


Unobservable Selection and Coefficient Stability: Theory and Evidence

Emily Oster


To cite this article: Emily Oster (2019) Unobservable Selection and Coefficient Stability: Theory and Evidence, *Journal of Business & Economic Statistics*, 37:2, 187-204, DOI: 10.1080/07350015.2016.1227711

To link to this article: <https://doi.org/10.1080/07350015.2016.1227711>

 [View supplementary material](#) 



 Published online: 01 Jun 2017.

 [Submit your article to this journal](#) 

 Article views: 39322

 [View related articles](#) 

 [View Crossmark data](#) 

 Citing articles: 1370 [View citing articles](#) 

Unobservable Selection and Coefficient Stability: Theory and Evidence

Emily OSTER

Department of Economics, Brown University, Providence, Rhode Island, and NBER (emily_oster@brown.edu)

A common approach to evaluating robustness to omitted variable bias is to observe coefficient movements after inclusion of controls. This is informative only if selection on observables is informative about selection on unobservables. Although this link is known in theory in existing literature, very few empirical articles approach this formally. I develop an extension of the theory that connects bias explicitly to coefficient stability. I show that it is necessary to take into account coefficient and R -squared movements. I develop a formal bounding argument. I show two validation exercises and discuss application to the economics literature. Supplementary materials for this article are available online.

KEY WORDS: Coefficient stability; Selection; Omitted variable bias.

1. INTRODUCTION

Concerns about omitted variable bias are common to most or all nonexperimental work in economics. (Despite recent trends, this still makes up the vast majority of results within economics: in 2012 the combination of the *American Economic Review*, the *Quarterly Journal of Economics*, and the *Journal of Political Economy* published 69 empirical, nonstructural articles, only 11 of which were randomized.) The most straightforward approach to such concerns is to include controls that can be observed. Angrist and Pischke (2010) argued that among the major advances in empirical economics in the past two decades is greater effort to identify the most important threats to validity, and to address them with appropriate selection of controls.

In some cases it is possible to argue that a control (or set of controls) fully captures a particular omitted variable. However, in many cases observed controls are an incomplete proxy for the true omitted variable or variables. For example, it is common in many applications to worry about confounding from socioeconomic status. Researchers often include controls to capture this, but typically with the acknowledgment that the controls observed in a typical dataset (e.g., education and income categories, race groups) do not perfectly capture overall socioeconomic status. Similarly, in many cross-country or cross-regional analyses authors seek to control for geographic differences across areas, but observed controls (extent of mountains, access to water) are incomplete proxies for the true omitted factor.

A common approach in these situations is to explore the *sensitivity* of treatment effects to the inclusion of observed controls. If a coefficient is stable after inclusion of the observed controls, this is taken as a sign that omitted variable bias is limited. (The next section and the final section of this article will discuss more explicitly the use of this approach within economics, but it is worth noting that it is the link between coefficient stability and omitted variable bias is often quite direct. For example, Chiappori, Orefice, and Quintana-Domeque (2012) stated: “It is reassuring that the estimates are very similar in the standard and the augmented specifications, indicating that our results are unlikely

to be driven by omitted variables bias.” Similarly, Lacetera, Pope, and Sydnor (2012) stated: “These controls do not change the coefficient estimates meaningfully, and the stability of the estimates from columns 4 through 7 suggests that controlling for the model and age of the car accounts for most of the relevant selection.”) The intuitive appeal of this approach lies in the idea that the bias arising from the observed (imperfect) controls is informative about the bias that arises from the full set including the unobserved components. This is not, however, implied by the baseline assumptions underlying the linear model.

Formally, using the observables to identify the bias from the unobservables requires making further assumptions about the covariance properties of the two sets. In particular, the case in which the omitted variable bias is fully identified by the observed controls corresponds to the extreme assumption that the relationship between treatment and unobservables can be fully recovered from the relationship between treatment and observables (Murphy and Topel 1990; Altonji, Elder, and Taber 2005a; Altonji et al. 2011).

Even under this most optimistic assumption, however, coefficient movements alone are not a sufficient statistic to calculate bias. To illustrate why, consider the case of a researcher estimating wage returns to education with individual ability as the only confounder, and where there are two orthogonal components of ability, one of which has a higher variance than the other. Assume wages would be fully explained if both ability components were observed but, in practice, the researcher sees only one of the two. The coefficient will appear much more stable if the observed ability control is the lower variance one, but this is not because the bias is smaller but simply because less of the wage outcome is explained by the controls.

© 2019 American Statistical Association
Journal of Business & Economic Statistics
April 2019, Vol. 37, No. 2

DOI: 10.1080/07350015.2016.1227711

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

This example is described in more detail in [Section 2](#). The key observation is that the quality of the control is diagnosed by how much of the variance in the outcome is explained by its inclusion or, equivalently, how much the R -squared moves when the controls are introduced. Omitted variable bias is proportional to coefficient movements, but only if such movements are scaled by the change in R -squared when controls are included. (This point is closely related to the partial R -squared logic in Imbens 2003.)

Recognizing this point, Altonji, Elder, and Taber (2005a, hence, AET) suggested a method for evaluating the robustness of results under the assumption that the relationship between treatment and unobservables can be recovered from the relationship between treatment and observables. Specifically, they provided a test statistic that is valid under the null of a zero treatment effect for how important the unobservables would have to be relative to the observables to eliminate the observed effect. The test statistic assumes that if one could observe the full set of unobservables the outcome variance would be fully explained, so the regression would have an R -squared of 1.

Empirical work in economics that invokes coefficient stability rarely either discusses R -squared movements or explicitly uses the AET method. I demonstrate this in [Section 2](#) using a sample of 76 results from 27 articles in top journals after AET and explicitly discuss coefficient stability in the context of imperfect controls. (The universe is articles from the *American Economic Review*, *Quarterly Journal of Economics*, *Econometrica* and the *Journal of Political Economy* for 2008–2010 with at least 20 citations in the ISI Web of Science, and those from 2011–2013 in the same journals with at least 10 citations. I further limit to articles with replication files available.) Only two of these articles mention the importance of their controls in explaining the variance of the outcome. Only one article in this set (Nunn and Wantchekon 2011) attempts to implement the method described in AET and they do not do it correctly. I show that these omissions are especially problematic as only a third of results considered would be robust by the AET standard.

At least three factors may have limited the take-up of the AET methodology. First, in their analysis of the linear case in AET they present a test that is valid only under the null of a zero treatment effect, and do not detail how a bias-adjusted treatment effect would be estimated. Among other things, this limits the possibility of evaluating the robustness of results for which the magnitude increases when controls are added. (Their working article, Altonji, Elder, and Taber (2002), does produce an implicit function of the bias-adjusted effect in the nonlinear case, although empirical researchers do not appear to have made use of this at all.) Second, their assumption that including the unobservables would produce an R -squared of 1 may understate the robustness of results, for example, in cases where there is measurement error in the outcome. Finally, perhaps most important, they do not provide any validation evidence for this approach. (A fourth factor, in addition to these, is that their article does not explicitly link their calculation to coefficient movements although the link is obvious econometrically.)

The aim of this article is to extend the methodology for evaluating robustness to omitted variable bias under the assumption that the relationship between treatment and unobservables can be recovered from the relationship between the treatment and observables. I link the bias explicitly to coefficient stability

and provide some validation suggesting the procedure performs well. Finally, I will return to analyzing the economics literature and suggest possible robustness standards using evidence from randomized data.

[Section 3](#) expands on the theory in AET. I focus on the case of a linear model in which an outcome is fully determined by a treatment, a set of observed, a set of unobserved covariates, and an iid error term. The coefficient of interest is the treatment effect. The observed and unobserved covariates are linked by shared covariance properties with the treatment. I show first that under a restrictive set of assumptions a consistent estimator of the treatment effect can be recovered through an intuitive function of coefficients and R -squared values. The inputs in this case are typically reported in standard regression output, making it easy to evaluate robustness of published results.

Following this, I show it is possible to generate a consistent, closed-form, estimator for omitted variable bias under less restrictive assumptions. This estimator also relies on coefficient and R -squared values, but in addition requires the researcher to observe the variance of the outcome and treatment, as well as information on the share of the *treatment* variance, which is explained by the observed controls. Although these values are not typically reported in standard regression tables, they are straightforward to calculate. Details of this estimator make clear the limits of coefficient stability: it is possible that the coefficient will be completely unchanged with the addition of controls but there may still be a large bias on the treatment effect. I show in this section how it is also possible to deliver these results in a GMM setup.

A key input into the estimator is the R -squared from a hypothetical regression of the outcome on treatment and both observed and unobserved controls; I denote this R_{\max} . If the outcome can be fully explained by the treatment and full controls set, then $R_{\max} = 1$ (this is the assumption in AET). In many empirical settings it seems likely (due, e.g., to measurement error) that the outcome cannot be fully explained even if the full control set is included. Knowledge about measurement error or expected idiosyncratic variation in the outcome can be used to develop intuition about this value.

The results allow researchers to calculate a consistent estimate of the bias-adjusted treatment effects under two assumptions: (1) a value for the relative degree of selection on observed and unobserved variables (δ) and (2) a value for R_{\max} . Given bounding values for both objects, researchers can calculate an identified set for the treatment effect. As in AET I suggest that equal selection (i.e., $\delta = 1$) may be an appropriate upper bound on δ .

The second contribution of the article, in [Section 4](#), is to perform some validation of the estimator and the bounding logic suggested in empirical settings. This section first uses NLSY data to construct a dataset relating education and wages; the data are constructed such that we know the true treatment effect. I evaluate the performance of this adjustment by excluding combinations of controls from the “observed” set. I estimate the relative degree of selection on observed and unobserved controls, which would be produced by each excluded set and calculate the bias-adjusted treatment effect. I show that in 89% of cases the bounds I propose would include the true effect; in only 62% of cases does the confidence interval of the naive coefficients estimated with controls include this effect. This may actually

undervalue the estimator performance as the control set in this exercise is selected at random rather than based on using the most important controls first, as would be common in practice.

In a second test, I estimate several relationships between maternal behavior and child outcomes; socioeconomic status confounders are a major concern. I match possibly biased observational estimates with external evidence on causal effects from randomized data or comprehensive meta-analyses (this is close in spirit to LaLonde 1986). I then ask whether the robustness tests described above would separate true from false associations. I find that the adjustment performs well: the approach identifies as robust only the two relationships for which external evidence confirms a link. Both of the validation exercises therefore suggest empirical support for this assumption.

The final contribution of the article, developed in Section 5, is to return to the application of the procedure to the economics literature. As in Section 2, I focus on a sample of well-cited articles from top journals in which coefficient stability is invoked in light of imperfect controls. I use the estimator developed to calculate bias-adjusted treatment effects under the assumption of $R_{\max} = 1$ and the assumption that $\delta = 1$. I am able to calculate these effects for both results where controls move the coefficient toward zero, and those that move it away from it. As in Section 2, I show that only a small share of results is robust to this adjustment.

I then consider how this conclusion changes with other bounds on R_{\max} , in particular focusing on bounds that are a function of the R -squared from the regression with controls. These capture the idea that there is variation in how predictable outcomes are, and this variation can be roughly inferred from how much is predicted by the observables. Denoting the R -squared from the regression with controls as \tilde{R} , I explore robustness to $R_{\max} = \Pi\tilde{R}$, with varying values of Π . Twenty-seven percent of results are robust to a value of $\Pi = 2$, and 57% to a value of $\Pi = 1.25$. I demonstrate that coefficient movements alone can be very misleading about the degree of robustness.

There is considerable variation across articles in the robustness of stability claims, but this does not suggest an appropriate general value for the bound on R_{\max} . There are many possible ways to calculate such a bound. In Section 5, I suggest one: that randomized results might provide a bounding value. I use a sample of randomized articles, also from top journals, to derive cut-off values of Π , which would allow at least 90% of randomized results to survive: this value is $\Pi = 1.3$. To the extent that this is an attractive methodology for generating bounds on R_{\max} it suggests that researchers might calculate a bias-adjusted treatment effect bound using a value of $R_{\max} = 1.3\tilde{R}$. In the sample of nonrandomized results considered, about 45% would survive this standard.

