Instrumental Variable Estimation in Demographic Studies:
The LATE interpretation of the IV estimator with heterogenous effects

Richard Breen

Department of Sociology and Nuffield College, Oxford

John Ermisch

Leverhulme Centre for Demographic Science and Nuffield College, Oxford

04 June 2021

Abstract

Heterogeneous effects of treatment on an outcome is a plausible assumption to make about the vast majority of causal relationships studied in the social sciences. In these circumstances the IV estimator is often interpreted as yielding an estimate of a Local Average Treatment Effect (LATE): a marginal change in the outcome for those whose treatment is changed by the variation of the particular instrument in the study. Our aim is to explain the relationship between the LATE parameter and its IV estimator by using a simple model which is easily accessible to applied researchers, and by relating the model to examples from the demographic literature. A focus of the paper is how additional heterogeneity in the instrument – treatment relationship affects the properties and interpretation of the IV estimator. We show that if the two kinds of heterogeneity are correlated, then the LATE parameter combines both the underlying treatment effects and the parameters from the instrument – treatment relationship. It is then a more complicated concept than many researchers realise.

KEYWORDS

Heterogenous effects, IV estimation, LATE, local average treatment effect.

Instrumental Variable Estimation in Demographic Studies:
The LATE interpretation of the IV estimator with heterogeneous effects

1. Introduction

In recent years, social scientists from several disciplines, including demography, have increasingly used Instrumental Variable (IV) estimation, thereby following a long tradition in economics and statistics going back to the work of Philip and Sewall Wright in the early decades of the 20th century (Stock and Trebbi 2003). Demographic studies using IV that have appeared in demography journals in the past two years include Andriano and Monden (2019), Billari, Giuntella and Stella (2019), Hart and Cools (2019), and undoubtedly more. IV estimation has become popular through its use in estimating causal effects when the exposure or treatment is believed to be ‘endogenous’. This may arise because unmeasured confounders are thought to affect both the treatment and outcome, or because of simultaneous determination of the two (bidirectional causality). An example of unmeasured confounding would be the effect of women’s educational attainment (the treatment) on fertility outcomes, because there may be unobserved traits such as women’s career aspirations which could affect both treatment and outcome. An example of simultaneity would be attitudes toward family and fertility outcomes, each affecting the other.

In these circumstances, Ordinary Least Squares (OLS) estimates are biased and inconsistent (i.e. they do not converge to the true value as the sample size approaches infinity), but IV estimates has the potential to be consistent. An instrumental variable is believed to be exogenous with respect to the outcome variable and only affects the outcome through its relationship with the treatment. The IV estimator is the ratio of the covariance between the instrument and the outcome to the covariance between the

instrument and the treatment. Following Imbens and Angrist (1994), when the treatment effect is heterogeneous the IV estimator is interpreted as yielding an estimate of a Local Average Treatment Effect (LATE). The LATE parameter refers to a marginal change in the outcome for those whose treatment is changed by the variation of the particular instrument in the study: in most applications this will not equal the average treatment effect in the population (ATE) or the average effect of treatment on the treated (ATT). The LATE interpretation is often used to explain why IV and OLS estimates differ substantially, especially when the size of the IV estimate of the effect of treatment greatly exceeds that of the OLS estimate, such as in two of the examples that we discuss below.

Our primary aim is to explain the relationship between the LATE parameter and its IV estimator by using a simple model with heterogeneous treatment effects which is easily accessible to applied researchers, and by relating the model to examples from the demographic literature. A particular focus is how additional heterogeneity in the instrument – treatment relationship affects the properties and interpretation of the IV estimator. As Heckman et al. (2006, p.393) write about the LATE parameter: ‘Responses to treatment are permitted to be heterogeneous in a general way. Responses of choices to instruments are not.’ We show that when we have both forms of heterogeneity and they are correlated the LATE parameter is a more complicated concept than many researchers realise, and the IV estimator does not necessarily identify it. We also discuss what researchers should report in applications of IV estimation so as to inform the reader about the internal validity, precision and external validity of the IV estimate.

2. Heterogeneous treatment effects and IV

In the discussion of IV estimation found in most econometric textbooks the treatment effect is a constant parameter, applicable to the entire population (see Goldberger (1964, pp. 284-287) for an early exposition). But heterogeneous effects of treatment on the outcome is a more reasonable assumption to make about the vast majority of causal relationships studied in the social sciences. For instance in the education - fertility example, women's fertility responses to their education may vary because women differ in their fertility preferences. When treatment effects are heterogeneous the IV estimator is interpreted as the LATE parameter.

The LATE interpretation of the treatment effect estimate from the IV estimator requires that, although the treatment status of individuals, or other units of study, can respond to the instrument in different ways, this heterogeneity is limited: some individuals may be unaffected by the instrument (i.e. their treatment status is independent of the instrument) while those who are affected must all be affected in the same direction. This assumption is usually called *monotonicity* or *uniformity*. The unaffected are often called 'always takers' or 'never takers', while the affected are *compliers*. In our education - fertility example, the instrument might be a reform to the educational system which increased the years of compulsory education, an example we take up below. Some women may attain the same level of education irrespective of the reform (non-compliers). The monotonicity assumption is that, while the sample may contain women unaffected by the reform, it does not contain women who *reduce* their education in response to the reform. Without monotonicity, the IV estimator is not guaranteed to estimate a weighted average of the underlying individual treatment effects among compliers.

One implication of these features of the LATE parameter is that different instruments have different complier sub-populations. As a consequence, different instruments define different LATE parameters, raising an issue of interpretation when more than one instrument is used in the analysis. Another implication is that the LATE parameter is not usually a parameter of the general population: this is important for assessing its value in demographic applications, such as how variation in women's education levels generates variation in fertility rates within a population (e.g. Lutz, Butz and Samir 2014).

In the paper we show that heterogeneity in the instrument – treatment relationship among compliers can have important implications for the interpretation of the LATE parameter and its IV estimate, even when the sample consists solely of compliers and unaffected. We demonstrate that if we allow for general heterogeneity in this relationship by considering different degrees of compliance, and if this heterogeneity is correlated with heterogeneous treatment effects, the LATE parameter will be a function not only of the treatment effects but also of deeper, 'structural' parameters from the instrument – treatment relationship. In these circumstances, the interpretation of the LATE parameter and its IV estimator is different from what many applied researchers in the social sciences may think. In particular, it is not the simple population-weighted average treatment effect among compliers.

