*The Economic Journal*, **114** (*March*), C52–C83. © Royal Economic Society 2004. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

# TREATMENT EFFECT HETEROGENEITY IN THEORY AND PRACTICE*

*Joshua D. Angrist*

Instrumental Variables (IV) methods identify internally valid causal effects for individuals whose treatment status is manipulable by the instrument at hand. Inference for other populations requires homogeneity assumptions. This paper outlines a theoretical framework that nests causal homogeneity assumptions. These ideas are illustrated using sibling-sex composition to estimate the effect of childbearing on economic and marital outcomes. The application is motivated by American welfare reform. The empirical results generally support the notion of reduced labour supply and increased poverty as a consequence of childbearing but evidence on the impact of childbearing on marital stability and welfare use is more tenuous.

Empirical research often focuses on causal inference for the purpose of prediction, yet it seems fair to say that most prediction involves a fair amount of guesswork. The relevance or 'external validity' of a particular set of empirical results is always an open question. As Karl Pearson (1911, p. 157) observed in an early discussion of the use of correlation for prediction, 'Everything in the universe occurs but once, there is no absolute sameness of repetition.' This practical difficulty notwithstanding, empirical research is almost always motivated by a belief that estimates for a particular context provide useful information about the likely effects of similar programmes or events in the future. Our investment of time and energy in often-discouraging empirical work reveals that empiricists like me are willing to extrapolate.

The basis for extrapolation is a set of assumptions about the cross-sectional homogeneity or temporal stability of causal effects. As a graduate student, I learned about parameter stability as 'the Lucas critique', while my own teaching and research focuses on the identification possibilities for average causal effects in models with heterogeneous potential outcomes. Applied micro-econometricians devote considerable attention to the question of whether homogeneity and stability assumptions can be justified and to the implications of heterogeneity for alternative parameter estimates. Regrettably, this sort of analysis sometimes comes at the expense of a rigorous examination of the internal validity of estimates, i.e., whether the estimates have a causal interpretation for the population under study. Clearly, however, even internally valid estimates are less interesting if they are completely local, i.e., have no predictive value for populations other than the directly affected group.

* Presented as the Sargan lecture. Denis Sargan developed much of the estimation theory that instrumental variables practitioners rely on today. For recent surveys of Sargan's contributions see Arellano (2002) and Phillips (2003). Special thanks go to Patricia Cortes and Francisco Gallego for outstanding research assistance and to Alberto Abadie, Victor Chernozhukov, Jerry Hausman, Guido Imbens, Jiemon Woutersen and seminar participants at Columbia, UCLA, UCSB and Yale for helpful discussions and comments. Thanks also go to the editors, Jon Temple and Ian Preston, for their detailed written comments. This is a revised version of NBER Working Paper 9708.

In this paper, I discuss the nature and consequences of homogeneity assumptions that facilitate the use of instrumental variables (IV) estimates for extrapolation.[1] To be precise, I am interested in the assumptions that link a Local Average Treatment Effect (LATE) tied to a particular instrument with the population Average Treatment Effect (ATE), which is not instrument-dependent. Implicitly, I have in mind prediction for populations defined by covariates. I focus on ATE because it answers the question: 'If we were to treat individuals with characteristics *X*, what would the likely change in outcomes be'? This allows me to sidestep variability due to changes in the process determining treatment status. For example, causal research often focuses on average treatment effects in the treated population. Overall average treatment effects are theoretically more stable than average effects on the treated, since the latter depend on who gets treated as well as on the distribution of potential outcomes.

The external validity of IV estimates is of special interest both because of the growing importance of IV methods in empirical work (see, e.g., Moffit, 1999), and because the *ex ante* generality of IV estimates is limited in a precise way by a number of well-known theoretical results. Except in special cases like constant treatment effects and certain types of randomised trials, the standard IV assumptions of exclusion and independence – analogous to the notion that the instrument induces a good experiment for the effect of interest – are not sufficient to capture the expected causal effect on a randomly selected individual or even in the population subject to treatment. Rather, basic IV assumptions identify causal effects on 'compliers', defined as the subpopulation of treated individuals whose treatment status can be influenced by the instrument. Although this limitation is unsurprising and the compliers are often of interest in themselves, the nature and plausibility of assumptions under which IV estimates have broader predictive power are worth exploring.

The next two Sections develop a theoretical framework linking alternative causal parameters to population subgroups defined by their response to an instrument. That is, I consider formal links between parameters like LATE and ATE. My agenda is to make this link using a range of assumptions, progressing from stronger (no selection bias) to weaker (a proportionality assumption). These theoretical ideas are then applied to the *same sex* instrument, used by Angrist and Evans (1998) to estimate the effects of childbearing on labour supply. This instrument arises from the fact that some parents prefer a mixed sibling sex composition. In particular, among parents who have at least two children, those with two boys or two girls are much more likely to go on to have a third child. Because child sex is virtually randomly assigned, a dummy for same sex sibling pairs provides an instrumental variable that can be used to identify the effect of childbearing on a range of economic and family outcomes.

My earlier work with Evans using the *same sex* instrument focused on the effects of childbearing on labour supply. While labour supply outcomes also appear

[1] Denis Sargan noted the difficulty of the core instrumental variables identification problem, i.e., identification in models with constant effects, in his seminal 1958 paper: 'It is not easy to justify the basic assumptions concerning these errors, namely that they are independent of the instrumental variables.' (p. 396; quoted in Arellano (2002)).

in this paper, the empirical work features an investigation of the effects of childbearing on marital stability. An inquiry into the effects of family size on marital stability can be motivated by American welfare reform, which penalises further childbearing by women receiving public assistance on the grounds that increases in family size make continued poverty and welfare receipt more likely. I therefore look at effects on marital status, poverty status, and welfare use, as well as labour supply. Estimates of ATE for the effects of childbearing are generally smaller than estimates of LATE. For example, while estimates of LATE for the effect of childbearing on welfare use and marital stability are mostly significantly different from zero, most (though not all) of the estimates of ATE for effects of childbearing on marital status and welfare use are small and insignificant. One tantalising result is that for teen mothers, LATE is identical to the population average treatment effect when the latter is imputed under two of the assumptions considered below.

The empirical results suggest a pattern of modest effects but the variability in parameter estimates across model specifications and samples, as well as the usual problem of more imprecise estimates under weaker identifying assumptions, reduces the predictive value of any findings. On balance it seems fair to say that the attempt to go from LATE to ATE weakens the evidence for an adverse effect of childbearing on marital stability and welfare use, but the estimates of ATE do not provide a sharp alternative to LATE. This is perhaps not surprising, given the difficulty of the underlying identification problem. As in the experimental sciences, the best evidence for predictive value is likely to come from new data sets and new experiments, which in the case of applied econometrics usually means new instruments.

## 1. Causality and Potential Outcomes in Research on Childbearing

The effects of children on marital stability have long been of interest to social scientists and are of course of more than academic interest to many married couples. Previous research (Becker *et al.*, 1977; Cherlin, 1977; Heaton, 1990 and Waite and Lillard, 1991) suggests the presence of young children increases marital stability, although many authors acknowledge serious selection problems that may bias results. A related issue is the connection between childbearing and women's standard of living. A large literature looks at the effect of teen childbearing on mothers' schooling, earnings, and welfare status, sometimes using instrumental variables (Bronars and Grogger, 1994). Interest in this question can be motivated by welfare reform, which includes 'family caps' in many US states. Family caps reduce or eliminate benefits paid for children born to welfare recipients, on the theory that further childbearing by welfare mothers increases the likelihood they will stay poor and therefore continue to receive benefits.[2]

Are children the glue that holds couples together or a burden that accelerates a fragile family's collapse? Does childbearing further impoverish poor women?

[2] See Maynard *et al.* (1998) for more on the motivation for family caps. The possibility of a link between childbearing and poverty notwithstanding, there is little evidence that family caps actually affect fertility. See, e.g., Grogger and Bronars (2001), Blank (2002) or Kearney (2002).

Implicit in these questions is the notion of potential outcomes, i.e., a contrast in circumstances with and without childbearing, for a given family. To represent this idea formally, let $D_i$ be an indicator for women with more than two children in a sample of women with at least two children. Because $D_i$ is binary, I will refer to it as a 'treatment', even though family size is not determined directly by a programme or policy. Let $Y_{1i}$ be a woman's circumstances if $D_i = 1$, and let $Y_{0i}$ be her circumstances otherwise. We imagine both of these potential outcomes are well-defined for everyone, though only one is ever observed for each woman. Formally, this can be expressed by writing the observed outcome, $Y_i$, as

$$Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i.$$

For both practical and substantive reasons, I focus here on fertility consequences defined with reference to the transition from two to more than two children. On the practical side, instruments based on sibling sex composition are available for this fertility increment. Angrist and Evans (1998) used parents' preferences for a mixed sibling sex composition to estimate the labour supply consequences of childbearing. On the substantive side, post-war reductions in marital fertility have been concentrated in the 2–3 child range (Westhoff *et al.*, 1963). While almost all couples want at least one child, the decision to have a third child may be due in part to a sense of whether this is good for long-term marital stability or, more generally, the economic welfare of the family. Finally, interest in the 2-to-3 child increment is supported by the fact that the population of welfare mothers in 1990 had an average of about 2.3 children and a median of 2 children.

Since $Y_{1i}$ and $Y_{0i}$ are never both observed for the same woman, research on causal effects tries to capture the average difference in potential outcomes for different subpopulations. For example, we may be interested in $\mathrm{E}(Y_{1i} - Y_{0i}|D_i = 1)$, which is the effect on women who have a third child. Note that $\mathrm{E}(Y_{1i}|D_i = 1)$ is an observable quantity, so estimating $\mathrm{E}(Y_{1i} - Y_{0i}|D_i = 1)$ is equivalent to estimating the counter-factual average, $\mathrm{E}(Y_{0i}|D_i = 1)$. Alternately, we may be interested in the unconditional average treatment effect (ATE), $\mathrm{E}(Y_{1i} - Y_{0i})$, which can be used to make predictive statements about the impact of childbearing on a randomly chosen woman (or a woman with a particular set of characteristics if the analysis conditions on covariates). Estimation of ATE is equivalent to estimation of both counterfactual averages, $\mathrm{E}(Y_{0i}|D_i = 1)$and $\mathrm{E}(Y_{1i}|D_i = 0)$.

Causal parameters are easy to describe but hard to measure. The observed difference in outcomes between those with $D_i = 1$ and $D_i = 0$ equals $\mathrm{E}(Y_{1i} - Y_{0i}|D_i = 1)$ plus a bias term:

$$\begin{aligned}\mathrm{E}(Y_i|D_i = 1) - \mathrm{E}(Y_i|D_i = 0) &= \mathrm{E}(Y_{1i}|D_i = 1) - \mathrm{E}(Y_{0i}|D_i = 0)\\ &= \mathrm{E}(Y_{1i} - Y_{0i}|D_i = 1)\\ &\quad + [\mathrm{E}(Y_{0i}|D_i = 1) - \mathrm{E}(Y_{0i}|D_i = 0)]. \qquad (1)\end{aligned}$$

The bias term disappears when childbearing is determined in a manner independent of a woman's potential outcomes. But this independence assumption seems unrealistic since childbearing decisions are made in light of information about family circumstances and earnings potential.

Two sorts of strategies are typically used to estimate causal effects in the presence of possible omitted variables bias. One assumes that conditional on covariates, $X_i$, the regressor of interest, $D_i$, is independent of potential outcomes. Then any causal effect of interest can be estimated from weighted conditional-on-$X$ comparisons. This is a strong assumption that seems most plausible when researchers have considerable prior information about the process determining $D_i$. Alternately, we might try to find an instrumental variable which, perhaps after conditioning on covariates, is related to $D_i$ but independent of potential outcomes. The instrument used here is a dummy variable indicating same-sex sibling pairs.

## 2. IV in Context

IV estimates capture the effect of treatment on the treated for those whose treatment status can be changed by the instrument at hand. This idea is easiest to formalise using a notation for potential treatment assignments that parallels the notation for potential outcomes. In particular, let $D_{0i}$ and $D_{1i}$ denote potential treatment assignments defined relative to a binary instrument. Suppose, for example, $D_i$ is determined by a latent-index assignment mechanism,

$$D_i = 1(\gamma_0 + \gamma_1 Z_i > \eta_i), \tag{2}$$

where $Z_i$ is a binary instrument, and $\eta_i$ is a random error independent of the instrument. Then the potential treatment assignments are $D_{0i} = 1(\gamma_0 > \eta_i)$ and $D_{1i} = 1(\gamma_0 + \gamma_1 > \eta_i)$, both of which are independent of $Z_i$.

The constant-effects latent-index assignment model is restrictive since it implies $D_{1i} > D_{0i}$ for all $i$, or *vice versa*. We can relax this restriction by allowing a random $\gamma_{1i}$ for each $i$, in which case the latent index model is just an alternative notation for $D_{0i}$ and $D_{1i}$. Whether linked to an index model or not, $D_{0i}$ tells us what treatment $i$ would receive if $Z_i = 0$, and $D_{1i}$ tells us what treatment $i$ would receive if $Z_i = 1$. The observed assignment variable, $D_i$, can therefore be written:

$$D_i = D_{0i}(1 - Z_i) + D_{1i}Z_i.$$

This notation makes it clear that, paralleling potential outcomes, only one potential assignment is ever observed for a particular individual.

The key assumptions supporting IV estimation are given below (for a model without covariates):

INDEPENDENCE. $\{Y_{0i}, Y_{1i}, D_{0i}, D_{1i}\} \perp\!\!\!\perp Z_i$.
FIRST STAGE. $\mathrm{P}(D_i = 1|Z_i = 1) \neq \mathrm{P}(D_i = 1|Z_i = 0)$.
MONOTONICITY. Either $D_{1i} \geq D_{0i}\ \forall i$ or *vice versa*; without loss of generality, assume the former.