The theoretical contribution of this article relates to a large literature on causal inference in the face of unobserved confounders (Rosenbaum and Rubin 1983). Imbens (2003) presented an analysis of sensitivity using a partial R -squared logic that is conceptually similar to the insights in this article. A number of methodological articles consider the approach of varying the covariate set as a sensitivity analysis (Heckman and Hotz 1989; Dehejia and Wahba 1999; Gelbach 2016). I also relate to the idea of specification and control set search (Leamer 1978; Pearl 2000; Angrist and Pischke 2010). It is worth noting that the approach in this article differs in some conceptual sense from

the latter set of references in that I am concerned with estimating a bias-adjusted treatment effect under an assumption about the full model rather than with searching for the appropriate full model.

The article follows most closely a series of articles that explore bias in treatment effect under proportional selection (Murphy and Topel 1990; Altonji, Elder, and Taber 2005a; Altonji et al. 2011). The derivation of closed-form estimators in the linear case represents an extension of the case of nonlinear estimation in Altonji, Elder, and Taber (2005a) and Altonji, Elder, and Taber (2002).

2. COEFFICIENT STABILITY: ILLUSTRATIVE EXAMPLE AND USE IN ECONOMICS

I motivate the analysis in the article with a simple illustration of the issue and with some data on coefficient stability within economics.

Illustrative Example. A central point of this article is to make clear that coefficient movements alone are not sufficient to evaluate bias, even under the strong assumption of related observed and unobserved variables. As an illustration, consider the case of a researcher estimating wage returns to education with individual ability as the only confounder. (This example is motivated by independent work by Pischke and Schwandt (2015), although their setting is focused on issues of measurement error in the ability measure.)

I consider the following simple setup. Assume that the model that determines wages is given by

$$Y = \beta X + W + C,$$

where W and C are two orthogonal components of “ability” and X is education. Assume that the variance of W is much larger than the variance of C but both relate to X in the same way. More specifically, a regression of X on either W or C yields the same coefficient.

Now consider the difference between the case where the researcher observes W (the high variance control) and the case where s/he observes C (the lower variance control). The key observation is that if the researcher observes the ability control with the lower variance, the coefficient will appear stable when the control is included. This is not, however, because the bias is small, but simply because the control is less important in explaining wages.

To see this precisely, consider Panel A of Table 1, which uses data constructed from the model above, with the assumption that $\beta = 0$. The first row shows controlled and uncontrolled coefficients when the observed control is the one with the larger variance; the second shows the coefficients when the observed control is the one with the smaller variance. The coefficient in the second row appears much more stable, even though the true effect is zero in both.

The key difference in the two rows is the change in R -squared, which diagnoses the poor quality of the proxy in the second row compared to the first. The uninformative control leaves the coefficient largely unchanged but also adds little to the R -squared. Omitted variable bias is proportional to coefficient movements, but only if such movements are scaled by movements in R -squared.

Table 1. Calibrated example

High versus low variance control			
Quality of observed control	Uncontrolled coefficient [R^2]	Controlled coefficient [R^2]	True effect
High variance control observed	0.202 [0.004]	0.002 [0.990]	0
Low variance control observed	0.202 [0.004]	0.200 [0.013]	0

NOTES: This table shows calculations based on simulated data. The true model is $Y = \beta X + W + C$, with $\beta = 0$. The data are constructed so the high variance control is W and $\text{var}(W) = 10$ and the low variance controls is C and $\text{var}(C) = 0.1$. $\text{var}(X) = 1$, $\text{cov}(X, W) = 0.2$ and $\text{cov}(X, C) = 0.002$. Note that $\text{cov}(X, C)$ is implied by the equal selection assumption, $\text{cov}(X, W)$, $\text{var}(C)$, and $\text{var}(W)$.

Coefficient Stability in Economics. Many economics articles use coefficient stability to argue for a causal treatment effect in the presence of imperfect controls. To get a sense of the formal robustness of these claims, I extract a sample of top journal publications. I begin with the universe of all articles in the *American Economic Review*, *Quarterly Journal of Economics*, *The Journal of Political Economy*, and *Econometrica* from 2008–2010 with at least 20 citations in the ISI Web of Science, and those from 2011–2013 in the same journals with at least 10 citations. I limit the sample to articles with replication files available so it is possible to do further robustness calculations. From these articles, I extract all results where the researcher explores the sensitivity of the result to a control set and, using a close reading of the articles, those in which this exercise appears to be designed to address an imperfect set of controls. (This later restriction does not eliminate many results; nearly all articles that comment on coefficient stability appear to do so to deal with this issue. However, there are exceptions. As an example, in their work on corruption in trucking routes, Olken and Barron (2009) commented on the stability of the coefficients when trip fixed effects are included. The inclusion of these effects is clearly an important robustness check, as it tests whether differences across trips drive the results. But there is no sense in which such fixed effects are a proxy for some unobserved omitted control. In a case like this it is worth noting that in fact the movement of the coefficient is notable only to the extent one wants to draw conclusions from the effect without controls.) The sample (full citation list in Appendix C) includes 27 articles with 76 total results.

The illustrative example above makes clear the importance of incorporating movements in R -squared in coefficient stability discussions. However, this importance is rarely acknowledged in these articles. Only two of the articles in the sample mention anything about movements in R -squared (or variance explained).

In principle, if coefficients and R -squared values typically move together, this will introduce less bias. If large coefficient movements were always accompanied by large R -squared movements, then the coefficient stability would be effectively a sufficient statistic. Similarly, if the R -squared values from the regression with controls were always very large—say, always close to 1—then the coefficient movements would be enough. In practice, neither of these is the case.

Figure 1 uses the results extracted from the 27 articles described above. I include only results where the controlled effect is significant. The figures graph the relationship between the percent movement in effect size and the absolute movement

in R -squared values. Figure 1(a) uses all results, and Figure 1(b) limits to cases where the inclusion of controls moves the coefficient toward zero.

It is not the case that the controlled regressions uniformly have a high R -squared. The range of values for the R -squared in the regression with controls is 0.029 to 0.992, with an average of 0.403. Moreover, there is at best a very weak relationship between coefficient movements and R -squared movements.

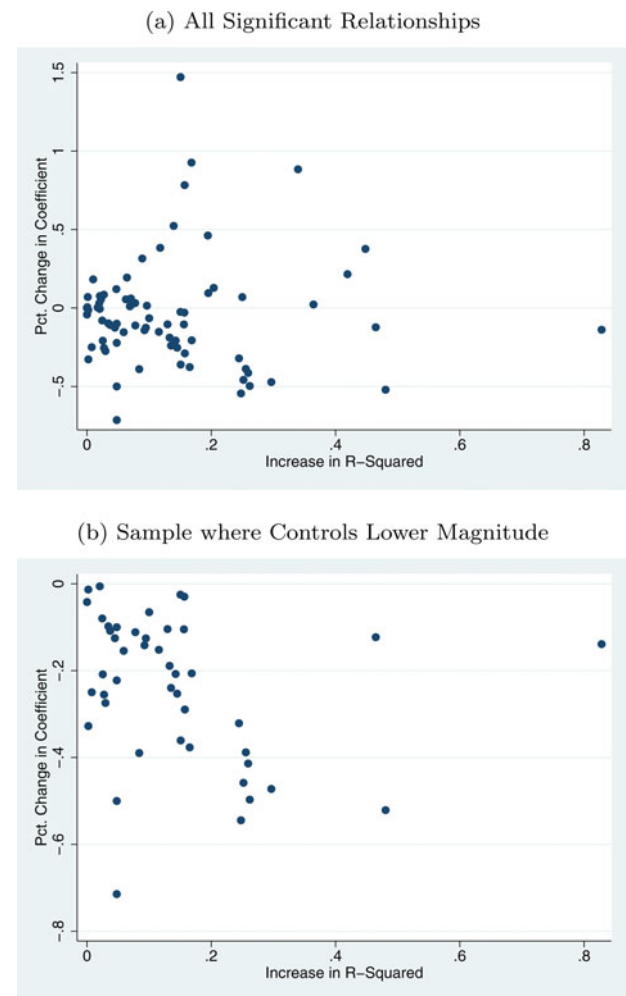


Figure 1. Coefficient stability and R -squared movements. (a) All significant relationships. (b) Sample where controls lower magnitude. These figures show the relationship between the percent change in coefficient and the increase in R -squared in sample of highly cited articles from top journals in economics. The sample is discussed in Section 2.

If we limit ourselves to results where the percent change in coefficient values is between -12% and -8% , for example, the range of changes in R -squared values is from 0.034 to 0.156. This observation suggests that the robustness of these results may vary widely even if they all have similarly small coefficient movements.

AET suggested a formal notion of robustness drawing on the assumption that the observed and unobserved controls have related covariance properties. In particular, they derived a consistent estimator for the ratio of the covariance between (a) the treatment and the unobserved controls and (b) the treatment and the observed controls, which would produce a treatment effect of zero. They suggested that results with a ratio above 1—that is, those in which the unobserved controls are more important in explaining the treatment than the observed controls—be viewed as robust. Their estimator is consistent under the null of a zero treatment effect and under the assumption that if both observed and unobserved controls were included in a regression the R -squared would be 1. Although AET do not explicitly derive the relation between their estimator and the idea of coefficient stability, the connection is theoretically direct.

The articles in the sample I consider were all published after AET. However, only one of them engages in a formal robustness calculation of the type AET suggest (Nunn and Wantchekon 2011). Further, even that article does not perform the exact calculation in AET. Instead, they follow Bellows and Miguel (2009) in implicitly assuming that the observed and unobserved controls are equally important in explaining the variance of the outcome. This ignores the fact that that coefficient movements must be scaled by R -squared movements.

Focusing on the 45 results in which the inclusion of controls moves the coefficient toward zero, I calculate the AET estimator for the articles in this sample. Table 2 shows some characteristics of the distribution. The median result would be eliminated if the ratio of the selection on unobservables to selection on observables was 0.73, less than the suggested robustness standard of 1. Indeed, only 30% of results are robust to a value of $\delta = 1$.

On its own, this observation makes clear that formalizing the coefficient stability procedure is important, and is likely to matter quantitatively in many settings. In practice, the specific

approach suggested by AET may be incomplete for several reasons. First, the estimator they detail is consistent only under the null of a zero treatment effect. It is not possible to evaluate robustness under other null assumptions, nor do they provide an explicit estimator for the bias-adjusted treatment effect in the linear case. This means it is not possible to use this to evaluate robustness of results where the inclusion of controls moves the coefficient away from zero—effectively, all results of this type are “robust.” However, if the coefficient is increased significantly by the included controls then the conclusions about magnitude may be radically biased even if the rejection of zero is retained. More generally, if one wants to make statements about the size of the “bias-adjusted treatment effect,” their estimator will not deliver this.

A second issue is that the calculation they suggest does not provide a direct way to allow for the possibility that the inclusion of the unobserved controls might not move the R -squared all the way to 1. Since in many settings there is likely to be some idiosyncratic variation in outcome—and, importantly, the degree of this is likely to vary—this will understate the robustness of results in an unpredictable way.

The theory section below addresses both of these issues while developing a consistent estimator for the bias-adjusted treatment effect in the linear version of the AET model. I then look at the empirical performance of this estimator in simulated data and, finally, return to the sample of articles described above in an attempt to develop a more tractable procedure for formalizing coefficient stability.

3. THEORY

I begin in this section by developing the estimator under a set of restrictive assumptions. Although these assumptions are unlikely to be met in practical settings, the resulting estimator has an intuitive form that lays bare many of the issues in this setting. I then develop the estimator under more general assumptions. Finally, I discuss implementation.

Before moving into the theory it is useful to briefly discuss the conceptual approach. I focus on an approach to estimating an unbiased treatment effect from a model in which there are some observed confounders and some unobserved confounders. Assuming the data-generating process specified below, the approach will generate bias-adjusted treatment effects. There may be numerous other threats to causality, including poorly specified functional form and others. (This analysis is appropriate only for a linear model. AET developed a similar estimator in the context of at least one type of nonlinear model (a bivariate probit model).) This approach will not remove the bias arising from those. Leamer (1978) and Angrist and Pischke (2010) (among others) provided more discussion on the related topic of model specification in search of a causal effect.

