Policy Research Working Paper 5804

On the Implications of Essential Heterogeneity for Estimating Causal Impacts Using Social Experiments

Martin Ravallion

The World Bank
Development Research Group
Director's office
September 2011

Policy Research Working Paper 5804

Abstract

Randomized control trials are sometimes used to estimate the aggregate benefit from some policy or program. To address the potential bias from selective take-up, the randomization is used as an instrumental variable for treatment status. Does this (popular) method of impact evaluation help reduce the bias when take-up depends on unobserved gains from take up? Such "essential heterogeneity" is known to invalidate the instrumental variable estimator of mean causal impact, though one still obtains another parameter of interest, namely mean impact amongst those treated. However, if essential heterogeneity is the only problem then the naïve (ordinary least squares) estimator also delivers this parameter; there is no gain from using randomization as an instrumental variable. On allowing the heterogeneity to also alter counterfactual outcomes, the instrumental variable estimator may well be more biased for mean impact than the naïve estimator. Examples are given for various stylized programs, including a training program that attenuates the gains from higher latent ability, an insurance program that compensates for losses from unobserved risky behavior and a microcredit scheme that attenuates the gains from access to other sources of credit. Practitioners need to think carefully about the likely behavioral responses to social experiments in each context.

This paper is a product of the Director's office, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at mravallion@worldbank.org.

Produced by the Research Support Team

On the Implications of Essential Heterogeneity for Estimating Causal Impacts Using Social Experiments

Martin Ravallion[1]

Development Research Group, World Bank
1818 H Street NW, Washington DC, 20433, USA

[1] These are the views of the author, and need not reflect those of the World Bank or any affiliated organization. For helpful comments the author is grateful to Pedro Carneiro, Emanuela Galasso, James Heckman, Aart Kraay, Michael Lokshin, David McKenzie and Berk Ozler.

I. Introduction

Any imaginable policy intervention is likely to have diverse impacts, possibly with losses as well as gains to those affected. Some forms of such impact heterogeneity are known to be ignorable when estimating the aggregate benefits, notably when the heterogeneity is orthogonal to the actual receipt of the intervention. However, that is not a particularly plausible form of heterogeneity. Naturally, people make rational choices about whether to participate in any offered treatment, and they almost certainly base their choices on things they know that are not available as data to the analyst. Take up will depend on latent gains from take up. This gives rise to what Heckman, Urzua and Vytlacil (HUV) (2006) term "essential heterogeneity"—an idea going back to Heckman (1992).

Under essential heterogeneity, we expect the instrumental variables (IV) estimator to be biased for the mean causal effect, as shown by HUV. HUV also question whether IV will perform any better than OLS; they write (p.392):[2]

> "If the analyst is interested in knowing the average response, the effect of the policy on the outcomes of [units] that adopt it or the effect of the policy if a particular [unit] adopts it, there is no guarantee that the IV estimator comes any closer to the desired target than the OLS estimator, and indeed, it may be more biased."

As Heckman and Vytlacil (2005, p. 671) put it: "The cure may be worse than the disease."

This paper explores this claim further in the context of a classic randomized experiment for which treatment is only possible for those randomly assigned the option of treatment but there is selective take-up. A popular estimator uses the randomized assignment as the IV for treatment. This is not valid for the mean causal effect, but still gives the mean impact on the treated, as demonstrated by Heckman and Vytlacil (2007, Section 9.2).[3] (The specific context determines which of these two parameters one is most interested in.) However, as section II shows, if the only way that the unobserved differences in returns to treatment generate bias is through their interaction with take up then the IV is not required to retrieve the mean impact on the treated in large samples; OLS delivers the same parameter. The cure is no worse, or better!

[2] I have dropped mathematic expressions from the quote, as these will not make sense out of context.

[3] Under certain conditions the IV estimator under essential heterogeneity does give another parameter of potential interest, namely the "local average treatment effect" (LATE) (Imbens and Angrist, 1994), given by the mean impact for those units induced to take up the treatment by a change in the IV. Imbens and Angrist identify conditions under which LATE can be identified by the standard IV estimator. Also see the discussion in HUV.

This comes with a potential bonus for the design of experiments: one does not need to know outcomes for those who were assigned the opportunity for treatment but chose not to participate.

However, the paper also argues that the standard formulation of the essential heterogeneity problem in terms of a latent interaction effect with treatment status is unduly restrictive. Section III examines the implications of allowing the latent factors creating essential heterogeneity to also influence counterfactual outcomes. Depending on the direction and strength of this extra effect, OLS may be less biased for the mean causal impact than the IV estimator—indeed, there is even a case in which OLS is unbiased for mean impact.