These assumptions capture the notion that the instrument is 'as good as randomly assigned' (independence), affects the probability of treatment (first-stage), and affects everyone the same way if at all (monotonicity). Imbens and Angrist (1994) show that together they imply:

$$\frac{\mathrm{E}(Y_i|Z_i = 1) - \mathrm{E}(Y_i|Z_i = 0)}{\mathrm{E}(D_i|Z_i = 1) - \mathrm{E}(D_i|Z_i = 0)} = \mathrm{E}(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}).$$

The left-hand side of this expression is the population analog of Wald's (1940) estimator for regression models with measurement error. The Local Average Treatment Effect (LATE) on the right hand side, $E(Y_{1i} - Y_{0i}|D_{1i} > D_{0i})$, is the effect of treatment on those whose treatment status is changed by the instrument, i.e., the population for which $D_{1i} = 1$ and $D_{0i} = 0$.[3]

A standard assumption invoked in most empirical studies is constant causal effects, i.e.,

$$Y_{1i} = Y_{0i} + \alpha,$$

for some constant $\alpha$. In the empirical study, below, this implies that IV consistently estimates the common effect of childbearing on all women, since, given constant effects, $E(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}) = \alpha$. The LATE result highlights the fact that in a more realistic world where this effect varies (and indeed it must vary if, for example, $Y_i$ is a binary outcome or other variable with limited support), then we can be sure only that IV captures the effect on individuals whose treatment status can be changed by manipulating $Z_i$. These are people with $D_{1i} = 1$ and $D_{0i} = 0$, or $D_{1i} - D_{0i} = 1$. Note also that since $D_{1i}$ and $D_{0i}$ are defined with reference to a particular instrument, we should expect different instruments to uncover different average causal effects. We might, for example, expect an IV strategy based on the *same sex* instrument to identify a different average effect from an instrument based on twin births. In fact, Angrist and Evans (1998) report IV estimates using twin birth instruments that are much lower than those using *same sex* instruments.

Angrist *et al.* (1996) refer to people with $D_{1i} - D_{0i} = 1$ as the population of *compliers*. This terminology is motivated by an analogy to randomised trials where $Z_i$ is a randomised offer of treatment and $D_i$ is actual treatment status. Since $D_{1i} - D_{0i} = 1$ implies $D_i = Z_i$, compliers are those who comply with an experimenter's intended treatment status (though not all those with $D_i = Z_i$ are compliers, as explained below). For compliers, the averages of $Y_{i1}$ and $Y_{0i}$ as well as the average difference are also identified. In particular, Abadie (2002) shows that

$$\frac{E(Y_iD_i|Z_i = 1) - E(Y_iD_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} = E(Y_{1i}|D_{1i} > D_{0i}) \tag{3a}$$

$$\frac{E[Y_i(1 - D_i)|Z_i = 1] - E[Y_i(1 - D_i)|Z_i = 0]}{E[(1 - D_i)|Z_i = 1] - E[(1 - D_i)|Z_i = 0]} = E(Y_{0i}|D_{1i} > D_{0i}). \tag{3b}$$

The entire (marginal) distributions of $Y_{1i}$ and $Y_{0i}$ are similarly identified, a fact used by Abadie *et al.* (2002) to estimate the causal effect of treatment on the quantiles of potential outcomes for compliers.

An important econometric result in the theory of causal effects is that when treatment is assigned by a mechanism like (2), population average treatment

[3] Proof of the LATE result: $E(Y_i|Z_i = 1) = E[Y_{0i} + (Y_{1i} - Y_{0i})D_i|Z_i = 1]$, which equals $E[Y_{0i} + (Y_{1i} - Y_{0i})D_{1i}]$ by independence. Likewise $E(Y_i|Z_i = 0) = E[Y_{0i} + (Y_{1i} - Y_{0i})D_{0i}]$, so the Wald numerator is $E[(Y_{1i} - Y_{0i})(D_{1i} - D_{0i})]$. Monotonicity means $D_{1i} - D_{0i}$ equals one or zero, so $E[(Y_{1i} - Y_{0i})(D_{1i} - D_{0i})] = E(Y_{1i} - Y_{0i}|D_{1i} > D_{0i})P(D_{1i} > D_{0i})$. A similar argument shows $E(D_i|Z_i = 1) - E(D_i|Z_i = 0) = E(D_{1i} - D_{0i}) = P(D_{1i} > D_{0i})$.

effects and the effect on the treated are not identified without assumptions such as constant effects or some other assumption beyond the three given above. This result or theorem appears in various forms; see, for example, Chamberlain (1986), Heckman (1990) and Angrist and Imbens (1991). The next Section develops a framework that highlights the limits to identification and the role played by alternative homogeneity assumptions in efforts to go beyond LATE. The *same sex* instrument offers an especially challenging proving ground for these ideas since at most 7% of American women had an additional child as a result of sex preferences. Causal effects on *same sex* compliers can therefore be quite far from overall average effects if the impact of childbearing on these women is not typical. Before turning to a general discussion of treatment effect heterogeneity, however, I briefly explore the relationship between LATE, ATE, and effects on the treated in a parametric model that mimics the *same sex* setup.

### 2.1. *A Parametric Example*

Following Heckman *et al.* (2001), I calculated average causal effects using a trivariate Normal model for the joint distribution of potential outcomes and the error term in the latent-index assignment mechanism given by (2). Assuming the distribution of $(Y_{1i}\ Y_{0i}\ \eta_i)'$ is joint standard Normal, ATE is zero by construction. Assume also that $\gamma_1 > 0$ so monotonicity is satisfied with $D_{1i} \geq D_{0i}$ and let $\rho_{10}$ be the correlation between $Y_{1i} - Y_{0i}$ and $\eta_i$. In this parametric model, LATE can be written:

$$\begin{aligned} \mathrm{E}(Y_{1i} - Y_{0i} | D_{1i} > D_{0i}) &= \mathrm{E}(Y_{1i} - Y_{0i} | \gamma_0 + \gamma_1 > \eta_i > \gamma_0) \\ &= \rho_{10}\Big\{[\varphi(\gamma_0) - \varphi(\gamma_0 + \gamma_1)][\Phi(\gamma_0 + \gamma_1) - \Phi(\gamma_0)]^{-1}\Big\}, \quad (4) \end{aligned}$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the Normal density and distribution functions. Similarly, we can use Normality to write the effect on the treated as:

$$\begin{aligned} \mathrm{E}(Y_{1i} - Y_{0i} | D_i = 1) &= \mathrm{E}[\mathrm{E}(Y_{1i} - Y_{0i} | \gamma_0 + \gamma_i Z_i > \eta_i, Z_i) | D_i = 1] \\ &= -\rho_{10}\{\lambda(\gamma_0 + \gamma_1)\mathrm{E}(Z_i | D_i = 1) \\ &\quad + \lambda(\gamma_0)[1 - \mathrm{E}(Z_i | D_i = 1)]\}. \quad (5) \end{aligned}$$

where $\lambda(\cdot)$ is the inverse Mill's ratio, $\varphi(\cdot)/\Phi(\cdot)$. This formula is useful for calculation, but the following expression better clarifies the difference between LATE and the effect on the treated:

$$\mathrm{E}(Y_{1i} - Y_{0i} | D_i = 1) = \mathrm{E}(Y_{1i} - Y_{0i} | \gamma_0 + \gamma_1 > \eta_i > \gamma_0)\omega + \mathrm{E}(Y_{1i} - Y_{0i} | \gamma_0 > \eta_i)(1 - \omega), \quad (6)$$

where $\omega = [\Phi(\gamma_0 + \gamma_1) - \Phi(\gamma_0)]\,[\mathrm{P}(Z_i = 1)/\mathrm{P}(D_i = 1)]$ and $1 - \omega = \Phi(\gamma_0)/\mathrm{P}(D_i = 1)$. Equation (6) shows the effect on the treated to be a weighted average of LATE and the average effect on those with $\gamma_0 > \eta_i$, with weights that depend on the first stage and the distribution of $Z_i$.

2.1.1. *LATE vs. the effect on the treated*

LATE and the effect on the treated both depend on the correlation between potential outcomes and the latent first-stage error, and on the first-stage coefficients. The effect on the treated also depends on the distribution of the instrument. The relationship between alternative causal parameters in the parametric model is sketched in Figure 1, which plots ATE (a constant equal to zero), LATE and the effect on the treated against $\Phi(\gamma_0)$ for a fixed first stage of 0.07 and an instrument that is Bernoulli (0.5). In other words, as with the *same sex* instrument

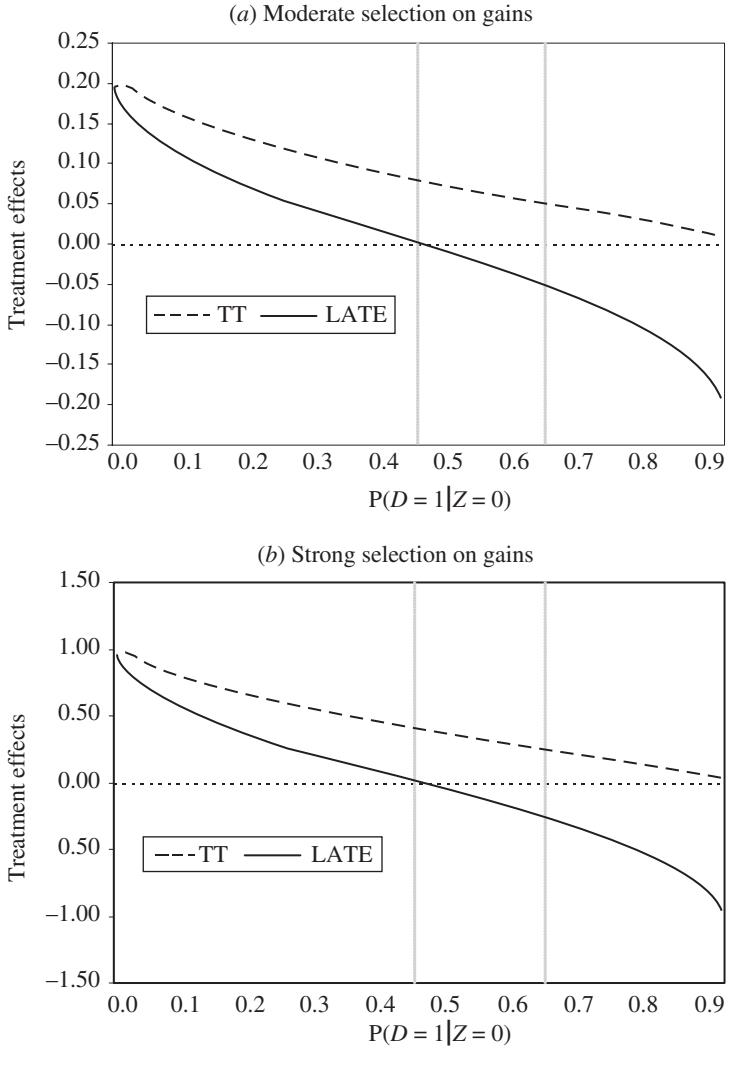


Fig. 1. *The Relationship Between LATE, ATE and the Effect on the Treated* (*TT*) *for Alternate First-stage Baseline Values.* The first-stage effect is fixed at 0.07 and ATE = 0. The top panel calculation sets the correlation between gains and the treatment index to −0.1, while the bottom panel sets this correlation to −0.5

in Angrist and Evans (1998), the simulated instrument is a dummy that equals one with probability $\frac{1}{2}$ and increases the probability that $D_i$ equals 1 by 7 percentage points. The calculation used for the top panel of Figure 1 sets $\rho_{10} = -0.1$, so that the probability of treatment increases with the gains from treatment, as in a Roy (1951) model, while the bottom panel calculation sets $\rho_{10} = -0.5$ for stronger selection on gains. With positive $\rho_{10}$, the Figure would be reflected through the horizontal axis.

The leftmost point in the Figure shows that LATE equals the effect on the treated when $\Phi(\gamma_0) = \mathrm{E}(D_i|Z_i = 0) = 0$. This is incompatible with the Normal latent-index model since it requires $\gamma_0 = -\infty$, but $\mathrm{E}(D_i|Z_i = 0) = 0$ is an important special case in practice, most commonly in randomised trials with partial compliance only in the treated group (Bloom, 1984; Angrist and Imbens, 1991).[4] At the other end of the Figure, the effect on the treated approaches the overall average effect when almost everyone gets treated. Finally, Figure 2 shows that increasing the size of the first stage effect from 0.07 to 0.30 pulls both LATE and the effect on the treated closer to the overall average effect.

The effect of treatment on the treated is above LATE for all first-stage baseline values, a consequence of the fact that selection on gains makes $\mathrm{E}(Y_{1i} - Y_{0i}|\gamma_0 > \eta_i)$ bigger than LATE. Moreover, LATE provides a better measure of the effect of treatment on a randomly chosen individual (ATE) than does the effect on the treated for most parameter values. A final important feature of the Figure (also apparent from (4)) is that LATE = ATE when $\gamma_1 = -2\gamma_0$ since $\varphi(\gamma_0) = \varphi(-\gamma_0)$ by symmetry of the Normal density. Thus, as noted by Heckman and Vytlacil (2000), a 'symmetric first stage' that changes the probability of treatment from $p$ to $1 - p$ implies LATE equals ATE in the Normal model, or in any latent variable model with jointly symmetric errors.[5]

## 3. Identification Problems and Prospects

Angrist *et al.* (1996) show that the potential-outcomes framework for IV divides a population into three groups, which I refer to below as 'potential-assignment subpopulations'. The first are compliers, i.e., those for whom $D_{1i} = 1$ and $D_{0i} = 0$. In the latent index model, compliers have $\gamma_0 + \gamma_1 > \eta_i > \gamma_0$. The other two groups include individuals whose treatment status is unaffected by the instrument. One consists of *never-takers*, with $D_{1i} = D_{0i} = 0$. Never-takers are never treated regardless

[4] A leading example is the randomised trial used to evaluate subsidised training programmes offered through the Job Training Partnership Act, one of America's largest Federally-sponsored training programmes. Subsidised training was offered but not compulsory in the randomly selected treatment group. About 60% of those offered treatment took up the offer, so $\mathrm{E}(D_i|Z_i = 1) = 0.6$, where $Z_i$ is the randomised offer of treatment and $D_i$ is actual training status. On the other hand, (virtually) no one in the control group received treatment, so $\mathrm{E}(D_i|Z_i = 0) \approx 0$. In this case, LATE is the effect on the treated because the set of always-takers is virtually empty. See Orr *et al.* (1996) for an IV analysis of the JTPA.