3.1 General Setup and Definitions

Throughout this section, I will focus on the following setup. Consider the regression model

$$Y = \beta X + \Psi \omega^0 + W_2 + \epsilon, \quad (1)$$

Table 2. Summary of results under AET adjustment

	Relative degree of selection to eliminate result (δ)
Min	0.0039
10%	0.099
25%	0.326
Median	0.739
75%	1.22
90%	8.45
Max	90.37

NOTES: This table uses data from published articles in economics to calculate the degree of selection on unobservables compared to observables that would be required to eliminate the result. The calculation used is that from Altonji, Elder, and Taber (2005). The calculation assumes that if unobservables were included in the regression all of the outcome variance would be explained.

where X is the (scalar) treatment and ω^0 is a vector of the observed controls, $\omega_1^0, \dots, \omega_j^0$. The index W_2 is not observed. Define $W_1 = \Psi\omega^0$ and assume that all elements of ω^0 are orthogonal to W_2 , so W_1 and W_2 are orthogonal. Without loss of generality, assume the elements of ω^0 are also orthogonal to each other. (All results go through identically if these elements are correlated.) Define the proportional selection relationship as $\delta \frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\sigma_2^2}$, where $\sigma_{iX} = \text{cov}(W_i, X)$, $\sigma_i^2 = \text{var}(W_i)$ for $i \in \{1, 2\}$, and δ is the coefficient of proportionality. Note that at this point we do not make any assumptions about δ so this relationship will always hold for some δ . Define $\text{var}(X) = \sigma_X^2$ and $\text{var}(Y) = \sigma_Y^2$.

This setup is drawn from AET, who assume $\epsilon = 0$ and that W_2 contains some error unrelated to X . These results go through in a straightforward way if $\epsilon = 0$.

As in AET, the orthogonality of W_1 and W_2 is central to deriving the results and may be somewhat at odds with the intuition that the observables and the unobservables are “related.” In practice, the weight of this assumption is in how we think about the proportionality condition. In Appendix A.1, I show formally that if we begin with a case in which the elements of W_1 are correlated with W_2 we can always redefine W_2 such that the results hold under some value of δ .

Denote the coefficient resulting from the short regression of Y on X as $\hat{\beta}$ and the R -squared from that regression as \hat{R} . Define the coefficient from the intermediate regression of Y on X and ω^0 as $\tilde{\beta}$ and the R -squared as \tilde{R} . Finally, define R_{\max} as the R -squared from a hypothetical regression of Y on X , ω^0 and W_2 . Note that other than R_{\max} these are values that are estimated in-sample based on the available data. R_{\max} is a (theoretical) population value.

The omitted variable bias of the OLS estimates $\hat{\beta}$ and $\tilde{\beta}$ are determined by auxiliary regressions of (1) each value $\omega_i^0, \dots, \omega_j^0$ on X ; (2) W_2 on X ; and (3) W_2 on X and ω^0 . Denote the in-sample estimated coefficient on X from regressions of each ω_i^0 on X as $\hat{\lambda}_{\omega_i^0|X}$ and the (unobservable) in-sample estimated coefficient on X from a hypothetical regression of W_2 on X as $\hat{\lambda}_{W_2|X}$. Finally, denote the coefficient on X from a regression of W_2 on X and ω^0 as $\hat{\lambda}_{W_2|X, \omega^0}$. Denote the population analogs of these values $\lambda_{\omega_i^0|X}$, $\lambda_{W_2|X}$ and $\lambda_{W_2|X, \omega^0}$.

The goal is to provide a consistent estimator for β from Equation (1). Throughout this section all estimates are implicitly indexed by n . Probability limits are taken as n approaches infinity. All observations are independent and identically distributed according to model (1).

3.2 Restricted Estimator

I develop a consistent estimator of the bias under two additional assumptions.

Assumption 1. $\delta = 1$, so $\frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\sigma_2^2}$.

Assumption 2. Consider a regression of X on ω^0 and denote the coefficient vector from this regression as (μ_1, \dots, μ_j) . The coefficients on these controls in the regression of Y on X and ω^0 are ψ_i . Assume that $\frac{\psi_i}{\mu_j} = \frac{\mu_i}{\mu_j} \forall i, j$.

The first of these assumptions implies an equal selection relationship—the unobservable and observables are equally

related to the treatment. The second assumption delivers the result that the coefficient on X in the intermediate regression of Y on X and controls is the same with the observed control set ω^0 as it would be if we could control for the index W_1 . The proof of this statement appears in Appendix A.2.

The intuition behind the second condition is straightforward: the relative contributions of each variable to X must be the same as their contribution to Y . It will be satisfied trivially if there is only one observed control. With multiple controls it is very unlikely to hold except in pathological cases. However, as long as the deviations from this condition are not extremely large, the estimator will provide an approximation to the consistent estimator.

Using the definitions above, and standard omitted variable bias formulas, I can express the probability limits of the short and intermediate regression coefficients in terms of the auxiliary regression coefficients.

$$\hat{\beta} \xrightarrow{p} \beta + \sum_{i=1}^J \psi_i \lambda_{\omega_i^0|X} + \lambda_{W_2|X}$$

$$\tilde{\beta} \xrightarrow{p} \beta + \lambda_{W_2|X, \omega^0}.$$

The asymptotic bias on $\tilde{\beta}$ (the coefficient on X with controls included) is $\lambda_{W_2|X, \omega^0}$. Under the second assumption above, $\tilde{\beta}$ is the same coefficient, which would be recovered from a regression of Y on X and the index W_1 . Given this, the bias is equal to $\frac{\sigma_{2X}^2}{\sigma_1^2(\sigma_X^2 - \frac{\sigma_{1X}^2}{\sigma_1^2})}$. Denote this bias as Π .

Define the following.

$$\beta^* = \tilde{\beta} - \left[\hat{\beta} - \tilde{\beta} \right] \frac{R_{\max} - \tilde{R}}{\tilde{R} - \hat{R}}$$

Proposition 1 summarizes the result.

Proposition 1. $\beta^* \xrightarrow{p} \beta$.

Proof. I outline the proof here, with details in Appendix A.3. Using the definition of coefficient and R -squared values and recalling the bias is denoted Π we have the following relationships:

$$\begin{aligned} (\hat{\beta} - \tilde{\beta}) &\xrightarrow{p} \left(\frac{\sigma_{1X}}{\sigma_X^2} \right) \left(1 - \frac{\sigma_{1X}}{\sigma_1^2} \Pi \right) \\ (\tilde{R} - \hat{R}) \hat{\sigma}_Y^2 &\xrightarrow{p} \sigma_1^2 + \Pi^2 \left(\sigma_X^2 - \frac{\sigma_{1X}^2}{\sigma_1^2} \right) \\ &\quad - \frac{1}{\sigma_X^2} \left(\sigma_{1X} + \Pi \left(\sigma_X^2 - \frac{\sigma_{1X}^2}{\sigma_1^2} \right) \right)^2 \\ (R_{\max} - \tilde{R}) \hat{\sigma}_Y^2 &\xrightarrow{p} \Pi \left(\frac{\sigma_1^2 \left(\sigma_X^2 - \frac{\sigma_{1X}^2}{\sigma_1^2} \right)}{\sigma_{1X}} - \Pi \left(\sigma_X^2 - \frac{\sigma_{1X}^2}{\sigma_1^2} \right) \right). \end{aligned}$$

These define a system of three equations in three unknowns (σ_1^2, σ_{1X} and Π). The system is identified and the solution is $\Pi = \left[\hat{\beta} - \tilde{\beta} \right] \frac{R_{\max} - \tilde{R}}{\tilde{R} - \hat{R}}$. \square

Some intuition for this result may be developed by observing that $\Pi = \tilde{\beta} - \beta$ so this result implies that $\frac{\hat{\beta} - \tilde{\beta}}{\tilde{\beta} - \hat{\beta}} = \frac{R_{\max} - \tilde{R}}{\tilde{R} - \hat{R}}$. That

is, under the equal selection assumption, the ratio of the movement in coefficients is equal to the ratio of the movement in R -squared. The objects W_1 and W_2 enter the equation for Y symmetrically in terms of coefficients, and equal selection implies they also are symmetric in their impact on X . The only way in which their impact may differ is if they have different variances. This possible difference will be captured in the differential contributions to R -squared. In the special case where the variances are equal, then $\frac{R_{\max} - \tilde{R}}{\tilde{R} - \hat{R}} = 1$ and the coefficient movement with inclusion of observed controls is equal to the expected coefficient movement with unobserved controls.

If we relax the assumption of equal selection, and return to the proportional selection relationship ($\delta \frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\sigma_2^2}$), it is straightforward to observe that we can calculate an approximation of the bias-adjusted treatment effect with

$$\beta^* \approx \tilde{\beta} - \delta \left[\hat{\beta} - \tilde{\beta} \right] \frac{R_{\max} - \tilde{R}}{\tilde{R} - \hat{R}}.$$

Given the restrictiveness of the assumptions required to generate this simple formulation it is not appropriate to suggest that researchers use this as an estimator directly. However, it is useful for developing intuition and, perhaps more so, because it can be calculated from objects included in standard regression tables. In many cases, this will be a close approximation to the consistent estimator for the bias developed below. It therefore provides a simple way to evaluate the robustness of published results.

3.3 Unrestricted Estimator

I now consider the estimator without the additional restrictions described above. I retain the notation and the proof method proceeds similarly.

The estimator now relies on observing an additional object from the data. Define \tilde{X} as the residual from a regression of X on ω^0 . Define the variance of this residual in sample as $\hat{\tau}_x$ and the population analog as τ_x . Effectively, the difference between τ_x and σ_X^2 captures how much of the variance of X can be explained by the controls. This is not an object that is reported in standard regression tables although it is straightforward to calculate (it would be sufficient to observe the R -squared from a regression of X on ω^0 along with the variance of X).

As above, I can express the probability limits of the short and intermediate regression coefficients in terms of the auxiliary regression coefficients:

$$\hat{\beta} \xrightarrow{p} \beta + \sum_{i=1}^j \psi_i^0 \lambda_{\omega_i^0|X} + \lambda_{W_2|X}$$

$$\tilde{\beta} \xrightarrow{p} \beta + \lambda_{W_2|X, \omega^0}.$$

The asymptotic bias on $\tilde{\beta}$ is $\frac{\delta \sigma_{1X} \sigma_2^2}{\sigma_1^2 \tau_x}$. Denote this bias Π .

Define the cubic function $f(v)$ as

$$f(v) = \delta \left((R_{\max} - \tilde{R}) \sigma_{yy} \right) \left(\hat{\beta} - \tilde{\beta} \right) \sigma_X^2 + v \left(\delta \left((R_{\max} - \tilde{R}) \sigma_y^2 \right) \left(\sigma_X^2 - \tau_x \right) - \left((\tilde{R} - \hat{R}) \sigma_y^2 \right) \tau_x \right.$$

$$\left. - \sigma_X^2 \tau_x \left(\hat{\beta} - \tilde{\beta} \right)^2 \right) + v^2 \left(\tau_x \left(\hat{\beta} - \tilde{\beta} \right) \sigma_X^2 (\delta - 2) \right) + v^3 (\delta - 1) (\tau_x \sigma_X^2 - \tau_x^2).$$

Proposition 2. The proposition has two cases depending on the roots of $f(v)$.

Case 1: $f(v)$ has a single real root, define this root as v_1 . Define

$$\beta^* = \tilde{\beta} - v_1. \beta^* \xrightarrow{p} \beta.$$

Case 2: $f(v)$ has three real roots, define them as $v_1, v_2,$ and v_3 .

Define a set $\beta^* = \{\tilde{\beta} - v_1, \tilde{\beta} - v_2, \tilde{\beta} - v_3\}$. One element of the set β^* converges in probability to β .

