Reconsidering the impact of family size on labour supply: The twin-problems of the twin-birth instrument

Nils Braakmann and John Wildman

Newcastle University*

[This version: January 16, 2015]

Abstract

We consider two econometric problems when investigating the impact of family size on labor market outcomes using the popular twin-birth instrument. The first is the potential for omitted variable bias caused by the fact that fertility treatments are linked to twin births and are typically unobserved. We present estimates corrected for this bias and find it to be comparatively small. Second, we show that the effects of twin-birth induced variation in family size as well as characteristics of the compliers vary substantially with time passed since birth, which has consequences for the interpretation of estimates across samples and time.

Keywords: Twin birth instrument, labor supply, fertility

JEL-classification: C26, J13, J22

Word count: 7694 (main text + references)

* Both Newcastle University, Business School – Economics, 5 Barrack Road, Newcastle upon Tyne, NE1 4SE, UK. Email: nils.braakmann@newcastle.ac.uk; john.wildman@newcastle.ac.uk.

All estimates used Stata 13.1. Do files are available from the first author on request. A previous version of this paper was circulated as "Fertility treatments and the use of twin births as an instrument for family size". We thank Marco Alfano, William H. Green, Victor Lavy, Christian Merkl, Steffen Mueller, Regina Riphahn, seminar participants in Hannover, Lueneburg, Newcastle and Nuremberg as well as participants at the 2014 EALE and ESPE conferences for comments.

1 Introduction

Estimating the impact of family size on labour supply outcomes is complicated by endogeneity problems that necessitate the use of instrumental variables or related strategies. A popular instrument in this literature is the occurrence of twin births.[1] The occurrence of a twin birth, on face value, looks like the perfect candidate for an instrument - it is clearly correlated with family size and it appears reasonable that it affects labour supply only through family size. However, there are two potential problems with using twin births as an instrument, both of which are related to the link between fertility treatments and multiple births often documented in the medical literature. The first is the potential of omitted variable bias caused by the fact that fertility treatments are typically unobserved. We present estimates corrected for this bias and find it to be comparatively small. The second issue is more subtle: Twin births are not uniformly distributed across time but are in fact increasing both in absolute number and as the share of all maternities since the early 1980s (see figure 1), a development which is likely to be partially caused by the changing prevalence of fertility treatments. We show that the impact of the twin-induced changes in family

[1] The instrument is also commonly used for the question whether larger families result in worse outcomes for children, i.e., the quantity-quality trade-off as predicted by models by Becker and Lewis (1973). In the case of the Becker and Lewis-quantity and quality model, the endogeneity issues arise as family size and children's outcomes are jointly determined in the same parental optimization process. Examples of this literature are Rosenzweig and Wolpin (1980), Black, Devereux and Salvanes (2005) and Angrist, Lavy and Schlosser (2010).

size vary substantially with the age of the twins. This implies that twin instrument-based estimates based on a single cross-section, such as a census, will depend to some extent on the age distribution of twins in the respective population, which has consequences for the comparability of results based on different samples.

(FIGURE 1 ABOUT HERE.)

In this paper we use data from the first 3 sweeps of the British Millennium Cohort Study (MCS) that follows a random sample of babies and their mothers born during late 2000 and 2001 (see section 2 for details on the data). In a first step, we consider a threat to the future, though not necessarily past, validity of this instrument, namely the increasing use of fertility treatments, such as in vitro fertilization (IVF) or drug treatment with Clomiphene citrate. It is a well-established fact in the medical literature (e.g., Callahan et al., 1994; Gleicher et al., 2000; Fauser, Devroey and Macklon, 2005) that fertility treatments greatly increase the risk of multiple births. In fact, in the dataset used in this paper, we find that the probability of having either twins or triplets[2] increases from around 1% for women without fertility treatment, to about 13% for women with fertility treatment. Even more worryingly, 24% of all the multiple births we observe in our sample are to women who have received fertility treatment, despite them forming only 2.6% of our sample.[3] Within the UK the use of fertility treatments has increased in most years since 1991. For IVF in 1991 there

[2] We will generally talk about multiple births. The vast majority (96%) of these in our sample are twins with the remaining ones being triplets.

[3] In principle, a similar risk could arise for instruments based on miscarriage as in Buckles and Munnich (2011) if either miscarriages induce people to seek fertility treatment or fertility treated women are more likely to miscarry. However, we do not have data that would allow us to look into this issue.

were around 8,000 cycles, by 2011 this had increased to just over 60,000 (Human Fertilisation and Embryology Authority, 2012). In 1992, in the UK, 0.3% of all babies born resulted from IVF treatment, by 2010 this had increased to 2% (Human Fertilisation and Embryology Authority, 2012).

The link between fertility treatments and twin births and the potential threat for the use of the latter as an instrumental variable has sometimes been discussed in the literature (see, e.g., Angrist, Lavy and Schlosser, 2010, p. 798, who discuss a potential bias arising from fertility treatments and then use a sample restricted to a time period before fertility treatments became common in Israel), but its actual impact has yet to be quantitatively analysed. The main theoretical concern is that while multiple births are probably still more or less random conditional on having received fertility treatment, they are unlikely to be unconditionally random. Even worse, deciding to undergo fertility treatment is a choice that is likely to be correlated with a number of characteristics that also influence labor supply – most prominently a very strong wish for children, but, as we demonstrate later in this paper, also with factors such as age, education, having worked before pregnancy, being white, marriage, family planning, complications during the pregnancy (i.e., health) and the birth weight of the first-born/only child. Given that we do not observe fertility treatments in most datasets commonly used by economists, these differences will introduce correlation between multiple births and (unobserved) determinants of fertility, which will render the instrument endogenous.[4]

[4] It is important to be clear that in our view these issues do not invalidate some of the earlier results in the literature and might in fact not invalidate the future use of this instrument in countries or time period where fertility treatments are relatively uncommon. However, fertility treatments might well pose a threat for the future use

Comparisons of labor supply and other characteristics in all sweeps of our data suggest mothers with and without fertility treatment are different, regardless of the number of children resulting from the pregnancy. We then compare first stages and labor supply regressions, i.e., second stages, for six models: Our base specification is one that could be estimated using most household datasets where information on fertility treatments is missing, i.e., we just use the birth of twins or triplets as an instrument for family size on a range of outcomes related to labor supply. In a second model, we additionally condition on having received fertility treatment. A comparison of these two models allows us to quantify (and correct for) the bias in the estimates in the base model. As fertility treatments are typically unobserved in most datasets, we estimate a third model that instead conditions on a set of commonly observed variables that we know to differ between women with and without fertility treatments. Results from this model allow us to make statements about whether this conditioning strategy might be a feasible approach when information on fertility treatments is lacking. In a fourth model, we condition on both fertility treatments and the same characteristics used in the previous model. This specification allows us to check whether the correlation between pre-pregnancy characteristics and multiple births arises exclusively because of fertility treatments. Finally, given that women with and without fertility treatment are different, it is also possible that their LATEs will differ. We investigate this question by estimating separate regressions for these two groups. Our findings suggest that the instrument generally becomes stronger in the first stages

---

of this instrument in countries where they occur regularly and, more importantly, where multiple births resulting from fertility treatments are quantitatively important. Attempts to reduce the occurrence of multiple births with fertility treatments that are under way in a number of countries including the UK might also help in the future.

after conditioning on fertility treatments, while the second stage results from the first four models are qualitatively identical, i.e., the estimates always have the same sign, with only small changes in magnitude. However, the results also suggest that family size has a stronger negative effect on women who underwent fertility treatment.

In a second contribution, we demonstrate that the impact of the twin-birth induced variation in family size on labour supply depends crucially on the time passed since the occurrence of the twin births. To show this we rely on the first three sweeps of the MCS with interviews conducted 9 months (sweep I), 3 years (sweep II) and 5 years (sweep III) after birth. We find that the impact of twin births on family size (the first stage) weakens over time, which is consistent with individuals adjusting their future fertility after the random shock of a multiple birth. First stages across all 3 sweeps continue to show a strong positive relationship between the occurrence of a multiple birth and family size.

However, we document that there are major changes in the composition of the complier group across the three sweeps. Furthermore, we can expect the reduced form, i.e., the impact of the twin birth on labour supply, to vary over time as the twins grow up, attend school and (at some stage) leave their parents' home. Consequently, second stages, i.e., the ratio of the reduced form and the first stage, differ substantially across the three stages. Specifically, there are strong and negative effects of the twin-birth induced variation in family size on the mother's employment probability after 9 months. These become weaker after 3 years and essentially disappear after 5 years, which coincides with the children entering schools.

These time-varying treatment effects would be comparatively innocuous if the share of twin births was constant over time. This is, however, unlikely to be the case for at least two reasons: First, there is a general trend towards giving birth later in life

in many societies. As older mothers are more likely to give birth to twins or triplets (e.g., Black, Devereux and Salvanes, 2005), this is likely to make twins more common in more recent years. Second, the availability and price of fertility treatments, as well as their link to twin births (due to improved medical treatments), also differs widely across time. These two factors will lead to problems of comparability when looking at labour supply estimates based on the twin-birth instrument coming from various cross-sections. Estimates based on a single cross-section such as a census (as in, e.g., Angrist and Evans, 1998) identify some weighted average of these time-varying treatment effects, where the weights depends effectively on the age distribution of twins in the respective dataset. As different cross-sections are likely to have different distributions of twins, it is possible that the effects will differ across papers and datasets, even in cases where individual-level effects are identical. This in turn makes comparisons between different papers using different samples complicated as it adds another source of heterogeneity.