Such 'double heterogeneity' is either explicit or implicit in papers by Heckman (1997), Heckman et al. (2006) and Wooldridge (1997, 2003). Unfortunately, the lessons of these papers do not appear to be widely known in the social sciences beyond economics. Our aim is to reinforce their work by discussing LATE in terms of a simple model and concrete

examples from the literature. In the next section we use a model of binary heterogeneity in both the instrument - treatment and treatment - outcome relationships as the foundation for a general consideration of heterogeneity.

3. Dichotomous treatment

In this section we advance our main argument: if the treatment effect is heterogeneous across the population and the relationship between the treatment and the instrumental variable is also heterogeneous, then: (1) if the two sources of heterogeneity are correlated, the LATE parameter involves parameters from the instrument – treatment relationship as well as the heterogeneous treatment effects; and (2) if the distribution of the instrument differs between groups in the population, the IV estimator depends on the instrument’s variance and the degree of compliance in each group, and may not be a consistent estimator of the LATE parameter. We demonstrate these propositions in a two-group heterogeneity model in which it is clear what are the ‘deeper’ or ‘structural’ parameters in the LATE parameter. We consider a treatment or exposure, an outcome, Y , and an instrumental variable, Z . In real-world applications one might condition on further variables, W , that also affect the outcome but are not affected by the treatment and inclusion of which may affect the validity of the instrument (see Deuchert and Huber(2017)), but here, for the sake of clarity, we leave such conditioning implicit. We show that these results easily extend to multiple discrete groups and continuous parameter heterogeneity; the two-group example is used only for simplicity.

3.1 The structural parameters

A dichotomous treatment variable, T , corresponds to the literature in which a treatment group is compared with a control group, and it is common in IV applications. Suppose for

illustration that there are two groups in the *complier population*, $g = 0,1$, for which the true model is:

$$\begin{aligned}
Y &= \beta_g T + v; \\
T^* &= \pi + \alpha_g Z + e; \\
T &= 1 \text{ if } T^* > 0, \text{ and } T = 0 \text{ if } T^* \leq 0; \\
\text{and } E(ev) &\neq 0, E(Zv) = 0, \alpha_g > 0, g = 0,1 \quad (1)
\end{aligned}$$

The β_g are the heterogeneous treatment effects, and the equation for the latent variable T^* determining treatment differs between the two groups (different α_g); this is the ‘double heterogeneity’. These parameters are the ‘deep’ or ‘structural’ parameters of the model. Endogeneity of T is indicated by the assumption that $E(ev) \neq 0$. The distribution of g in the complier population is of a discrete form with two mass points such that the probability of being in group 0 is s_0 , and $s_1 = 1 - s_0$ for group 1; e and v are drawn from the same continuous joint distribution for both groups. We assume that the analyst has no information at all about these groups. If $\alpha_0 = \alpha_1$, then the model corresponds to that of Angrist and Pischke (2009, p. 157) in their discussion of the foundations of the LATE approach, and in which the double heterogeneity issue does not exist.

To put some flesh on the simple model, we draw on a recent study by Billari, Giuntella and Stella (2019) in this journal (hereafter BGS), which uses German data to estimate the impact of the availability of broadband on the probability of having a birth during a year. Their treatment variable is whether or not a household has a digital subscriber line (DSL), and they are concerned that unobserved influences on having a birth are correlated with having a DSL (i.e. $E(ev) \neq 0$ in the model in equation (1)).

3.2 The instrumental variable

The most difficult issue when attempting to address treatment endogeneity is finding an instrument. In order for Z to serve as an instrumental variable, two basic conditions must hold: (a) it is not correlated with v (Z is *valid*), which means $E(Zv) = 0$ in our model; and (b) it is correlated with T (Z is *relevant*), which means $\alpha_g \neq 0$ in equation (1). The stronger the correlation of Z and T in the data the more powerful is the instrument. Although it may not be immediately evident, condition (a) entails that Z only affects Y through T (the *exclusion restriction*). Even if Z is essentially random (e.g. birth of twins), Z may fail to satisfy condition (a) because it could also affect Y in ways other than through T . Randomness is sufficient for a causal interpretation of the ‘reduced form’ relating Y to Z , but not to identify the LATE parameter (Angrist and Pischke 2009, pp.152-3). If there is only one instrument, the data reveals information about the second condition, but the first condition is a maintained hypothesis driven by a theoretical model and by other knowledge suggesting that the exclusion restriction holds.

To simplify, let Z be dichotomous, which is another common case in the literature and is true in BGS, and let e and v have a joint standard normal distribution. The *probability of compliance* is the probability that $T = 1$ when $Z = 1$ and $T = 0$ when $Z = 0$, and for group g it is given by $c_g = \Phi(\alpha_g + \pi) - \Phi(\pi)$, where $\Phi(\cdot)$ is the standard normal distribution function.

In BGS, there are three instruments, all of which are dichotomous. The first and main instrument is equal to one for households at a distance from the main distribution frame (MDF) of more than 4.2 km; zero otherwise. It is primarily responsible for the correlation with the treatment (having a DSL line), and we focus on it. The second instrument equals one for households with a distance above the 4.2 km threshold who cannot be connected

to another MDF at a distance below 4.2 km, zero otherwise; and the third indicates whether or not an original East German household adopted a particular telephone technology after reunification that made it more expensive to obtain a DSL connection. All of these instruments are related negatively to having a DSL connection, and they have reasonable power as indicated by an F-test of the hypothesis that the three excluded instruments are all zero in the first stage regression for DSL (instrument relevance). We return to the issue of multiple instruments later. For comparison with our simple model, define Z so that unity means being *close to* a MDF (0.89 of the sample live at a closer distance than 4.2 km) rather than far; thus, we expect $c_g > 0$.

3.3 The LATE parameter

Given that the estimated treatment effects in BGS are found to vary by age, sex and parent's education it is reasonable to assume they may also be heterogeneous in other ways. Thus, it appears likely that the treatment effect estimates should be interpreted as LATE, as BGS suggests. The nature of the LATE parameter depends on whether there is also heterogeneity in the relationships between the distance to the MDF and having a DSL connection and whether it is correlated with heterogeneity in the treatment effects.