Proof. This is an outline of the proof, with details in Appendix A.4. Using the definition of coefficient and R -squared values and recalling that the bias is denoted Π we have the following relationships:

$$\begin{aligned} (\hat{\beta} - \tilde{\beta}) &\xrightarrow{p} \frac{\sigma_{1X}}{\sigma_X^2} - \Pi \left(\frac{\sigma_X^2 - \tau_x}{\sigma_X^2} \right) \\ (\tilde{R} - \hat{R}) \hat{\sigma}_{yy} &\xrightarrow{p} \sigma_1^2 + \Pi^2 (\tau_x) - \frac{1}{\sigma_X^2} (\sigma_{1X} + \Pi (\tau_x))^2 \\ (R_{\max} - \tilde{R}) \hat{\sigma}_{yy} &\xrightarrow{p} \Pi \left(\frac{\sigma_1^2 \tau_x}{\sigma_{1X}} - \Pi \tau_x \right). \end{aligned}$$

These define a system of three equations in three unknowns (σ_1^2, σ_{1X} , and Π). Solving recursively leaves us with Π as the root of the equation $f(v)$ given above. This is a cubic with all real coefficients so it has either one or three real roots. If it has a single real root, that is the solution. If it has multiple real roots, one of the three will be the solution. \square

Proposition 2 parallels Theorem 2 in the Altonji, Elder, and Taber (2002) working article; they consider the case where X is binary.

The corollary below develops the case of $\delta = 1$.

Corollary 1. Define

$$v_1 = \frac{-\Theta - \sqrt{\Theta^2 + 4 \left((R_{\max} - \tilde{R}) \sigma_y^2 \right) \left(\hat{\beta} - \tilde{\beta} \right)^2 \left(\sigma_X^2 \right)^2 \tau_x}}{-2 \tau_x \left(\hat{\beta} - \tilde{\beta} \right) \left(\sigma_X^2 \right)}$$

$$v_2 = \frac{-\Theta + \sqrt{\Theta^2 + 4 \left((R_{\max} - \tilde{R}) \sigma_y^2 \right) \left(\hat{\beta} - \tilde{\beta} \right)^2 \left(\sigma_X^2 \right)^2 \tau_x}}{-2 \tau_x \left(\hat{\beta} - \tilde{\beta} \right) \left(\sigma_X^2 \right)},$$

where $\Theta = \left(\left((R_{\max} - \tilde{R}) \sigma_y^2 \right) \left(\sigma_X^2 - \tau_x \right) - \left((\tilde{R} - \hat{R}) \sigma_y^2 \right) \tau_x - \left(\sigma_X^2 \right) \tau_x \left(\hat{\beta} - \tilde{\beta} \right)^2 \right)$. Define a set $\beta^* = \{\tilde{\beta} - v_1, \tilde{\beta} - v_2\}$. One element of the set β^* converges in probability to β .

Proof. This follows immediately from *Proposition 2*, with $\delta = 1$. \square

In either case—regardless of whether $\delta = 1$ —this problem may have multiple solutions. Only one element of the set will converge in probability to the true β . I discuss solution selection below.

Proposition 3 shows a result related to δ . In particular, I solve for the value of δ to match a particular treatment effect. This will be central to implementation since it allows us to ask how large the relative selection on observables and unobservables would need to be to produce a treatment effect of zero.

Proposition 3. Define some value $\hat{\beta}$. Define $\hat{\delta}$ as the coefficient of proportionality for which $\beta = \hat{\beta}$. Define

$$\delta^* = \frac{(\tilde{\beta} - \hat{\beta}) (\tilde{R} - \hat{R}) \hat{\sigma}_y^2 \hat{\tau}_x + (\tilde{\beta} - \hat{\beta}) \hat{\sigma}_x^2 \hat{\tau}_x (\hat{\beta} - \tilde{\beta})^2 + 2 ((\tilde{\beta} - \hat{\beta}))^2 (\hat{\tau}_x (\hat{\beta} - \tilde{\beta}) \hat{\sigma}_x^2) + ((\tilde{\beta} - \hat{\beta}))^3 ((\hat{\tau}_x \hat{\sigma}_x^2 - \hat{\tau}_x^2))}{((R_{\max} - \tilde{R}) \hat{\sigma}_y^2 (\hat{\beta} - \tilde{\beta}) \hat{\sigma}_x^2 + (\tilde{\beta} - \hat{\beta}) (R_{\max} - \tilde{R}) \hat{\sigma}_y^2 (\hat{\sigma}_x^2 - \hat{\tau}_x) + ((\tilde{\beta} - \hat{\beta}))^2 (\hat{\tau}_x (\hat{\beta} - \tilde{\beta}) \hat{\sigma}_x^2) + ((\tilde{\beta} - \hat{\beta}))^3 ((\hat{\tau}_x \hat{\sigma}_x^2 - \hat{\tau}_x^2)))}$$

Under this definition, $\delta^* \xrightarrow{P} \hat{\delta}$.

Proof. The proof follows from setting $\Pi = \tilde{\beta} - \hat{\beta}$, substituting into the $f(v)$ function and solving for δ . \square

Proposition 3 shows that there is a single value of δ to match any targeted treatment effect—for example, a single value of δ will match a treatment effect of zero.

Together, **Proposition 2**, **Corollary 1**, and **Proposition 3** define the details of the estimator. The following subsections discuss some issues in interpretation and extensions.

3.3.1 Interpretation of ϵ . In discussing empirical applications, it will be crucial to take a stand on the value of R_{\max} , which is influenced by ϵ . Conceptually, ϵ represents an error, which is uncorrelated with X , W_1 , or W_2 . One interpretation of ϵ is that it captures the degree of measurement error in the outcome. Another interpretation is that ϵ captures the influence of anything, which is determined after X , W_1 , and W_2 are determined. Both of these interpretations may be useful in choosing a value for R_{\max} in a particular context.

3.3.2 Solution Selection. This estimator may deliver multiple solutions for β . One of these will be the true β under the proportional selection relationship. With an added assumption we can typically eliminate at least one solution and, in the case where $\delta = 1$, always produce a single solution.

Using the notation in Assumption 2, note that can define $W_1 = \psi_1 \omega_1^o + \dots + \psi_J \omega_J^o$. Now define $\hat{W}_1 = \hat{\psi}_1 \omega_1^o + \dots + \hat{\psi}_J \omega_J^o$, where $\hat{\psi}_i$ is the coefficient on ω_i^o , which is estimated in the observable regression of Y on X and the observed controls.

Assumption 3. $\text{Sign}(\text{cov}(X, \hat{W}_1)) = \text{Sign}(\text{cov}(X, W_1))$.

Effectively, this assumes that the bias from the unobservables is not so large that it biases the *direction* of the covariance between the observable index and the treatment. Under Assumption 3, if $\delta = 1$ there is a unique solution.

In the case where $\delta \neq 1$ there may be multiple solutions, one closer to the controlled treatment effect and one further. The obvious procedure of selecting the treatment effect closest to the controlled coefficient will be appropriate if one is willing to assume that the bias is fairly small.

I argue below that in empirical settings a value of $\delta = 1$ is a good bounding value; this is consistent with arguments in AET. For the purposes of implementation, therefore, it may be appropriate to consider either (a) calculating the bias-adjusted effect under the assumption of $\delta = 1$, with Assumption 3 active or (b) calculating the value of δ such that $\beta = 0$. Either of these will provide a unique solution.

3.3.3 Additional Controls. It is useful to consider a simple extension in which there is an additional observed set of controls.

Formally, consider the case where the full model is

$$Y = \beta X + \Psi \omega^o + W_2 + m + \epsilon, \tag{2}$$

where m is orthogonal to ω^o , W_2 and ϵ and the assumptions about orthogonality with ϵ are as above. Assume that the covariance between m and X is unrelated to the covariance between X and ω^o and W_2 . It is straightforward to observe in this case that if we simply regress all other variables on m and take the residuals, we return to the setup above and the results go through as stated there. In practice, this means that the controls m are included in both controlled and “uncontrolled” regressions, and X is residualized with respect to m when generating σ_x^2 and τ_x .

3.3.4 Inference. Standard errors around β^* could be generated using a bootstrap approach. Such an approach depends on the estimator displaying asymptotic normality. I show evidence for this using simulation. I simulate data from two populations with varying data-generating processes. (The inputs are described in the figure notes.) The populations are of size 1,000,000 and I run 1000 Monte Carlo simulations of the estimator, drawing 10,000 observations each time.

The distributions of estimated β^* in the two cases are shown in **Figure 2**. A normal distribution is overlaid. The distributions appear normal and a Shapiro–Wilk test does not reject normality in either case. This suggests that a bootstrap approach may be an acceptable way to generate standard errors if that is of interest.

3.3.5 Relation to Coefficient Stability. All else equal, coefficient stability correlates with a smaller amount of bias. However, it is crucial to note that it is possible for coefficients to be stable—indeed, to be completely unchanged—even in the presence of very large bias.

To see this in theory, assume $\delta = 1$ and consider the conditions under which the uncontrolled coefficient $\hat{\beta}$ is exactly equal to the controlled coefficient $\tilde{\beta}$. Using the notation above, this occurs if and only if $\frac{\sigma_{1X}}{\sigma_x^2} + \frac{\sigma_{1X}\sigma_2^2}{\sigma_1^2\tau_x} \frac{\tau_x}{\sigma_x^2} = \frac{\sigma_{1X}\sigma_2^2}{\sigma_1^2\tau_x}$. One condition that will cause this to hold is if $\sigma_{1X} = 0$. The formula for the bias is $\frac{\sigma_{1X}\sigma_2^2}{\sigma_1^2\tau_x}$ so if $\sigma_{1X} = 0$, then there is no bias and $\beta = \tilde{\beta}$.

However, this condition will also hold if $\sigma_1^2 = \frac{\sigma_x^2 - \tau_x}{\tau_x} \sigma_2^2$. Under this assumption, the movement in R -squared is

$$(\sigma_x^2 - \tau_x) \left(\frac{\sigma_2^2}{\tau_x} - \left(\frac{\sigma_{1X}}{\sigma_x^2 - \tau_x} \right)^2 \right),$$

which will be nonzero as long as $\sigma_x^2 > \tau_x$ and $\sigma_2^2 > 0$. In this way, the coefficient movement is zero and the R -squared movement is positive, which would appear to suggest limited (or zero) bias. However, the bias in this case is actually $\frac{\sigma_{1X}}{\sigma_x^2 - \tau_x}$, which is nonzero.

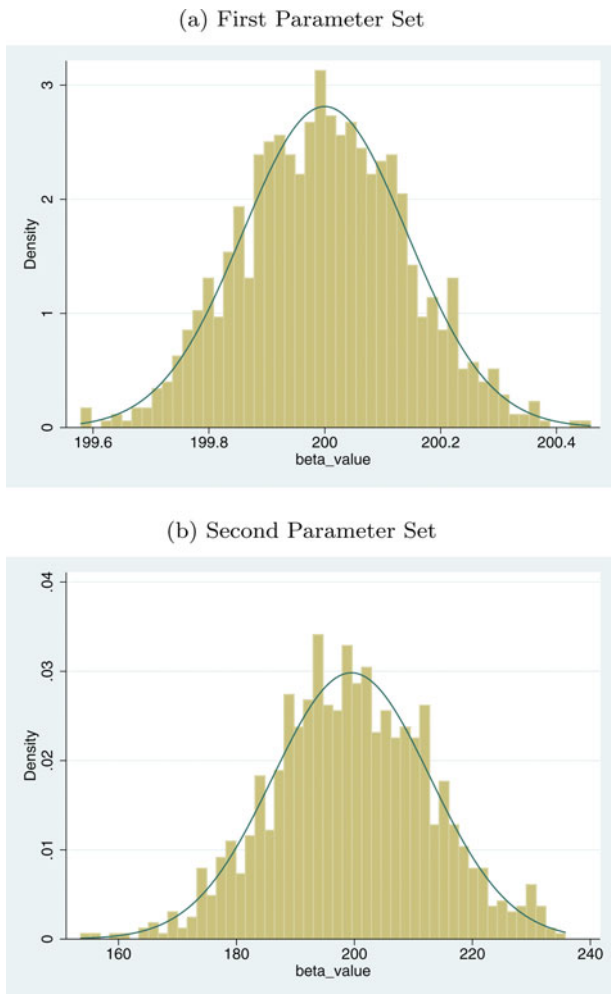


Figure 2. Distribution of estimated treatment effects. (a) First parameter set. (b) Second parameter set. These figures show the distribution of estimated bias-adjusted treatment effects under two parameter sets. The figures are generated by drawing 1000 samples of size 10,000 from a population of 1,000,000. The data generating values for the first set are: $\beta = 200, \delta = 1, \gamma_1 = 100, \gamma_2 = 200, \text{cov}(X, w_1^o) = 0.1, \text{cov}(X, w_2^o) = 0.1, \text{var}(W_2) = 20,000, \text{var}(w_1^o) = \text{var}(w_2^o) = 1$. The second set uses the same inputs but with $\text{var}(W_2) = 250,000$. In both cases, I add an iid error with mean 100 and standard deviation 1.

