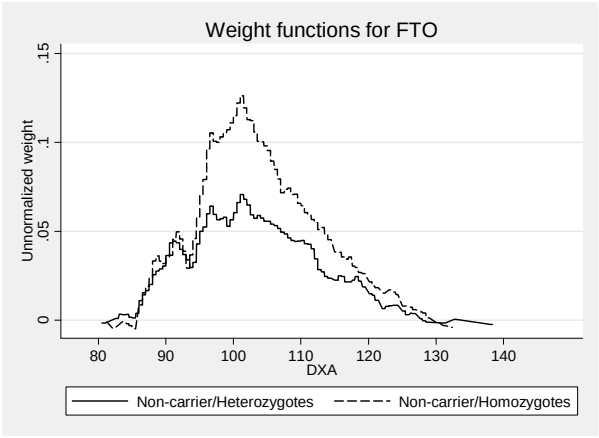


Figure 2: Distribution and weight functions of DXA for MC4R genotypes

Figure 2a

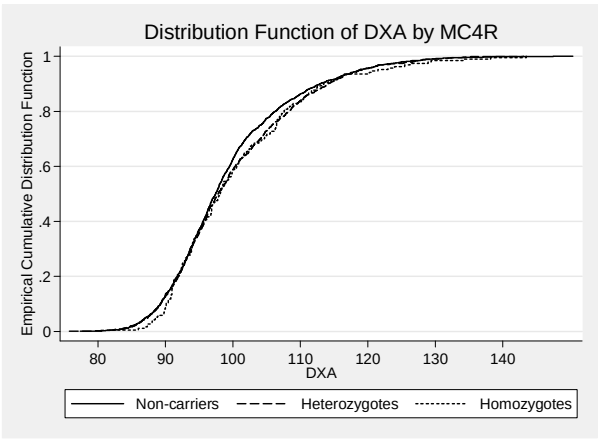


Figure 2b

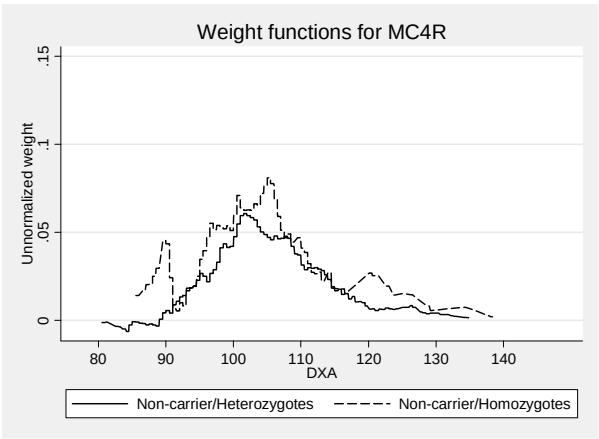
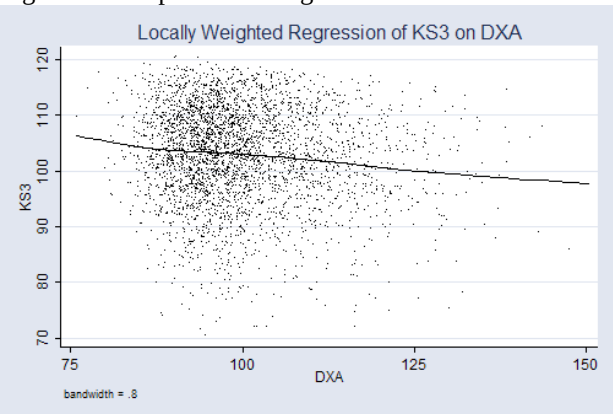


Figure 3: Non-parametric regression of KS3 on DXA



Appendix A: Directed Acyclic Graphs

Figure A1: A directed acyclic graph (DAG) of the effect of child weight *W* on educational outcomes *S*, using instruments *Z*

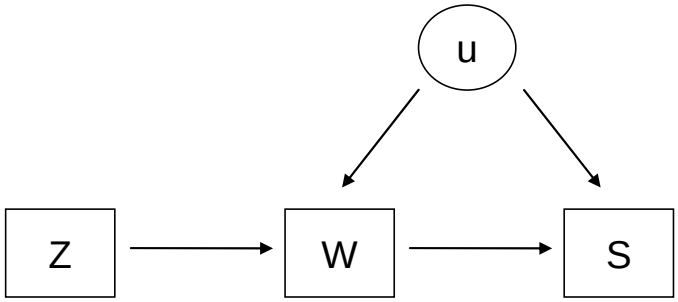


Figure A2: A DAG of the effect of weight *W* on educational outcomes *S*, using lagged weight *L* as instruments