Whether the use of randomization as an IV is to be preferred for estimating mean impact can thus be seen to depend on the behavioral assumptions made in modeling outcomes for the specific program being evaluated. Section III describes stylized examples of the various cases, which point to important differences between types of programs in the biases to be expected even with a randomized assignment of the option for take up.

II. Biases in Standard Impact Estimators under Essential Heterogeneity

Selective take-up is to be expected in almost any randomized experiment with human subjects. So we have a potential source of bias, as is well-recognized in the literature. In practice, the near universal fix for this problem is to use a dummy variable for the randomized assignment as the instrumental variable (IV) for treatment status.[4]

The assignment to treatment will naturally be correlated with receiving the treatment. But can it be legitimately excluded from the main regression, as also required for a valid IV? It is often argued that the fact of randomized assignment means that the other conditions for a valid IV will invariably hold, namely that the IV only affects outcomes via treatment. At first glance it sounds reasonable to assume that being randomly assigned to some treatment only matters to its outcomes if one actually receives that treatment. However, on closer inspection this assumption is far from reasonable. As HUV point out, plausible behavioral responses to the option for

[4] The theoretical conditions for this to work are laid out by Angrist, Imbens and Rubin (1996). The IV estimator is identical to the method of correcting for selective compliance proposed by Bloom (1984) in which the intent-to-treat estimate is deflated by the proportion who take up the assigned option of treatment. This assumes no "crossovers" in that treatment is only possible for units assigned to treatment. With crossovers, one obtains instead a local average treatment effect (Imbens and Angrist, 1994).

treatment invalidate the exclusion restriction. This section elaborates on this point and explores further its implications for impact evaluation.

The randomized assignment is denoted *Z*, which takes the value 1 if a unit is assigned to treatment and 0 otherwise.[5] The actual treatment status is *D*, taking the value 1 if treated and 0 otherwise. At least some units take up treatment, but not all ($0 < E(D) < 1$). Treatment is only possible if one is assigned to the treatment group, but take up is voluntary. The unit-specific impact is β, which is the difference between the outcome under treatment and that when untreated. Since we cannot observe someone in two states of nature at the same time, β is unobserved. The mean impact is $\bar{\beta}$ and η represents the variation in impact around the mean, i.e., $\beta = \bar{\beta} + \eta$. The mean impact on those treated is $E(\beta | D = 1) = \bar{\beta} + E(\eta | D = 1)$. Given that treatment (conditional on assignment) is a choice variable, a natural assumption is that those who take up the treatment tend to have higher η's, i.e., $E(\eta | D = 1) > E(\eta | D = 0)$, implying that $E(\eta | D = 1) > 0$, given that η has zero (unconditional) mean.

The standard regression specification for estimating the mean impact on outcome *Y* is:

$$Y = \alpha + \beta D + \varepsilon = \alpha + \bar{\beta} D + (\varepsilon + \eta D) \qquad (1)$$

The heterogeneity in impact is swept into the error term, such that the coefficient on *D* gives the mean causal effect. Now consider the standard IV estimator in which the randomized assignment is used as the IV for treatment. This converges in large samples to:

$$\text{Plim}\,\hat{\bar{\beta}}_{IV} = \bar{\beta} + \frac{Cov(\varepsilon + \eta D, Z)}{Cov(D, Z)} = \bar{\beta} + \frac{Cov(\eta D, Z)}{Cov(D, Z)} \qquad (2)$$

Here I use the fact that randomization implies that $Cov(\varepsilon, Z) = 0$. Randomization also implies that $Cov(\eta, Z) = 0$. However, under essential heterogeneity, it does not imply that $Cov(\eta D, Z) = 0$ since there will be sorting on the η's; amongst those assigned the program, those who choose to take it up will tend to have higher η's. Thus the randomized assignment is not a valid IV for identifying mean impact, as was pointed out by HUV.

[5] Here and elsewhere I use the same notation as HUV.