In practice, this will become a serious issue when the variance of W_2 is large *and* the proportionality assumption underlying the restricted estimator in Section 3.2 is seriously violated.

3.3.6 GMM Representation. It is useful to observe that there is a GMM representation of this problem. To observe this consider Equation (3) below, which explicitly includes a set of many observed controls $j \in \{1 \dots J\}$ and excludes the error term.

$$Y = \beta X + \sum_{j=1}^J \Psi_j \omega_j^o + W_2. \tag{3}$$

By assumption, W_2 in this setup is orthogonal to all of the ω_j^o , which means we can define the normal OLS moments on ω_j^o with W_2 treated as the error. (If W_2 is not mean 0 then

we will also need to include a constant that will add another moment.)

Additional moments are generated by the combination of the proportional selection assumption and the auxiliary regressions of X on the index of observed and unobserved controls. More specifically, we have the following GMM setup

$$E \begin{pmatrix} (Y - \beta X - \sum_{j=1}^J \Psi_j \omega_j^o) \omega^o \\ (X - \alpha_1 - \pi (\sum_{j=1}^J \Psi_j \omega_j^o)) (\sum_{j=1}^J \Psi_j \omega_j^o) \\ (X - \alpha_1 - \pi (\sum_{j=1}^J \Psi_j \omega_j^o)) \\ (X - \alpha_1 - \delta \pi (Y - \beta X - \sum_{j=1}^J \Psi_j \omega_j^o)) \\ (Y - \beta X - \sum_{j=1}^J \Psi_j \omega_j^o) \\ (X - \alpha_1 - \delta \pi (Y - \beta X - \sum_{j=1}^J \Psi_j \omega_j^o)) \end{pmatrix} = 0,$$

where the first moment is a vector of moments for each observed control. The value of δ is provided by the user.

There are $J + 4$ moments and $J + 4$ parameters to estimate (the J elements of $\Psi_j, \alpha_1, \alpha_2, \pi,$ and β .) The system can be estimated by GMM.

The case of primary concern above is represented in Equation (4), which replicates Equation (3) but includes an orthogonal error term

$$Y = \beta X + \sum_{j=1}^J \Psi_j \omega_j^o + W_2 + \epsilon. \tag{4}$$

Note that if we consider $W_2 + \epsilon$ as a single object, we are back in the case above without error and the moments are the same.

The proportionality condition is given by $\delta \frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\sigma_2^2}$ with $R_{\max} < 1$. Note, however, that we can rewrite this as $\delta \frac{\sigma_2^2}{\text{var}(W_2 + \epsilon)} \frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\text{var}(W_2 + \epsilon)}$. We now have this condition defined with $R_{\max} = 1$ and $\delta_{\text{mod}} = \delta \frac{\sigma_2^2}{\text{var}(W_2 + \epsilon)}$. The object $\frac{\sigma_2^2}{\text{var}(W_2 + \epsilon)}$ can be summarized as the share of the total unobserved variance explained by W_2 . This will be approximated by the ratio $\frac{R_{\max} - \bar{R}}{1 - \bar{R}}$ but will not be exact.

3.4 Implementation

In empirical work in economics, discussions of coefficient stability are typically used in establishing robustness. The estimator above suggests two related ways that such robustness statements might be made. I detail these below.

Statements About δ . One approach to robustness is to assume a value for R_{\max} and calculate the value of δ for which $\beta = 0$. This can be interpreted as the degree of selection on unobservables relative to observables that would be necessary to explain away the result (under the full model hypothesized). A value of $\delta = 2$, for example, would suggest that the unobservables would need to be twice as important as the observables to produce a treatment effect of zero.

This approach is akin to the robustness statements suggested by AET. (The calculation will be different since their test produces a value of δ under the null that $\beta = 0$, whereas the calculation in this section is correct for the true β .) They suggest that a value of $\delta = 1$ may be an appropriate cutoff. A value of $\delta = 1$

suggests the observables are at least as important as the unobservables. One reason to favor this is that researchers typically focus their data collection efforts (or their choice of regression controls) on the controls they believe *ex ante* are the most important (Angrist and Pischke 2010). A second is that W_2 is residualized with respect to ω^0 so, conceptually, we want to think of the omitted variables as having been stripped of the portion related to the included variables.

This suggested robustness leaves open the question of what is a reasonable R_{\max} to assume in describing the identified set. I discuss this in two specific empirical contexts in Section 4 and in more detail in the context of the economics literature in Section 5.

Bounding Statements About β . A second approach to robustness is to use bounds on R_{\max} and δ to develop a set of bounds for β . Such bounds could then be compared to, for example, a value of zero or some other value of interest.

I consider this with language similar to partial identification (Manski 2003; Tamer 2010). Consider the estimator $\beta^*(R_{\max}, \delta)$ defined as above. Without any additional assumptions, I note that R_{\max} is bounded between \bar{R} (the R -squared in the regression with controls) and 1. I assume that the proportional selection is positive, that is, that the covariance between X and the observables has the same sign as the correlation between X and the unobservables. This bounds the value of δ below at 0 and it is bounded above at some arbitrary upper bound $\bar{\delta}$.

We can then define bounds for β . One bound is $\tilde{\beta}$, the value of β delivered when $R_{\max} = \bar{R}$ or $\delta = 0$. The other bound is $\beta^*(1, \bar{\delta})$. Without more assumptions, this is either positive or negative infinity, since $\bar{\delta}$ is unbounded. The insight of partial identification is that it may be possible to use additional intuition from the problem to further bound both R_{\max} and δ .

Consider first the issue of bounding δ . I argue that for many problems $\delta = 1$ is an appropriate bound, for the reasons discussed above. Ultimately, this is an empirical issue, and I will discuss at least some evidence for this bound in Section 4.

In the case of R_{\max} , it may be possible to generate a bound smaller than 1 by, for example, considering measurement error in Y or evaluating variation in Y that cannot be related to X because it results from choices made after X is determined. Define an assumed upper bound on R_{\max} as \bar{R}_{\max} , with $R_{\max} \leq 1$.

With these two bounding assumptions, I can define a bounding set as $\Delta_s = [\tilde{\beta}, \beta^*(\bar{R}_{\max}, 1)]$.

Empirically, the question of interest in considering Δ_s is whether the conclusions based on the full set are similar to what we would draw based on observing the controlled coefficient $\tilde{\beta}$. If inclusion of controls moves the coefficient toward zero, one natural question is whether the set includes zero. Regardless of the direction of movement one could ask whether the bounds of the set are outside the confidence interval on $\tilde{\beta}$ —this effectively asks whether the conclusions based on the controlled coefficient are robust.

As above, an assumption about \bar{R}_{\max} is necessary for generating this bounding statement.

Statements About R_{\max} . If the question of interest is whether the bias-adjusted β is different from zero, and one assumes that the bound $\delta = 1$ holds, a third approach is to calculate bounds on R_{\max} . That is, researchers could report the value

of R_{\max} for which $\beta = 0$ if $\delta = 1$. This value could then be discussed in terms of whether it is plausible that the unobservables explain more of the variance than implied by this value.

Stata Code. These calculations can be performed using STATA code that accompanies this article. The command is *psacalc*.

4. EMPIRICAL VALIDATION

The results above provide a way to recover an estimate of unbiased treatment effects under the assumption that selection on observables and unobservables is proportional. However, the theoretical discussion does not provide any insight as to how this is likely to perform in empirical settings.

In this section, I explore this issue using two approaches. In the first subsection, I ask how this adjustment performs in simulated data where, by definition, we know the treatment effect. I construct the data with a full set of controls and then explore coefficient bias when various sets of controls are excluded. This allows for a test of whether the proportional selection relationship would lead to better inference in this setting, and allows for direct estimation of values of δ . The latter is helpful in evaluating the empirical validity of the bounding assumption suggested above. I perform this exercise in the familiar setting of wage returns to education.

In the second subsection, I use observational data on the relationship between maternal pregnancy and early life behaviors and child outcomes. I compute possibly biased treatment effects, perform the adjustment, and compare the resulting conclusions to external evidence on causal impacts. I ask whether the adjusted coefficients generate more accurate conclusions than the simple controlled estimates.

4.1 Simulated Data: Returns to Education

In this section, I consider validation of the estimator in real data, which is constructed such that we know the treatment effect. I use the canonical example of estimating wage returns to education.

Estimation of this relationship starts with standard Mincer regressions of wages on education, experience, and experience-squared. One central confounder is family background: people whose mothers have more education, for example, are more likely to be highly educated but also have higher wages for other reasons. (A second obvious confounder is ability. It would be possible to do an exercise similar to this one with that confounder. Since the exercise is not about finding the causal effect of education on wages, but is simply about exploring this adjustment, there is no loss to ignoring the issue of ability.) Using data from the NLSY, I construct a dataset in which I define the “true” return to education as the impact of education controlling for a full set of family background characteristics. I then consider the bias—both in simple controlled regressions and after this adjustment is performed—in hypothetical cases in which I do not observe the full set of controls. This exercise will allow me to see how the adjustment performs, to compare the performance of the simple and the general estimator and to estimate values of δ and ask how they compare to the bounds suggested in Section 3.4.

4.1.1 Data and Empirical Strategy. I use data from the NLSY-79 cohort. I am concerned with the impact of years of education on log wages, and I begin by considering the standard Mincer regression of log wages on educational attainment. I use the higher of the two educational levels recorded in 1981 and 1986 and the higher of the two wage values recorded in 1996 and 1998. Experience and experience-squared are calculated in the typical way (experience = age – education years – 6). I also control for sex.

My concern is with confounding by demographics and family background. I capture this with eight variables: region of residence, race, marital status, mother’s education, father’s education, mother’s occupation, father’s occupation, and number of siblings. All variables are controlled for fully flexibly, with dummies. Summary statistics for these data appear in Appendix B.

I construct a dataset by regressing log wages on education, experience, sex, and the full set of family background data. I generate fitted values, and then take these as the “true” effects in the model—that is, the effect on education we see in this regression is the unbiased treatment effect in the simulated data. (Clearly, this is not to suggest that this is the causal impact of education on wages. I mean only to assume that this is the true effect in the simulated data, against which I will evaluate estimates that exclude some of the controls used in constructing the effect.)

The regression of this fitted value on the full set of controls has an R -squared of 1 by construction. In practice, however, wages are not fully predicted by family background or individual characteristics. I therefore add an orthogonal error term to this fitted value. To generate a magnitude for this term I regress the log wage measure I use on log wages in 1992 or 1994 (again, I take the higher of the two). This regression has an R -squared of 0.45. I argue that family background, education, etc., should not explain more of the outcome than the previous year’s wages, since these variables all contribute to that wage. I therefore add an orthogonal error term to the fitted value such that the ultimate regression R -squared is about 0.45.

It is important to note that the addition of this error term is done largely for realism; it will be instructive to explore errors that may be introduced by incorrectly assuming that $R_{\max} = 1$. However, the calculations of δ is not sensitive to this addition.