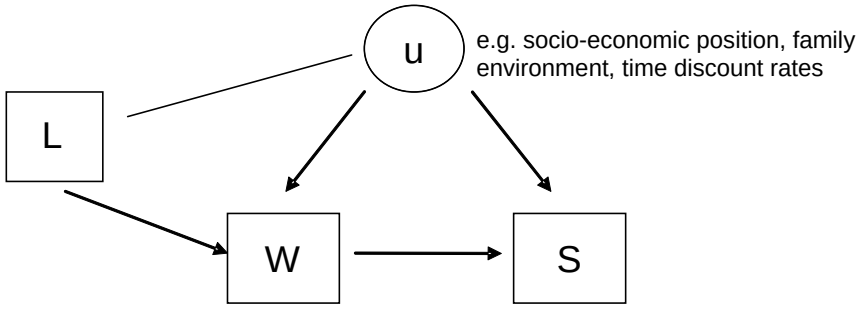


Figure A3: A DAG with maternal lagged weight *M* as part of, or proxy for, parental inputs *P*

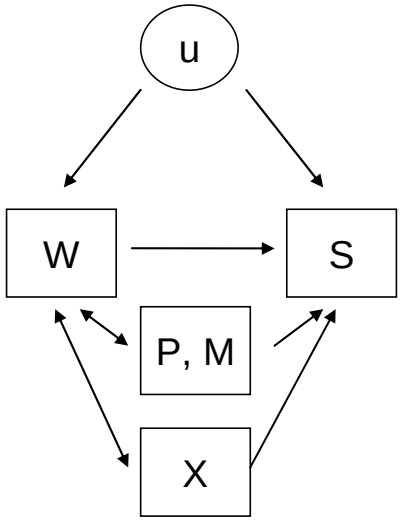
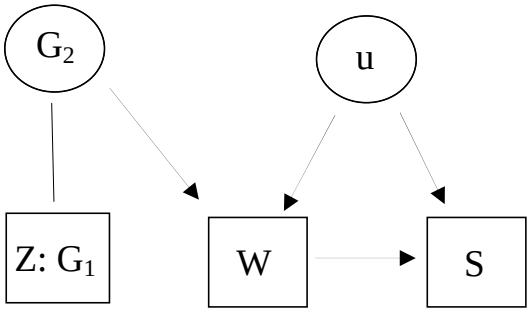
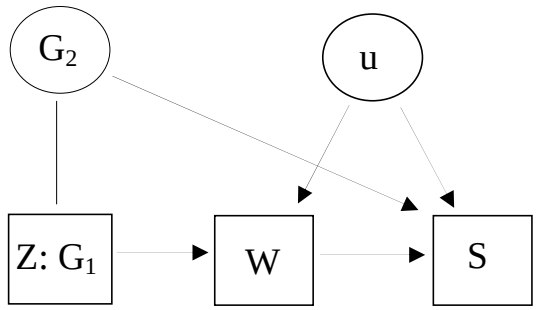


Figure A4: Two DAGs, showing a Mendelian randomization experiment with linkage disequilibrium where (in Figure A4-a) the IV assumptions are not violated, and (in Figure A4-b) the IV assumptions are violated

(a)



(b)



Appendix B: A Brief Introduction to Genetics

Each cell in the human body contains a nucleus in which most DNA (99.9995%) is kept.[34] DNA is stored in structures called chromosomes, where each chromosome contains a single continuous piece of DNA. All cells in the human body apart from gametes (i.e. germ cells) contain 46 chromosomes, organised into 23 chromosome pairs: one copy of chromosome 1-22 from each parent, plus an X-chromosome from the mother and either an X or a Y chromosome from the father.