[5] Joint symmetry means that if $f(y_{ji}, \eta_i)$ is the joint density of $y_{ji} = Y_{ji} - \mathrm{E}(Y_{ji})$ and $\eta_i$, then $f(-y_{ji}, -\eta_i) = f(y_{ji}, \eta_i)$. A weaker parametric restriction with the same result (a symmetric first stage ranging from $p$ to $1 - p$ gives LATE = ATE) is that $\mathrm{E}(y_{ji}|\eta_i)$ is an odd function (as for a linear model) and that $\eta_i$ has a symmetric distribution. Angrist (1991) somewhat more loosely noted that IV estimates should be close to ATE when the first stage changes the probability of treatment at values centered on one-half, as is required for the first stage to be symmetric.

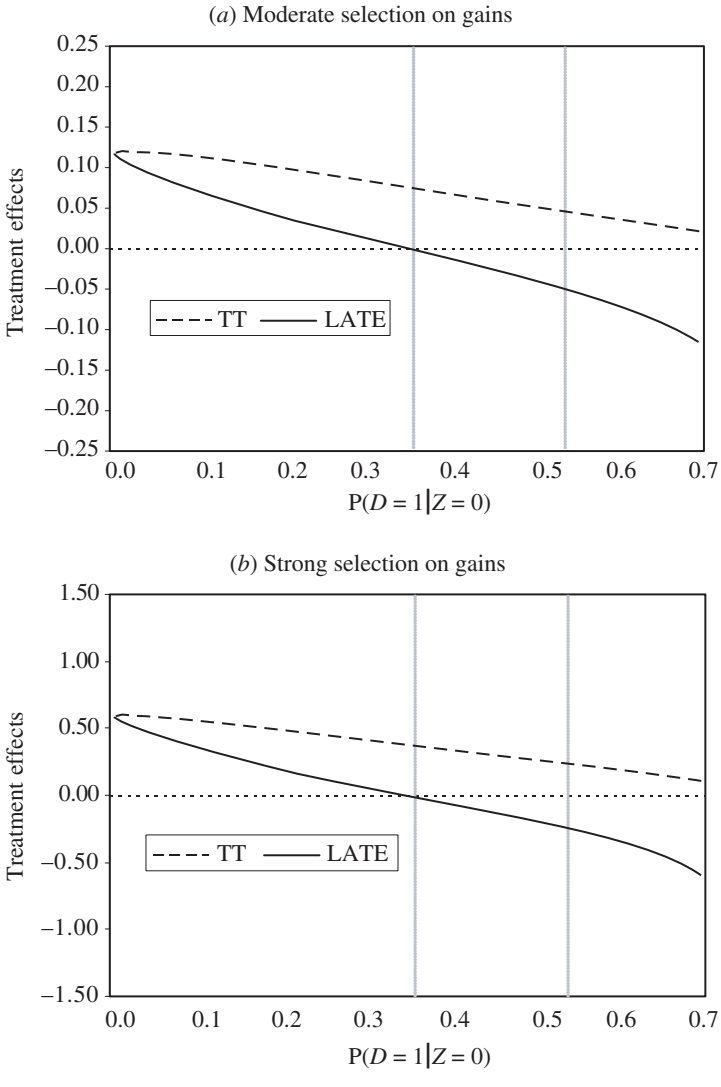


Fig. 2. *The Relationship Between LATE, ATE and the Effect on the Treated* (*TT*) *for Alternate First-stage Baseline Values.* The first-stage effect is fixed at 0.30 and ATE = 0. The top panel calculation sets the correlation between gains and the treatment index to −0.1, while the bottom panel sets this correlation to −0.5

of the value of $Z_i$ to which they might be exposed. In the latent index model, never-takers have $\eta_i > \gamma_0 + \gamma_1$. The second unaffected group consists of *always-takers*, with $D_{1i} = D_{0i} = 1$. Always-takers are always treated regardless of the value of $Z_i$ to which they might be exposed. In the latent-index model, always-takers have $\gamma_0 > \eta_i$. A possible fourth group with $D_{0i} = 1$ and $D_{1i} = 0$ is empty by virtue of the monotonicity assumption.

The set of the treated is the union of the disjoint sets of always-takers and compliers with $Z_i = 1$. This provides an interpretation for the following identity:

$$D_i = D_{0i} + (D_{1i} - D_{0i})Z_i,$$

since $D_{0i} = 1$ indicates always-takers and $(D_{1i} - D_{0i})Z_i$ indicates compliers with $Z_i = 1$. Because $Z_i$ is independent of complier status, compliers with $Z_i = 1$ are representative of all compliers. Causal effects on the treated can therefore be decomposed as:

$$\begin{aligned}\mathrm{E}(Y_{1i} - Y_{0i}|D_i = 1) = {} & \mathrm{E}(Y_{1i} - Y_{0i}|D_{0i} > D_{1i})[1 - \mathrm{P}(D_{0i} = D_{1i} = 1|D_i = 1)] \\ & + \mathrm{E}(Y_{1i} - Y_{0i}|D_{0i} = D_{1i} = 1)\,\mathrm{P}(D_{0i} = D_{1i} = 1|D_i = 1). \quad (7)\end{aligned}$$

Equation (7) generalises (6), which gives the same decomposition for the Normal model. Because an instrumental variable provides no information about average treatment effects in the set of always-takers, LATE is identified while $\mathrm{E}(Y_{1i} - Y_{0i}|D_i = 1)$ is not.

To pinpoint the identification challenge in this context further, note that $\mathrm{E}(Y_{1i}|D_{0i} = D_{1i} = 1)$ and $\mathrm{E}(Y_{0i}|D_{0i} = D_{1i} = 0)$ can be estimated using the following relations:

$$\mathrm{E}(Y_{1i}|D_{0i} = D_{1i} = 1) = \mathrm{E}(Y_{1i}|D_{0i} = 1) = \mathrm{E}(Y_i|D_i = 1,\ Z_i = 0) \quad (8a)$$

$$\mathrm{E}(Y_{0i}|D_{0i} = D_{1i} = 0) = \mathrm{E}(Y_{0i}|D_{1i} = 0) = \mathrm{E}(Y_i|D_i = 0,\ Z_i = 1). \quad (8b)$$

The missing pieces of the identification puzzle are therefore the fully counterfactual averages, $\mathrm{E}(Y_{1i}|D_{0i} = D_{1i} = 0)$ and $\mathrm{E}(Y_{0i}|D_{0i} = D_{1i} = 1)$.

### 3.1. *Restricting Potential-Assignment Subpopulations*

The conditional expectation functions (CEFs) of $Y_{1i}$ and $Y_{0i}$ given potential assignments provide a framework for the discussion of alternative identification strategies. These CEFs can be written:

$$\mathrm{E}(Y_{1i}|D_{0i}, D_{1i}) = \alpha_1 + \beta_{10}D_{0i} + \beta_{11}D_{1i} \quad (9a)$$

$$\mathrm{E}(Y_{0i}|D_{0i}, D_{1i}) = \alpha_0 + \beta_{00}D_{0i} + \beta_{01}D_{1i}. \quad (9b)$$

Equations (9*a*) and (9*b*) impose no restrictions since there are three potential-assignment subpopulations and three parameters in each CEF. The 6 conditional means, $\mathrm{E}(Y_{ji}|D_{0i}, D_{1i})$, are uniquely determined by (9*a*,*b*) as follows:

| Group | Definition | Indicator | CEF for $Y_{0i}$ | CEF for $Y_{1i}$ |
|---|---|---|---|---|
| Compliers | $D_{1i} = 1,\ D_{0i} = 0$ | $D_{1i} - D_{0i}$ | $\alpha_0 + \beta_{01}$ | $\alpha_1 + \beta_{11}$ |
| Always-takers | $D_{1i} = D_{0i} = 1$ | $D_{0i}$ | $\alpha_0 + \beta_{00} + \beta_{01}$ | $\alpha_1 + \beta_{10} + \beta_{11}$ |
| Never-takers | $D_{1i} = D_{0i} = 0$ | $1 - D_{1i}$ | $\alpha_0$ | $\alpha_1$ |

The CEF for observed outcomes, $\mathrm{E}(Y_i|D_i, Z_i)$, has a distribution with 4 points of support, while the CEFs of $Y_{0i}$ and $Y_{1i}$ given $D_{0i}$ and $D_{1i}$ depend on 6 parameters. This suggests the latter are not identified from the former without additional restrictions, a result implied by the theorem below.

THEOREM. *Suppose the Independence, First-Stage, and Monotonicity assumptions hold and that $Y_{0i}$ and $Y_{1i}$ have multinomial distributions. Let $f_0(y|D_{1i}, D_{0i})$ and $f_1(y|D_{1i}, D_{0i})$ denote the conditional distribution functions for potential outcomes given potential assignments and let $f_{YDZ}(y, d, z)$ denote the joint distribution of $Y_i$, $D_i$, and $Z_i$. Then $f_0(y|D_{1i}, D_{0i})$ and $f_1(y|D_{1i}, D_{0i})$ are not identified from $f_{YDZ}(y, d, z)$.*

*Proof.* Factor the d.f. using $f_{YDZ}(y, d, z) = f_{Y|DZ}(y|d, z)g_{DZ}(d, z)$. The second term is unrestricted. Let

$$f_j(y|D_{1i}, D_{0i}) = \alpha_j(y) + \beta_{j0}(y)D_{0i} + \beta_{j1}(y)D_{1i},$$

substitute into $f_{Y|DZ}(y, d, z)$, and iterate expectations to obtain the multinomial likelihood solely as a function of the parameters determining $f_0(y|D_{1i}, D_{0i})$ and $f_1(y|D_{1i}, D_{0i})$. Finally, substitute for $f_0(y|D_{1i}, D_{0i})$ and $f_1(y|D_{1i}, D_{0i})$ to show the likelihood is invariant to the choice of $\beta_{00}(y)$ and $\beta_{11}(y)$ as long as $\alpha_1(y) + \beta_{11}(y)$ is constant. Non-identification of $\beta_{00}(y)$ implies non-identification of the marginal distribution of $Y_{0i}$ while non-identification of $\beta_{11}(y)$ implies non-identification of the marginal distribution of $Y_{1i}$.

The multinomial distributional assumption raises the question of how general the theorem is. It seems general enough for practical purposes since, as noted by Chamberlain (1987), any distribution can be approximated arbitrarily well by a multinomial. Moreover, I would like to rule out identification based on continuity or support conditions to avoid paradoxes such as 'identification at infinity'.[6]

### 3.2. *A Menu of Restrictions*

A variety of restrictions on (9*a*,*b*) are sufficient to identify ATE. I briefly discuss four cases that strike me as being of special interest. The simplest is ignorable treatment assignment or 'no selection bias.'

RESTRICTION 1 (*No Selection Bias*).

$$\beta_{00} = \beta_{01} = \beta_{10} = \beta_{11} = 0.$$

This implies LATE $= \alpha_1 - \alpha_0 =$ ATE. Under Restriction 1, ATE can be estimated from simple treatment-control comparisons.

Because the assumption of no selection bias involves four restrictions while two would be sufficient, ATE is over-identified in this case.[7] A standard Hausman (1978) test for endogeneity exploits over-identification by comparing IV and OLS estimates, equivalent here to a comparison of Wald estimates with simple

[6] See Chamberlain (1986). The multinomial assumption has some content since it implies that potential outcomes have bounded support, so that ATE and effects on the treated are bounded. See Manski (1990) or Heckman and Vytlacil (2000).

[7] A weaker version of Restriction 1 with $\beta_{11} = \beta_{00} = 0$ is also sufficient to identify ATE since this equates never-takers with compliers for the CEF of $Y_{1i}$ and always-takers with compliers for the CEF of $Y_{0i}$. This seems no easier to motivate than Restriction 1, so I limit the discussion to the over-identified case.

treatment-control differences. A modified and potentially more powerful test can be based on the fact that under Restriction 1, $E(Y_{1i}|D_{0i} = 1) = \alpha_1$ and $E(Y_{0i}|D_{1i} = 0) = \alpha_0$. Using (8*a*,*b*), this suggests the following specification test:

*Test for Selection Bias.*

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} = [E(Y_i|D_i = 1, Z_i = 0) - E(Y_i|D_i = 0, Z_i = 1)]. \quad (T1)$$

In the Appendix, I show how a test statistic based on T1 can be computed using regression software.