Given this dataset, the empirical exercise is straightforward. I iterate through excluding all sets of controls (up to 6 of the 8). In each case I: (1) calculate the δ implied by the included and excluded control set; (2) calculate β^* with this δ and the true R_{\max} ; and (3) calculate whether the set bounded by $\tilde{\beta}$ and $\beta^*(R_{\max}, 1)$ contains the true effect.

4.1.2 Results. Figure 3(a) shows the distributions of the true β and the estimated $\tilde{\beta}$ and the values of β^* . The true effect in the simulated data is 0.087, with a standard error of 0.003. The β^* values cluster at the true effect value. This is a simple numerical check of the procedure: if we know the true R_{\max} and the true δ the adjustment works as it should. Not surprisingly, the estimates of $\tilde{\beta}$ are shifted substantially to the right of the true β . The estimates of β from the regressions with controls are systematically biased upward.

Figure 3(b) shows the values of δ calculated in this exercise. This value is not mechanical: nothing in the setup constrains any

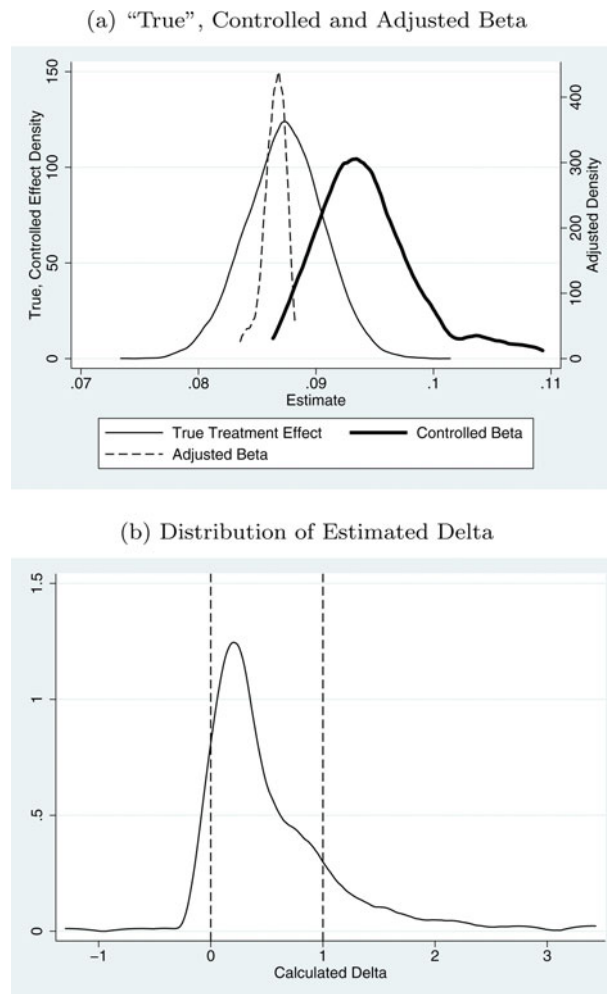


Figure 3. NLSY wage data simulation. (a) “True,” controlled, and adjusted beta. (b) Distribution of estimated delta. These figures show results from the validation using the constructed NLSY wage dataset. The analysis is described in Section 5.

particular value of δ . In the figure, I show the full distribution of δ and the $[0,1]$ bounds that I suggest would be appropriate in many settings.

The average δ is 0.545 and 86% of values fall within the $[0,1]$ range. Only 2 (of 211) values are negative. The cases with values of $\delta > 1$ are instructive. These are combinations of controls where the index of the omitted variables are more important in explaining education than the included ones. Of the 28 cases with $\delta > 1$, 92% of them excluded either maternal or paternal education. This makes clear that these variables are among the most important confounders; this should not be surprising and, indeed, it seems likely that researchers would think to include these first, before considering data on (e.g.) parental occupation or number of siblings. Put differently, if we consider control set selection not at random as I do in this example but with the idea that the most important controls are selected first, it is likely that the $[0, 1]$ bound would fit in an even larger share of cases. The fact that the average δ is less than 1 supports the idea of 1 as an upper bound on δ , rather than as an average value.

I can comment on the bounding logic described in Section 4. Given the δ values, it is straightforward to observe that if we

calculate the set $[\tilde{\beta}, \beta^*(R_{\max}, 1)]$, in 89% of cases this will include the true value. This is an improvement over the regression with observed controls. The naive estimate with controls captures the true value of β only 62% of the time. It is worth saying that if I were to use a value of $R_{\max} = 1$ to do these calculations the adjustment would be too large.

A final approach to evaluating performance is to look more directly at the likely control set in a more limited data source. The full set of demographic controls used in the analysis above are region of residence, race, marital status, mother's education, father's education, mother's occupation, father's occupation, and number of siblings. The data on parents and siblings is available in the NLSY because of the panel nature of the data. Such data would not be available in many cross-sectional datasets, or even in panel datasets with more limited information on early life.

It is simple to consider what we would conclude if we saw only the more standard controls of region, marital status, and race (these would, e.g., be available in datasets like the SIPP or the CPS). The coefficient estimate with these controls only is 0.107, versus the true effect of 0.087. The value of δ which matches the true effect is 0.79 and if we assume a value of $\delta = 1$ the adjusted coefficient is 0.082. The $\delta = 1$ bound would, therefore, include the true estimate of δ .

4.2 Observational Data: Maternal Behavior and Child Outcomes

A second approach to validation is to take a setting in which we have some possibly biased observational relationships and we think we have a sense of the causal effect from external sources. Given this, the question is whether this approach can separate causal from noncausal associations. (Altonji, Elder, and Taber (2008) did a version of this for the relationship between survival and catheterization.)

In this section, I undertake this type of validation exercise in the context of the link between maternal behaviors, infant birth weight, and child IQ. These relationships are of some interest in economics, and of wider interest in public health and public policy circles. A literature in economics demonstrates that health shocks while children are in the womb can influence early outcomes and later cognitive skills (e.g., Almond and Currie 2011; Almond and Mazumder 2011). A second literature, largely in epidemiology and public health, suggests that even much smaller variations in behavior—occasional drinking during pregnancy, not breastfeeding—could impact child IQ and birth weight. These latter studies, in particular, are subject to significant omitted variable concerns, largely associated with omitted socioeconomic status and family background. I consider five relationships in all: the relationship between child IQ and breastfeeding, drinking during pregnancy, and low birth weight/prematurity, and the relationship between birth weight (as the outcome) and maternal drinking and smoking in pregnancy.

4.2.1 Data. I use NLSY data, this time from the Children and Young Adult sample, which has information on the children of NLSY participants. I measure IQ with PIAT test scores for children 4 to 8 and birth weight (in grams) as reported

by the mother. In the latter analysis, I include all children. In all cases I control for child sex and, with IQ, for their age. These are not considered as part of the confounding set.

The IQ treatments are: months of breastfeeding, any drinking of alcohol in pregnancy, and an indicator for being low birth weight and premature (<2500 g and <37 weeks of gestation). The birth weight treatments are maternal smoking and drinking intensity during pregnancy. I measure family background, the confounding category, with child race, maternal age, maternal education, maternal income, and maternal marital status. Summary statistics for these data appear in Appendix B.

4.2.2 Empirical Strategy. I run regressions with and without the controls to extract $\hat{\beta}$, \hat{R} , $\tilde{\beta}$, and \tilde{R} . I adopt a bounding value for R_{\max} drawn from within sibling correlations (Mazumder 2011). In theory, R_{\max} should reflect how much of the variation in child IQ and birth weight could be explained if we had full controls for family background; I argue this is the thought experiment approximated by the sibling fixed effect R -squared. The figures are 0.61 for IQ and 0.53 for birth weight.

Given this R_{\max} bound, I first calculate the set $[\tilde{\beta}, \beta^*(R_{\max}, 1)]$. I also find the value of δ that would produce $\beta = 0$ under the assumed R_{\max} and compare this to $\delta = 1$. These two analyses effectively contain the same information.

The conclusions from these robustness calculations are compared to the conclusions we would expect to get if we were able to estimate the full model. To ask whether the adjusted coefficient leads to the correct answer, it is necessary to know what this answer is.

I use two types of evidence. First, I consider external evidence from randomized trials (where available) and meta-analyses. Randomized evidence suggests that breastfeeding is not linked with full-scale IQ (Kramer et al. 2008) and most evidence does not suggest an impact of occasional maternal drinking on child IQ (see, e.g., O'Callaghan et al. 2007; Falgreen-Eriksen et al. 2012). (Although the question of whether occasional maternal drinking lowers IQ is a controversial issue, as I show below the observational data actually estimates *positive* impacts of maternal drinking on IQ, and the fact that those effects are not causal is not a subject of much debate.) In contrast, low birth weight and prematurity do seem to be consistently linked to low IQ (Salt and Redshaw 2006), a link that also has a biological underpinning (de Kieviet et al. 2012). Occasional maternal drinking is typically not thought to impact birth weight (Henderson, Gray, and Brocklehurst 2007), but there is better evidence that smoking does (e.g., from trials of smoking cessation programs as in Lumley et al. 2009).

Second, I consider the conclusions one would draw from sibling fixed effects regressions in the NLSY data described above, which provides a more "within sample" test of fully controlling for family background. Of course, sibling fixed effects estimates may be subject to their own concerns about causality, so it is perhaps comforting that the conclusions are the same using either approach.

4.2.3 Results. Table 3 reports the results: Panel A shows results on IQ, Panel B on birth weight.

The first column shows treatment effects, standard errors, and R -squared values without the socioeconomic status controls. Column 2 shows similar values with the full control set.

Table 3. Maternal behavior, child IQ, and birth weight

Panel A: Child IQ, standardized (NLSY) ($R_{\max} = 0.61$)						
Treatment variable	(1) Baseline effect (Std. error), [R^2]	(2) Controlled effect (Std. error), [R^2]	(3) Null reject? (extrnl. evid.)	(4) Sibling FE estimate	(5) Identified set	(6) $\tilde{\delta}$ for $\beta = 0$ given R_{\max}
Breastfeed (Months)	0.045*** (0.003) [0.045]	0.017***(0.002) [0.256]	No	-0.007 (0.005)	[-0.033,0.017]	0.37
Drink in Preg. (Any)	0.176***(0.026) [0.008]	0.050** (0.023) [0.249]	No	0.026 (0.036)	[-0.146,0.050]	0.26
LBW + Preterm	-0.188*** (0.057) [0.004]	-0.125*** (0.050) [0.251]	Yes	-0.111 (0.070)	[-0.124,-0.033] [†]	1.37
Panel B: Birth weight in grams (NLSY) ($R_{\max} = 0.53$)						
Treatment variable	Baseline effect (Std. error), [R^2]	Controlled effect (Std. error), [R^2]	Null reject? (extrnl. evid.)	Sibling FE estimate	Identified set	$\tilde{\delta}$ for $\beta = 0$ given R_{\max}
Smoking in Preg	-183.1*** (12.9) [0.31]	-172.5*** (13.3) [0.35]	Yes	-94.3*** (27.6)	[-172.5,-30.3] [†]	1.08
Drink in Preg. (Amt)	-16.7*** (5.15) [0.30]	-14.1*** (5.06) [0.34]	No	-1.53 (7.48)	[-14.1,0.49]	0.96

NOTES: This table shows the validation results for the analysis of the impact of maternal behavior on child birth weight and IQ. Baseline effects include only controls for child sex and (1) age dummies in the case of IQ and (2) gestation week in the case of birth weight. Full controls: race, age, education, income, marital status. Sibling fixed effects estimates come from NLSY in all panels. The identified set in Column (5) is bounded below by β and above by β^* calculated based on R_{\max} given in the top row of each panel and $\tilde{\delta} = 1$. Column (6) shows the value of $\tilde{\delta}$ which would produce $\beta = 0$ given the values of R_{\max} reported in the title of each Panel. * significant at 10% level, ** significant at 5% level, *** significant at 1% level. [†]identified set excludes zero.

More breastfeeding is associated with higher IQ in these regressions, and low birth weight is associated with lower child IQ. More maternal drinking appears in these data to be associated with *higher* child IQ later, a finding that is likely to be due to selection given limited biological mechanisms. Both samples show smoking and drinking are associated with lower birth weight. All analyses reported show significant effects with the full set of controls. Interpreting these results in a naive way, one would conclude that each has a significant link with child outcomes.