Locations (or loci) where DNA varies between people are called polymorphisms. The most commonly studied form of polymorphism is a Single Nucleotide Polymorphism (SNP): a single base-pair variation in a DNA locus. As chromosomes come in pairs, humans have two variations at each locus, called alleles. These alleles can either be the same or different. The term genotype is used to describe the specific set of alleles inherited at a particular chromosome locus. For example, individuals can have one of three genotypes of the FTO SNP (one of the genetic variants used here): they can be homozygous for the common allele (TT), heterozygous (AT), and homozygous for the rare allele of *FTO* (AA).[35] The visible or measurable effect of a particular genotype is called the phenotype.

The phenotype we examine in this paper is child weight. Studies that examine the heritability of weight generally report large proportions of the variance that are due to genetics: between 0.4 and 0.7 (Plomin, 1986).[36] A high heritability however, does not imply that any individual genetic variant has large phenotypic effects. For example, there are many

[34] A small amount of DNA exists in the mitochondria, structures that supply the cell with energy. The remainder of this section refers only to nuclear DNA, which is the DNA used to obtain genetic variation in this (and most genetic epidemiology) studies.

[35] Conventionally, italics are used to indicate the name of a genetic variant (e.g. *FTO*); not italicising indicates the protein influenced by a particular genetic variant (e.g. the FTO SNP). We use the same convention.

[36] The heritability of a characteristic is defined as the proportion of the total variance that is explained by genetic factors. It is most commonly calculated from twin studies by comparing intra-pair correlations for the characteristic in monozygotic (MZ) twins with intra-pair correlation in dizygotic (DZ) twins. The implicit assumption is that the effects of shared environmental factors are similar for MZ and DZ twins. The heritability is of a characteristic is calculated as twice the difference between MZ and DZ intra-pair correlations ($h^2 = 2*(r_{MZ}-r_{DZ})$). The fact that heritability is a relative measure means that it will differ between populations that have a different prevalence of non-genetic causes of the characteristic.

different SNPs that affect human weight, though all with small effects: so-called ‘polygenes’. Together, these variants will have a large phenotypic effect.

Mendelian randomization refers to the random assortment of genes from parents to children that takes place at conception. It uses Mendel’s second law that states that the inheritance of one trait is independent from the inheritance of another. Despite the random allocation of alleles being at the level of parents to offspring, this randomness seems also to hold at the population level: genetic variants are generally unrelated to confounders that often plague studies in the economics and observational epidemiologic literature, such as socio-economic position and lifestyle factors (Bhatti et al., 2005; Davey Smith et al., 2008; Kivimäki et al., 2008; Lawlor et al., 2008). This therefore suggests that the genetic markers are exogenous to behavioural or environmental factors that may affect the outcome of interest. Instrumenting the phenotype with the genetic variant will therefore isolate the causal effect from any confounding factors, such as the choices made by children and parents. We discuss this in more detail in section II.

Appendix C: The Hardy-Weinberg Equilibrium (HWE)

The HWE states that genotype frequencies in a population remain constant from generation to generation unless specific disturbing influences are introduced. These include non-random mating, mutation, gene flow, genetic drift and natural selection.[37] Consider the two *FTO* alleles, A and T, with population proportions p and q respectively. There are no other alleles, so $p + q = 1$. The proportions of AA, AT and TT genotypes in the population are given by $(p+q)^2 = p^2 + 2pq + q^2$, i.e. p^2, $2pq$, and q^2 respectively. This is also referred to as the Hardy-Weinberg proportions, which hold under the assumption of the absence of the above mentioned influences. Hence under the HWE, p and q are independent.

Deviations from the HWE can be tested by comparing the observed genotype frequencies in the data with the expected frequencies based on the HWE, using a χ^2-test. The null hypothesis states that the data are in Hardy-Weinberg proportions.[38] The results for the child's and mother's genotypes are presented in Table C.

Table C1: Testing the Hardy-Weinberg Equilibrium

	Child		Mother	
	FTO	*MC4R*	*FTO*	*MC4R*
% Non-carrier	36.9	57.8	37.1	59.3
% Heterozygous	47.4	36.9	47.9	35.5
% Homozygous	15.7	5.3	15.0	5.2
Total observations	3513	3513	2266	2253
% observed allele 1	60.6	76.3	61.1	77.1
% observed allele 2	39.4	23.7	38.9	23.0
$\chi^2(1)$:	0.177	1.095	0.145	0.038
p-value:	0.672	0.306	0.724	0.905

[37] Many of these do not generally refer to humans. For example, gene flow increases the variability of the gene pool, as members of local populations with a distinct gene pool mate with occasional immigrants from an adjacent population of the same species, introducing new genetic variants or altering gene frequencies in the residents. Genetic drift may occur if the population is very small. Chance may then drift an allele frequency to higher or lower values, ultimately causing the entire population to be homozygous.