The Hausman test for selection bias replaces $E(Y_i|D_i = 1, Z_i = 0) - E(Y_i|D_i = 0, Z_i = 1)$ on the right hand side of T1 with $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$. The Hausman test will also work in the causal framework outlined here since under Restriction 1 both OLS and IV estimate ATE. The difference between T1 and a Hausman test arises from the fact that the Hausman test implicitly compares $E(Y_{ji}|D_{1i} > D_{0i})$ with $E(Y_{ji}|D_i = j)$ for $j = 0,1$, while T1 implicitly compares $E(Y_{ji}|D_{1i} > D_{0i})$ with $E(Y_{ji}|D_{1i} = D_{0i} = j)$ for $j = 0,1$. These two pairs of comparisons are the same under monotonicity but not in general. The empirical results below suggest that T1, which uses monotonicity, indeed provides a more powerful specification test.[8]

While pivotal for specification testing, the assumption of no selection bias is an unattractive basis for causal inference here, since the use of IV is motivated by the possibility of selection bias. An alternative assumption that allows for selection bias amounts to the claim that the difference between $Y_{1i}$ and $Y_{0i}$ is mean-independent of potential treatment assignments. I refer to this as 'conditional constant effects'. Formally, this means:

RESTRICTION 2 (*Conditional Constant Effects*).

$$\beta_{00} = \beta_{10}; \quad \beta_{01} = \beta_{11}.$$

This pair of restrictions is just sufficient to identify ATE. In particular, we again have LATE $= \alpha_1 - \alpha_0 =$ ATE, or, equivalently, $E(Y_{1i} - Y_{0i}|D_{1i}, D_{0i}) = E(Y_{1i} - Y_{0i})$. While Restriction 2 allows for selection bias in the sense that $Y_{1i}$ and $Y_{0i}$ are correlated with potential treatment assignments, the correlation is restricted to be the same for both potential outcomes, so that the difference between $Y_{1i}$ and $Y_{0i}$ is orthogonal to potential treatment assignments.

In the *same sex* example, Restriction 2 amounts to saying that average treatment effects, while not constant, are nevertheless the same regardless of a woman's likelihood of having children. Restriction 2 rules out Roy (1951) type selection, where treatment status is determined at least in part by the gains from treatment. In the case of childbearing, for example, a woman's childbearing decision must be independent of individual-level variation in the labour-supply consequences of childbearing. On the plus side, Restriction 2 is weaker than the usual constant-

[8] Abadie (2002) develops a number of related bootstrap specification tests.

effects assumption in that it does not require a deterministic link between $Y_{1i}$ and $Y_{0i}$.[9]

A third restriction, which I call 'linearity', is appealing because it is not fundamentally inconsistent with a benchmark Roy-type selection model. The linearity condition is:

RESTRICTION 3 (*Linearity*).

$$\beta_{00} = \beta_{01}; \quad \beta_{10} = \beta_{11}.$$

In this case, the potential-outcomes CEFs can be written:

$$\mathrm{E}(Y_{1i}|D_{0i}, D_{1i}) = \alpha_1 + \beta_{11}(D_{0i} + D_{1i}) \tag{10a}$$

$$\mathrm{E}(Y_{0i}|D_{0i}, D_{1i}) = \alpha_0 + \beta_{01}(D_{0i} + D_{1i}). \tag{10b}$$

Restriction 3 requires the potential-outcomes CEF to be linear in $D_i^* \equiv D_{0i} + D_{1i}$, where $D_i^*$ is a summary measure of the desire or suitability of an individual for treatment. If the restriction is false, we can nevertheless think of (10*a*) and (10*b*) as providing a minimum mean-squared error approximation to the unrestricted model, (9*a*) and (9*b*).

To see how average causal effects are identified under Restriction 3, write the probabilities of being an always-taker and never-taker as

$$\mathrm{P}(D_{0i} = D_{1i} = 1) = \mathrm{E}(D_{0i}) = p_a$$

$$\mathrm{P}(D_{0i} = D_{1i} = 0) = \mathrm{E}(1 - D_{1i}) = p_n$$

and note that

$$\mathrm{E}(D_{0i} + D_{1i}) = 1 + (p_a - p_n).$$

Substitute into (10*a*) and (10*b*) and difference to obtain

$$\begin{aligned}\mathrm{E}(Y_{1i} - Y_{0i}) &= [(\alpha_1 + \beta_{11}) - (\alpha_0 + \beta_{01})] + (\beta_{11} - \beta_{01})(p_a - p_n) \\ &= \mathrm{E}(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}) + \{[\mathrm{E}(Y_{1i}|D_{0i} = 1) - \mathrm{E}(Y_{1i}|D_{1i} > D_{0i})] \\ &\quad -[\mathrm{E}(Y_{0i}|D_{1i} > D_{0i}) - \mathrm{E}(Y_{0i}|D_{1i} = 0)]\}(p_a - p_n). \end{aligned} \tag{11}$$

The components on the right hand side of (11) are easily estimated; details are given in the Appendix.

A calculation similar to that used to derive (11) shows that the effect of treatment on the treated can be constructed using

$$\begin{aligned}\mathrm{E}(Y_{1i} - Y_{0i}|D_i = 1) &= [(\alpha_1 + \beta_{11}) - (\alpha_0 + \beta_{01})] + (\beta_{11} - \beta_{01})(p_a/p_d) \\ &= \mathrm{E}(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}) \\ &\quad + \{[\mathrm{E}(Y_{1i}|D_{0i} = 1) - \mathrm{E}(Y_{1i}|D_{1i} > D_{0i})] \\ &\quad - [\mathrm{E}(Y_{0i}|D_{1i} > D_{0i}) - \mathrm{E}(Y_{0i}|D_{1i} = 0)]\}(p_a/p_d), \end{aligned} \tag{12}$$

[9] Note that the first part of Restriction 2 is sufficient to identify the effect of treatment on the treated, while the second part is sufficient to identify the effect of treatment on the non-treated. Although conditional constant effects is the basis of much empirical work and may be a reasonable approximation for practical purposes, as a theoretical matter this is typically implausible unless treatment is exogenous; see, e.g., Wooldridge (1997, 2003).

where $p_d$ is the probability of treatment. From (12), we can immediately derive the Bloom (1984) result that if there are no always-takers, the Wald estimator is the effect on the treated.[10]

### 3.3. *Symmetry Revisited*

Restriction 3 is closely related to the symmetry property discussed in the parametric example. To see this, note that as a consequence of linearity we can interpolate the CEF for compliers by averaging as follows:

$$\mathrm{E}(Y_{ji}|D_{1i} > D_{0i}) = [\mathrm{E}(Y_{ji}|D_{0i} = 1) + \mathrm{E}(Y_{ji}|D_{1i} = 0)]/2. \tag{13}$$

This means that expected outcomes for compliers can be obtained as the average of expected outcomes for always and never-takers. What distributional assumptions support a relation like (13)? Suppose treatment is determined by a latent-index assignment mechanism, as in (2). Then,

$$\mathrm{E}(Y_{ji}|D_{1i} = D_{0i} = 0) = \mathrm{E}(Y_{ji}|\eta_i > \gamma_0 + \gamma_1)$$

$$\mathrm{E}(Y_{ji}|D_{1i} = D_{0i} = 1) = \mathrm{E}(Y_{ji}|\eta_i < \gamma_0),$$

and

$$\mathrm{E}(Y_{ji}|D_{1i} > D_{0i}) = \mathrm{E}(Y_{ji}|\gamma_0 + \gamma_1 > \eta_i > \gamma_0).$$

If in addition, $\gamma_1 = -2\gamma_0$, then (13) holds as long as $(Y_{ji}, \eta_i)$ is jointly symmetric, as in the Normal model. The restriction $\gamma_1 = -2\gamma_0$ implies

$$\mathrm{P}(D_i = 1|Z_i = 0) = \mathrm{P}(\eta_i < \gamma_0) = 1 - p \tag{14}$$

$$\mathrm{P}(D_i = 1|Z_i = 1) = \mathrm{P}(\eta_i < -\gamma_0) = p$$

for some $p \in (0,1)$ so the first stage is also symmetric (e.g, a first stage effect of 0.1 that shifts the probability of treatment from $1 - p = 0.45$ to $p = 0.55$).

The upshot of the previous discussion is that a symmetric latent error distribution and a symmetric first stage imply the interpolating property, (13), or, equivalently, Restriction 3. Moreover, we again have LATE equals ATE since $p_a = p_n$ given the first stage described in (14).[11] Intuitively, a symmetric first-stage with symmetrically distributed latent errors equates LATE with ATE because average treatment effects for individuals with characteristics that place them in the middle of the $\eta_i$ distribution (compliers) are representative of average treatment effects for individuals over the entire distribution of $\eta_i$.

A first-stage relationship may be fortuitously symmetric, as for the 1990 Census sample of teen mothers using the *same sex* instrument. In such cases, it seems

[10] With no always-takers, we have $D_{0i} \equiv 0$, so Restriction 3 is not binding.

[11] To see this, note that $\mathrm{P}(D_i = 1|Z_i = 0) = 1 - p$ implies $p_a = 1 - p$. Since $p_a + p_n + [\mathrm{P}(D_i = 1|Z_i = 1) - \mathrm{P}(D_i = 1|Z_i = 0)] = (1 - p) + p_n + (2p - 1) = 1$, this implies $p_n = 1 - p$. As noted in the discussion of the parametric model, LATE = ATE given (14) also results when $\mathrm{E}(Y_{ji}|\eta_i)$ is an odd function and the marginal distribution of $\eta_i$ is symmetric. This makes it possible to have a relation like (13) with, say, a binary or otherwise limited dependent variable for which a symmetric distribution is implausible.

reasonable to invoke Restriction 3 and proceed under the assumption that LATE equals ATE. But what if, as seems more typical, the first stage shifts the probability of treatment asymmetrically? In the empirical Section, I describe a simple scheme for using covariates to construct a subsample with a symmetric first stage. IV should estimate average treatment effects in this specially constructed sample. This approach naturally raises the question of how to use average treatment effects for one sample to make inferences about average effects in another. For a recent attack on this question, see Hotz *et al.* (2000), who outline a procedure designed to extrapolate the results from randomised trials across sites with different populations. Here I rely on the fact that if effects differ little between two samples with and without a symmetric first-stage, then given Restriction 3, the extrapolation problem is solved under the maintained assumption that average treatment effects would be similar in the symmetric sample and its complement.

### 3.4. *Weakening Restriction 3*

Suppose again that treatment assignment can be modelled using (2) and that the potential-outcomes CEFs are linear in $\eta_i$ (as would be the case under joint Normality). Then we can write,

$$\mathrm{E}(Y_{1i}|D_{0i}, D_{1i}) = \alpha_1 + \rho_1 \mathrm{E}(\eta_i|D_{0i}, D_{1i}) \tag{15a}$$

$$\mathrm{E}(Y_{0i}|D_{0i}, D_{1i}) = \alpha_0 + \rho_0 \mathrm{E}(\eta_i|D_{0i}, D_{1i}) \tag{15b}$$

where

$$\begin{aligned}\mathrm{E}(\eta_i|D_{0i}, D_{1i}) = {} & \mathrm{E}(\eta_i|D_{1i} = 0) + [\mathrm{E}(\eta_i|D_{0i} = 1) - \mathrm{E}(\eta_i|D_{1i} = 1, D_{0i} = 0)]D_{0i} \\ & + [\mathrm{E}(\eta_i|D_{1i} = 1, D_{0i} = 0) - \mathrm{E}(\eta_i|D_{1i} = 0)]D_{1i}.\end{aligned} \tag{16}$$

Substituting (16) into (15*a*) and (15*b*) generates an expression for the coefficients in (9*a*), (9*b*). This leads to the following generalisation of Restriction 3:

RESTRICTION 4 (*Proportionality*).

$$\beta_{00} = \theta\beta_{01}; \quad \beta_{10} = \theta\beta_{11}, \quad \text{for } \theta > 0.$$

The first part of the proportionality restriction comes from (15 *a,b*) alone. Using (16), we have

$$\theta = \frac{[\mathrm{E}(\eta_i|D_{0i} = 1) - \mathrm{E}(\eta_i|D_{1i} = 1, D_{1i} = 0)]}{[\mathrm{E}(\eta_i|D_{1i} = 1, D_{0i} = 0) - \mathrm{E}(\eta_i|D_{1i} = 0)]}, \tag{17}$$

which shows why $\theta$ is positive.

Restriction 4 leads to a generalisation of the interpolation formula for average potential outcomes. In particular, we now have

$$\mathrm{E}(Y_{ji}|D_{1i} > D_{0i}) = [1/(1+\theta)]\mathrm{E}(Y_{ji}|D_{0i} = 1) + [\theta/(1+\theta)]\mathrm{E}(Y_{ji}|D_{1i} = 0), \tag{18}$$

so that if $\theta = 0$, compliers have the same expected potential outcomes as always-takers, while as $\theta$ approaches infinity, compliers have the same expected potential outcomes as never-takers.

The linearity assumption used to motivate Restriction 4 seems most plausible in the context of a model for continuous outcomes. It may be more of stretch, however, for binary outcomes such as marital status. On the other hand, without covariates the distribution of $\eta_i$ is arbitrary. We can therefore define $\eta_i$ as the latent error term in an assignment mechanism like (2), after transformation to a uniform distribution on the unit interval.[12] This guarantees that (15*a*,*b*) can generate fitted values for outcome CEFs that also fall in the unit interval. Alternately, the weighted average in (18) can be motivated directly as a natural generalisation of equally-weighted interpolation using (13).