The remainder of this paper is organised as follows: Section describes the data, section 3 explains some of the methodological points in our paper. Results can be found in section 4. Section 5 concludes.

2 Data

We use data from three waves of the Millennium Cohort Study (MCS), which tracks a random sample of children (and their families) born during late 2000 and 2001 in the UK. Interviews were conducted at wave one when the children were around 9 months old, subsequent waves took place when the children were 3 and 5 years old. Details on the design and sampling in the MCS can be found in Dex and Joshi (2005) and Hansen and Joshi (2007). The dataset is one of the few that we are

aware of that covers fertility treatments alongside information on the mother and the development of the child. The dataset only contains mothers with at least one child, which is the group where the instrument has predictive power and which is also close to the sample restriction used by, e.g., Angrist and Evans (1998). [5]

Our estimation sample is based on the following restrictions: First, we use only cases where the mother conducted the parent interview, leading to the loss of 28 observations where the father was interviewed. Second, the MCS tracks the children born during the sampling week, not necessarily the parents, i.e., the main respondent can change in each sweep, either because the partner was interviewed or because the main carer for the child changed, for example because of adoption or death. For sweeps 2 and 3 we only use cases where the same person as in sweep 1 was interviewed, resulting in the loss of 881 (from 15,590) observations in sweep 2 and 226 (from 12,984) observations in sweep 3. We also lose some observations in each sweep due to missing values (around 150 observations each in sweeps 1 and 2 and around 100 in sweep 3). Following these restrictions we have 18340 observations for sweep 1, 14460 for sweep 2 and 12581 for sweep 3.

Our main outcomes of interest are i) various dummies for employment status, mainly whether the mother is working, self-employed, a student or at home to care for the family, ii) the mother's weekly working hours, calculated in two ways, either with zeros or with missing values for people not working, and finally, iii) whether she has

[5] The multiple birth instrument has no predictive power for the question whether someone has one vs. no child, as everyone who gives birth to twins or triplets will have decided to have at least one child. It has predictive power for the number of children beyond one as someone who planned to have one child will end up with two or three instead.

a partner who is working. In sweep 1, we additionally have information on whether she is currently on maternity leave.

We have two variables of interest: The first is whether the mother gave birth to twins or triplets. Almost all multiple births in the dataset are twins with only 10 cases of triplets. The latter are split equally between women with and without fertility treatment. Our sample contains 254 multiple births (i.e., twins or triplets) in sweep 1, of these 193 appear in sweep 2 and 170 appear in sweep 3. Our second key variable is whether the pregnancy was preceded by fertility treatment. In sweep 1, we have 478 women with fertility treatments, of these 394 remain in sweep 2 and 348 in sweep 3. The most common fertility treatment in the data is drug therapy with Clomiphene citrate, followed by various forms of in vitro fertilization. The treatment we look at in all second stage regressions is the number of children each woman has at each sweep. Note that women can have other children than the one tracked by the MCS.

Table 1 presents descriptive information on the estimation sample.

[TABLE 1 ABOUT HERE.]

3 Twin births as an instrument for fertility

A The basic identification strategy and the twin-birth instrument

To illustrate the basic identification problems we use a causal diagram (or directed acyclic graph (DAG)) (Pearl, 2000; see Morgan and Winship, 2007, for a textbook treatment). In Figure 1 each directed edge (i.e., single headed arrow) such as the one from *family size* to *Y* represents a cause-effect-relationship between variables in the model, in the sense that the variable at the origin of the edge (start of the arrow) causes the variable at the terminus. A bidirected edge, such as the one between X_1 and X_2, represents common causes of the two factors that are not part of the model.

(FIGURE 2 ABOUT HERE.)

In figure 1 we are interested in the link between *family size* and *Y* or written as a linear equation

$$Y_i = \alpha + \tau * family\ size_i + \varepsilon_{i,}. \quad (1)$$

where τ is the parameter of interest. In female labor supply regressions Y_i would typically either be a dummy for labor force status or some other measure of labor supply such as desired or actual working hours, while $family\ size_i$ would typically be the number of children the mother gave birth to, or the number of children that live in the same household as her.

A direct estimation of this link is hindered by the presence of (potentially unobserved) sets of confounding variables, X_1 and X_2.[6] In equation (1) these would be part of ε and would render $family\ size_i$ endogenous. For example, in female labor supply models, both family size and the propensity to work will be influenced by (typically unobserved) preferences for work and family size. Furthermore, a woman's work opportunities will to some extent determine the opportunity costs of childrearing.

If, initially, we ignore the issues caused by fertility treatments, one way to proceed is to use *multiple births* as an instrument for *family size*. This appears to be an attractive strategy because the biological process governing whether a pregnancy results in a singleton or multiple births is outside of the control of the respective

[6] If both X_1 and X_2 were observed, it would be possible to condition on them and use OLS, matching or other selection-on-observables estimators to look at the link between family size and the outcome.

parents and thus uncorrelated with any unobserved preferences for family life, any parental optimization process, or the opportunity costs of childrearing.[7]

In figure 1, this situation is depicted in panel (a). In this scenario, *multiple births* lead to quasi-random variation in *family size* that is unrelated to the confounders X_1 and X_2 (or equivalently to ε). In this case the probability limit of the IV estimate of τ can be written as:

$$\hat{\tau} = \tau + \frac{Cov(multiple\ birth, \varepsilon)}{Cov(multiple\ birth, family\ size)} \quad (2)$$

B. Omitted variable bias through fertility treatments

Equation (2) makes it clear that if multiple births and the unobservables, ε_i, from (1) are uncorrelated, the IV estimate will be consistent as *Cov(multiple births, ε)* would be zero and the bias term in equation (2) would disappear. A central condition for this to be plausible is that twin births are (more or less) random. However, with fertility treatments this is unlikely to be the case: Fertility treatments are known to cause multiple births and fertility treatments are likely to be correlated with at least some of the confounders: In many countries, fertility treatment is expensive and not fully covered by (state) health insurance, which implies that it is likely to be correlated with parental resources. These in turn matter for labor supply and parental investment into children as they determine the budget constraint and the (non-labor) income a parent can expect when not working. Furthermore, pregnancies preceded by fertility treatment are by definition always planned. They are also likely to be

[7] There has been some debate about the quality of this instrument (e.g., Black, Devereux and Salvanes, 2005) as it is known that multiple births become more likely for older mothers. However, it is usually comparatively easy to account for this by conditioning on age in a flexible way, for example through age dummies.

correlated with a strong desire for children as fertility treatments are generally preceded by a number of attempts to conceive naturally, i.e., they are generally not the first thing someone tries when trying to become pregnant.

Panel (b) of figure 1 illustrates the resulting problem: Fertility treatments create an association between *multiple births* and the confounders in X_1, i.e., multiple births are not randomly assigned. This in turn opens a backdoor path $Y \leftarrow X_1 \rightarrow fertility\ treatments \rightarrow multiple\ births \rightarrow family\ size \rightarrow Y$ between *multiple births* and the outcome. In more standard econometric terms, we can consider fertility treatments as an omitted variable. This means that the error term for equation (1) can be re-written as:

$$\varepsilon_i = \delta_1 * fertility\ treatment_i + v_i \quad (3)$$

where δ_1 is the marginal effect of fertility treatment on labor market decisions and v_i is a new error term that is still correlated with family size, i.e., it is likely that family size will still be endogenous after conditioning on having received fertility treatment. From (3) we can see that the covariance between multiple birth and ε_i is:

$$Cov(multiple\ birth, \varepsilon) = \delta_1 * Cov(multiple\ birth, fertility\ treatment) \quad (4)$$

Using (4) we can write the *plim* of τ as:

$$\hat{\tau} = \tau + \delta_1 \frac{Cov(multiple\ birth, fertility\ treatment)}{Cov(multiple\ birth, family\ size)} \quad (5)$$

Equation (5) demonstrates that bias of the IV estimate will depend on two elements: Firstly, the strength of the relationship between fertility treatments and the respective outcome (δ_1), i.e., how strongly the differences between mothers with and without fertility treatment affect the outcome of interest, and secondly, the importance of fertility treatments for the occurrence of multiple births, i.e., the covariance between multiple births and fertility treatments. This covariance is likely to be positive as the use of fertility treatments is consistently linked to multiple births in the

medical literature (e.g., Callahan et al., 1994; Gleicher et al., 2000; Fauser, Devroey and Macklon, 2005). Indeed, in our sample the likelihood of having multiple births is 1% for women without fertility treatment and 13% for women who had fertility treatment and 24% of all multiple births observed in the data are preceded by fertility treatments.

As an increasing number of women use fertility treatments, the second part of the bias term in (5) will become stronger as *Cov(multiple birth, fertility treatment)* will increase. It is also possible that δ_1 will change as the composition of the group of women who undergo fertility treatment changes.[8] Furthermore, it is not possible, *a priori*, to sign δ_1. For example, in labor supply regressions, it could be positive because fertility treatments are used by individuals with a higher propensity to work, or it could be negative as the use of fertility treatments will be correlated with a desire for children and that may be correlated with fewer individuals choosing employment.