Column 3 reports whether external evidence, summarized above, suggests a causal impact. As noted, external evidence supports a relationship between low birth weight and IQ and between smoking and low birth weight but the other relationships do not have broad support. Column 4 shows sibling fixed effects regressions, which result in similar conclusions. The only difference is in the impact of low birth weight on child IQ, where the NLSY regression coefficient is significant only at the 11% level.

Column 5 shows the bounding set, using the R_{\max} estimates in the top row of each panel and $\tilde{\delta} = 1$. This procedure performs well. The two cases in which the identified set does not include zero are those where the external evidence suggest significant results. Put differently, if one were to use the rule of accepting the effect as causal only if the identified set excluded zero, this would lead to the same conclusions as the external evidence. In all cases the identified set includes the sibling fixed effect estimates. In Column 6, I calculate the values of $\tilde{\delta}$ such that $\beta = 0$. I show that the effects confirmed in external data are those for which $\tilde{\delta} > 1$ is necessary to produce $\beta = 0$.

There are two final points to make about this analysis. First, similar to the wage analysis above, the average value of $\tilde{\delta}$ which matches the adjusted effects to the sibling fixed effect values is less than 1—it is 0.47—pointing to the value of 1 as a bound. Second, doing these calculations with a value of $R_{\max} = 1$ as the bound would lead us to reject all the associations—including the two which are confirmed in outside data.

The results in this section suggest that the robustness framework performs well. It also makes clear the importance of doing formal bias calculations. In this latter example, if we based our analysis only on the size (say, in percent terms) of the coefficient movements we would conclude that the link between drinking and low birth weight is much more robust than the link between low birth weight and IQ, since the former moves only by 10% and the latter by 30%. In fact, the low birth weight and IQ link has more external support. This is confirmed by the identified set conclusions, and mechanically it is reflective of the much larger change in R -squared in the low birth weight—IQ relationship.

5. APPLICATION TO ECONOMICS LITERATURE

I now return to the application of this approach within economics, using the data described in Section 2. I undertake two exercises. First, I return to the sample of articles discussed in Section 2 and the question of how robust these claims are to the formal adjustment. I calculate bias-adjusted treatment effects under the assumption that $R_{\max} = 1$ but also under varying values of R_{\max} . The discussion in Section 4 suggests that $R_{\max} = 1$ may lead to over-adjustment in many cases.

In the second subsection, I discuss how R_{\max} may be chosen in practice. I detail one approach based on an analysis of results from randomized treatment effects.

5.1 Robustness of Coefficient Stability Claims

The data in this section is the same as that used in Section 2. Recalling that discussion: I begin with the universe of all articles in the *American Economic Review*, *Quarterly Journal of Economics*, *The Journal of Political Economy*, and *Econometrica* from 2008–2010 with at least 20 citations in the ISI Web of Science, and those from 2011–2013 in the same journals with at least 10 citations. I limit the sample to articles with replication files available so it is possible to do further robustness

calculations. From these articles, I extract all results where the researcher explores the sensitivity of the result to a control set and, using a close reading of the articles, those in which this exercise appears to be designed to address an imperfect set of controls. The sample (full citation list in Appendix C) includes 27 articles with 76 total results.

The empirical exercise is as follows. I extract the relevant inputs from replication files. Note that in cases where controls are included sequentially, I compare the fewest-controls to the most-controls set. For each result, I calculate the bias-adjusted treatment effect with $\delta = 1$ and varying values of R_{max} .

I consider $R_{max} = 1$ as one bound. I also consider a parameterization of R_{max} as a function of \tilde{R} : $R_{max} = \min\{\Pi\tilde{R}, 1\}$ with varying values of Π . This function assumes that the degree of variation accounted for by the observables (including the treatment) is informative as to the degree accounted for by the unobservables.

Having calculated the identified set using these R_{max} values, I consider two standards for robustness. First, focusing on the subset of results for which the inclusion of controls moves the coefficient toward zero, I ask whether the set includes zero. I also consider whether the bounds of the set fall within ± 2.8 standard errors of the controlled estimate, an analysis that can be done by including results where controls move the coefficient away from zero (± 2.8 standard errors is the bounds of the 99.5% confidence interval). This second standard captures a test of whether the size of the estimate from the regression with controls is similar to the bias-adjusted estimate.

I summarize the robustness of a given relationship with the largest value of Π for which the result survives the robustness standard. The results appear in Figure 4(a) and 4(b). Figure 4(a) shows the primary robustness with rejection of zero; Figure 4(b) uses all results and shows the magnitude test. These graphs show the share of relationships, which would survive varying values of Π , with $R_{max} = \min\{\Pi\tilde{R}, 1\}$. In either case, I find only about 9% to 16% of results would survive $R_{max} = 1$. This share is smaller than the share implied by the AET calculation in Section 2 since the analysis does not use the null of $\beta = 0$ but, instead, calculates the δ value with the true β . Within the set of results which would not survive this standard, there is a wide range of robustness. Some of these results would not survive even quite small differences between \tilde{R} and R_{max} .

To quantify this, Panel A of Table 4 shows the share of results which would survive $R_{max} = 1$ and three values of Π . About 40% of results would not survive $\Pi = 1.25$. Considering the rejection-of-zero robustness criteria, within this set that is not robust to $\Pi = 1.25$, the average study fails at a value of $\Pi = 1.15$ or, in point estimate terms, a predicted increase in R -squared of 0.06 with inclusion of unobserved controls.

In nearly all of the analyses discussed in this section, the authors discuss only coefficient movements. As noted, this is potentially misleading for two reasons. First, it fails to take into account the R -squared movements. Second, as I note in Section 3.3.5, it is possible that coefficient stability is misleading even in the context of large R -squared movements in some cases, particularly if the assumption of proportionality that underlies the restricted estimator described is seriously violated. I explore the importance of these two issues empirically.

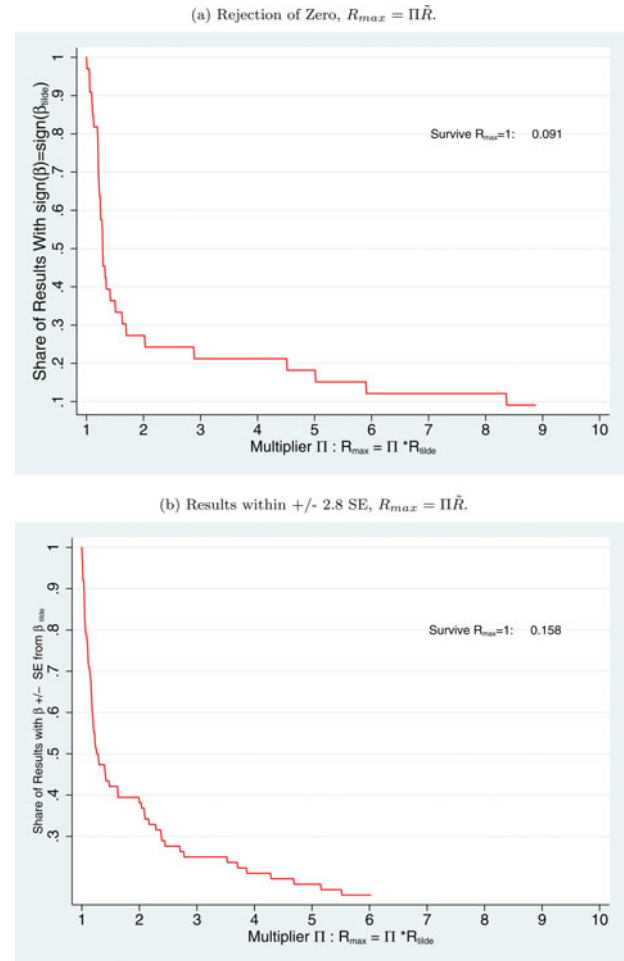


Figure 4. Robustness of stability results in Economics literature. (a) Rejection of zero, $R_{max} = \Pi\tilde{R}$. (b) Results within ± 2.8 SE, $R_{max} = \Pi\tilde{R}$. These graphs show the performance of nonrandomized results under the proportional selection adjustment. Each figure graphs the share of results that would survive varying parameterization of R_{max} , in all cases assuming $\tilde{\delta} = 1$. Subfigure (a) indicates the share of results that would survive $R_{max} = \Pi\tilde{R}$ for varying values of Π , with survival in this case meaning the identified set does not include zero. This figure contains only relationships where the effect is significant with controls and adding the controls moves the coefficient toward zero. Subfigure (b) indicates the share of results for which the full identified set would be within 2.8 standard errors of the controlled coefficient. This subfigure includes all relationships.

First, consider how the conclusions would differ from those which rely *only* on coefficient movements. To explore this, I choose an $R_{max} = 1.3\tilde{R}$ cutoff and compare the percent reduction (in absolute value) in coefficients for results that do and do not survive this cutoff. I choose this value because it will be the cutoff I identify later in the analysis of randomized data. Figure 5 shows these results. There is full overlap in the distributions of coefficient movements between robust and nonrobust results, illustrating the fact that coefficient movements alone do not provide much insight about these.

Second, I calculate the bias implied by the restricted estimator in each case. The restricted estimator accounts for movements in R -squared values but makes the strong proportionality

Table 4. Robustness of stability results

Panel A: Nonrandomized data, share of results, which survive $\tilde{\delta} = 1$, varying R_{\max}				
	(1) $R_{\max} = 1$	(2) $R_{\max} = \min(2\tilde{R}; 1)$	(3) $R_{\max} = \min(1.5\tilde{R}; 1)$	(4) $R_{\max} = \min(1.25\tilde{R}; 1)$
Share with adjusted β same sign as $\tilde{\beta}$ <i>Sample: Add controls, moves toward zero</i>	9.1%	27%	36%	57%
Share with adjusted $\beta \pm 2.8$ SE of $\tilde{\beta}$ <i>Sample: All</i>	16%	38%	42%	51%
Panel B: Randomized data, share of results, which survive $\tilde{\delta} = 1$, varying R_{\max}				
	$R_{\max} = 1$	$R_{\max} = \min(2\tilde{R}; 1)$	$R_{\max} = \min(1.5\tilde{R}; 1)$	$R_{\max} = \min(1.25\tilde{R}; 1)$
Share with adjusted β same sign as $\tilde{\beta}$ <i>Sample: Add controls, moves toward zero</i>	42%	82%	91%	97%
Share with adjusted $\beta \pm 2.8$ SE of $\tilde{\beta}$ <i>Sample: All</i>	37%	82%	86%	91%

NOTES: This table describes the survival of nonrandomized (Panel A) and randomized (Panel B) results under the proportional selection adjustment. Both panels show the share of results that would survive $\tilde{\delta}$ with varying R_{\max} values. I consider two definitions of survival: (1) the identified set does not include zero and (2) the outer bound of the set is within 2.8 standard errors of $\tilde{\beta}$. The first of these is considered only for results that move toward zero when controls are added.

assumption, so the difference in the effect implied by this restricted version and the true bias-adjusted effect gives a sense of how misleading conclusions may be if a simpler version of the adjustment is used. I do the calculation assuming, again, that $R_{\max} = 1.3\tilde{R}$. In about 80% of cases one would draw the correct conclusion about robustness from the restricted estimator. However, the restricted version generally understates the bias, on average by about 40%. The error is larger in cases where much of the treatment variance is explained by the controls.

5.1.1 Example. To be more concrete on the issues above, I illustrate with an example.

Nunn and Wantchekon (2011) analyzed the impact of the slave trade on mistrust in Africa. This is a salient example because the authors worry explicitly about unobserved

differences across areas, and present a number of arguments to support the interpretation of their results as causal. In contrast to most articles in this literature, they undertake direct calculations based on the theory in AET. They use coefficient movements in their regressions to calculate the value of δ which would be required to produce $\beta = 0$. They argue the results are robust because all the calculated values of δ are greater than 1. Equivalently, the adjusted treatment effects have the same sign as the effects in regressions with controls if $\delta = 1$.