To develop an estimator using (18), substitute Restriction 4 into (9*a*) and (9*b*) to obtain:

$$\mathrm{E}(Y_{1i}|D_{0i}, D_{1i}) = \alpha_1 + \beta_{11}(\theta D_{0i} + D_{1i})$$

$$\mathrm{E}(Y_{0i}|D_{0i}, D_{1i}) = \alpha_0 + \beta_{01}(\theta D_{0i} + D_{1i}).$$

Differencing and averaging, we have

$$\begin{aligned}\mathrm{E}(Y_{1i} - Y_{0i}) &= [(\alpha_1 + \beta_{11}) - (\alpha_0 + \beta_{01})] + (\beta_{11} - \beta_{01})(\theta p_a - p_n) \\ &= \mathrm{E}(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}) + \{\theta^{-1}[\mathrm{E}(Y_{1i}|D_{0i} = 1) \\ &\quad - \mathrm{E}(Y_{1i}|D_{1i} > D_{0i})] - [\mathrm{E}(Y_{0i}|D_{1i} > D_{0i}) - \mathrm{E}(Y_{0i}|D_{1i} = 0)]\} \\ &\quad \times (\theta p_a - p_n). \end{aligned} \tag{19}$$

We can map out the values of ATE consistent with the data by evaluating (19) for alternative choices of $\theta$. This sensitivity analysis is subject to the caveat that at the extremes where $\theta$ equals zero or infinity, ATE is not identified, a fact apparent from (18).[13]

An alternative to sensitivity analysis is to try to estimate $\theta$ using (17). Although $\theta$ is not identified without further assumptions, it clearly depends in large part on the first stage coefficients, $\gamma_0$ and $\gamma_1$. This suggests a strategy for estimating $\theta$ using information on these coefficients only. Suppose that (15*a*,*b*) holds for a latent error transformed to Uniform as discussed above, or that the CDF of $\eta_i$ can be approximated by a uniform distribution on the unit interval. Then a straightforward calculation gives

$$\theta = (\gamma_0 + \gamma_1)/(1 - \gamma_0) = \mathrm{P}(D_i = 1|Z_i = 1)/[1 - \mathrm{P}(D_i = 1|Z_i = 0)]. \tag{20}$$

This has the property that $\theta = 1$ when $\mathrm{P}(D_i = 1|Z_i = 1) = 1 - \mathrm{P}(D_i = 1|Z_i = 0)$, while capturing deviations from symmetry in a straightforward manner. The value of $\theta$ calculated using (20) in the 1990 Census sample analysed here is 0.61, close to the value calculated using Normality (0.58).

3.4.1. *Specification tests for homogeneity restrictions*

Because ATE is – by definition – invariant to the particular instrument used to estimate it, Restrictions 2, 3, and 4 can be partly checked by comparing alternative

[12] This requires that the underlying error have a continuous distribution.

[13] To compute the effect of treatment on the treated under Restriction 4, replace $\theta p_a - p_n$ with $\theta p_a / p_d$ in (19).

estimates using different instruments. In the case of Restriction 2, this amounts to a Sargan (1958) over-identification test comparing alternative IV estimates of the same structural coefficient. Under Restrictions 3 and 4, the relevant comparison should use (19) to convert estimates of LATE into estimates of ATE. A final set of specification tests is suggested by the fact that under Restrictions 3 or 4,

$$\mathrm{E}(Y_{1i} - Y_{0i}|D_{1i}, D_{0i}) = (\alpha_1 - \alpha_0) + (\beta_{11} - \beta_{01})(\theta D_{0i} + D_{1i}).$$

A test of whether $\beta_{11} - \beta_{01}$ equals zero is therefore a test of conditional constant effects, while a test of whether $\beta_{11} - \beta_{01}$ is positive is a test for Roy-type selection on the gains from treatment.

## 4. Childbearing, Marital Status, and Economic Welfare

The *same sex* instrument is a dummy for having two boys or two girls at first and second birth. Angrist and Evans (1998) showed this instrument increases the likelihood mothers with at least two children go on to have a third child by about 6–7 percentage points but is otherwise uncorrelated with mothers' demographic characteristics. The data set used here is the 1990 Census extract used in the Angrist and Evans paper. This sample includes mothers aged 21–35 with two or more children, the oldest of whom was less than 18 at the time of the Census.

Descriptive statistics are reported in Table 1 for the full sample, for a subsample of ever-married women and for four subsamples defined by mothers' education and age at first birth. The division into subsamples was motivated by earlier results showing markedly different effects of childbearing by maternal education and because of the policy interest in teen mothers. The probability of having a third child ranges from a low of 0.33 in the sample of women with some college, to a high of 0.5 in the sample of teen mothers. The probability of having a same-sex sibling pair is more or less constant at 0.505. Some of the estimates control for the demographic covariates listed in Table 1 using linear models.[14] Means for the outcome variables of interest appear at the bottom of Table 1.

### 4.1. *OLS, IV, and 2SLS Estimates*

The effect of *same sex* on the probability of having a third child varies from a low of 5.9 percentage points in the some-college sample to a high of 6.5 percentages points in the no-college sample. This can be seen in the first row of Table 2, which reports first-stage estimates. The first-stage effect without covariates, $\mathrm{E}(D_i|Z_i = 1) - \mathrm{E}(D_i|Z_i = 0) = \mathrm{E}(D_{1i} - D_{0i})$, is also an estimate of the proportion of the population in the compliers group.[15] As a benchmark, the next two rows of

[14] See Abadie (2003) and Frolich (2002) for nonlinear causal models with covariates.

[15] Although we cannot identify individual compliers in any sample and tabulate their characteristics directly, it is possible to describe the distribution of complier characteristics and to compare this to the unconditional distribution. In particular, the difference in first-stage estimates across samples defined by covariates characterises the distribution of covariates among compliers. To see this, note that for a binary covariate, $x_i$, $\mathrm{E}(x_i|D_{1i} > D_{0i})/\mathrm{E}(x_i) = \mathrm{E}(D_{1i} - D_{0i}|x_i = 1)/\mathrm{E}(D_{1i} - D_{0i})$. Table 2 therefore also shows *same sex* compliers to be less educated and more likely to have been married than the overall average.

Table 1

*Descriptive Statistics, women aged 21–35 with 2 or more children*

| Variable | Means and Standard Deviations | | | | | |
|---|---|---|---|---|---|---|
| | All women | Ever married | No college | Some college or + | Teen mothers | Adult mothers |
| *Children ever born* | 2.50 | 2.49 | 2.55 | 2.41 | 2.72 | 2.41 |
| | (0.76) | (0.75) | (0.81) | (0.68) | (0.90) | (0.68) |
| *More than 2 children* (= 1 if mother had more than 2 children) | 0.375 | 0.370 | 0.405 | 0.328 | 0.500 | 0.324 |
| | (0.484) | (0.483) | (0.491) | (0.470) | (0.500) | (0.468) |
| *Boy 1st* (= 1 if 1st child was a boy) | 0.512 | 0.512 | 0.510 | 0.514 | 0.509 | 0.513 |
| | (0.500) | (0.500) | (0.500) | (0.500) | (0.500) | (0.500) |
| *Boy 2nd* (= 1 if 2nd child was a boy) | 0.511 | 0.511 | 0.509 | 0.512 | 0.508 | 0.511 |
| | (0.500) | (0.500) | (0.500) | (0.500) | (0.500) | (0.500) |
| *Two boys* (= 1 if first two children were boys) | 0.264 | 0.264 | 0.262 | 0.266 | 0.262 | 0.264 |
| | (0.441) | (0.441) | (0.440) | (0.442) | (0.440) | (0.441) |
| *Two girls* (= 1 if first two children were girls) | 0.241 | 0.241 | 0.242 | 0.240 | 0.245 | 0.240 |
| | (0.428) | (0.427) | (0.428) | (0.427) | (0.430) | (0.427) |
| *Same sex* (= 1 if first two children were the same sex) | 0.505 | 0.505 | 0.504 | 0.506 | 0.507 | 0.504 |
| | (0.500) | (0.500) | (0.500) | (0.500) | (0.500) | (0.500) |
| *Age* | 30.4 | 30.6 | 29.92 | 31.3 | 28.8 | 31.1 |
| | (3.5) | (3.4) | (3.60) | (3.0) | (3.89) | (3.0) |
| *Age at first birth* (mother's age when first child was born) | 21.8 | 22.0 | 20.85 | 23.4 | 17.9 | 23.5 |
| | (3.5) | (3.5) | (3.16) | (3.50) | (1.13) | (2.8) |
| *Black Mother* | 0.131 | 0.092 | 0.141 | 0.115 | 0.237 | 0.088 |
| | (0.337) | (0.289) | (0.348) | (0.319) | (0.425) | (0.283) |
| *Hispanic Mother* | 0.113 | 0.112 | 0.144 | 0.065 | 0.156 | 0.096 |
| | (0.317) | (0.315) | (0.351) | (0.246) | (0.363) | (0.294) |
| *Never Married* | 0.068 | – | 0.089 | 0.036 | 0.144 | 0.037 |
| | (0.252) | – | (0.285) | (0.186) | (0.351) | (0.190) |
| *Married Now* | 0.798 | 0.857 | 0.768 | 0.846 | 0.649 | 0.859 |
| | (0.401) | (0.350) | (0.422) | (0.361) | (0.477) | (0.348) |
| *Divorced* | 0.081 | 0.087 | 0.083 | 0.079 | 0.122 | 0.065 |
| | (0.273) | (0.282) | (0.276) | (0.270) | (0.327) | (0.246) |
| *Divorced or Separated* | 0.127 | 0.137 | 0.136 | 0.114 | 0.197 | 0.099 |
| | (0.333) | (0.344) | (0.343) | (0.317) | (0.398) | (0.299) |
| *High School Graduate* (= 1 if high school diploma and no further education) | 0.420 | 0.423 | 0.685 | – | 0.426 | 0.417 |
| | (0.494) | (0.494) | (0.465) | – | (0.495) | (0.493) |
| *Some College* (= 1 if some college, but no degree) | 0.264 | 0.269 | – | 0.682 | 0.174 | 0.301 |
| | (0.441) | (0.444) | – | (0.466) | (0.379) | (0.458) |
| *College Graduate* (= 1 if bachelor's degree or higher) | 0.123 | 0.131 | – | 0.318 | 0.018 | 0.166 |
| | (0.329) | (0.338) | – | (0.466) | (0.131) | (0.372) |
| *In Poverty* (= 1 if family income below the poverty line) | 0.197 | 0.158 | 0.256 | 0.103 | 0.347 | 0.136 |
| | (0.398) | (0.364) | (0.437) | (0.303) | (0.476) | (0.343) |
| *Welfare Recipient* (= 1 if public assistance income > 0) | 0.098 | 0.065 | 0.130 | 0.047 | 0.187 | 0.062 |
| | (0.297) | (0.247) | (0.336) | (0.212) | (0.390) | (0.241) |
| *Worked for pay* (= 1 if worked for pay in 1989) | 0.662 | 0.674 | 0.623 | 0.723 | 0.650 | 0.667 |
| | (0.473) | (0.469) | (0.485) | (0.447) | (0.477) | (0.471) |
| *Weeks worked* (weeks worked in 1989) | 26.2 | 26.9 | 24.27 | 29.3 | 25.0 | 26.7 |
| | (22.9) | (22.9) | (23.00) | (22.4) | (22.81) | (22.9) |
| Number of observations | 380,007 | 357,063 | 236,418 | 143,589 | 110,156 | 269,851 |

*Notes*: Data are from the 1990 PUMS. The sample includes women with 2 or more children whose 2nd child was at least age 1 and who had their first birth at age 15 or later. The no-college sample includes women with no college or with an associate occupational degree. The some-college sample includes women with an associate academic degree, some college but no degree, or a college degree. Teen mothers are those who had their first birth at age 19 or younger. Standard deviations are reported in parentheses. All calculations use sample weights.

Table 2

*First-stage and Labour Supply Estimates*

| | All women | | Ever married | | No college | | Some college or + | | Teen mothers | | Adult mothers | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | OLS | IV | OLS | IV | OLS | IV | OLS | IV | OLS | IV | OLS | IV |
| (*a*) No covariates | | | | | | | | | | | | |
| First Stage | | | | | | | | | | | | |
| *More than 2 children* | 0.0628 | – | 0.0663 | – | 0.0652 | – | 0.0594 | – | 0.0638 | – | 0.0616 | – |
| | (0.002) | – | (0.002) | – | (0.002) | – | (0.003) | – | (0.003) | – | (0.024) | – |
| Outcomes | | | | | | | | | | | | |
| *Worked for pay* | −0.132 | −0.084 | −0.126 | −0.083 | −0.132 | −0.105 | −0.112 | −0.057 | −0.140 | −0.026 | −0.130 | −0.109 |
| | (0.002) | (0.027) | (0.002) | (0.026) | (0.002) | (0.034) | (0.003) | (0.044) | (0.003) | (0.051) | (0.002) | (0.032) |
| *Weeks worked* | −7.34 | −5.15 | −7.12 | −5.09 | −7.22 | −6.52 | −6.59 | −3.21 | −7.47 | −4.76 | −7.19 | −5.26 |
| | (0.08) | (1.30) | (0.09) | (1.27) | (0.11) | (1.60) | (0.14) | (2.18) | (0.15) | (2.40) | (0.10) | (1.57) |
| (*b*) With covariates | | | | | | | | | | | | |
| First Stage | | | | | | | | | | | | |
| *More than 2 children* | 0.0623 | – | 0.0658 | – | 0.0644 | – | 0.0592 | – | 0.0633 | – | 0.0623 | – |
| | (0.0017) | – | (0.0017) | – | (0.0022) | – | (0.0027) | – | (0.0033) | – | (0.0019) | – |
| Outcomes | | | | | | | | | | | | |
| *Worked for pay* | −0.148 | −0.097 | −0.148 | −0.097 | −0.144 | −0.120 | −0.151 | −0.060 | −0.132 | −0.071 | −0.154 | −0.108 |
| | (0.002) | (0.027) | (0.002) | (0.026) | (0.002) | (0.034) | (0.003) | (0.043) | (0.003) | (0.049) | (0.002) | (0.031) |
| *Weeks worked* | −8.33 | −5.93 | −8.43 | −5.92 | −7.92 | −7.43 | −8.84 | −3.45 | −7.51 | −7.55 | −8.62 | −5.30 |
| | (0.09) | (1.27) | (0.09) | (1.24) | (0.11) | (1.57) | (0.14) | (2.15) | (0.15) | (2.28) | (0.10) | (1.52) |

*Notes*: The Table reports OLS and IV estimates of the coefficient on the *More than 2 children variable* . The IV estimates use *Same sex* as an instrument. The covariates included in panel (*b*) are *Age, Age at first birth,* and dummies for *Boy 1st, Boy 2nd, Black, Hispanic, Other race, High school graduate, Some college* and *College graduate.* The samples are the same as in Table 1. Standard errors are reported in parentheses. All calculations use sample weights.

Table 2 show estimates of the effect of childbearing on two of the labour supply variables studied by Angrist and Evans (1998). These are IV and OLS estimates from models without covariates, i.e., Wald estimates and simple treatment-control contrasts.