The *LATE parameter* in our simple model is the effect of the treatment on the outcome conditional on being a complier ($c_g > 0, g = 0,1$):

$$LATE_p = \frac{s_0\beta_0c_0 + s_1\beta_1c_1}{s_0c_0 + s_1c_1} \quad (2)$$

The LATE parameter is thus a weighted average of the group-specific treatment effects (β_g 's) with weights given by $\frac{s_0c_0}{s_0c_0 + s_1c_1}$ and $\frac{s_1c_1}{s_0c_0 + s_1c_1}$. The weights therefore depend on the probability of compliance in each group (c_g) relative to the complier population average

compliance rate as well as on the complier population shares in each group (s_g). A low probability of compliance in a group reduces the weight of its treatment parameter in $LATE_p$. There may be non-compliers in the wider population, but $LATE_p$ provides no information about their causal response because the instrument does not affect their treatment status.

If $\alpha_0 = \alpha_1$, then $c_0 = c_1$ and the LATE parameter does not depend on c_g : it equals $s_0\beta_0 + s_1\beta_1$, given that $c_g > 0$; i.e. it is an average treatment effect *for compliers only*. This is the standard case in the LATE literature, which is easy to interpret. If the entire population were compliers in these circumstances, then the LATE parameter would coincide with the average treatment effect (ATE). More generally, $LATE_p$ differs from the ATE, which is the average for the entire population, including non-compliers: $ATE = (s_0\beta_0 + s_1\beta_1)s_C + (1 - s_C)\beta_{NC}$, where s_C is the proportion of the population who are compliers and β_{NC} is the average treatment effect among non-compliers.

It is immediately clear that the LATE parameter may not be informative for some demographic questions because it is not a population parameter, in contrast to the ATE (assuming the sample is representative). It may, however, answer questions for a section of the population which may be of particular interest and can provide some evidence about the causal impact of a treatment (e.g. availability of broadband).

3.4 The IV estimator

The IV estimator of the LATE parameter is

$$b_{IV} = \frac{\sum_{g=0}(Y - \bar{Y})Z + \sum_{g=1}(Y - \bar{Y})Z}{\sum_{g=0}(T - \bar{T})Z + \sum_{g=1}(T - \bar{T})Z} \quad (3)$$

where $\sum_{g\cdot}$ indicates summation over the group sample and the 'bar' over the variable represents a sample mean. Its *probability limit* as the sample size (N) increases to infinity is

$$plim_{N \rightarrow \infty} b_{IV} = \frac{s_0 \beta_0 c_0 \sigma_{z0}^2 + s_1 \beta_1 c_1 \sigma_{z1}^2}{s_0 c_0 \sigma_{z0}^2 + s_1 c_1 \sigma_{z1}^2} \quad (4)$$

Comparing equations (2) and (4) we see that although the IV estimator is also a weighted average of the group-specific treatment effects (the β_g 's), the weights depend on σ_{zg}^2 , which is the variance of Z in group g , making it an inconsistent estimator of $LATE_p$, unless $\sigma_{z0}^2 = \sigma_{z1}^2$, in which case $plim_{N \rightarrow \infty} b_{IV} = LATE_p$. Note that because Z is dichotomous, $\sigma_{zg}^2 = p_{zg}(1 - p_{zg})$ where p_{zg} is the probability that $Z=1$. The expressions in equations (2) and (4) easily generalise to many groups. If $\beta_0 = \beta_1 = \beta$, then the IV estimator converges to β ; that is, we estimate the homogenous causal effect with the conventional IV estimator straightforwardly despite the two T - Z equations and any heterogeneity in the distribution of Z .

3.5 Correlated heterogeneity

The LATE parameter and its IV estimate can be substantially different from the population-weighted average treatment effect among compliers. For example, assume that the treatment effect parameters are $\beta_0 = 1.25$ and $\beta_1 = 0.25$, and that $\alpha_0 = 1, \alpha_1 = 0.15, \pi = 0, \sigma_{z0}^2 = \sigma_{z1}^2$ and $s_0 = 0.5$. Then the probabilities of compliance are $c_0 = 0.34$ and $c_1 = 0.06$, implying from equation (2) that the LATE parameter is 1.10. It is well above the average effect for compliers, $s_0 \beta_0 + s_1 \beta_1 (= 0.75)$, because the group with a higher treatment effect also has a higher probability of compliance. If the values for the treatment effects were reversed between groups, the LATE parameter would be 0.40.

It is the correlation between c_g and β_g which creates the tendency for the LATE parameter to differ from the average treatment effect for compliers. To explore this further, we generate β_g for 20 groups by sampling from a uniform distribution over the range 0.10 to 1.5. We then generate a variable u in the same way where the range is 0.05 to 0.35. Now let $c_g = 0.15 + q\beta_g + u$. If $q = 0$ and there are equal shares of each group in the complier population, then the average β_g for compliers and the LATE parameter are both 0.88. If, however, $q = 0.1$ (making the correlation between c_g and β_g equal 0.43), the LATE parameter is 0.92, and if $q = -0.1$, it is 0.81. Thus, the important issue is not whether or not there is heterogeneity in c_g , but whether it is correlated with heterogeneity in the treatment effects. When it is not, then it is as if there were a common value of c_g across groups equal to the average c_g .

A natural extension is to allow continuous variation in c_g and β_g . Assume that among compliers heterogeneity in β and c follow a joint continuous distribution (with $c > 0$, as all are compliers), making them random coefficients. Then

$$LATE_p = \frac{E(c\beta)}{E(c)} = \frac{E(c)E(\beta) + cov(c, \beta)}{E(c)} \quad (5)$$

where $cov(c, \beta)$ indicates covariance between c and β . Notice that equation (2) for the two-group case is the discrete version of (5). When $cov(c, \beta) = 0$, the LATE parameter is $E(\beta)$; i.e. the average treatment effect among compliers. Otherwise, it depends on the joint distribution of the two types of heterogeneity. One situation in which correlation between c and β is inherent in the research problem is the so called ‘Roy model’ (Roy 1951) (e.g. that of Bjorklund and Moffitt (1987)), in which the choice about the treatment T (e.g. education) varies with the gains from treatment, β (e.g. earnings). In our example with the continuous uniform distributions in the preceding

paragraph, $E(\beta) = 0.8, E(c) = qE(b) + E(u) + 0.15$ and $cov(c, \beta) = qvar(\beta)$, from which we can calculate the LATE parameter from equation(5) (assuming $q > -0.1$ to ensure $c > 0$). The LATE parameter lies above 0.8 when $q > 0$ and below it when $q < 0$.