Faced with these problems there are two ways to block the backdoor path $Y \leftarrow X_1 \rightarrow fertility\ treatments \rightarrow multiple\ births \rightarrow family\ size \rightarrow Y$ opened by the relationship between X_1, *fertility treatments* and *multiple births*. Firstly, if we observe fertility treatment, as we do, then it is possible to condition on it directly. This closes the backdoor path and removes any association between the confounders in X_1 and *multiple births*. Secondly, if all elements in X_1 were observed, one could condition on those directly, which would have an equivalent effect. A problem with this second strategy is that it is unlikely that all elements of X_1 are observed in any given dataset. However, as the first option is only available when the use of fertility treatments is observed, conditioning on variables that may be part of X_1 may be the

[8] Note that δ_1 would be zero if either no or all multiple births are due to fertility treatments.

only option when using datasets lacking this information. This strategy has its own risk as it may introduce further bias, rather than ameliorating the bias present: Theoretically, it is only clear that conditioning on the full set of confounders in X_1 would cause δ_1 to be zero and eliminate the bias. Conditioning on a subset of confounders can attenuate the problem if δ_1 shrinks towards zero as a result. However, it could also aggravate the problem: Consider a case where X_1 consists of only two variables, *A* and *B*, whose effects cancel each other out, so that δ_1 would be zero without conditioning. Conditioning on either one of them in this case would cause δ_1 to be non-zero and would actually increase bias.

C. Time varying treatment effects

A second issue with the use of twin birth instruments concerns the timing of the twin births or (equivalently) the age of twins in the sample. There are three issues to consider i) The impact of a twin birth on family size if people have the opportunity to adjust their fertility over time (the first stage), ii) the impact of the twin birth on labor supply, which might change over time as the twins age (the reduced form) and iii) the age distribution of twins in the respective population, which will determine the overall effect in an IV labor supply regression, as it determines the weights in the aggregation of individual-level effects to an overall effect.

Some of these issues have been considered previously in the literature, but the problems in their entirety, and their possible link to IVF, has not been fully discussed. Jacobsen et al. (1999) highlight the fact that many families will adjust their subsequent fertility decisions to compensate for the presence of twins. To illustrate this point, consider first the first stage:

$$Family\ size_i = \pi + \gamma * multiple\ birth_i + \mu_i, \qquad (6)$$

The logic behind the instrument is that the birth of a pair of twins leads to a larger-than-planned family size. In other words, the instrument only works if families cannot (fully) adjust to the arrival of an additional child. An example where this condition would be fulfilled is a woman giving births to twins at the last planned birth, i.e., a case where a woman who wanted one further, last, child receives two instead. However, it is important to note that there will be a substantial number of women for whom realised fertility in the long term is unaffected by twin births. Whenever a twin birth occurs at any birth before the last, it is, in principle possible, to adjust fertility over the following years. Say a woman always wanted two children. At her first planned pregnancy she gives birth to twins. This twin birth will have different effects in the short and the long term. In the short term, she has one more child than she planned to have at this point in time. In the long term, however, she can simply decide not to have another child and can end up with her originally planned family size. This suggests that the first stage could be written as

$$Family\ size_{it} = \pi + \gamma_t * multiple\ birth_{it} + \gamma_{t-1} * multiple\ birth_{it-1} + \gamma_{t2} * multiple\ birth_{it2} + \ldots + \gamma_{t-k} * multiple\ birth_{itk} + \mu_{it},, \quad (7)$$

i.e., we allow the effect of a multiple birth to be different dependent on when it occurred in relation to the point in time family size is measured.

This specification highlights the fact that individuals can adjust their family size post multiple birth. For households we expect the impact of multiple births on family size to fall over time as found by Rosenzweig and Wolpin (1980), Bronars and Groggar (1994) and Jacobsen et al. (1999).[9] A direct implication of this adjustment of fertility

[9] Note that the exact value of γ_k depends on the share of the multiple births being twins, triplets, quadruplets, etc. If there were only twins, γ_k would start at a value of 1. As the vast majority of multiple births tend to be twins (96% in our sample), the estimate of γ_k should start at a value close to 1 directly after birth.

is that composition of the complier group, i.e., those individuals who have a larger-than-planned family at each point in time, might change over time.

Similarly, we would expect the reduced form, i.e., the impact of a multiple birth on labour market outcomes to weaken over time as children grow up, become more independent and finally leave the household. This suggests that the reduced form could be written as

$$Y_{it} = \alpha + \lambda_t * multiple\ birth_{it} + \lambda_{t-1} * multiple\ birth_{it-1} + \lambda_{t-2} * multiple\ birth_{it-3} + ... + \lambda_{t-k} * multiple\ birth_{it-k} + \varepsilon_{it,}. \quad (8)$$

This model captures the fact that the impact of the multiple birth-induced variation in family size may change over time, for example as the children become less dependent on their mother as they grow up.

In a cross sectional model a potential problem arises when estimating and comparing labor supply regressions across different cross-sectional samples without accounting for the time passed since the multiple birth: If we estimate the first stage as in (6), $\hat{\gamma}$ is a weighted average of the $\hat{\gamma}_t$ one would get by estimating equation (7). Correspondingly, $\hat{\lambda}$ is a weighted average of the $\hat{\lambda}_t$ from equation (8). The weights in both cases depend on the age distribution of the children born in multiple births. If the age distribution of these children was constant over time, comparisons between estimates based on different samples would not be problematic as the weighting of the first stage and reduced form coefficients would be identical in the different samples. However, if, as we observe, multiple births are increasing over time then γ may be larger in later cohorts than earlier cohorts, not because the impact of multiple births on family size at the individual level is changing, but because the number of younger twins is increasing in the population. Such differences mean that comparing results from cross-sections would be affected by the distribution of twins.

(FIGURE 3 ABOUT HERE.)

Consider, for example, a case where a researcher has a single cross-sectional dataset, say a census as in Angrist and Evans (1998). Figure 3 illustrates such a situation: Say, a researcher has access to microdata from the 2000 and 2010 UK census. Realistically, twins in a census can be identified as long as they live in their parents' house, for simplicity assume that this occurs up to the age of 20. The estimates based on the 2000 census would then effectively rely on twin births that occurred during the period 1980 to 2000. This situation is depicted in panel 3(a), where the dashed lines mark this period. For the 2010 census, estimates would be based on twin births from 1990 to 2010. This is illustrated in panel 3(b). If the effects of twin births vary over time, either because effects genuinely vary with the time passed since birth or because the composition of compliers in each birth cohort changes over time, the estimates in the first and second stages will depend partially on the distribution of twins across birth cohorts and time. If more twin births occurred relatively close to the census date, the estimated effects would likely be dominated by the short-term effects, i.e., a combination of relatively fewer families being able to adjust their families and relatively young children in the families affected by multiple births. If a larger proportion of the multiple births in the population occurred earlier, however, first stages would likely be weaker as more families had time to adjust their fertility. Similarly, as the children born in the multiple births would be older, the reduced form coefficients might also be closer to zero. As the second stage is simply the reduced form divided by the first stage, i.e.

$$\tau = \lambda / \gamma, \tag{9}$$

the estimated treatment effect in the latter case could be larger or smaller than the one in the first case.

Now consider a situation where the twin-based estimates based on the 2000 and 2010 samples differ. There are in principle several explanations for this difference. First, the effect of family size on female labour supply might have changed, be it because of changes on the individual level, such as attitudes, or be it because of changes to public policy, such as child care. A second explanation would be that the distribution of twin births over time (i.e., the age distribution of twins) in the two samples is different, leading to a different weighting of the time-varying effects of the twin-birth induced fertility. A third possible explanation is changes in the composition of the compliers in both samples. Furthermore, if the frequency of multiple births in the population is related to IVF decisions, the endogeneity problem we discussed earlier might also be more or less severe in one of the two samples.

It is important to be clear that while these arguments do not necessarily point towards a "bias" in the conventional definition, they are definitely another source of heterogeneity that hinders the comparison of results across papers using different samples.

D. Modelling

In the following we estimate and compare six models across our three samples collected at different intervals after birth.[10] The first model uses information that

[10] We have framed the discussion in this section in terms of a continuous outcome Y as most of the literature uses linear models. We have also estimated instrumental variable probits for the binary outcomes that we use, such as whether the individual is employed. The magnitude of the results is comparatively similar to the 2SLS results that we present. More importantly, the relative pattern of results across the different models, which matters for this paper, is practically identical. In other words, using an

would be available in most datasets and ignores the availability of information on fertility treatments, i.e., we just instrument for family size using a dummy for whether the woman gave birth to twins or triplets. The second includes a control for whether she also received fertility treatment. Estimates from this model are consistent as conditioning on fertility treatments is sufficient for the multiple births instrument to be valid. A comparison of these two models provides a picture of the size of the bias caused by unobserved fertility treatments. As a third model we condition on a set of variables that should be available in most datasets lacking information on fertility treatments, variables that could plausibly be part of X_1. These include the education of the mother, whether she worked before the pregnancy, age at birth, ethnicity and marital status.[11] A comparison of this model with the two previous models allows us to judge whether this conditioning strategy helps to attenuate any eventual bias. In a fourth model, we condition on both fertility treatments and the previously mentioned pre-pregnancy characteristics. Our discussion suggests that the only link between X_1 and *multiple birth* arises due to fertility treatments. If this is indeed the case, conditioning on pre-pregnancy characteristics and fertility treatments should not lead to different results than conditioning on fertility treatments alone. Finally, as women with and without fertility treatments are clearly different, we also evaluate whether the first and second stages for them are different. To do this we estimate separate

---

IV probit instead of 2SLS (unsurprisingly) does not help at all with an eventual bias caused by fertility treatments being unobserved.