The Wald (IV) estimates of the effect of a third child on employment status and weeks worked suggest mothers reduced their labour supply as a consequence of childbearing, though not by as much as indicated by the OLS estimates. For example, women who had a third child were about 13 percentage points less likely to work, but the corresponding IV estimate suggests a causal effect of only 8 percentage points. The OLS and IV estimates for weeks worked are about −7 and −5. The IV estimates of labour supply effects are larger for less-educated women than for those with some college; in fact, the labour supply estimates are not significant in the some-college sample. In contrast, the IV estimates are smaller for women who had their first birth as teenager than for women who had their first birth as an adult.

The last two rows in Table 2 show first-stage, OLS and two-stage least squares (2SLS) estimates after adding controls for age, age at first birth, dummies to indicate first-born and second-born boys, race dummies, and dummies for three schooling groups. Since *same sex* is uncorrelated with these covariates, including them has little effect on the 2SLS estimates. Moreover, in spite of the fact that some of the covariates are good predictors of outcomes, estimates with covariates are only slightly more precise than those without. Perhaps more surprisingly, the OLS estimates of labour supply effects also change little in response to the addition of covariates.

Estimates of the effect of having a third child on marital status, poverty status, and welfare use are reported in Table 3 for models with and without covariates. In the sample of all women, those with more children are less likely to be married. But this is at least in part due to uncontrolled demographic factors such as age at first birth, since OLS estimates with controls show that additional childbearing is associated with an increase in the likelihood of being married. In contrast to the OLS estimates, IV estimates with or without covariates suggest that the causal effect of childbearing is a reduced probability of being married. Thus, an important finding is that when the effect of childbearing is estimated in models with demographic controls, IV and OLS estimates have opposite signs.

The most important change in marital status caused by childbearing appears to be an increase in the likelihood of being divorced or separated. The estimated effects of childbearing on the probability of being ever-married or divorced (but not separated) are not significantly different from zero. Consistent with an increase in marital breakup, the birth of a third child also appears to lead to a marked increase in the likelihood a woman lives in a family with total family income below the poverty line. Here we should expect at least a mechanical effect since the poverty threshold falls as family size increases. Although OLS estimates are larger than IV estimates in models without covariates, OLS and IV estimates in models with covariates both indicate that a third child increases the likelihood a woman is poor by 9–10 percentage points.

Given the elevated rates of marital breakup and the increase in poverty rates that appear to be caused by childbearing, it seems reasonable to expect that the birth of

Table 3

*Effects on Marital Status, Poverty Status and Welfare Use*

| | All women | | Ever married | | No college | | Some college or + | | Teen mothers | | Adult mothers | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Outcomes | OLS | IV | OLS | IV | OLS | IV | OLS | IV | OLS | IV | OLS | IV |
| (*a*) No covariates | | | | | | | | | | | | |
| *Ever Married* | −0.021 | −0.010 | – | – | −0.026 | −0.023 | −0.002 | 0.007 | −0.016 | −0.010 | 0.0004 | −0.003 |
| | (0.001) | (0.153) | – | – | (0.001) | (0.021) | (0.001) | (0.020) | (0.002) | (0.0391) | (0.0009) | (0.014) |
| *Married Now* | −0.025 | −0.062 | −0.0075 | −0.055 | −0.035 | −0.082 | 0.008 | −0.035 | −0.015 | −0.066 | 0.018 | −0.048 |
| | (0.002) | (0.024) | (0.001) | (0.020) | (0.002) | (0.030) | (0.002) | (0.036) | (0.003) | (0.051) | (0.002) | (0.025) |
| *Divorced* | −0.013 | 0.011 | −0.012 | 0.012 | −0.012 | 0.0044 | −0.017 | 0.022 | −0.024 | −0.010 | −0.022 | 0.016 |
| | (0.001) | (0.016) | (0.001) | (0.016) | (0.001) | (0.0195) | (0.002) | (0.027) | (0.002) | (0.035) | (0.001) | (0.017) |
| *Divorced or Separated* | 0.0023 | 0.053 | 0.0056 | 0.055 | 0.0070 | 0.057 | −0.011 | 0.046 | −0.0030 | 0.048 | −0.018 | 0.049 |
| | (0.0013) | (0.019) | (0.0014) | (0.020) | (0.0016) | (0.024) | (0.002) | (0.032) | (0.0027) | (0.043) | (0.001) | (0.021) |
| *In Poverty* | 0.143 | 0.095 | 0.124 | 0.082 | 0.167 | 0.107 | 0.070 | 0.088 | 0.178 | 0.143 | 0.083 | 0.062 |
| | (0.002) | (0.029) | (0.002) | (0.020) | (0.002) | (0.031) | (0.002) | (0.030) | (0.003) | (0.050) | (0.002) | (0.024) |
| *Welfare Recipient* | 0.067 | 0.033 | 0.050 | 0.032 | 0.079 | 0.028 | 0.030 | 0.049 | 0.091 | 0.018 | 0.030 | 0.032 |
| | (0.001) | (0.018) | (0.001) | (0.014) | (0.002) | (0.024) | (0.002) | (0.022) | (0.003) | (0.042) | (0.001) | (0.017) |
| (*b*) With covariates | | | | | | | | | | | | |
| *Ever Married* | 0.0026 | −0.0051 | – | – | 0.0038 | −0.025 | 0.0061 | 0.023 | 0.0087 | −0.031 | 0.0043 | 0.0025 |
| | (0.0009) | (0.0136) | – | – | (0.0013) | (0.019) | (0.0012) | (0.018) | (0.0022) | −0.034 | (0.0009) | (0.0126) |
| *Married Now* | 0.037 | −0.052 | 0.037 | −0.046 | 0.028 | −0.080 | 0.057 | −0.011 | 0.020 | −0.078 | 0.047 | −0.041 |
| | (0.001) | (0.021) | (0.001) | (0.019) | (0.002) | (0.028) | (0.002) | (0.033) | (0.003) | (0.047) | (0.002) | (0.023) |
| *Divorced* | −0.037 | 0.0071 | −0.039 | 0.0067 | −0.029 | 0.0010 | −0.048 | 0.018 | −0.026 | −0.021 | −0.040 | 0.017 |
| | (0.001) | (0.0157) | (0.001) | (0.0159) | (0.001) | (0.0196) | (0.002) | (0.026) | (0.002) | (0.035) | (0.001) | (0.017) |
| *Divorced or Separated* | −0.033 | 0.048 | −0.036 | 0.047 | −0.023 | 0.054 | −0.049 | 0.038 | −0.013 | 0.040 | −0.041 | 0.048 |
| | (0.001) | (0.019) | (0.001) | (0.019) | (0.002) | (0.025) | (0.002) | (0.031) | (0.003) | (0.043) | (0.001) | (0.020) |
| *In Poverty* | 0.093 | 0.097 | 0.087 | 0.085 | 0.113 | 0.113 | 0.055 | 0.074 | 0.148 | 0.186 | 0.065 | 0.061 |
| | (0.001) | (0.021) | (0.001) | (0.019) | (0.002) | (0.028) | (0.002) | (0.029) | (0.003) | (0.047) | (0.002) | (0.022) |
| *Welfare Recipient* | 0.039 | 0.033 | 0.031 | 0.032 | 0.048 | 0.032 | 0.020 | 0.041 | 0.071 | 0.043 | 0.022 | 0.030 |
| | (0.001) | (0.017) | (0.001) | (0.014) | (0.002) | (0.023) | (0.001) | (0.021) | (0.003) | (0.040) | (0.001) | (0.016) |

*Notes*: The samples and models are as in Table 2, with different dependent variables. Standard errors are reported in parentheses. All calculations use sample weights.

a third child also increases the likelihood a woman is on welfare. Both the IV and OLS estimates tend to support this, though the IV estimates are imprecise. The OLS estimate of the effect on welfare use range from 6.7 percentage points without covariates to 3.9 percentage points with covariates. The IV estimate is a marginally significant 3.3% with or without covariates. While small in levels, an effect of this magnitude represents a roughly one-third increase in the number of women on welfare.

The IV estimates show no relationship between childbearing and the probability a woman has ever been married, so estimates limited to the sample of ever-married women are unlikely to be affected by selection bias. Not surprisingly, therefore, the IV estimates in the sample of ever-married women are almost identical to those in the full sample. On the other hand, while the IV estimate of the reduction in marriage rates is a significant 8 percentage points (s.e. = 0.028) for women with no college, it is close to zero and insignificant for women with some college. The effects of childbearing on poverty are also larger in the no-college sample, though the difference in effects on welfare use by college status is reversed and much smaller than the difference in effects on poverty rates.

The difference in estimates by mothers' age at first birth also suggest a pattern of larger effects with decreasing socioeconomic status, though the contrast is not as clear cut as the differences by schooling group. While the increases in marital dissolution and welfare receipt are larger for teen mothers than for adult mothers, the estimates are significant only in the latter group. Estimates of effects on divorce/separation are similar in the two groups, though again much more precise for the sample of adult mothers. This difference in precision undoubtedly reflects the smaller sample of teen mothers. One clear contrast, however, is the higher likelihood that a third birth pushes a teen mother into poverty. The impact on poverty status is significant regardless of mothers' age at first birth but it is roughly three times larger for teen mothers.

### 4.2. *Heterogeneity across Potential-assignment Subpopulations*

The first-stage estimates imply that 6–7% of each sample consists of compliers, i.e., mothers who had a child in response to a homogenous sibling-sex mix. Because the overall probability of treatment ranges upwards from about 0.32, the overwhelming majority of treated individuals are always-takers. This can be seen in Table 4, which gives the distribution of potential-assignment subpopulations. In the sample of all women, for example, 6.3% are compliers, 34% are always-takers (i.e., have a third child without regard to sibling-sex composition) and 59% are never-takers (i.e., would never have a third child regardless of sibling-sex composition). The proportion of treated who are compliers is $1 - (p_a/p_d)$, or about 8%. Given the relatively small proportion of compliers, the scope for differences in average causal effects across potential-assignment subpopulation is substantial.

Table 4 also reports the estimate of $(p_a - p_n)$, the multiplier that determines how far LATE is from ATE when the latter is calculated using Restriction 3 and (11), or in models with covariates as described in the Appendix. The estimate of

Table 4

*Potential-Assignment Subpopulations*

| | | No Covariates | | | | | With Covariates | |
|---|---|---|---|---|---|---|---|---|
| Sample | $P(D = 1)$ (1) | $p_c$ (2) | $p_a$ (3) | $p_n$ (4) | $p_a - p_n$ (5) | $\theta$ (6) | $p_c$ (7) | $p_a - p_n$ (8) |
| All Women | 0.375 | 0.063 | 0.344 | 0.594 | −0.250 | 0.619 | 0.062 | −0.250 |
| | | (0.0018) | | | (0.0018) | | (0.0017) | (0.0018) |
| Ever Married | 0.370 | 0.066 | 0.337 | 0.597 | −0.261 | 0.607 | 0.066 | −0.260 |
| | | (0.0018) | | | (0.0018) | | (0.0017) | (0.0018) |
| No College | 0.405 | 0.065 | 0.372 | 0.563 | −0.191 | 0.696 | 0.064 | −0.191 |
| | | (0.0023) | | | (0.0023) | | (0.0022) | (0.0023) |
| Some College | 0.328 | 0.059 | 0.298 | 0.642 | −0.344 | 0.510 | 0.059 | −0.344 |
| | | (0.0027) | | | (0.0027) | | (0.0027) | (0.0027) |
| Teen Mothers | 0.500 | 0.064 | 0.468 | 0.468 | −0.0006 | 0.999 | 0.063 | −0.0005 |
| | | (0.0034) | | | (0.0034) | | (0.0033) | (0.0034) |
| Adult Mothers | 0.324 | 0.062 | 0.293 | 0.645 | −0.352 | 0.502 | 0.062 | −0.351 |
| | | (0.0020) | | | (0.0020) | | (0.0019) | (0.0020) |

*Notes*: The first column reports the proportion treated. The second column shows the proportion of compliers in the sample, which is given by the first-stage effect of *same sex*. The estimates of the proportion of always-takers and never-takers and the parameter $\theta$ were calculated as described in the text. Estimates with covariates were calculated as described in the Appendix. Standard errors are reported in parentheses. All calculations use sample weights.

$(p_a - p_n)$ is −0.25 in the full sample, and ranges from 0 for teen mothers to −0.352 in the sample of adult mothers.

4.2.1. *Symmetric subpopulations*

The value of zero for $(p_a - p_n)$ in the teen mother sample is noteworthy because it means that LATE is the same as ATE under Restriction 3. This is a consequence of the fact that the first stage for teen mothers is almost perfectly symmetric: the *same sex* instrument shifts the probability of further childbearing from about 0.47 to 0.53. Moreover, because $\theta$ for teen mothers is about 1 when estimated using (20), estimates of ATE for teen mothers under Restriction 4 are also close to LATE.