3.6 Comparing IV and OLS estimates

The OLS estimator would estimate the ATE if the treatment were exogenous. Thus, when it is consistent, OLS estimates a population parameter. This needs to be considered when comparing the OLS and IV estimators and assessing their external validity for demographic applications of the parameter estimates at the population level.

The main results of BGS's IV and OLS estimation of the linear probability model used in the paper are shown in Table 1 for the demographic groups in which the IV estimates are precise enough to suggest an effect of a DSL connection. As Table 1 shows, the OLS estimate is much lower than the IV estimate. In a constant treatment effect model, this finding would suggest that women with access to a DSL connection tend to have much lower fertility for unmeasured reasons (in reference to equation (1), $E(ev) < 0$). When treatment effects of a DSL connection are heterogeneous, however, the interpretation of the IV estimate may be different.

Table 1: Estimates of the effect of a DSL connection on annual birth probability for women aged 25-45 (standard errors in parentheses)

	All	Highly Educated
IV estimate ¹	0.113 (0.055)	0.133 (0.062)
OLS estimate ²	0.005 (0.005)	0.005 (0.006)

¹Tables 2 and 3, respectively, Billari, Giuntella and Stella (2019).

²Tables A1 and A4, respectively, Billari, Giuntella and Stella (2019).

As equation (5) indicated, interpretation depends on whether the impact of being close to a MDF on the probability of having a DSL connection is correlated with the impact of the connection on fertility. It could be argued that families who wish to have another child are more likely to obtain a DSL connection when they are close to the main distribution frame (i.e. respond more to Z). Then the LATE parameter would exceed the average effect of a connection on fertility among compliers.

Although the difference between the OLS and IV estimates is relatively clear in the BGS study, in other applications it may be less so. The Wu-Hausman test (e.g. see Wooldridge 2002) was designed for a constant treatment effect, but, with qualifications, it remains a helpful diagnostic in the context of heterogeneous effects. The test's null hypothesis is that the treatment is exogenous and it compares two estimators of a single parameter: one which is consistent under both the null and the alternative hypothesis and another that is consistent and more efficient under the null, but is inconsistent under the alternative. The idea is that as the sample size increases to infinity the two estimators should estimate the same parameter if the null hypothesis is correct; if they are significantly different, the null should be rejected. If all of the population are compliers and the two types of heterogeneity are uncorrelated, IV is a consistent estimator of the ATE under both hypotheses; under the alternative (i.e. endogeneity) OLS is inconsistent for the ATE. Under these two premises, failure to reject the null is consistent with exogeneity of the treatment, indicating that both OLS and IV estimate the ATE, a population parameter, but OLS is more efficient. In that case OLS provides a simple unbiased and efficient estimator of the ATE.

Unfortunately, if the two types of heterogeneity are correlated, the Hausman test may also fail to reject the null even though the treatment is endogenous, and it may also indicate rejection when the treatment is exogenous. This can occur because, even though everyone is a complier, both estimators are inconsistent for the ATE, or the IV is inconsistent. When there are non-compliers, the Hausman test is likely to indicate rejection because the two estimators are estimating different parameters and so it is not testing exogeneity. Nevertheless, although the interpretation of the test is more nuanced when there are heterogeneous effects its result is worth reporting and interpreting.

How might we account for the very large difference between the OLS and IV estimates reported by BGS ? The most likely reason for it is that LATE only applies to compliers. The OLS estimate of the ATE depends on the treatment effects among non-compliers and the proportion of non-compliers in the population, neither of which are observed of course. The share of non-compliers does not matter for the LATE estimate because they do not contribute to identifying it. Thus, an OLS estimate of 0.005 would be consistent with a LATE estimate of 0.133 if there were only a small proportion of compliers, and the non-compliers had small treatment effects. In other words, the LATE estimate would be much higher because very few women respond to the instrument. But then the implication of the study for the impact of a DSL connection on overall fertility would be that a DSL connection did not change overall fertility by much.

3.7 Reduced form for the outcome variable

There is an equivalent expression for the IV estimator when there is one instrument, which might be more transparent than equation (3). Denote the coefficient of Z in the regression equation between T and Z (the 'first stage') as b_{First} . The 'reduced form' is a

regression equation relating the outcome to the instrument; denote the coefficient of Z in it as b_{RF} . Then, dividing the numerator and denominator of the IV estimator in equation (3) by the product of the sample size and the sample variance of Z , we can write $b_{IV} = b_{RF}/b_{First}$. Thus, the reduced form regression coefficient is proportional to the IV estimate of the LATE parameter (in our model its probability limit is $\frac{s_0\beta_0c_0\sigma_{Z0}^2+s_1\beta_1c_1\sigma_{Z1}^2}{s_0\sigma_{Z0}^2+s_1\sigma_{Z1}^2}$, making it a variance weighted average of the group-specific reduced form parameters, $\beta_g c_g$).

It would be of considerable concern if the causal effect was not visible from the reduced form estimate b_{RF} . Furthermore, as an OLS estimator, the estimate of b_{RF} is consistent and unbiased, in contrast to the IV estimator, which is biased, although it is ‘approximately unbiased’ (or ‘median unbiased’) when the model is just identified, as in our equation (1) model (Angrist and Pishke 2009, pp. 205-218). For this reason, reporting the estimated reduced form is helpful for readers of applied studies, and of course reporting the first stage regression equation between T and Z is necessary for assessing instrument relevance (in our model the probability limit of b_{First} is $\frac{(s_0c_0\sigma_{Z0}^2+s_1c_1\sigma_{Z1}^2)}{(s_0\sigma_{Z0}^2+s_1\sigma_{Z1}^2)}$). BGS do not report the estimated reduced form equation for fertility.