[38] These tests have relatively low power (Lewontin and Cockerham, 1959).

Appendix D: *FTO, MC4R* and a Random Set of Additional Variables

Table D1: Coefficients (std err) of the indicators presented in the first column regressed on *FTO* and *MC4R*

	FTO		*MC4R*	
Sleep variables				
Length of night's sleep (school day), 81 months	0.014	(0.017)	-0.011	(0.020)
Length of night's sleep (Saturday), 81 months	0.011	(0.022)	-0.031	(0.025)
Chid has regular sleeping routine, 81 months[1]	-0.003	(0.005)	0.001	(0.005)
Child has difficulty sleeping, 81 months[1]	-0.016	(0.011)	0.007	(0.013)
Sleeping problem anxiety score, 81 months	0.029	(0.034)	-0.010	(0.036)
Behaviour / self-esteem child				
Child is picked on / bullied, 9 years[1]	-0.009	(0.009)	-0.015	(0.011)
Depression score child, 9 years	0.325	(0.225)	0.164	(0.258)
Anti-social score child, 9 years	0.216	(0.206)	-0.386	(0.239)
Child locus of control, 8 years	0.023	(0.244)	-0.264	(0.291)
Child's scholastic competence score, 8 years	0.063	(0.241)	-0.110	(0.283)
Child's global self worth score, 8 years	-0.012	(0.237)	-0.178	(0.279)
Child's total self esteem, 8 years	0.019	(0.205)	-0.123	(0.240)
Strength & Difficulties Questionnaire (SDQ)				
Anti-social behaviour (mother-reported), 9 years	-0.044	(0.224)	0.268	(0.256)
Hyperactive behaviour (mother-reported), 9 years	-0.262	(0.217)	0.407	(0.257)
Emotional problems (mother-reported), 9 years	-0.152	(0.218)	0.027	(0.254)
Conduct problems (mother-reported), 9 years	-0.056	(0.220)	0.305	(0.260)
Peer problems (mother-reported), 9 years	-0.300	(0.216)	0.036	(0.259)
Learning difficulties				
Freq. to special class due to learning difficulties, 81 months	0.013	(0.014)	0.021	(0.019)
Freq. to special class due to learning difficulties, 9 years	0.003	(0.020)	-0.004	(0.023)
Freq. to special class due to learning difficulties, 11 years	0.033*	(0.018)	0.022	(0.022)
Child ever had speech/language therapy, 91 months[1]	-0.004	(0.007)	0.002	(0.008)
Child has dyslexia (mother-reported)[1]	-0.002	(0.004)	0.003	(0.005)
Child is autistic (mother-reported)[1]	0.001	(0.002)	0.001	(0.002)
Mother's health and behaviour				
Mother's self-esteem (Bachman score)	0.310*	(0.184)	0.056	(0.220)
Mother's depression score, 18 weeks gestation	0.037	(0.032)	0.004	(0.038)
Mother's somatic problems score, 18 weeks gestation	-0.001	(0.035)	0.033	(0.040)
Financial situation of the household				
House is owner-occupied, 21 months[1]	-0.005	(0.007)	0.008	(0.008)
House is rented or via housing association, 21 months[1]	0.001	(0.005)	0.001	(0.006)
Council housing, 21 months[1]	0.002	(0.005)	-0.010*	(0.005)
Indicators at birth of child				
Month of birth (1=September, 12 = October)	0.016	(0.039)	0.012	(0.043)
Admission to special care birth unit[1]	-0.008	(0.005)	0.005	(0.006)
Multiple births (twins or triplets)[1]	-0.001	(0.003)	0.002	(0.004)
Gestational age at delivery	0.027	(0.032)	0.025	(0.039)
Caesarean section[1]	-0.006	(0.007)	0.008	(0.008)
Different measures of child weight / fat mass				
Waist circumference, 11 years	1.14***	(0.198)	0.85***	(0.242)
Hip circumference, 11 years	0.95**	(0.172)	0.43**	(0.210)

Notes: [1] Binary indicator. * p<0.10; ** p<0.05; *** p<0.01