[11] Given the relative richness of information in the MCS we could condition on additional variables. However, we deliberately restrict our choice to variables that are realistically available to researchers trying to use the multiple birth instrument with standard household data.

models for the two groups and compare the results. All of these estimates include dummy variables for the current age of the mother in years to control for the earlier discussed age differences between single and multiple birth mothers.

We also test for differences in the characteristics of the compliers within a single birth cohort over time. Compliers are generally unobservable in the data, however, there are ways to characterize them (Angrist and Pischke, 2009, pp. 166-172). In particular, for discrete characteristics x_i, we can describe the likelihood of a complier having that characteristic relative to the population by dividing the first stage for the sub-sample with $x_i = 1$ by the overall first stage. The resulting complier-population-ratios should be interpreted as relative likelihoods, i.e., a value of 2 indicates that compliers are twice as likely to have the respective characteristic than the general population. Values above 1 indicate that the characteristic is more common among the compliers than in the population and values below 1 indicate the opposite. All the characteristics we consider are based on pre-pregnancy characteristics, i.e., they are by construction unaffected by a later multiple or singleton birth. We repeat this exercise for all three sweeps of our data and compare results.

E. Descriptive comparisons

[TABLE 2 ABOUT HERE.]

Table 2 compares the pre-pregnancy characteristics of women based on sweep 1 of the MCS. There are a range of statistically significant and economically large differences in the table that give reason for concern: Women with fertility treatment are more likely to have a (higher or first) degree, are less likely to have no qualification, are on average 4 years older at birth, are 20 percentage points more likely to have worked before the pregnancy or to be married, are a lot less likely to be single, are 6 percentage point less likely to be non-white, have somewhat smaller

families at sweep 1 (despite the higher likelihood of multiple births), are 13 percentage points more likely to have experienced complications during pregnancy and are 47 percentage points less likely to have an unplanned pregnancy. For most of these factors it is easy to imagine a link with labor supply. Importantly, running a regression of a dummy for having received fertility treatment on these variables results in an R^2 of about 0.04, suggesting that these variables are by far not the only thing in which women with and without fertility treatment differ.

[TABLE 3 ABOUT HERE.]

As stated before it should be possible to use multiple births as an instrument after conditioning on fertility treatments, as multiple births are probably still conditionally random. Table 3 provides some evidence on this conjecture. We compare the same characteristics as in table 2 between women with singleton and multiple births conditional on having received fertility treatment. The picture painted in this table is a lot rosier than the one in table 2: While there are still some significant differences between women with single and multiple births in each group, these are generally a lot smaller and often not statistically significant. These suggest that using multiple births as an instrument for family size might be possible as long as we are able to condition on having undergone fertility treatment.

4 Female labor supply

We begin by documenting differences in the outcomes between women with and without fertility treatments conditional on having had a single or a multiple births.

[TABLE 4 ABOUT HERE.]

Table 4 documents these differences: In general, single-birth women with and without fertility treatment appear to be quite different. Women with fertility treatment

are more likely to have a working partner in all sweeps and are also significantly more likely to be working in both sweeps 1 and 2. They are also more likely to use paid childcare. The differences in employment appear to disappear by sweep 3 when most, i.e., 99%, of the children in our data attend school. For those who work, working hours do not appear to be too different. Women with multiple births in the two groups appear to be much more similar. While there are still differences in the probability of having a working partner in all sweeps, the gap in employment probabilities is much smaller than among single-birth women and only significantly different from zero in sweep 1. These results suggest that there are some differences between the groups that are not related to variations in family size caused by multiple births. We now evaluate whether these also lead to differences in the first and second stages of standard labor supply regressions.

[TABLE 5 ABOUT HERE.]

Table 5 begins with the first stage regressions. Consider first the two models in columns (i) and (ii). The inclusion of a control for fertility treatment clearly strengthens the relationship between multiple births and family size: The coefficient on multiple births increases by around 20% in sweeps 1 and 2 and by about 25% in sweep 3. At the same time, the first stage F-value increases substantially. Conditioning on pre-pregnancy characteristics in column (iii) strengthens the first-stage relationship, but does very little to the first-stage coefficient on multiple births relative to column (i). The results from column (iv) where we condition on both fertility treatments and pre-pregnancy characteristics leads to results that are very similar to column (ii), but with a slightly higher F-value. The latter is simply the familiar result that IV estimates improve in precision after conditioning on other exogenous variables.

Comparing the first stages for women with and without fertility treatment in columns (v) and (vi) reveals that the instrument is a much better predictor of family size for women with fertility treatments with much higher first stage R^2-values and equal F-values despite a much smaller sample size.

Finally, the evidence in table 5 suggests that the time passed since the twin birth matters for the results: 1 year after the birth the impact of a multiple birth on family size (measured at the respective survey) are substantially larger than in later sweeps. In fact, in the models that are likely to be unbiased the impact on family size is slightly above 1, which is sensible given that women did not have time to adjust their future fertility in response to the multiple birth and a twin would result in one extra child, while the few triplets in our data would result in 2 extra children. In later sweeps, women had time to make adjustments to their fertility, which should enable some of them to go back to their target family size. However, the instrument remains strong with a positive on family size, suggesting that a substantial share of mothers end up with more children than they originally wanted.

[TABLE 6 ABOUT HERE.]

Table 6 presents results from the characterisation of compliers in each of the three samples. In sweep 1, compliers appear to be more likely to have had a surprising pregnancy and either no or relatively high qualifications and less likely to medium qualification like O and A-levels or an undergraduate degree. They are also less likely to experienced problems during the pregnancy, while compliers and the general population appear to be quite similar in terms of ethnicity, marriage and employment before the pregnancy. Over time, we can see marked changes in the composition of the compliers: In sweep 3, compliers are more likely to come from low and medium

qualification up to A-level, while those with diplomas or degress become relatively less frequent over time. In sweep 3, compliers are also much more likely to be non-white than the general population and have about the same share of people who experienced problems during the pregnancy. Furthermore, individuals for with surprising pregnancies become more frequent among the compliers relative to sweep 1. Overall, the evidence suggests that the composition of compliers changes quite markedly with time passed since the multiple birth.

[TABLES 7, 8 AND 9 ABOUT HERE.]

A bigger question is to what extent these differences matter for second stage results? Tables 7 to 9 present evidence for sweeps 1, 2 and 3 respectively. The first thing to notice is that results in columns (i) and (ii) are generally similar. Having more children lowers the propensity to be working in favour of staying at home and caring for the family. These effects also appear to be stronger when at least one of the children is young, and decline as the child ages (across sweeps 1 to 3). There also does not appear to be any effect on the working hours for those who are working. The relatively similarity of the results in these two columns suggest that the bias from omitting fertility treatments might be negligible.

The results from columns (iii) suggest that conditioning on pre-pregnancy characteristics also does not lead to substantial changes in results. However, there are several cases where the size of coefficients in column (iii) is different from those in both columns (i) and (ii). This finding highlights that conditioning on a subset of potential confounders might sometimes make matters worse. The results in column (iv) generally suggest that adding pre-pregnancy characteristics does not change the results if we also condition on fertility treatments. This result again suggests that the

only source of correlation between multiple births and mothers' characteristics arises because of fertility treatments.

The third thing to note from columns (v) and (vi) is that the magnitude of the effects seems to differ between women with and without fertility treatment. In general, it appears that the negative effects are much larger for women who received fertility treatment. This result is plausible as one might expect that women who underwent the trouble and (potentially considerable) cost to undergo fertility treatment are also more likely to sacrifice part of their career to look after these children. In sum, the results suggest that despite existing behavioural differences between women with and without fertility treatment the bias in labor supply regressions relying on a multiple birth instrument appears to be comparatively small. There are however differences in the magnitude of the effects of an additional child in the two groups with the family penalty appearing to be larger for women who underwent fertility treatment.

Comparing results over the three sweeps suggest very different effects on labour supply: For sweep 1, the effect of the twin-birth induced variation in family size on female employment is strongly negative and both economically and statistically significant. We also see a that most of these women, remain at home to look after their family. Three years after the birth the effects are still similar in magnitude, even though they have become weaker in terms of statistical significance. After 5 years, however, the picture changes substantially: Point estimates are a much closer to zero and are always insignificant. This pattern of result implies that one might get very different results from a dataset where most of the twin births occurred several years before the sampling period than from one where most twin births are relatively recent. It also suggests that estimates from any cross-sectional dataset will

always depend on the distribution of birth dates for the twins (or triplets) in the sample.

5 Conclusion

This paper evaluated the rise of fertility treatments as a threat to the commonly used multiple birth instrument for family size. Fertility treatments might threaten this identification strategy as they are linked to the occurrence of multiple births as well as to a range of characteristics that might influence labor supply. Using the British Millennium Cohort Study, which allows us to distinguish between women with and without fertility treatment, we investigate the consequences of usually not being able to control for fertility treatment in labor supply regressions.