3.8 Precision of IV estimate

The estimator of the variance of the IV estimator is inversely related to the correlation between the instrument and the treatment variable (Z and T), and therefore to the size of b_{First} . Less powerful instruments therefore create less precise IV estimates and thus larger confidence intervals. When there are multiple instruments some entail more precise estimates than others, namely those in which Z and T are more highly correlated.

Using all instruments available may increase the sensitivity of the estimate to weak instruments. IV estimates often have large confidence intervals which may include the OLS point estimate. Despite the fact that imprecision of IV estimates is a common problem, researchers often only appear to care about whether it is significantly different from zero.

3.9 Over-identifying restrictions tests

The simple model here has one endogenous variable and one instrument, meaning the model is ‘just identified’. As noted, it is usually difficult to find even one convincing instrument. But if there are more instruments than endogenous variables, as in BGS, the model is ‘over-identified’, making it possible to test the ‘validity of the over-identifying restrictions.’ The appropriate tests (e.g. see Wooldridge (2010) or Bollen (2012)) were derived for the case of a constant treatment parameter. They essentially validate multiple instruments by whether or not they estimate the same parameter (for a discussion see Angrist & Pischke 2009, pp. 141-147). But when the treatment effects are heterogeneous, which is the focus of this paper, such tests are no longer strictly appropriate, because each instrument may capture a different causal effect (Angrist & Pischke 2009, pp. 166-7). Rejection of the restrictions may just imply different causal effects associated with each instrument.

BGS reports tests of overidentifying restrictions indicating acceptance of them. Failure to reject could reflect a couple of situations: (a) imprecision of the underlying multiple IV estimates (effectively, one for each IV) underlying the test, making it difficult to conclude one was significantly different from another, and therefore difficult to reject the over-identifying restrictions; or (b) the different instruments may have essentially the same

complier populations. If the instruments reflected very different complier populations, then failure to reject might indicate a homogeneous treatment effect, making the test indicative of external validity. In BGS however, both (a) and (b) above appear likely, suggesting that all three instruments effectively estimate a common LATE parameter. But that does not mean that the treatment effect is constant across the population.

Thus, although the tests are still useful, their interpretation is more nuanced when the treatment effects are heterogeneous. Angrist and Pischke (2009, p.213) suggest picking the best instrument and reporting just-identified estimates with it. Just-identified estimates are median-unbiased and less likely to be sensitive to weak instruments than over-identified ones.

3.10 Point and interval estimates of causal effects using IV

Finding even one credible instrument is often difficult. As well as providing another example of a binary treatment and instrument, the following example demonstrates that in some circumstances it is possible to obtain informative bounds for the causal effect when there are concerns about whether the instrumental variable satisfies the validity (exclusion) condition. It addresses a popular issue concerning the impact of being a teenage mother on her subsequent life.

Ermisch and Pevalin (2005), hereafter EP, apply an idea developed by Hotz et al. (1997) to estimate the causal effect of early motherhood (before the age of 20) on future partnership outcomes, and as an outcome we focus here on the 'quality' of the partner she has at age 30 in terms of his employment. In this case, early motherhood is the binary treatment, quality of partner is the outcome and the instrument is a miscarriage as a

teenager. The point estimates of the causal effect assume that miscarriages are random and all pregnancies are reported. It is an estimate of the LATE parameter for pregnant women who would have had a teen birth in the absence of the miscarriage. Although it is possible for a teenager to have a miscarriage, become pregnant again and have a teen birth, this is relatively rare, and EP ignore it. Then, because a woman cannot have a teen birth despite having a miscarriage (no 'always takers'), the LATE parameter is also an estimate of the ATT – the average effect of the treatment (teen birth) on those women who gave birth. This measure should be of particular interest to policy-makers because policies that seek to reduce teenage childbearing are likely to target women who would have become teen mothers in the absence of the policy.

EP use information on the cohort of British women born in 1970, collected at birth and ages 5, 10 and 30, and pregnancy histories at age 30 in the 1970 British Cohort Study. The reduced form is the impact of not having a miscarriage on whether the woman's partner at age 30 was employed (not all women had a partner at age 30 but this was not related to having had a teen birth). EP estimate this as -0.156. Because the probability of having a teen birth given the woman has a miscarriage is zero, the first stage is just the average probability of a pregnant teenager having a birth (as opposed to an abortion) if she does not have a miscarriage, which is 0.774 in the study. The potential for correlated heterogeneity here amounts to whether pregnant women less likely to have an abortion are also more (less) likely to have an employed partner. The point estimate of the LATE parameter in EP for the probability that the woman's partner at age 30 is employed is $-\frac{0.156}{0.774} = -0.202$ (SE=0.086). For comparison, the OLS estimate is -0.155 (SE=0.026), putting it within one standard error of the IV estimate. Similarly, the effect on the

probability that her partner has no post-16 education is 0.220 (SE=0.095), with the corresponding OLS estimate being 0.169 (SE=0.026).

There is an alternative OLS estimator to compare with the IV estimator with these data: one based on the sample of women becoming pregnant as a teenager. It is preferable to OLS over the entire sample because women who do not become pregnant as teenagers provide little information about the effect of a teen-birth on women having one. It is however subject to selection bias: among women becoming pregnant as a teenager, women having the child may differ in mean outcomes from women who do not, even if they did not have the child. For the employment and post-compulsory education outcomes, the OLS estimates on this sample are -0.145 (SE=0.035) and 0.137 (SE=0.049), respectively. Because of the relatively large standard errors of the IV estimates, the Wu-Hausman test indicates that the OLS estimate is not significantly different from the IV estimate in each case, and is more precisely estimated. One could argue that it is worth trading off some possible bias for more precision and take this as the central estimate of the ATT.

Smoking and drinking during pregnancy, which are unobserved in the data, significantly increase the probability of a miscarriage, and this behaviour may be correlated with subsequent partner-outcomes. Thus, the instrumental variable may not be valid. Furthermore, pregnancies are likely to be under-reported in the pregnancy history. By using external estimates of the proportion of miscarriages that are random and of the proportion of pregnancies that are under-reported it is possible to obtain bounds on the causal effect under some assumptions. They may be informative in the sense of not containing zero.