The first two columns of Table 5 focus on the comparison between estimates for all women and teen mothers only, repeating earlier estimates for these samples from Table 3 without covariates, including the first-stage coefficient and intercept. For the most part, IV estimates for teen mothers are similar to those for the sample of all women. While the estimated effect on employment is considerably lower at −0.026 (s.e. = 0.051) versus −0.084 (s.e. = 0.027) in the full sample, the effect of childbearing on weeks worked is −5.2 (s.e. = 1.3) in the full sample and −4.8 (s.e. = 2.4) for teen mothers. Similarly, the effect on marital status is −0.062 (s.e. = 0.024) in the full sample and −0.066 for teen mothers (0.051). Note that we can view the parameters estimated in the full sample as estimates of $\mathrm{E}(Y_{1i} - Y_{0i} | D_{1i} > D_{0i}, X = \textit{all women})$, while the estimates in column 2, for teen mothers, can be interpreted as measuring $\mathrm{E}(Y_{1i} - Y_{0i} | X = \textit{teen mothers})$ under Restriction 3 or 4. A test of equality across columns 1 and 2 is therefore a joint test of the invariance of average treatment effects to conditioning both on $X$ and on the compliers potential-outcomes subpopulation. The fact that these are similar is

Table 5

*Symmetric First Stage Samples*

| Variables | All Women (1) | Teen Mothers (2) | Symmetric sample I: $\pi_0(X) \geq 0.4$ & $\pi_0(X) \leq 0.6$ (3) | Symmetric sample I: $\pi_0(X) < 0.4$ or $\pi_0(X) > 0.6$ (4) | Symmetric sample II: $\pi_0(X) \geq 0.35$ (5) | Symmetric sample II: $\pi_0(X) < 0.35$ (6) |
|---|---|---|---|---|---|---|
| First Stage (OLS estimates) | | | | | | |
| *Coefficient* | 0.063 | 0.064 | 0.071 | 0.059 | 0.068 | 0.058 |
| | (0.002) | (0.003) | (0.003) | (0.002) | (0.003) | (0.002) |
| *Constant* | 0.344 | 0.468 | 0.471 | 0.296 | 0.465 | 0.253 |
| | (0.001) | (0.002) | (0.002) | (0.001) | (0.002) | (0.001) |
| Outcomes (IV Estimates) | | | | | | |
| *Worked for pay* | −0.084 | −0.026 | −0.038 | −0.109 | −0.080 | −0.092 |
| | (0.027) | (0.051) | (0.045) | (0.034) | (0.038) | (0.039) |
| *Weeks worked* | −5.15 | −4.76 | −3.72 | −6.03 | −5.90 | −4.71 |
| | (1.30) | (2.40) | (2.21) | (1.62) | (1.83) | (1.86) |
| *Ever Married* | −0.010 | −0.0098 | −0.016 | −0.0031 | −0.0051 | −0.0092 |
| | (0.015) | (0.0391) | (0.032) | (0.0170) | (0.0267) | (0.0162) |
| *Married Now* | −0.062 | −0.066 | −0.033 | −0.066 | −0.075 | −0.039 |
| | (0.024) | (0.051) | (0.045) | (0.027) | (0.038) | (0.028) |
| *Divorced* | 0.011 | −0.010 | −0.029 | 0.025 | 0.012 | 0.0057 |
| | (0.016) | (0.035) | (0.031) | (0.018) | (0.026) | (0.0189) |
| *Divorced or Separated* | 0.053 | 0.048 | 0.020 | 0.063 | 0.068 | 0.033 |
| | (0.019) | (0.043) | (0.038) | (0.022) | (0.0032) | (0.023) |
| *In Poverty* | 0.095 | 0.143 | 0.095 | 0.087 | 0.136 | 0.048 |
| | (0.023) | (0.050) | (0.044) | (0.027) | (0.036) | (0.028) |
| *Welfare Recipient* | 0.033 | 0.018 | 0.021 | 0.034 | 0.027 | 0.032 |
| | (0.018) | (0.042) | (0.035) | (0.020) | (0.029) | (0.020) |
| Number of Observations | 380,007 | 110,156 | 103,803 | 276,204 | 162,264 | 217,743 |

*Notes*: Columns 1 and 2 repeat estimates from Tables 2 and 3, for models without covariates. Estimates using samples with a symmetric first stage are reported in columns 3 and 5. Estimates for complementary samples are reported in columns 4 and 6. Standard errors are shown in parentheses. All calculations use sample weights.

evidence against substantial treatment effect heterogeneity in both dimensions, though of course there are scenarios where this test has no power.[16]

There is some evidence for a difference in effects on poverty status between the teen mother and all-women samples. For all women, the IV estimate of the effect of childbearing on poverty status is 0.095 (s.e. = 0.023), while the corresponding estimate is 0.143 (s.e. = 0.05) in the teen mother sample. The comparison across samples is weakened, however, by the fact the estimates in the teen mother sample are much less precise than in the full sample. This raises the question of whether we can construct a larger sample with a symmetric first stage. I attempted to construct such a sample by estimating a Probit first-stage allowing interactions with covariates and then selecting the sample based on covariate-specific fitted values.[17]

[16] As with an over-identification test, the power of the test turns on maintaining the validity of a benchmark. Here, we maintain $E(Y_{1i} - Y_{0i}|X = \textit{teen mothers}) = E(Y_{1i} - Y_{0i})$.

[17] A maintained assumption here is that the distribution of $Y_{ji}$ and $\eta_i$ is jointly symmetric conditional on the covariates used to select the sample with a symmetric first stage.

The details of the symmetric sample selection are as follows. The idea is to use a parametric model to capture the variation in the first-stage effect of *same sex* on childbearing with demographic covariates. The model allows for a large set of interaction terms with covariates. I then look for covariate values where the predicted first-stage effect is symmetric in the sense required by Restriction 3. I began with a Probit first-stage equation:

$$\mathrm{P}(D_i = 1 | Z_i, X_i) = \Phi[\boldsymbol{\kappa}_0' \mathbf{X}_i + (\boldsymbol{\kappa}_1' \mathbf{X}_i)\mathbf{Z}_i], \tag{21}$$

where $\mathbf{X}_i$ is a vector of covariates that includes age, age at first birth, Black and Hispanic dummies, and dummies indicating women with some college and college graduates. The main effects, $\boldsymbol{\kappa}_0' \mathbf{X}_i$, and interaction terms, $\boldsymbol{\kappa}_1' \mathbf{X}_i$, use the same parameterisation of covariate effects (in particular, they both allow for linear terms in the age variables plus main effects for the dummies). In practice, $\boldsymbol{\kappa}_0' \mathbf{X}_i$ takes on about 1,700 distinct values. For each of these values, I calculated

$$\hat{\pi}_0(\mathbf{X}_i) \equiv \Phi(\hat{\kappa}_0' \mathbf{X}_i),$$

the distribution of which is plotted in Figure 3. This gives the distribution of the probability of childbearing for women with different *X*-characteristics and $\mathbf{Z}_i$ equal to zero, i.e. the probability of being an always-taker. The distribution of $\hat{\pi}_0(\mathbf{X}_i)$ is concentrated around the overall average of about 0.34, though there is considerable spread.

By definition, a symmetric first stage shifts the probability of treatment across the value of one-half. To identify a sample where this is most likely, I initially selected women with $\hat{\pi}_0(\mathbf{X}_i)$ between 0.4 and 0.6. Column 3 of Table 5 reports estimates for this sample, which has about 104,000 observations. The estimated first-stage in this sample shifts the probability of treatment from 0.47 to 0.54, i.e., approximately from $p$ to $1 - p$, as required by symmetry. For most outcomes, the IV estimates in this symmetric sample are smaller in absolute value than in the full sample and smaller than in the sample complementary to the symmetric sample, for which results are reported in column 4. For example, the estimated effect on weeks worked in the symmetric sample is −3.7 (s.e. = 2.2), while the corresponding estimate in the complementary sample is −6 (s.e. = 1.6). Again, however, the comparison is handicapped by a lack of precision.

The long right tail of the distribution of first-stage base values plotted in Figure 3 suggests that an even larger symmetric sample can be constructed simply by dropping values of $\hat{\pi}_0(\mathbf{X}_i)$ beginning from the left and working up. As it turns out, limiting the sample to individuals with values of $\hat{\pi}_0(\mathbf{X}_i)$ greater than or equal to 0.35 leads to a first stage that shifts the probability of treatment from 0.465 to 0.533, virtually perfectly symmetric. This can be seen in column 5 of Table 5, which reports first-stage and IV estimates for the resulting sample of 162,264 observations. Most of the estimates in this symmetric sample are close to those in the full sample. For example, the effect on weeks worked is −5.9 (s.e. = 1.8) and the effect on divorce or separation is 0.068 (s.e. = 0.032). Perhaps surprisingly, the estimated effect on poverty status differs markedly between this sample and its complement (0.136 versus 0.048) but the estimated effect is still significantly different from zero in the complementary sample.

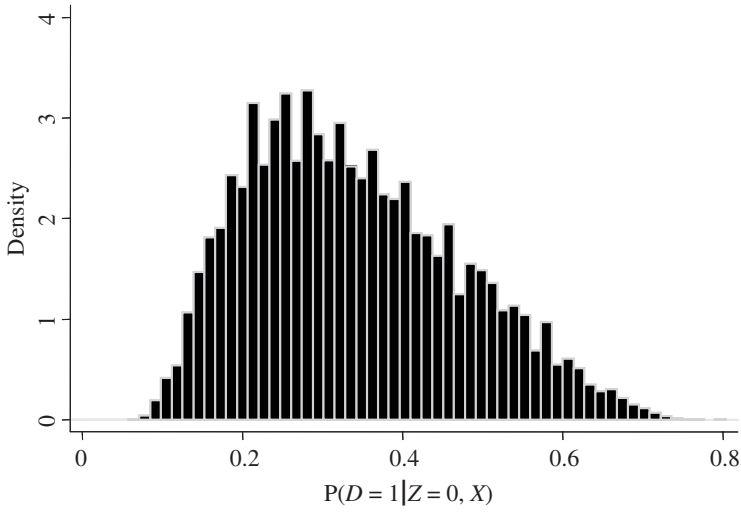


Fig. 3. *Distribution of First-stage Base Probability as a Function of Covariates.* The covariates are age, age at first birth, Black and Hispanic dummies for some college and college graduates. There are about 1,700 values in the histogram

4.2.2. *Imputation of ATE*

The results in Table 5 reflect an attempt to identify or construct samples where LATE = ATE. Alternately, we can use (11) or (19) to impute a value of ATE for the various subsamples analysed in Table 3. The results of this effort are presented in Table 6 for four outcomes; this Table also reports the no-selection alternative used to construct the specification test discussed at the beginning of Section 3.2. The estimates of the no-selection alternative are all slightly farther from the estimates of LATE than the corresponding OLS estimates. For example, the OLS estimate of the effect on weeks worked in the full sample is −7.34 (s.e. = 0.08), while the no-selection alternative is −7.56 (s.e. = 0.12). This suggests, as noted earlier, that the contrast between IV and the no-selection alternative provides a more powerful specification test than a conventional IV/OLS comparison.

Estimates of ATE constructed using (11) for the effect of childbearing on weeks worked are similar to the estimates of LATE, even in samples where the first-stage is not symmetric. For example, the estimate of ATE for the sample of non-teen (i.e., adult) mothers is −4.1 (s.e. = 1.5), in comparison with an estimate of LATE of −5.3 (s.e. = 1.6). Using the estimates of $\theta$ shown in Table 4 and (19) generates somewhat smaller estimates for the effect on weeks worked other than in the teen mother sample, though again mostly still significant.

Estimates of ATE for outcomes other than weeks worked are mostly insignificantly different from zero. This contrasts with the mostly significant estimates of LATE. Again, this is partly a problem of precision. But the estimates of ATE outside the teen mother sample move substantially closer to zero than the estimates of LATE. For example, while LATE suggests the probability of divorce or separation increases by 0.053 (s.e. = 0.019), the corresponding estimates of ATE are 0.028 (s.e. = 0.019) when $\theta$ equals one and 0.009 (s.e. = 0.021) when $\theta$ is

Table 6

*Imputation of ATE*

| Outcome | Sample | OLS (1) | No Selection Alternative (2) | LATE (3) | $\theta = 1$ (4) | ATE Estimated $\theta$ (5) |
|---|---|---|---|---|---|---|
| *Weeks Worked* | All Women | −7.34 | −7.56 | −5.15 | −4.31 | −3.19 |
| | | (0.08) | (0.12) | (1.30) | (1.27) | (1.45) |
| | Ever Married | −7.12 | −7.33 | −5.09 | −4.41 | −3.45 |
| | | (0.09) | (0.13) | (1.27) | (1.23) | (1.43) |
| | No College | −7.22 | −7.33 | −6.52 | −5.73 | −4.94 |
| | | (0.11) | (0.16) | (1.60) | (1.57) | (1.70) |
| | Some College or + | −6.59 | −6.90 | −3.21 | −2.38 | −0.60 |
| | | (0.14) | (0.21) | (2.18) | (2.08) | (2.69) |
| | Teen Mothers | −7.47 | −7.66 | −4.76 | −4.76 | −4.76 |
| | | (0.15) | (0.22) | (2.40) | (2.40) | (2.40) |
| | Adult Mothers | −7.19 | −7.43 | −5.26 | −4.06 | −2.19 |
| | | (0.10) | (0.15) | (1.57) | (1.48) | (1.91) |
| *Divorced or Separated* | All Women | 0.0023 | 0.0005 | 0.053 | 0.028 | 0.0092 |
| | | (0.0013) | (0.0019) | (0.019) | (0.019) | (0.0216) |
| | Ever Married | 0.0056 | 0.0043 | 0.055 | 0.024 | −0.0002 |
| | | (0.0014) | (0.0020) | (0.020) | (0.019) | (0.0221) |
| | No College | 0.0070 | 0.0053 | 0.057 | 0.032 | 0.011 |
| | | (0.0016) | (0.0024) | (0.024) | (0.024) | (0.026) |
| | Some College or + | −0.011 | −0.014 | 0.046 | 0.034 | 0.034 |
| | | (0.002) | (0.003) | (0.032) | (0.029) | (0.037) |
| | Teen Mothers | −0.0030 | −0.0063 | 0.048 | 0.048 | 0.048 |
| | | (0.0027) | (0.0040) | (0.043) | (0.043) | (0.043) |
| | Adult Mothers | −0.018 | −0.021 | 0.049 | 0.017 | −0.0024 |
| | | (0.001) | (0.002) | (0.021) | (0.019) | (0.024) |
| *In Poverty* | All Women | 0.143 | 0.150 | 0.095 | 0.049 | −0.0023 |
| | | (0.002) | (0.002) | (0.023) | (0.024) | (0.029) |
| | Ever Married | 0.124 | 0.129 | 0.082 | 0.054 | 0.020 |
| | | (0.002) | (0.002) | (0.020) | (0.021) | (0.026) |
| | No College | 0.167 | 0.175 | 0.107 | 0.061 | 0.014 |
| | | (0.002) | (0.003) | (0.031) | (0.032) | (0.036) |
| | Some College or + | 0.070 | 0.071 | 0.088 | 0.059 | 0.031 |
| | | (0.002) | (0.003) | (0.030) | (0.031) | (0.042) |
| | Teen Mothers | 0.178 | 0.181 | 0.143 | 0.143 | 0.143 |
| | | (0.003) | (0.005) | (0.050) | (0.051) | (0.051) |
| | Adult Mothers | 0.083 | 0.088 | 0.062 | 0.017 | −0.039 |
| | | (0.002) | (0.003) | (0.024) | (0.025) | (0.033) |
| *Welfare Recipient* | All Women | 0.067 | 0.072 | 0.033 | 0.0058 | −0.026 |
| | | (0.001) | (0.002) | (0.018) | (0.0185) | (0.022) |
| | Ever Married | 0.050 | 0.052 | 0.032 | 0.015 | −0.0051 |
| | | (0.001) | (0.002) | (0.014) | (0.015) | (0.0182) |
| | No College | 0.079 | 0.085 | 0.028 | −0.001 | −0.031 |
| | | (0.002) | (0.003) | (0.024) | (0.025) | (0.029) |
| | Some College or + | 0.030 | 0.030 | 0.049 | 0.037 | 0.029 |
| | | (0.002) | (0.002) | (0.022) | (0.022) | (0.030) |
| | Teen Mothers | 0.091 | 0.096 | 0.018 | 0.018 | 0.018 |
| | | (0.003) | (0.004) | (0.042) | (0.043) | (0.043) |
| | Adult Mothers | 0.030 | 0.032 | 0.032 | 0.006 | −0.024 |
| | | (0.001) | (0.002) | (0.017) | (0.017) | (0.023) |