We find that there are indeed differences, both in pre-pregnancy characteristics and outcomes, between women with and without fertility treatments. Conditional on having undergone fertility treatment, the birth of twins or triplets appears to be a random event. Fortunately, first stage results usually do not change much between specifications with and without controls for fertility treatments, but including fertility treatment controls appears to strengthen the first stage relationship. The bias in the second stages that arises from omitting fertility treatment controls appears to be comparatively small in magnitude and does not affect qualitative results. In all specifications, conditioning instead of a set of typically observed pre-pregnancy characteristics does not appear to help very much and might in fact cause a different type of bias. We find evidence that effects differ between women with and without fertility treatments (or their respective children), which might be because of higher resources among women with fertility treatments or because this group is more strongly selected in terms of the desire to have children. We also find evidence that

effects depend strongly on the time passed since the birth of the twins: First stages become weaker over time even though the instrument remains strong throughout. We also observe that the composition of compliers is changed as individuals adjust their fertility over time. Second stages change considerably between regressions at 9 months, 3 and 5 years after the births with point estimates getting closer to zero and weaker statistical significance. This pattern of result implies that one might get very different results from a dataset where most of the twin births occurred several years before the sampling period than from one where most twin births are relatively recent. It also suggests that estimates from any cross-sectional dataset will always depend on the distribution of birth dates for the twins (or triplets) in the sample.

References

Angrist, Joshua D. and William N. Evans (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review 88(3)*, pp. 450-477.

Angrist, Joshua D. and Joern-Steffen Pischke (2009). *Mostly harmless econometrics – an empiricist's companion*. Princeton University Press.

Angrist, Joshua, Victor Lavy and Analia Schlosser (2010). Multiple experiments for the causal link between the quantity and quality of children. *Journal of Labor Economics 28(4)*, pp. 773-823.

Becker, Gary S. and H. Gregg Lewis (1973). On the interaction between the quantity and quality of children. *Journal of Political Economy 81(2)*, pp. S279-288.

Black, Sandra, Paul Devereux and Kjell G. Salvanes (2005). The more the merrier? The effect of family composition on children's outcomes. *The Quarterly Journal of Economics* 120(2), pp. 669-700.

Bronars, Stephen G and Jeff Grogger (1994). The economic consequences of unwed motherhood: Using twin births as a natural experiment. *American Economic Review 84(5)*, pp. 1141-56.

Buckles, Kasey S. and Elizabeth L. Munnich (2012). Birth spacing and sibling outcomes. *Journal of Human Resources 47(3)*, pp. 613-642.

Callahan, Tamara L., Janet E. Hall, Susan L. Ettner, Cindy L. Christiansen, Michel F. Greene and William F. Crowley (1994). The economic impact of multiple-gestation pregnancies and the contribution of assisted-reproduction techniques on their incidence. *New England Journal of Medicine 331(4)*, pp. 244-249.

Dex, S. and Joshi, H. (2005) *Children of the 21st century: from birth to nine months*, Policy Press, Bristol, UK

Fauser, Bard C.J.M., Paul Devroey and Nick S. Macklon (2005). Multiple birth resulting from ovarian stimulation for subfertility treatment. *The Lancet 365(9473)*, pp. 1807-1816.

Gleicher, Norbert, Denise M. Oleske, Ilan Tur-Kaspa, Andrea Vidali and Vishvanath Karande (2000). Reducing the risk of high-order multiple pregnancy after ovarian stimulation with gonadotropins. *New England Journal of Medicine 343(1)*, pp. 2-7.

Hansen, K. and Joshi, H. (2007) *Millennium Cohort Study Second Survey: a user's guide to initial findings*, Institute of Education, London, UK

Human Fertilisation and Embryology Authority (2012) *Fertility Treatment in 2012: Trends and Figures*, HFEA, UK

Jacobsen, Joyce P., James Wishart Pearce III and Joshua L. Rosenbloom (1999). The Effects of Childbearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment. *Journal of Human Resources 34(3)*, pages 449-474.

Morgan, Stephen L. and Christopher Winship (2007). *Counterfactuals and causal inference*. Cambridge University Press, Cambridge.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Rosenzweig, Mark and Kennth I. Wolpin (1980). Testing the quantity-quality fertility model: The use of twins as a natural experiment. *Econometrica 48(1)*, pp. 227-240.

Qualifications and Curriculum Authority (2003). *Foundation Stage Profile Handbook*. London.

Table 1: Descriptive statistics labor supply sample

Variable	Observations	Mean	Std.dev.	Min.	Max.
Twin birth	18340	0.013	0.11	0	1
Triplet birth	18340	0.001	0.02	0	1
Multiple birth	18340	0.014	0.12	0	1
Had fertility treatment	18340	0.026	0.16	0	1
Pregnancy was surprising	18340	0.460	0.50	0	1
No qualification	18340	0.195	0.40	0	1
Qualification up to O-level/GCSE or equivalent	18340	0.335	0.47	0	1
A-level	18340	0.093	0.29	0	1
Higher education diploma	18340	0.084	0.28	0	1
First degree	18340	0.124	0.33	0	1
Higher degree (Master, PhD)	18340	0.033	0.18	0	1
Age at birth	18340	28.326	5.95	14	51
Had job before pregnancy	18340	0.023	0.15	0	1
Non-white ethnicity	18340	0.159	0.37	0	1
Married (1st marriage)	18340	0.555	0.50	0	1
Remarried (2nd or higher marriage)	18340	0.041	0.20	0	1
Single	18340	0.335	0.47	0	1
Divorced or separated	18340	0.068	0.25	0	1
Illness or problems during pregnancy	18340	0.378	0.48	0	1
Fertility and outcomes at time of sweep 1 interview (within 1 year of birth)					
Number of children	18340	1.953	1.09	1	10
Age	18340	29.137	5.95	14	52
Employed	18340	0.400	0.49	0	1
On maternity leave	18340	0.018	0.13	0	1
Self-employed	18340	0.026	0.16	0	1
Student	18340	0.009	0.09	0	1
At home to care for family	18340	0.542	0.50	0	1
Weekly working hours (includes 0)	18340	11.745	14.66	0	86
Weekly working hours (excludes 0)	8669	24.848	11.35	1	86
Has working partner	18340	0.724	0.45	0	1
Fertility and outcomes at time of sweep 2 interview (3 years after birth)					
Number of children	14460	2.221	1.08	1	13
Age	14460	31.854	5.85	17	54
Employed	14460	0.477	0.50	0	1
Self-employed	14460	0.008	0.09	0	1
Student	14460	0.012	0.11	0	1
At home to care for family	14460	0.437	0.50	0	1
Weekly working hours (includes 0)	14460	12.447	14.38	0	114
Weekly working hours (excludes 0)	7558	23.814	11.19	1	114
Has working partner	14460	0.752	0.43	0	1
Uses childcare by conducted by relatives/friends	14460	0.282	0.45	0	1
Uses paid childcare	14460	0.126	0.33	0	1
Fertility and outcomes at time of sweep 3 interview (5 years after birth)					
Number of children	12581	2.394	1.06	1	13
Age	12581	34.124	5.81	18	58
Employed	12581	0.527	0.50	0	1
Self-employed	12581	0.011	0.11	0	1
Student	12581	0.012	0.11	0	1
At home to care for family	12581	0.371	0.48	0	1
Weekly working hours (includes 0)	12581	13.955	14.46	0	100
Weekly working hours (excludes 0)	7390	23.758	11.10	0	100
Has working partner	12581	0.751	0.43	0	1
Child attends school	12581	0.988	0.11	0	1

Table 2: Comparison of pre-pregnancy characteristics of women with and without fertility-treatment

	Without fertility treatment		With fertility treatment		P-Value means
Variable	Mean	Std.dev.	Mean	Std.dev.	different[a]
Twin birth	0.01	0.10	0.12	0.32	0.0000
Triplet birth	0.00	0.02	0.01	0.10	0.0294
Multiple birth	0.01	0.10	0.13	0.33	0.0000
Pregnancy was surprising	0.47	0.50	0.00	0.00	0.0000
Birth weight 1st child (kg)	3.36	0.57	3.19	0.65	0.0000
Number of children at sweep 1 interview	1.96	1.09	1.54	0.75	0.0000
No qualification	0.20	0.40	0.12	0.32	0.0000
Qualification up to O-level/GCSE or equivalent	0.34	0.47	0.33	0.47	0.9835
A-level	0.09	0.29	0.11	0.31	0.3286
Higher education diploma	0.08	0.28	0.09	0.29	0.4430
First degree	0.12	0.33	0.18	0.38	0.0026
Higher degree (Master, PhD)	0.03	0.18	0.08	0.27	0.0002
Age at birth	28.22	5.94	32.29	4.94	0.0000
Had job before pregnancy	0.62	0.49	0.81	0.39	0.0000
Non-white ethnicity	0.16	0.37	0.10	0.30	0.0000
Married (1st marriage)	0.55	0.50	0.76	0.43	0.0000
Remarried (2nd or higher marriage)	0.04	0.20	0.06	0.25	0.0302
Single	0.34	0.47	0.12	0.32	0.0000
Divorced or separated	0.07	0.25	0.06	0.24	0.4015
Illness or problems during pregnancy	0.37	0.48	0.50	0.50	0.0000
Observations	17862		478		

[a] Based on two sample t-test with unequal variances.