When adjustment is made for non-random miscarriages EP obtain bounds on the impact of a teen-birth on the probability that the woman's partner at age 30 is employed. These are -0.218 and -0.163 ignoring the standard errors around the bounds, and -0.266 and -0.043 using their 95% confidence interval. Allowing also for under-reported pregnancies (assuming all teen-births are reported, but not all abortions and miscarriages) widens the bounds, and although there are data to suggest that about 40% of abortions were unreported, it is less clear what to assume about the proportion of pregnancies ending in a miscarriage. If it is assumed that it is 10% or less, the bounds are -0.218 and another negative number; thus, the bounds exclude zero.

3.11 Summary

Compared to the IV estimator in the standard, constant-treatment-effect case, the IV estimator of the LATE parameter depends on structural parameters in the instrument – treatment relationship when they differ between groups and they are correlated with the corresponding treatment effect parameters. The estimator also depends on the variance of the instrument in each group, and this is unknown because we do not observe the groups. It is doubtful that most applied researchers are aware of the nature of the weights for the β_g 's that arise because of these features of heterogeneity.

4. Continuous treatment

We begin with a discrete group model in which treatment variable X is continuous, similar to the model in equation (1), except the treatment variable now has a linear relationship with the instrument: $X = \gamma_g Z + e, \gamma_g > 0, g = 0,1$ (we assume that X and Y are measured as differences from their means). The counterpart of the LATE parameter

in the continuous treatment case is what Angrist and Pischke (2009, p.187) call the *Average Causal Response* (ACR), which is the same as the LATE parameter in equation (2) when c_g is replaced by γ_g .

The probability limit of the IV estimator in this model is analogous to the discrete treatment case, replacing c_g by γ_g in equation (4). As with LATE, the IV estimator of the ACR parameter depends not only on γ_g and β_g but also on the distributions of Z in each group (σ_{zg}). If $\sigma_{z0} = \sigma_{z1}$, then the IV estimate converges to the ACR, but otherwise it depends on the realisations of Z and therefore X . This simple model illustrates the statement by Masten and Torgovitsky (2016, p.1001) that as the treatment variable goes from 'binary to multivalued discrete to continuous, the 2SLS estimand becomes an increasingly complicated weighted average of LATEs between different X [treatment] realizations (Angrist & Imbens, 1995; Angrist, Graddy, & Imbens, 2000). This type of weighted average of LATEs can be difficult to interpret.'

When the treatment is continuous, there are, under certain conditions, IV estimators which provide a consistent estimate of an average causal effect in the presence of correlated double heterogeneity (e.g. Florens et al (2008), Imbens and Newey (2009) and Masten and Torgovitsky (2016)), although other restrictions are required for identification. These estimators have rarely been applied. We now consider four recent examples of IV estimation with a continuous treatment.

4.1 Women's education and completed fertility

There are many papers studying the relationship between education and fertility which exploit compulsory schooling reforms as instrumental variables. A recent one by Fort,

Schneeweis and Winter-Ebmer (2016), hereafter FSW, uses reforms in England and Continental Europe implemented between 1936 and 1975 to assess the causal effect of education on the number of biological children and the incidence of childlessness, both of which are measures of completed fertility, not period measures like the BGS study. Their treatment variable (X) is the number of years of education and their instrumental variable (Z) is the number of mandatory schooling years given by law. With F-statistics across their various samples ranging from 19 to 29, their instrument is powerful. FSW assume that each additional mandatory year of education exerts the same effect on the actual number of years of schooling in all the Continental European countries included in the study (the 1947 reform in England is treated separately), and they indeed do not find significantly different parameter estimates in either the first stage, reduced form or IV regressions. Table 2 reports their main estimates.

Table 2: Estimates of the effect of years of women's education on completed family size (standard errors in parentheses)

	England	Continental Europe
IV estimate	-0.307 (0.088)	0.291 (0.081)
OLS estimate	-0.047 (0.014)	-0.038 (0.005)
Reduced form	-0.214 (0.087)	0.090 (0.020)
First stage: impact of IV on Yrs. Educ.	0.698 (0.159)	0.311 (0.062)
F-statistic	19.20	25.31

Source: Fort, Schneeweis and Winter-Ebmer (2016), Table 3; the sample of women centred 7 years before and after the pivotal reform cohort.

Data: The Survey of Health, Ageing and Retirement in Europe (SHARE) for Continental Europe and the English Longitudinal Study of Ageing (ELSA) for England.

What is striking is that the OLS estimate is always much smaller in size than the IV estimate, and in the Continental European countries has the opposite sign to the IV estimate. Further analysis suggests that marriage played an important role in these

results: ‘more education induced by compulsory schooling laws increased marriage rates in Continental Europe but not in England’ (FSW, p.1841). FSW also report the reduced form for the number of children, which indicates relatively precisely estimated relationships between number of compulsory years of education and completed fertility, which is reassuring. They subject the reduced forms in individual countries to considerable analysis.

Regarding correlated double heterogeneity, is it likely that women whose fertility is more strongly affected by their education are also more (or less) likely to increase their education in response to the reform? The question is not addressed in the study and it is difficult to say *a priori* whether uncorrelatedness is a good working assumption for this application. It is conceivable, for example, that the fertility of women whose education responds strongly to the reform (i.e. it adds more years to their education) is not very sensitive to their education because their desired family size is low. Suppose for example that there are two groups for which $\beta_0 = 0$ and $\beta_1 = 0.5$, and that $\gamma_0 = 0.8, \gamma_1 = 0.5, \sigma_{z0}^2 = \sigma_{z1}^2$ and $s_0 = 0.5$. Then the LATE parameter is 0.19 while the average effect among compliers is 0.25.

Despite the fact that compliers cannot be identified because they are defined by counterfactual outcomes (those women who changed their education decisions due to the mandatory schooling reforms), FSW characterise the population of compliers with respect to pre-treatment variables by using Bayes Theorem (Angrist 2004). In addition to being women at the lower end of the education distribution, FSW find that the compliers in their study are more likely to have grown up in poorer and larger households and have more siblings relative to the average woman in the sample. Thus, FSW clarify

to whom their causal effect estimate applies and therefore the limitations of the study's external validity. For example, the estimated effects could tell us little about the impact of education on fertility at higher levels of education.