*Notes*: Columns 1 and 3 repeat estimates from Tables 2 and 3. Column 2 shows the no-selection alternative under Restriction 1 and for the selection-bias test. Column 4 reportes estimates of ATE under Restriction 3 and column 5 reports estimates of ATE under Restriction 4.

estimated using (20). The evidence that further childbearing increases divorce or separation for the typical woman with two children is therefore weaker than the estimates of LATE would suggest. Except for the sample of teen mothers, the estimates of ATE for effects on poverty status are also smaller than the corresponding estimates of LATE.

## 5. Summary and Conclusions

The framework outlined here provides a strategy for modelling treatment effect heterogeneity across potential-assignment subpopulations. I focused initially on restrictions that make IV estimates of causal effects on compliers representative of the overall population average treatment effect. This framework also leads to procedures that can be used to impute average treatment effects from information on average outcomes for compliers, always-takers, and never-takers. An illustration of these ideas using *same sex* instruments suggests this approach may be useful in applied work.

On the empirical side, estimates of LATE for teen mothers are close to the corresponding average treatment effects for this population, when the latter are inferred using linearity or proportionality assumptions. While estimates of the overall average effect of childbearing are smaller than the corresponding IV estimates, most of the estimated effects on labour supply and poverty status remain substantial and significant. On the other hand, most (though not all) of the estimated average effects on marital status and welfare use are small and insignificant.

Estimates of the effects of childbearing on marital stability and welfare participation using the *same sex* instrument suggest the outline of a coherent picture but many features remain unresolved. In this application, the theory of parameter heterogeneity runs quickly into the sandpile of sampling variance and specification uncertainty. On balance, I think extrapolation efforts of the sort implemented here are more likely to weaken the case for the predictive value of a particular causal estimate than to provide a concrete and precise alternative to traditional IV. For example, the evidence for an adverse effect of childbearing on marital stability and welfare use is clearly weakened by the attempt to go from LATE to ATE. This sort of destructive evidence seems to me to be a prominent feature of life in the empirical world. The external validity of IV estimates is ultimately established less by new econometric methods than by replication in new data sets and, of course, by new instruments.

*Massachusetts Institute of Technology and NBER*

## Appendix: Computation

1. *The Test for Selection Bias*

Drop individual subscripts from the notation. Consider the following two-equation system:

$$Y = \Delta_0 + \Delta_1 D + \mu \qquad (A1)$$

$$Y = \delta_0 + \delta_1[1(D = Z)] + \delta_2[(D - Z)/2] + v. \quad (A2)$$

The test for selection bias is a test of whether $\Delta_1 = \delta_2$ when (A1) is estimated by IV using $Z$ as an instrument and (A2) is estimated by OLS. The two coefficients and the asymptotic standard error for their difference can be estimated by stacking (A1) and (A2) and allowing for heteroscedastic and correlated residuals. In practice, for sample sizes on the order of that used here, it seems reasonable to treat the estimate of $\delta_2$ as non-stochastic and use the standard error of the estimate of $\Delta_1$ to construct a t-test.

2. *Estimates under Restriction 3*

Use (A2) to write:

$$\mathrm{E}(Y_1|D = 1, Z = 0) = \mathrm{E}(Y_1|D_1 = D_0 = 1) = \delta_0 + \delta_2/2$$

$$\mathrm{E}(Y_0|D = 0, Z = 1) = \mathrm{E}(Y_1|D_1 = D_0 = 0) = \delta_0 - \delta_2/2.$$

Estimates of $\mathrm{E}(Y_0|D_1 > D_0)$ and $\mathrm{E}(Y_1|D_1 > D_0)$ can be obtained as IV estimates of the coefficients $\Delta_{01}$ and $\Delta_{11}$ in (A3) and (A4), below:

$$DY = \Delta_{10} + \Delta_{11}D + \mu_1 \quad (A3)$$

$$(1 - D)Y = \Delta_{00} + \Delta_{01}(1 - D) + \mu_0. \quad (A4)$$

Estimates of ATE under Restriction 3 are a linear combination of $\delta_0$, $\delta_2$, $\Delta_{01}$, and $\Delta_{11}$ . These coefficients and the standard error for any linear combination of them can be estimated by stacking (A2), (A3), and (A4).

To further simplify, rewrite (11) in terms of the parameters in (A2)–(A4) as

$$\mathrm{E}(Y_1 - Y_0) = \Delta_{11}[1 - (p_a - p_n)] - \Delta_{01}[1 + (p_a - p_n)] + 2\delta_0(p_a - p_n). \quad (A5)$$

To accommodate models with covariates, it is convenient to use a regression set-up to estimate $p_a - p_n$. Define a dependent variable $d^* = D(2Z - 1) - Z$. Regress $d^*$ on $Z$; the coefficient on $Z$ is an estimate of $p_a - p_n$. Note that (without covariates) the standard error for the estimated $p_a - p_n$ is the same as the standard error for the first-stage coefficient since the latter can be written $1 - p_n - p_a$. To estimate $\mathrm{E}(Y_1 - Y_0|D = 1)$, replace $p_a - p_n$ with $p_a/p_d$ in (A5).

Models with covariates were estimated by adding covariates to the relevant first-stage equations, and to (A1)–(A4). As a shortcut for inference for estimates of ATE using (A5), it seems reasonable to treat $(p_a - p_n)$ and $\delta_0$ as known since these are estimated much more precisely than $\Delta_{11}$ and $\Delta_{01}$, which are themselves instrumental variables estimates. Note also that IV estimates of $\Delta_{11}$ and $\Delta_{01}$ are independent.

3. *Estimates under Restriction 4*

Substitute parameters from (A1)–(A4) into (19) and simplify to obtain

$$\begin{aligned}\mathrm{E}(Y_1 - Y_0) = {} & \Delta_{11}[1 - (p_a - \theta^{-1}p_n)] - \Delta_{01}[1 + (\theta p_a - p_n)] \\ & + [\delta_0(1 + \theta^{-1}) + (\delta_2/2)(\theta^{-1} - 1)](\theta p_a - p_n).\end{aligned} \quad (A6)$$

Standard errors were calculated treating $p_a$, $p_n$, $\delta_0$, $\delta_2$, and $\theta$ as known.

## References

Abadie, A. (2002). 'Bootstrap tests for distributional treatment effects in instrumental variables models', *Journal of the American Statistical Association*, vol 97, pp. 284–92.

Abadie, A. (2003). 'Semi-parametric instrumental variables estimation of treatment response models', *Journal of Econometrics*, vol. 113, pp. 231–63.
Abadie, A., Angrist, J. and Imbens, G. (2002). 'Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings', *Econometrica*, vol. 70, pp. 91–117.
Angrist, J. (1991). 'Instrumental variables estimation of average treatment effects in econometrics and epidemiology', NBER Technical Working Paper No. 115 (November).
Angrist, J. and Evans, W. E. (1998). 'Children and their parents' labor supply: evidence from exogenous variation in family size', *American Economic Review*, vol. 88, pp. 450–77.
Angrist, J. and Imbens, G. (1991). 'Sources of identifying information in evaluation models', NBER Technical Working Paper No. 117 (December).
Angrist, J., Imbens, G. and Rubin, D. B. (1996). 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association*, vol. 91, pp. 444–55.
Arellano, M. (2002). 'Sargan's instrumental variables estimation and the generalized method of moments', *Journal of Business and Economic Statistics*, vol. 20, pp. 450–9.
Becker, G., Landes, E. and Michael, R. T. (1977). 'An economic analysis of marital instability', *Journal of Political Economy*, vol. 85, pp. 1141–87.
Blank, R. M. (2002). 'Evaluating welfare reform in the United States', NBER Working Paper 8983 (June).
Bloom, H. S. (1984). 'Accounting for no-shows in experimental evaluation designs', *Evaluation Review*, vol. 8, pp. 225–46.
Bronars, S. G. and Grogger, J. (1994). 'The economic consequences of unwed motherhood: using twins as a natural experiment', *American Economic Review*, vol. 84, pp. 1141–56.
Chamberlain, G. (1986). 'Asymptotic efficiency in semiparametric models with censoring', *Journal of Econometrics*, vol. 32, pp. 189–218.
Chamberlain, G. (1987). 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Econometrics*, vol. 34, pp. 305–34.
Cherlin, A. (1977). 'The effect of children on marital dissolution', *Demography*, vol. 14, pp. 265–72.
Frolich, M. (2002). 'Nonparametric IV estimation of local average treatment effects with covariates', IZA Discussion Paper No. 588 (September).
Grogger, J. and Bronars, S. B. (2001). 'The effect of welfare payments on the marriage and fertility behavior of unwed mothers: results from a twins experiment', *Journal of Political Economy*, vol. 109, pp. 529–45.
Hausman, J. (1978). 'Specification tests in econometrics', *Econometrica*, vol. 46, pp. 1251–71.
Heaton, T. B. (1990). 'Marital stability throughout the child-rearing years', *Demography*, vol. 27, pp. 55–63.
Heckman, J. J. (1990). 'Varieties of selection bias', *American Economic Review*, vol. 80, pp. 313–8.
Heckman, J. J., Tobias, J. and Vytlacil, E. (2001). 'Four parameters of interest in the evaluation of social programs', *Southern Economic Journal*, vol. 68, pp. 210–23.
Heckman, J. J. and Vytlacil, E. (2000). 'Local instrumental variables', NBER Technical Working Paper 252 (April).
Hotz, V. J., Imbens, G. and Klerman, J. A. (2000). 'The long-term gains from GAIN: a re-analysis of the impacts of the California GAIN program', NBER Working Paper 8–7 (November).
Imbens, G. W. and Angrist, J. (1994). 'Identification and estimation of local average treatment effects', *Econometrica*, vol. 62, pp. 467–75.
Kearney, M. (2002). 'Is there an effect of incremental welfare benefits on fertility behavior? A look at the family cap', NBER Working Paper 9093 (August).
Manski, C. (1990). 'Nonparametric bounds on treatment effects', *American Economic Review*, vol. 80, pp. 319–22.
Maynard, R., Boehnen, E., Corbett, T., Sandefur, G. and Mosley, J. (1998). 'Changing family formation behavior through welfare reform', in(R. Moffit, ed.), *Welfare, The Family, and Reproductive Behavior: Research Perspectives*, Washington, DC: National Academy Press.
Moffit, R. A. (1999). 'New developments in econometric methods for labor market analysis', Chapter 24 in (O. Ashenfelter and D. Card, eds.), *The Handbook of Labor Economics*, vol. 3A, Amsterdam: North-Holland.
Orr, L. L., Bloom, H. S., Bell, S. H., Doolittle, F., Lin, W. and Cave, G. (1996). *Does Training for the Disadvantaged Work?*, Washington, DC: The Urban Institute.
Phillips, P. C. B. (2003). Vision and influence in econometrics: John Denis Sargan', *Econometric Theory* (forthcoming).
Pearson, K. (1911). *The Grammar of Science*, London: Adam and Charles Black.
Roy, A. (1951). 'Some thoughts on the distribution of earnings', *Oxford Economic Papers*, vol. 3, pp. 135–46.
Sargan, D. (1958), 'The estimation of economic relationships using instrumental variables', *Econometrica*, vol. 26, pp. 393–415.

Waite, L. J. and Lillard, L. A. (1991). 'Children and marital disruption', *American Journal of Sociology*, vol. 96, pp. 930–53.
Wald, A. (1940). 'The fitting of straight lines if both variables are subject to error', *Annals of Mathematical Statistics*, vol. 11, pp. 284–300.
Westhoff, C. F., Potter, R. G., and Sagi, P. L. (1963). *The Third Child: A Study in the Prediction of Fertility*, Princeton NJ: Princeton University Press.
Wooldridge, J. M. (1997). 'On two-stage least squares estimation of the average treatment effect in a random coefficient model', *Economics Letters*, vol. 56, pp. 129–33.
Wooldridge, J. M. (2003). 'Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model', *Economics Letters*, vol. 79, pp. 185–91.