Table 3: Comparison of pre-pregnancy characteristics of women with single and multiple births by fertility treatment

	Women without fertility treatment					Women with fertility treatment				
	Single birth		Multiple birth		P-Value means	Single birth		Multiple birth		P-value means
	Mean	Std.dev.	Mean	Std.dev.	different	Mean	Std.dev.	Mean	Std.dev.	different
Pregnancy was surprising	0.47	0.50	0.49	0.50	0.5801	0.00	0.00	0.00	0.00	
Birth weight 1st child (kg)	3.37	0.56	2.44	0.52	0.0000	3.30	0.58	2.42	0.59	0.0000
No qualification	0.20	0.40	0.23	0.42	0.2319	0.12	0.33	0.10	0.30	0.5671
Qualification up to O-level/GCSE or equivalent	0.34	0.47	0.31	0.46	0.4655	0.33	0.47	0.38	0.49	0.4690
A-level	0.09	0.29	0.06	0.24	0.0804	0.10	0.30	0.16	0.37	0.1934
Higher education diploma	0.08	0.28	0.14	0.35	0.0250	0.10	0.30	0.07	0.25	0.3539
First degree	0.12	0.33	0.13	0.34	0.6152	0.18	0.38	0.16	0.37	0.7929
Higher degree (Master, PhD)	0.03	0.18	0.03	0.16	0.5826	0.08	0.27	0.07	0.25	0.6464
Age at birth	28.20	5.94	30.12	5.70	0.0000	32.21	4.92	32.87	5.09	0.3433
Had job before pregnancy	0.62	0.49	0.64	0.48	0.5791	0.81	0.39	0.80	0.40	0.8603
Non-white ethnicity	0.16	0.37	0.12	0.32	0.0799	0.11	0.31	0.07	0.25	0.2613
Married (1st marriage)	0.55	0.50	0.59	0.49	0.2460	0.75	0.44	0.85	0.36	0.0376
Remarried (2nd or higher marriage)	0.04	0.20	0.06	0.24	0.2043	0.07	0.25	0.03	0.18	0.1630
Single	0.34	0.47	0.28	0.45	0.0871	0.12	0.33	0.05	0.22	0.0212
Divorced or separated	0.07	0.25	0.06	0.24	0.7486	0.06	0.23	0.07	0.25	0.8138
Illness or problems during pregnancy	0.37	0.48	0.46	0.50	0.0238	0.49	0.50	0.54	0.50	0.4754
Observations	17,669		193			417		61		

Table 4: Comparisons of outcomes for women with and without fertility treatment with same number of children born

	Single births					Multiple births				
	No FT		FT		P-value means different	No FT		FT		P-value means different
	Mean	Std.dev.	Mean	Std.dev.		Mean	Std.dev.	Mean	Std.dev.	
Sweep I outcomes										
Employed	0.40	0.49	0.52	0.50	0.0000	0.31	0.46	0.43	0.50	0.1127
On maternity leave	0.03	0.16	0.06	0.23	0.0078	0.02	0.14	0.03	0.18	0.6332
Self-employed	0.02	0.13	0.04	0.19	0.0222	0.04	0.19	0.08	0.28	0.2315
Student	0.01	0.09	0.00	0.07	0.2716	0.01	0.07	0.02	0.13	0.5164
At home to care for family	0.54	0.50	0.38	0.49	0.0000	0.63	0.48	0.44	0.50	0.0132
Weekly working hours (includes 0)	11.63	14.61	16.78	15.62	0.0000	9.98	14.49	15.36	15.18	0.0166
Weekly working hours (excludes 0)	24.81	11.33	25.82	11.90	0.1726	24.39	12.69	26.77	9.61	0.2744
Has working partner	0.72	0.45	0.90	0.29	0.0000	0.74	0.44	0.89	0.32	0.0047
Observations	17,669		417			193		61		
Sweep II outcomes										
Employed	0.47	0.50	0.59	0.49	0.0000	0.46	0.50	0.43	0.50	0.6834
Self-employed	0.01	0.09	0.00	0.05	0.0810	0.00	0.00	0.00	0.00	n/a
Student	0.01	0.11	0.01	0.08	0.1192	0.01	0.12	0.06	0.24	0.2026
At home to care for family	0.44	0.50	0.31	0.47	0.0000	0.46	0.50	0.41	0.50	0.5159
Weekly working hours (includes 0)	12.38	14.36	15.45	15.01	0.0002	11.73	14.35	12.41	14.23	0.7711
Weekly working hours (excludes 0)	23.83	11.14	23.34	12.49	0.5595	23.14	11.89	24.35	10.13	0.6217
Has working partner	0.75	0.43	0.92	0.28	0.0000	0.77	0.42	0.82	0.39	0.4496
Uses childcare by conducted by relatives/friends	0.28	0.45	0.29	0.46	0.7216	0.23	0.42	0.16	0.37	0.2744
Uses paid childcare	0.12	0.33	0.22	0.42	0.0000	0.08	0.27	0.18	0.39	0.0947
Observations	13,942		343			142		51		
Sweep III outcomes										
Employed	0.53	0.50	0.57	0.50	0.1298	0.54	0.50	0.52	0.50	0.8149
Self-employed	0.01	0.11	0.00	0.06	0.0195	0.01	0.09	0.00	0.00	0.3193
Student	0.01	0.11	0.01	0.08	0.2225	0.01	0.09	0.02	0.14	0.5745
At home to care for family	0.37	0.48	0.30	0.46	0.0067	0.34	0.48	0.31	0.47	0.6931
Weekly working hours (includes 0)	13.92	14.48	15.32	13.71	0.0817	13.68	14.32	15.48	14.16	0.4594
Weekly working hours (excludes 0)	23.80	11.11	22.53	10.67	0.0950	22.86	11.47	23.97	10.20	0.6284
Has working partner	0.75	0.43	0.89	0.32	0.0000	0.80	0.41	0.88	0.33	0.1902
Child attends school	0.99	0.11	0.99	0.11	0.8490	0.98	0.13	0.96	0.20	0.4232
Observations	12,111		300			122		48		

Table 5: First stage results, labor supply sample

	(i) All women	(ii) All women, controls for fertility treatment	(iii) All women, controls for pre-pregnancy characteristics	(iv) All women, controls for pre-pregnancy characteristics & fertility treatment	(v) Only women with fertility treatment	(vi) Only women without fertility treatment
			Sweep I			
Multiple birth (1 = yes)	0.882*** (0.072)	1.042*** (0.071)	0.885*** (0.063)	1.018*** (0.062)	1.101*** (0.094)	1.028*** (0.086)
Fertility treatment (1 = yes)		-0.784*** (0.034)		-0.653*** (0.032)		
R^2	0.010	0.024	0.227	0.237	0.248	0.011
Kleinbergen-Paap F-stat	149.13	215.69	194.46	265.26	136.49	141.21
Observations	18,340	18,340	18,340	18,340	478	17,862
			Sweep II			
Multiple birth (1 = yes)	0.685*** (0.077)	0.830*** (0.077)	0.694*** (0.067)	0.814*** (0.067)	1.007*** (0.121)	0.787*** (0.093)
Fertility treatment (1 = yes)		-0.643*** (0.042)		-0.535*** (0.041)		
R^2	0.006	0.015	0.191	0.198	0.180	0.006
Kleinbergen-Paap F-stat	78.94	116.27	107.70	147.70	69.08	71.46
Observations	14,460	14,460	14,460	14,460	394	14,066
			Sweep III			
Multiple birth (1 = yes)	0.573*** (0.083)	0.715*** (0.084)	0.605*** (0.073)	0.725*** (0.074)	0.903*** (0.127)	0.665*** (0.103)
Fertility treatment (1 = yes)		-0.580*** (0.046)		-0.491*** (0.046)		
R^2	0.004	0.012	0.157	0.163	0.138	0.004
Kleinbergen-Paap F-stat	48.04	73.06	68.05	95.37	50.48	41.57
Observations	12,581	12,581	12,581	12,581	348	12,233

Coefficient, robust standard errors in parentheses. */**/*** denote statistical significance on the 10%, 5% and 1% level respectively. All estimates include age in years as dummies. Column (iii) also contains dummies for various completed qualifications, age at birth, a dummy for having worked before the pregnancy, a dummy for non-white ethnicity and dummy variables for marital status.

Table 6: Analysis of compliers characteristics

	Sweep 1		Sweep 2		Sweep 3	
	First stage	Relative frequency compliers	First stage	Relative frequency compliers	First stage	Relative frequency compliers
			Full sample			
Multiple birth	0.882***		0.685***		0.573***	
	(0.072)		(0.077)		(0.083)	
			Pregnancy was surprising			
Multiple birth	1.013***	1.149	0.979***	1.429	0.928***	1.620
	(0.138)		(0.159)		(0.188)	
			No qualification			
Multiple birth	0.983***	1.115	0.801***	1.169	0.868***	1.515
	(0.224)		(0.259)		(0.315)	
			Highest qualification O-level or equivalent			
Multiple birth	0.780***	0.884	0.708***	1.034	0.612***	1.068
	(0.103)		(0.117)		(0.131)	
			Highest qualification A-level			
Multiple birth	0.735***	0.833	0.613***	0.895	0.577***	1.007
	(0.141)		(0.159)		(0.181)	
			Highest qualification diploma			
Multiple birth	1.012***	1.147	0.659***	0.962	0.572***	0.998
	(0.185)		(0.186)		(0.180)	
			Highest qualification degree			
Multiple birth	0.726***	0.823	0.413***	0.603	0.370***	0.646
	(0.108)		(0.120)		(0.123)	
			Highest qualification higher degree (Master and PhD)			
Multiple birth	1.018***	1.154	0.675***	0.985	0.606***	1.058
	(0.172)		(0.189)		(0.196)	
			Had job before pregnancy			
Multiple birth	0.886***	1.005	0.678***	0.990	0.572***	0.998
	(0.070)		(0.076)		(0.081)	
			Non-white			
Multiple birth	0.886***	1.005	0.780***	1.139	0.706**	1.232
	(0.235)		(0.219)		(0.276)	
			Married			
Multiple birth	0.886***	1.005	0.634***	0.926	0.558***	0.974
	(0.086)		(0.090)		(0.096)	
			Illness or problems during pregnancy			
Multiple birth	0.698***	0.791	0.594***	0.867	0.582***	1.016
	(0.089)		(0.099)		(0.110)	

Coefficient, robust standard errors in parentheses. */**/*** denote statistical significance on the 10%, 5% and 1% level respectively. All estimates include age in years as dummies.