But what does the IV estimator give us? We can write the bias term on the RHS of (2) as:

$$\frac{Cov(\eta D,Z)}{Cov(D,Z)}=\frac{E(\eta DZ)-E(\eta D)E(Z)}{E(DZ)-E(D)E(Z)} \qquad (3)$$

Evaluating these terms further by exploiting the fact that both *D* and *Z* are binary, with *D*=1 implying *Z*=1 (since assignment to treatment is necessary for receiving treatment), we have:[6]

$$E(\eta DZ)=E(\eta D)=E(\eta | D=1)E(D) \qquad (4.1)$$

$$E(DZ)=E(D) \qquad (4.2)$$

Substituting (4) into (3) and then (3) into (2) we have:

$$\text{Plim}\,\hat{\bar{\beta}}_{IV}=\bar{\beta}+E(\eta | D=1) \qquad (5)$$

Thus the IV estimator still gives the mean impact on the treated, but overestimates the overall mean impact under essential heterogeneity.[7]

However, if essential heterogeneity is the only source of bias—specifically, if $Cov(D,\varepsilon)=0$ —then the “naïve” OLS estimator also gives mean impact on the treated. To see this, note first that for the OLS estimator (in obvious notation):

$$\text{Plim}\,\hat{\bar{\beta}}_{OLS}=\bar{\beta}+\frac{Cov(\eta D,D)}{Var(D)} \qquad (6)$$

The bias term is:[8]

$$\frac{Cov(\eta D,D)}{Var(D)}=\frac{E(\eta D^2)-E(\eta D)E(D)}{E(D)(1-E(D))}=\frac{E(\eta D)}{E(D)}=E(\eta | D=1) \qquad (7)$$

Thus OLS also converges to $\bar{\beta}+E(\eta | D=1)$. While the presence of essential heterogeneity naturally biases both estimators of overall mean impact, it turns out to be exactly the same bias.

[6] Note that $E(\eta DZ)=E(\eta DZ | D=1)E(D)+E(\eta DZ | D=0)(1-E(D))=E(\eta | D=1)E(D)$ given that $E(\eta DZ | D=0)=0$.

[7] This result can be found in Heckman and Vytlacil (2007, Section 9).

[8] I use the fact that $E(\eta D^2)=E(\eta D)=E(\eta | D=1)E(D)$.

Of course, the reason is different. For the IV method, the bias stems from the violation of the exclusion restriction, while for OLS, it stems directly from the endogeneity of treatment.

Notice that if essential heterogeneity is the only concern then one does not need to know the randomized assignment (as required by the IV estimator) to obtain a consistent estimate of the mean impact for those treated; the data for OLS—outcomes and treatment status—are sufficient. Outcomes for those treated can then be collected alongside the receipt of treatment. The control group need only represent those for whom treatment is not an option.

III. Allowing Essential Heterogeneity to Alter Counterfactual Outcomes

The above formulation has followed the literature on essential heterogeneity in postulating that it only matters through the implied interaction effect between take-up and the gains from treatment. This is restrictive. More generally, one can allow the same latent factors causing variability in the gains from treatment to have bearing on counterfactual outcomes.

To introduce this feature, let the error term in (1) now take the form:

$$\varepsilon = \gamma\eta + \upsilon \qquad (8)$$

We can identify "pure" essential heterogeneity as when $\gamma = 0$. More generally, latent characteristics that enhance impact may be associated with higher ($\gamma > 0$) or lower ($\gamma < 0$) counterfactual outcomes. (In the special case $\gamma = -1$ we have a constant impact amongst those treated.) To keep the focus on the implications of essential heterogeneity, I assume that the innovation error term (υ) in (8) is orthogonal to treatment ($Cov(D,\upsilon) = 0$). One can weaken this assumption to conditional exogeneity, by adding controls (or a control function).

For interpreting the econometric model in (1) and (8), it is helpful to consider a more explicit model of outcomes as returns to a primary individual characteristic χ, which is observed by each experimental unit but not by the analyst, and where the return to higher χ can be systematically altered by the intervention. We can write this model as:[9]

[9] Note that the error terms are the same in these equations given that (by assumption) $Cov(D,\upsilon) = 0$.

$$\begin{aligned} Y &= a_0 + b_0\chi + \upsilon \; if \; D = 0 \\ Y &= a_1 + b_1\chi + \upsilon \; if \; D = 1 \end{aligned} \tag{9}$$

This is equivalent to the model in (1) and (8) where the correspondence is as follows:

$$\begin{aligned} \alpha &= a_0 + b_0 E(\chi) \\ \overline{\beta} &= a_1 - a_0 + (b_1 - b_0)E(\chi) \\ \eta &= (b_1 - b_0)(\chi - E(\chi)) \\ \gamma &= b_0 /(b_1 - b_0) \end{aligned} \tag{10}$$

Rational take up requires that $E(\chi|D=1) > E(\chi|D=0)$ if the gain ($\overline{\beta} + \eta$) is increasing in χ (i.e., if $b_1 > b_0$), while $E(\chi|D=1) < E(\chi|D=0)$ if $b_1 < b_0$. Either way, $E(\eta|D=1) > 0$.