4.2 Maternal and child mortality

Andriano and Monden (2019) study the extent to which the correlation between maternal schooling and child survival is causal. The outcome variable is under-5 child mortality and the treatment variable is years of mother's education. Their instrument comes from schooling reforms introducing universal primary education (UPE) in the second half of the 1990s in Malawi and Uganda. In particular, official district-level statistics (27 and 38 districts in Malawi and Uganda, respectively) on the number of primary schools are used to measure a district's reform intensity (proportional change in the number of primary schools between 1994 and 1996 in Malawi, and between 1995 and 2001 for Uganda) for the birth cohorts affected by the reform. Thus, exposure to the UPE reforms depends on both the year and district of birth.

The F-test for the set of coefficients of the interactions between year of birth and program intensity for the youngest cohorts (the excluded instruments) are 57 for Malawi and 23 for Uganda (their Table 3), indicating substantial power of the instruments in predicting maternal education. As the main equation is a Cox survival model, its estimation uses the 'two-stage residual inclusion' (2SRI) approach, which is identical to IV in a linear setting.

The LATE parameter is the impact of maternal education on child mortality for those women whose years of education were increased by the UPE reform. But the IV estimates (i.e. those using 2SRI) of the odds ratio for child death associated with years of

maternal education are only slightly lower (i.e. the reduction in mortality is larger) than those that treat maternal education as exogenous (the analogue of OLS in this setting): the non-2SRI and 2SRI estimates with confidence intervals (CI) are 0.97 (CI: 0.958–0.984) and 0.90 (CI: 0.811–0.998), respectively, for Malawi, and 0.95 (CI: 0.934–0.969) and 0.83 (CI: 0.687–1.012), respectively, for Uganda. Despite the power of the instrument, the confidence intervals for the 2SRI estimates are relatively wide, with the non-2SRI point estimate well within them in each case, so that they are not significantly different from one another according to the Hausman test. This could be because maternal education is exogenous, but it also could indicate that the two types of heterogeneity are correlated.

Thus, in this study the conventional Cox estimator produces similar results and is more efficient. *If heterogeneity is not correlated*, we could conclude that maternal education is exogenous and the estimated causal effect is a population parameter, the average treatment effect, which enhances its external validity, at least for each particular country. The authors note that their findings are broadly in line with those of similar studies of change in child survival in education in developing countries related to mother's education, which obtained IV estimates that were greater than OLS estimates (Chou et al. 2010; Grépin and Bharadwaj 2015; Makate and Makate 2016).

4.3 Siblings and socio-demographic outcomes

Correlated double heterogeneity will occur when whatever is responsible for heterogeneity in the instrument – treatment relationship is also responsible (or is correlated with what is responsible) for heterogeneity in the treatment – outcome relationship. Consider the use of the sex mix of the first two children as an instrument

for number of siblings in models that seek to estimate the effect of number of siblings on educational attainment (Angrist and Evans 1998; Conley and Glauber 2006). The LATE parameter is the impact on education of having a third child among those whose first two children are the same sex. A variable such as wealth or income might affect the probability that parents whose first two children were both boys or both girls would have a third child, with better-off parents more likely to respond positively. But the same variables could also act to suppress the effect of having more siblings on any of their children's educational attainment. In this case the average negative effect of siblings on attainment among compliers will be underestimated by the IV estimator. A variable such as wealth or income is relatively straightforward to measure, but others, which might play a similar role in generating heterogeneous responses to the instrument and the treatment, might be less so. Parents who care more about the 'quality' of children would probably respond less strongly to the instrument but would invest more in the children they have. This would induce a positive correlation between the two kinds of heterogeneity, given that the effect of the number of siblings on educational attainment is expected to be negative.

Siblings could effect each other's fertility, through for example social influence or learning. But they also are likely to share underlying characteristics (such as fertility preferences) that cannot be measured. Hart and Cools (2019) use the sex mix to induce exogenous variation in siblings' fertility with an aim to estimating the causal effect of a sibling's fertility on their own. Unfortunately their IV estimates are very imprecise, which is probably more common than published studies would indicate (because such results are not published as often).

4.4 Alcohol consumption and health

Similar issues might arise when genes, or functions of genes such as polygenic scores, are used as instruments. In an early example of “Mendelian randomization” Davey-Smith et al. (2008) discuss the use of the ALDH2 gene as an instrument for alcohol consumption to assess the effect of alcohol on a range of health outcomes, such as high blood pressure or coronary heart disease. In these and other cases, differential response to the instrument might arise through unmeasured factors. Imagine that some people, for whatever reason, adopt a healthy lifestyle. On the one hand, this means that, even if they have the genetic variant that allows them to drink, they do so moderately. On the other, it means that the negative effect of alcohol on health outcomes is weakened by their healthy lifestyle. In this case, heterogeneity in the instrument – treatment relationship is positively correlated with heterogeneity in the treatment – outcome relationship.

5. Conclusion

Heterogeneous treatment effects are likely the norm in demographic and other social science applications. In many applications, demographers are looking for population parameters, such as an average causal effect across the population. But in many cases we cannot estimate them under credible assumptions. The exogenous variation that IV estimation exploits usually only allows us to estimate a local average effect (LATE) that applies to a population, whom we do not observe directly but whose behaviour is changed by the instrumental variable: ‘compliers’ in the treatment effect language. These estimates can nevertheless provide important information about how some people’s demographic behaviour, such as their fertility, varies with attributes of the person or their environment (e.g. their education or their access to an internet connection), and, as in the education-fertility example in this paper, it is sometimes possible to obtain a good

idea about who has been affected. Several authors have contended that the LATE at best identifies an effect which does not correspond to a parameter of a theoretical model or a quantity of relevance to policy (Deaton 2010; Heckman 1997, 1999; Heckman et al. 2006; Heckman and Urzua 2010), while Imbens (2010) has argued 'better LATE than nothing' (also see Angrist and Imbens 1999). We not believe the issue can be settled in the abstract, but depends on the particular case being studied, as some of our examples suggest.