Table 7: Outcomes Sweep I interview (within 1 year of birth)

	(i) All women	(ii) All women, controls for fertility treatment	(iii) All women, controls for pre-pregnancy characteristics	(iv) All women, controls for pre-pregnancy characteristics & fertility treatment	(v) Only women with fertility treatment	(vi) Only women without fertility treatment
			Employed (1 = yes)			
Number of children	-0.107***	-0.106***	-0.122***	-0.105***	-0.096	-0.107***
	(0.033)	(0.028)	(0.031)	(0.027)	(0.061)	(0.031)
			Self-employed (1= yes)			
Number of children	-0.010	-0.012	-0.010	-0.011	-0.018	-0.009
	(0.011)	(0.009)	(0.011)	(0.010)	(0.023)	(0.010)
			On maternity/parental leave (1 = yes)			
Number of children	0.025*	0.018	0.025*	0.019	0.029	0.016
	(0.015)	(0.012)	(0.014)	(0.012)	(0.032)	(0.013)
			Fulltime student (1 = yes)			
Number of children	0.003	0.002	0.004	0.003	0.007	-0.001
	(0.006)	(0.005)	(0.006)	(0.005)	(0.014)	(0.005)
			At home and caring for family (1 = yes)			
Number of children	0.094***	0.101***	0.110***	0.099***	0.078	0.106***
	(0.033)	(0.029)	(0.030)	(0.026)	(0.060)	(0.032)
			Weekly working hours (includes 0 for those not working)			
Number of children	-1.997**	-2.340***	-2.297***	-2.151***	-1.917	-2.452**
	(0.998)	(0.851)	(0.890)	(0.781)	(1.833)	(0.958)
			Weekly working hours (excludes those not working)			
Number of children	-0.020	-0.185	0.110	-0.047	0.391	-0.606
	(1.275)	(1.118)	(1.269)	(1.109)	(1.702)	(1.385)
			Has a working partner (1= yes)			
Number of children	-0.005	-0.025	-0.025	-0.028	-0.032	-0.022
	(0.029)	(0.024)	(0.024)	(0.021)	(0.039)	(0.029)
Observations (all but second working hours regression)	18,340	18,340	18,340	18,340	478	17,862
Observations (second working hours regression)	8669	8669	8669	8669	306	8363

Coefficient, robust standard errors in parentheses. */**/*** denote statistical significance on the 10%, 5% and 1% level respectively. All estimates include age in years as dummies. Column (ii) additionally contains a dummy for having received fertility-treatment. Column (iii) also contains dummies for various completed qualifications, age at birth, a dummy for having worked before the pregnancy, a dummy for non-white ethnicity and dummy variables for marital status.

Table 8: Outcomes Sweep II interview (3 years after birth)

	(i) All women	(ii) All women, controls for fertility treatment	(iii) All women, controls for pre-pregnancy characteristics	(iv) All women, controls for pre-pregnancy characteristics & fertility treatment	(v) Only women with fertility treatment	(vi) Only women without fertility treatment
			Employed (1 = yes)			
Number of children	-0.082	-0.083*	-0.100**	-0.081*	-0.194***	-0.044
	(0.052)	(0.044)	(0.049)	(0.042)	(0.073)	(0.053)
			Self-employed (1 = yes)			
Number of children	-0.009***	-0.007***	-0.008***	-0.007***	-0.002	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
			Fulltime student (1 = yes)			
Number of children	0.025	0.019	0.025	0.020	0.057*	0.005
	(0.017)	(0.013)	(0.017)	(0.014)	(0.033)	(0.013)
			At home and caring for family (1 = yes)			
Number of children	0.073	0.081*	0.091*	0.078*	0.118	0.065
	(0.052)	(0.043)	(0.048)	(0.041)	(0.073)	(0.053)
			Weekly working hours (includes 0 for those not working)			
Number of children	-2.468*	-2.418**	-2.698**	-2.153*	-4.004**	-1.827
	(1.466)	(1.228)	(1.350)	(1.151)	(2.037)	(1.491)
			Weekly working hours (excludes those not working)			
Number of children	-0.813	-0.551	-0.471	-0.237	-1.077	-0.909
	(1.613)	(1.428)	(1.590)	(1.407)	(2.217)	(1.762)
			Has a working partner (1= yes)			
Number of children	-0.012	-0.035	-0.036	-0.039	-0.078	-0.014
	(0.043)	(0.036)	(0.039)	(0.034)	(0.056)	(0.044)
Observations (all except below)	14,460	14,460	14,460	14,460	394	14,066
Observations (second working hours regression)	7558	7558	7558	7558	253	7305

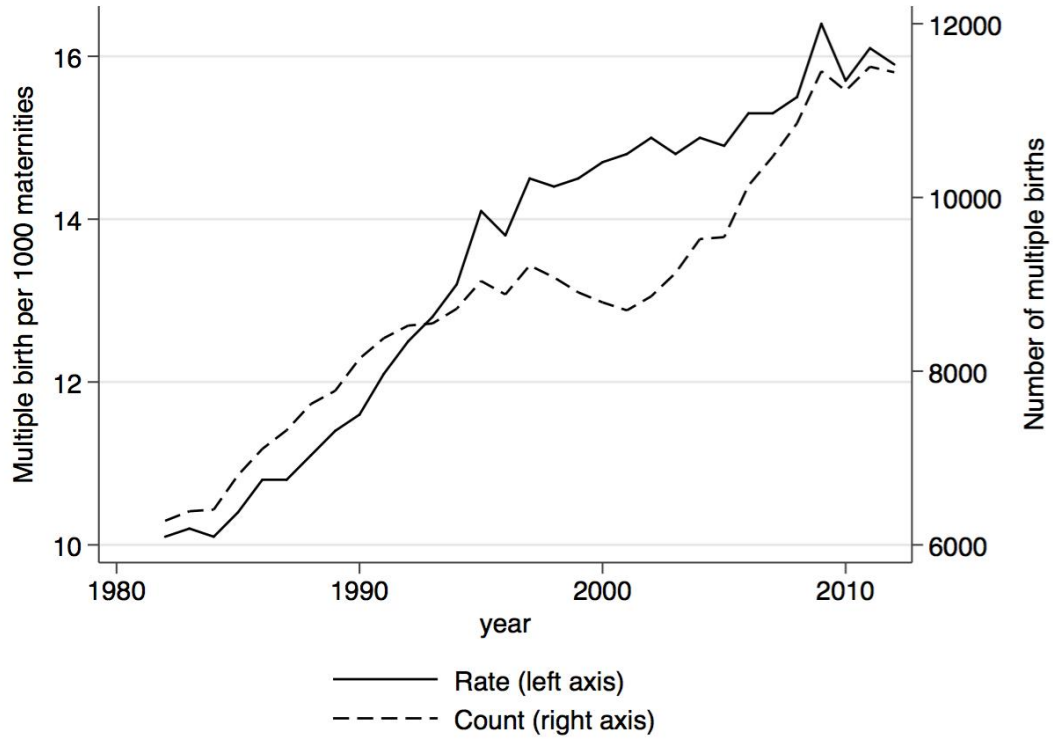
Coefficient, robust standard errors in parentheses. */**/*** denote statistical significance on the 10%, 5% and 1% level respectively. All estimates include age in years as dummies. Column (ii) additionally contains a dummy for having received fertility-treatment. Column (iii) also contains dummies for various completed qualifications, age at birth, a dummy for having worked before the pregnancy, a dummy for non-white ethnicity and dummy variables for marital status.

Table 9: Outcomes Sweep III interview (5 years after birth)

	(i) All women	(ii) All women, controls for fertility treatment	(iii) All women, controls for pre-pregnancy characteristics	(iv) All women, controls for pre-pregnancy characteristics & fertility treatment	(v) Only women with fertility treatment	(vi) Only women without fertility treatment
			Employed (1 = yes)			
Number of children	-0.042	-0.032	-0.078	-0.043	-0.084	-0.013
	(0.066)	(0.054)	(0.061)	(0.051)	(0.086)	(0.067)
			Self-employed (1 = yes)			
Number of children	-0.005	-0.003	-0.003	-0.002	-0.002	-0.002
	(0.011)	(0.009)	(0.010)	(0.009)	(0.002)	(0.012)
			Fulltime student (1 = yes)			
Number of children	0.002	0.002	0.003	0.003	0.026	-0.004
	(0.015)	(0.012)	(0.014)	(0.012)	(0.025)	(0.012)
			At home and caring for family (1 = yes)			
Number of children	0.001	0.008	0.041	0.020	0.029	-0.000
	(0.063)	(0.051)	(0.055)	(0.047)	(0.084)	(0.064)
			Weekly working hours (includes 0 for those not working)			
Number of children	-1.586	-1.282	-2.370	-1.417	-0.545	-1.618
	(1.879)	(1.532)	(1.688)	(1.422)	(2.452)	(1.912)
			Weekly working hours (excludes those not working)			
Number of children	-1.272	-0.646	-1.002	-0.394	0.804	-1.424
	(1.721)	(1.461)	(1.697)	(1.441)	(2.335)	(1.800)
			Has a working partner (1= yes)			
Number of children	0.048	0.011	0.006	-0.003	-0.030	0.025
	(0.052)	(0.042)	(0.048)	(0.040)	(0.058)	(0.054)
Observations (all but second working hours regression)	12,581	12,581	12,581	12,581	348	12,233
Observations (second working hours regression)	7390	7390	7390	7390	235	7155

Coefficient, robust standard errors in parentheses. */**/*** denote statistical significance on the 10%, 5% and 1% level respectively. All estimates include age in years as dummies. Column (ii) additionally contains a dummy for having received fertility-treatment. Column (iii) also contains dummies for various completed qualifications, age at birth, a dummy for having worked before the pregnancy, a dummy for non-white ethnicity and dummy variables for marital status.