There are three possible regimes for how heterogeneity alters counterfactual outcomes: $\gamma > 0$, $-1 \le \gamma < 0$ and $\gamma < -1$. These relate to differences in the types of programs being evaluated, given likely behavior responses. I provide stylized examples of each regime.

Regime 1: $\gamma > 0$. Consider the following training program. The source of latent heterogeneity is learning ability, which is unobserved by the analyst but known individually. People choose whether to participate in the (randomly assigned) program on the basis of their ability, as this determines their expected benefits. (As usual, there is some cost of participation, including forgone income.) Labor market earnings are the outcomes of interest. The program imparts skills that are complementary to ability, so that the returns to ability are greater under treatment ($b_1 > b_0$). (For example, an accountancy course enhances the returns to numeracy.) Absent the program, higher ability yields higher income ($b_0 > 0$). Thus $\gamma > 0$.

Regime 2: $-1 \le \gamma < 0$. Consider instead a public insurance scheme, providing support for those suffering (say) ill-health or a crop failure. Participants are compensated for income losses stemming from some unobserved risky behavior on their part, denoted by χ. The program attracts those with high χ. Expected income is the outcome variable. In the absence of the program, those who undertake the risky activity are assumed to have higher expected utility but lower expected income ($b_0 < 0$). However, with the program in place, the risk-takers are largely compensated for any loss, leaving a net gain in expected income ($b_1 > 0$). Thus: $-1 \leq \gamma < 0$.

Regime 3: $\gamma < -1$. A variation on the example for Regime 1 is to suppose that the training program provides skills that substitute for latent ability (rather than the two being complements). Thus the scheme dulls the benefits from higher innate ability and is more attractive to those with lower ability. In this case, we have $b_1 < b_0$ and $\gamma < -1$. To give another example of Regime 3 for a different type of program, consider a microcredit scheme, which provides extra credit to some target group and χ denotes access to credit from other sources. Take up is higher for those with lower χ; $E(\chi|D=1) < E(\chi|D=0)$. (For example, self-targeting mechanisms in the scheme's design discourage participation for those with high χ.) Greater access to credit from alternative sources increases counterfactual incomes ($b_0 > 0$) as well as participants' incomes ($b_1 > 0$). However, the scheme attenuates the gain enjoyed by those with greater access to credit from alternative sources, i.e., $b_1 < b_0$. So, again, $\gamma < -1$. And, given that take up is greater for those with lower χ, we have $E(\eta|D=1) > 0$.[10]

What biases can be expected in standard estimates of the mean impact? It is readily verified that the IV estimator still gives the mean impact on those treated (but not the overall mean impact) when the essential heterogeneity matters to counterfactual outcomes. The role of the IV estimator is not to remove the bias stemming from the interaction effect between treatment and gains from treatment, but rather to remove any bias in how the same source of essential heterogeneity alters counterfactual outcomes. For the OLS estimator we now have:

$$\text{Plim}\,\hat{\bar{\beta}}_{OLS} = \bar{\beta} + \frac{Cov(\eta D, D)}{Var(D)} + \frac{\gamma\, Cov(\eta, D)}{Var(D)} \qquad (11)$$

Similarly to the derivation above for the bias in the IV estimator, we have:

[10] Notice that positive returns to the latent characteristic with or without treatment ($b_0 > 0$ and $b_1 > 0$) are consistent with both $\gamma < 0$ and $E(\eta|D=1) > 0$ as long as the selection yields $E(\chi|D=1) < E(\chi)$, as is implied by the assumption that those with low access to credit tend to take up the scheme.

$$\text{Plim}\,\hat{\bar{\beta}}_{OLS} = \bar{\beta} + \left[1 + \frac{\gamma}{1 - E(D)}\right] E(\eta | D = 1) \qquad (12)$$

In Regime 1 ($\gamma > 0$), while both the OLS and IV estimators overestimate mean impact, the use of the randomized assignment as the IV reduces the bias in the OLS estimate.

In Regime 2 ($-1 \leq \gamma < 0$), the outcome depends on the program participation rate. When the participation rate is relatively low—specifically $E(D) < 1 + \gamma$ (≥ 0)—OLS overestimates mean impact, but is less biased than the IV estimator. With a sufficiently high participation rate, $E(D) > 1 + \gamma$, OLS underestimates mean impact but will have a lower (higher) absolute bias than the IV estimate if $2(E(D) - 0.5)$ is less than (greater than) $1 + \gamma$. A program participation rate less than 0.5 (common in practice) is sufficient for OLS to have lower bias than the IV estimator in Regime 2. OLS is unbiased when $E(D) = 1 + \gamma$, implying that the odds of program take-up, $E(D)/(1 - E(D))$, equal the relative returns to the latent characteristic, b_1 / b_0. This is clearly a knife-edge property.