Our main point is that what is being estimated may be difficult to interpret because of the correlated 'double heterogeneity' discussed in the paper. Unless we have strong reasons to believe that the relationship between the instrument and treatment, which in many models represents a choice or selection mechanism for treatment, is the same for everyone, or is uncorrelated with the heterogeneous treatment effect parameters, we should interpret IV estimates with caution. They do not necessarily estimate an average effect among compliers. Studies using IV estimation should address this point directly, and present arguments about why the correlated double heterogeneity issue may not apply and explain how the results may be affected if it does apply. In addition, applied studies should report the first stage estimation results, the reduced form for the outcome variable, comparison with the OLS estimates and overidentification tests when there is more than one instrument, and discuss the precision of the IV estimates and the likely external validity of the estimates.

In preferring IV to OLS we are, in effect, preferring one set of assumptions over another and often restricting ourselves to an effect which is not a population parameter. Also, valid and powerful instruments are hard to fine. In this paper we have pointed to an issue

that is seldom considered: heterogeneity in the effect of the instrument on treatment and its correlation with variation in treatment effects. If independence between the two seems unlikely to hold we might consider other approaches that combine methods that control only for observables along with assessments of the sensitivity of the causal estimates to the assumptions on which they rest.

References

Andriano, L. and Monden, C.W.S. (2019), The causal effect of maternal education on child mortality: evidence from a quasi-Experiment in Malawi and Uganda. *Demography* 56:1765–1790.

Angrist, J.D. (2004). ‘Treatment effect heterogeneity in theory and practice’, *Economic Journal*, vol. 114 (March), pp. C52–C83.

Angrist, J. D. and W.N. Evans (1998) Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review*, 88 (3): 450-477.

Angrist, J. D. and Imbens, G.W. (1999) Comment on James J. Heckman, “Instrumental variables: A study of implicit behavioural assumptions in making program evaluations.” *Journal of Human Resources* 34(4): 824-827.

Angrist, J. D. and Pischke, J-S. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Billari, F.C., Giuntell, O. and Stella, L. (2019), Does broadband affect fertility? *Population Studies* 73(3): 297-316.

Bjorklund, A. and R. Moffitt (1987) The estimation of wage gains and welfare gains in self-selection. *The Review of Economics and Statistics* 69(1): 42-49.

Bollen, KA (2012) Instrumental Variables in Sociology and the Social Sciences. *Annual Review of Sociology* 38: 37-72

Chou, S, -Y., Liu, J. -T., Grossman, M. and Joyce, T. (2010). Parental education and child health. *American Economic Journal: Applied Economics*. 2(1): 33-36.

Conley, D. and R. Glauber (2008). Parental educational investment and children's academic risk estimates of the impact of sibship size and birth order from exogenous variation in fertility. *Journal of Human Resources* 41 (4): 722-737

Davey Smith, G., N. Timpson and S. Ebrahim (2008). Strengthening causal inference in cardiovascular epidemiology through Mendelian randomization. *Annals of Medicine*, 40(7): 524-41.

Deaton, A. S. 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* 48(2): 424-55.

Ding, P. and Vander Weele, T.J. (2016), Sensitivity Analysis Without Assumptions, *Epidemiology* 27(3):368-377.

Florens, J-P., Heckman, J.J., Meghir, C., and Vytlacil, E. (2008) Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica* 76:1191–1206.

Deuchert, E., and Huber, M. (2017). A cautionary tale about control variables in IV estimation. *Oxford Bulletin of Economics & Statistics*, 79(3), 411–425.

Grépin, K.A. and Bharadwaj, p. (2015). Maternal education and child mortality in Zimbabwe. *Journal of Health Economics*. 44:97-117.

Hart, R. K. and Cools, S (2019) Identifying interaction effects using random fertility shocks. *Demographic Research* 40: 261–278

Heckman, J. J. and Robb, R. (1985) Alternative methods of evaluating the impact of interventions (pp. 156-245) in J.J. Heckman and B. Singer (Eds) *Longitudinal Analysis of Labor Market Data Volume 10*. New York: Cambridge University Press.

Heckman, J. J. and Robb, R. (1985) Alternative methods for solving the problem of selection bias in evaluating the effect of treatments on outcomes (pp. 63-107) *Drawing Inferences from Self-selected Samples*. New York: Springer Verlag.

Heckman, J. J. (1997) Instrumental variables: A study of implicit behavioural assumptions in making program evaluations. *Journal of Human Resources* 32(3): 441-462.

Heckman, J. J. (1999) Instrumental variables: response to Angrist and Imbens. *Journal of Human Resources* 34(4): 828-837.

Heckman, J. J. and S. Urzua (2010) Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics* 156: 27-37.

Heckman, J. J., Urzua, S. and E. Vytlacil (2006) Understanding instrumental variable in models with essential heterogeneity. *The Review of Economics and Statistics* 88(3): 389-432.

Hotz, V.J., Mullins, C.H. and Sanders, S.G. (1997). Bounding causal effects using data from a contaminated natural experiment: analyzing the effects of teenage childbearing. *The Review of Economic Studies* 64:575-603.

Imbens, G. W. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48(2): 399-423.

Imbens, G. W. and Angrist, J.D. (1994) Identification and estimation of local average treatment effects, *Econometrica* 62(4):467-476.

Imbens, G. W., and Newey, W.K. (2009) Identification and estimation of triangular simultaneous equations models without additivity, *Econometrica* 77:1481-1512.

Lutz, W., Butz, W. P. and Samir, K.C. (2014) *World Population and Human Capital in the Twenty-First Century*. Oxford University Press.

Makate, M. and Makate, C. (2016) The causal effect of increased primary schooling on child mortality in Malawi: Universal primary education as a natural experiment. *Social Science and Medicine*. 168-72-83.

Masten, M.A. and Torgovitsky, A. (2016) Identification of instrumental variable random coefficients models. *The Review of Economics and Statistics*. 98(5): 1001–1005.

Roy, A. (1951). Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers* 3 (2): 135–146.

Stock, J. and Trebbi, F. (2003) Who Invented Instrumental Variable Regression? *Journal of Economic Perspectives*. 17(3):177–194.

Wooldridge, J. M. (1997), On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model. *Economics Letters* 56:129–133.

Wooldridge JM. 2002. *Econometric Analysis of Cross Section and Panel Data*. 2nd Edition. Cambridge, MA: MIT Press.

_____ (2003) Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model. *Economics Letters* 79: 185–191.