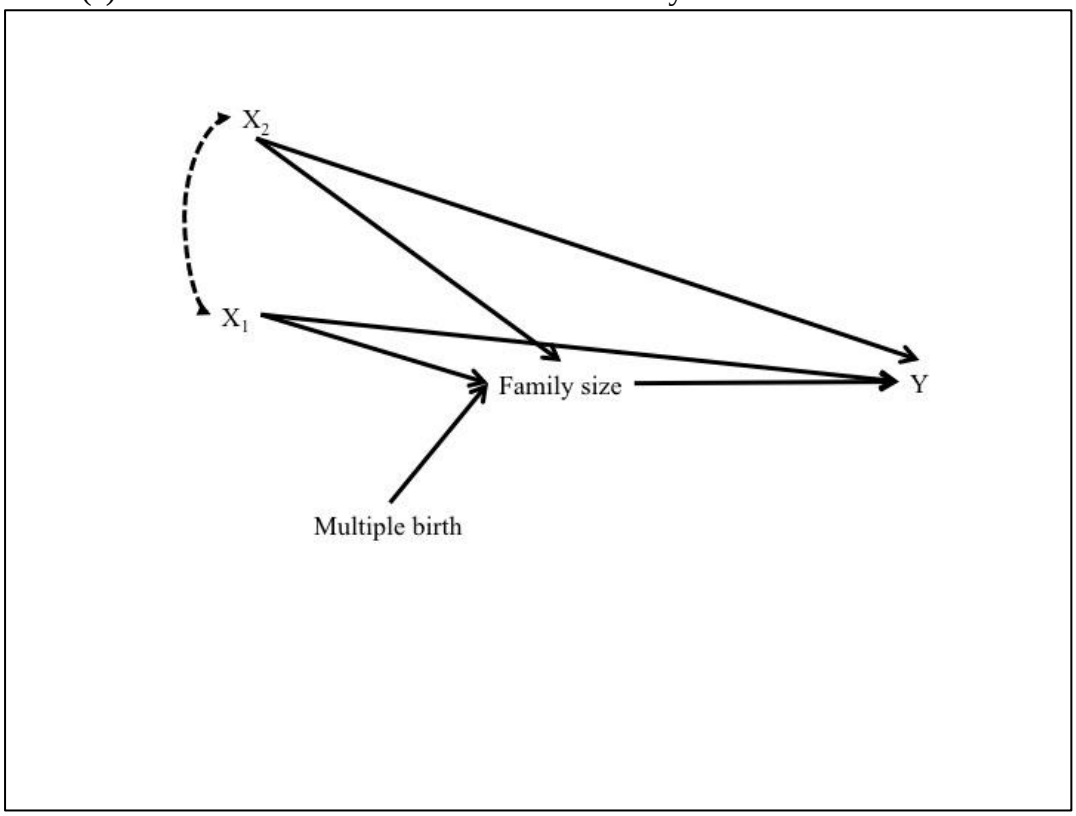
Figure 1: Multiple birth over time, UK, 1982 to 2012



Notes: Data is from the *Characteristics of Birth 2* series of the Office for National Statistics. We begin the series in 1982 as data from 1981 is missing due to a registrars' strike.

Figure 2: Causal diagram for the multiple birth instrument with and without fertility treatments

Panel (a): The twin births instrument without fertility treatments



Panel (b): The twin births instrument with fertility treatments

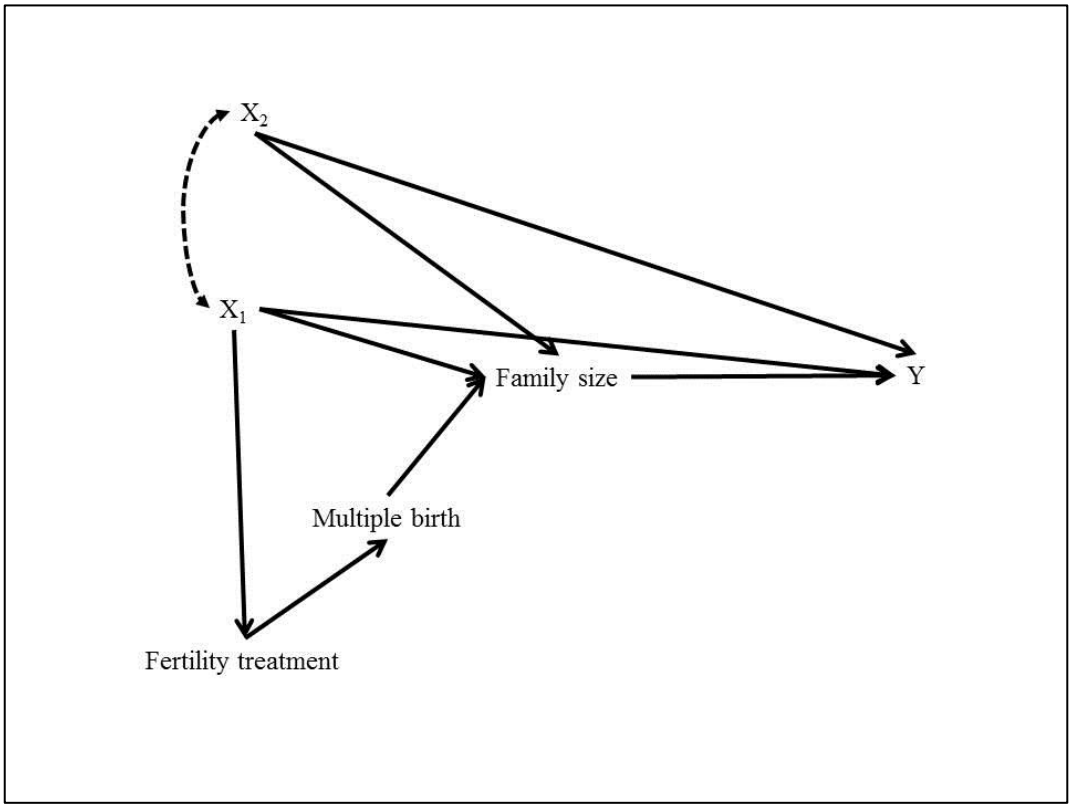
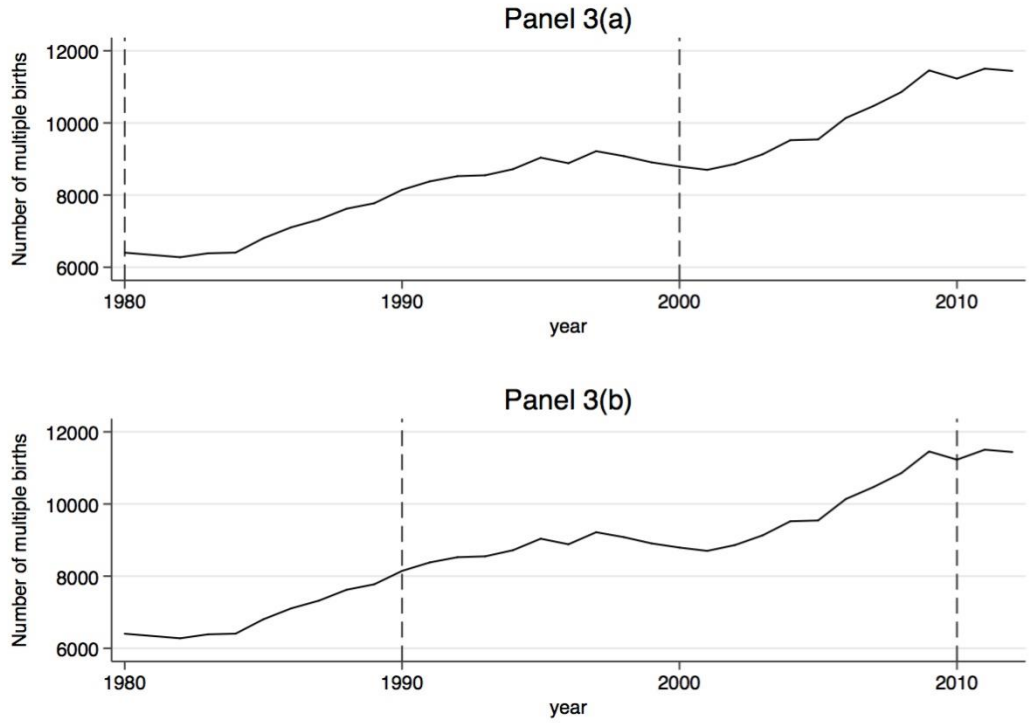


Figure 3: The distribution of multiple births over time and cross-sections drawn at various points



Notes: Data is from the *Characteristics of Birth 2* series of the Office for National Statistics. Data for 1981, which missing due to a registrars' strike, is linearly extrapolated between 1980 and 1982 for the sake of the example.