In Regime 3 ($\gamma < -1$), OLS underestimates the true mean impact, while IV overestimates it. As in Regime 2, with $E(D) > 1 + \gamma$, the (absolute) OLS bias will be less than (greater than) the IV bias if $2(E(D) - 0.5)$ is less than (greater than) $1 + \gamma$. (A necessary condition for the OLS estimate to be less biased is that $E(D) < 0.5$, but this is not sufficient.) The weighted mean of the two estimates will be unbiased when the weight on the OLS estimate is $(1 - E(D))/(-\gamma)$. The lower the treatment rate, the higher the weight on the OLS estimate. The greater the effect of the impact heterogeneity on counterfactual outcomes the higher the weight on the IV estimate.

IV. Conclusions

Essential heterogeneity is such an intuitively plausible idea that the onus on analysts should be to establish *a priori* grounds why it does not exist. Short of such grounds, we can expect a bias in the estimate of mean causal impact obtained by using the randomized assignment as the instrumental variable for treatment status. The bias stems from a failure of the exclusion restriction, even with a perfectly randomized assignment. This much is clear from the theoretical literature. But will the IV estimator still help in reducing the bias in the naïve OLS estimator, as obtained by ignoring the endogeneity of treatment and simply subtracting the mean of the outcomes for the control group from that for the treatment group?

The answer depends on what impact parameter one is interested in, the type of program one is evaluating and the behavioral responses to that program. If one is only interested in the mean impact for those actually treated then the IV estimator is still unbiased, even though the randomized assignment is not a valid IV. However, if the latent interaction effect between take up and the gains from take up is the only source of bias then the OLS estimator also delivers the mean treatment effect on the treated. Under the "pure" form of essential heterogeneity studied in the literature, the IV and OLS estimates converge asymptotically.

The two estimators only differ in large samples when there is some other source of bias, on top of that from the interaction effect between the take up and the unobserved variation in the impact of the treatment. A natural extension to the standard formulation of the essential heterogeneity problem is to allow the same factors creating the heterogeneity in impact to also matter to counterfactual outcomes. If these work in the same direction—such that higher counterfactual outcomes due to the latent factor come hand-in-hand with higher returns to treatment—then the IV estimator can be trusted to reduce the OLS bias in mean impact. A training program providing complementary skills to latent ability is probably a good example.

However, there is no *a priori* reason to expect the two sources of bias to work in the same direction. That depends on the type of program and the behavioral responses to that program. If the latent factors leading to higher returns to treatment are associated with lower counterfactual outcomes then the "IV cure" for endogenous treatment can be worse than the disease. The paper has described examples, based on a stylized public insurance program and a microcredit scheme for fighting poverty. Indeed, the OLS estimator may even be unbiased, despite the selective take-up. And even when it is not, averaging the IV and OLS estimates can reduce the bias under certain conditions.

Reference

Angrist, Joshua, Guido Imbens and Donald Rubin, 1996, “Identification of Causal Effects Using Instrumental Variables,” *Journal of the American Statistical Association*, Vol. XCI, pp. 444-455.

Bloom, Howard S. 1984. “Accounting for No-shows in Experimental Evaluation Designs,” *Evaluation Review* 8: 225-246.

Heckman, James. 1992. “Randomization and Social Policy Evaluation.”In *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel, 201–30. Cambridge and London: Harvard University Press.

Heckman, James, Serio Urzua and Edward Vytlacil. 2006. “Understanding Instrumental Variables in Models with Essential Heterogeneity,” *Review of Economics and Statistics* 88(3): 389-432.

Heckman, James and Edward Vytlacil. 2005. “Structural Equations, Treatment Effects and Econometric Policy Evaluation,” *Econometrica* 73(3): 669-738.

________________________________. 2007. “Econometric Evaluation of Social Programs, Part ii: Using the Marginal Ttreatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments,” in James Heckman and Edward Leamer (eds) *Handbook of Econometrics Vol. 6B*, Amsterdam: North Holland.

Imbens, Guido and Joshua Angrist. 1994. “Identification of Local Average Treatment Effects,” *Econometrica* 62(2): 467-475.