Although it is not made explicit, the calculations they undertake in the article assume that $R_{\max} = \tilde{R} + (\tilde{R} - \hat{R})$. (They draw this from Bellows and Miguel (2009).) In other words, they assume that the unobservable controls explain as much of the outcome as the observable controls. In practice, the R -squared values in their regressions do not move much; as an example, in the first row of their Table 4, considering the “Trust Relatives” measure, adding controls increases R -squared from 0.115 to 0.133. Their adjustment assumes that the R -squared in the regression with controls would be 0.151. This change is quite small, and it seems reasonable to explore the impact on the results from changing it.

A set of results from their Table 4 are reported in Table 5. The first two columns show their estimated effects, and the third column shows the estimated β using their implicit assumption on R_{\max} and $\delta = 1$. Three of the five results have adjusted β values the same sign as the estimated effects; two differ in sign. This result differs from the result in the article because the robustness in the article is done under the null of $\beta = 0$. On its face, this suggests that the results are less robust than implied in the article even with the implicit assumption used.

Column (4) estimates these bias-adjusted β s using the assumption that $R_{\max} = 1.3\tilde{R}$. As noted above, this cutoff is derived from randomized data later, so it serves as a focal point. Although the R_{\max} values implied by this cutoff are still fairly small (e.g., in the first row the implied R_{\max} is just 0.169) only one of the five results are robust to this assumption. Column (5) calculates the value of R_{\max} at which each of the results fail. In

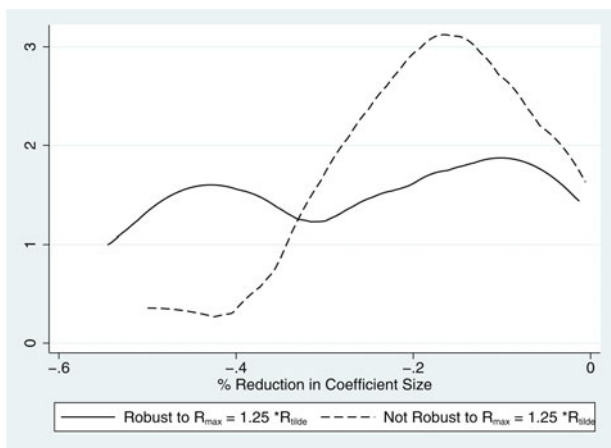


Figure 5. Relationship between full robustness and coefficient movement. This graph shows the range of coefficient movements in nonrandomized studies. These studies are divided into those that are robust to the proportional selection adjustment with $R_{\max} = 1.3\tilde{R}$ (solid line) and those that are not (dotted line). This includes only relationships in which the inclusion of controls moves the coefficient toward zero.

Table 5. Example: Nunn and Wantchekon (2011)

Result description	(1) Baseline effect (Std. error)[R^2]	(2) Controlled effect (Std. error)[R^2]	(3) Bias-adjusted β $R_{\max} = \tilde{R} + (\tilde{R} - \hat{R})$	(4) Bias-adjusted β $R_{\max} = 1.3\tilde{R}$	(5) Max R_{\max} for $\beta < 0$
Trust relatives	-0.193 (0.043) [0.106]	-0.178 (0.031) [0.130]	-0.103	0.352	0.161
Trust neighbors	-0.238 (0.044) [0.115]	-0.202 (0.029) [0.159]	-0.069	-0.044	0.210
Trust local council	-0.177 (0.027) [0.175]	-0.128 (0.021) [0.205]	0.046	0.821	0.230
Intragroup trust	-0.208 (0.041) [0.121]	-0.187 (0.032) [0.155]	-0.091	0.100	0.197
Intergroup trust	-0.145 (0.031) [0.093]	-0.115 (0.030) [0.119]	0.010	0.194	0.142

NOTES: This table shows the results from Nunn and Wantchekon (2011, Table 4). The first columns show the baseline and controlled effects. Columns (3) through (5) show the bias-adjusted β under various assumptions on R_{\max} . Column (3) uses the assumption from their article. Column (4) uses an alternative assumption, which is based on the conclusions from the randomized data. Column (5) estimates the maximum value of R_{\max} for which the result would survive in each case.

general, these values are fairly low and suggest that even if the unobservables play a fairly small role in explaining the outcome, the results may not be robust by this test.

In addition to showing the importance of the formal adjustment in analyzing coefficient stability, this example illustrates one way that robustness might be explored in this context. Rather than assuming a value for R_{\max} it would be feasible to explore a range of values and report, for example, the value of R_{\max} for which the result is no longer robust (if $\delta = 1$). Comparing this to the R -squared in the regression with controls, authors and readers can discuss the concept of robustness more concretely.

An alternative approach is to try to generate some general guidelines about R_{\max} . Below, I suggest one approach to this, using data from randomized trials.

5.2 Evidence on Stability Cutoffs from Randomized Data

The evidence above makes clear that, even within a sample of articles which argue for coefficient stability, there is a lot of variation in the robustness of results depending on R_{\max} . A natural following question is whether we can suggest any guidance about where one might draw the line—specifically, is there some value of Π (where $R_{\max} = \Pi\tilde{R}$) above which we should consider a result robust?

I argue that one place to look for such guidance is in reports from randomized data. Randomized experiments are becoming increasingly common within economics and articles reporting results of these experiments often include regressions with and without controls. Sometimes these are explicitly used to test balance in the experiment, although it is also commonly done to increase precision. Assuming that the data are correctly randomized, if the sample size were infinite, the effects would not be expected to move at all. In practice, with finite data, coefficients can move a bit simply due to very small differences across groups.

When nonrandomized articles invoke a coefficient stability heuristic to argue the results they observe are causal, they are (perhaps implicitly) suggesting that the treatment is as good as random. Including controls does not change the coefficient because there is no confounding; this is exactly the argument we know holds in randomized cases. Given this, I argue we

can use the stability of randomized data as a guide to how much stability we would expect in nonrandomized data *if the treatment were assigned exogenously*: is the coefficient stability within the range the researcher would expect with a randomly assigned treatment?

The approach in this section is to assume that the effects estimated in randomized data are causal and to therefore assume that they should survive this adjustment procedure. (An obvious concern is that, perhaps, these articles are not correctly randomized. This would lead me to a standard which was too lax. I address this in two ways. First, I have focused on articles published in highly ranked journals, increasing the chance that the randomization was of high quality. Second, I will draw guidelines that fit nearly all but not all articles, thus accepting that a small share of randomized articles may suffer from true lack of balance and should not be used to guide this approach.) I then ask what value of Π in the R_{\max} parameterization would make this true.

The baseline set of articles for this analysis is all randomized articles (lab or field) published in the *American Economic Review*, *Quarterly Journal of Economics*, *Journal of Political Economy*, *Econometrica*, and the *American Economic Journal – Applied Economics* in the period 2008 through 2013. (I include AEJ-Applied because it has published a large number of experimental articles. This journal began publishing in 2009.) I extract from these all articles that report sensitivity of a treatment effect to controls. In cases where there are multiple effects reported (i.e., multiple outcomes), I include all effects. I use replication files or researcher requests to extract the estimator inputs where possible. The final sample includes 65 results.

The full set of references is in Appendix C.

I undertake the same analysis as in the nonrandomized data: calculate the bias-adjusted treatment effect assuming $\delta = 1$ and varying R_{\max} and compare the results to the two standards for robustness.

Figure 6(a) and 6(b) shows the distributions of sensitivity for the randomized data. A first thing to note is that these results are more robust than the nonrandomized results. About 40% of randomized results would survive a cutoff of $R_{\max} = 1$. Nearly all would survive a cutoff of $R_{\max} = 1.25\tilde{R}$, much greater than for the nonrandomized results. Panel B of Table 4 shows the survival shares for this dataset explicitly under the varying R_{\max} cutoffs.

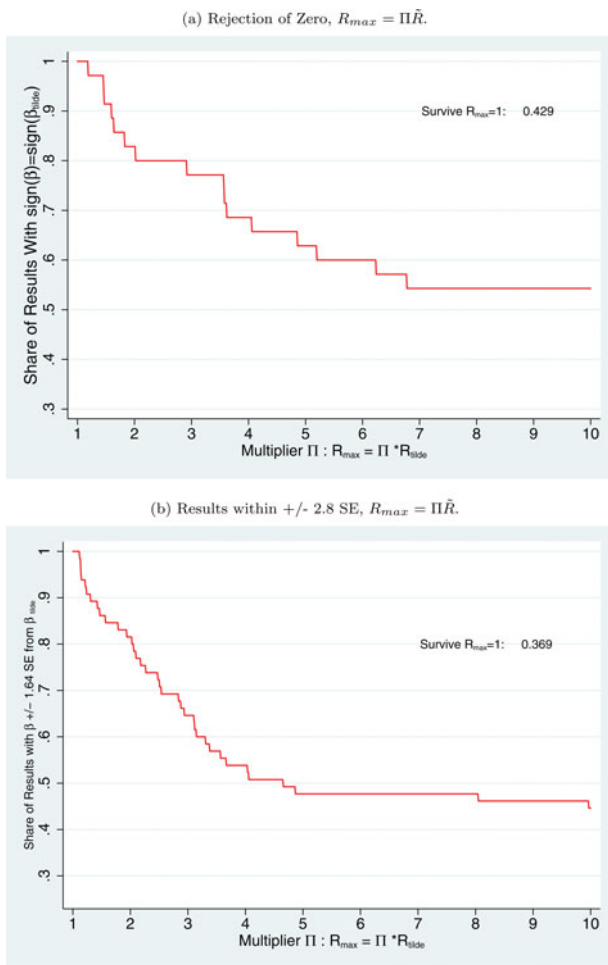


Figure 6. Results from randomized data. (a) Rejection of zero, $R_{\text{max}} = \Pi \tilde{R}$. (b) Results within ± 2.8 SE, $R_{\text{max}} = \Pi \tilde{R}$. These graphs show the performance of randomized results under the proportional selection adjustment. Each figure graphs the share of results that would survive varying parameterizations of R_{max} , in all cases assuming $\tilde{\delta} = 1$. Subfigure (a) indicates the share of results that would survive $R_{\text{max}} = \Pi \tilde{R}$ for varying values of Π , with survival in this case meaning the identified set does not include zero. This figure contains only relationships where the effect is significant with controls and adding the controls moves the coefficient toward zero. Subfigure (b) indicates the share of results for which the full identified set would be within 2.8 standard errors of the controlled coefficient. This subfigure includes all relationships.

It is not surprising that the randomized results are more robust. The fact that they do not all survive $R_{\text{max}} = 1$ is because even small changes in coefficients can be blown up with this assumption. I use these data to develop robustness cutoff values. I base these on the value of Π which would allow 90% of results to survive in both the confidence interval and the rejection of zero test. This leads to the bounding values of $\Pi = 1.3$. This value suggests a bound where the unobservables explain somewhat less than the observables (where the latter includes the treatment). This has some intuitive appeal if we think that the observables are chosen with an eye to those which are most important in explaining the outcome.

Under this approach, to argue for a level of stability consistent with randomized treatment, researchers should consider

whether the set $[\tilde{\beta}, \beta^*(\min\{1.3\tilde{R}, 1\}, 1)]$ excludes zero or, equivalently, that the δ which produces $\beta = 0$ with $R_{\text{max}} = 1.3\tilde{R}$ exceeds 1. Applying this to the nonrandomized data above, I find that 45% of results would survive this standard. This standard could be valuable to explore even in cases where the controls cause the coefficient to move away from zero; in that case the question would be whether considering the full set would lead to very different conclusions than the controlled estimate.

6. CONCLUSION

Coefficient stability is a commonly invoked argument against omitted variable bias. In fact, such stability is informative only if authors also consider the importance of the controls in explaining the variance of the outcome. I connect the heuristic to the idea of a proportional selection relationship on observed and unobserved variables. I describe a tractable strategy for generating bounds on treatment effects and show validation in empirical contexts. Importantly, I show that most existing work does not consider these issues explicitly, and I argue that many results with “stable coefficients” are not very robust.

I suggest a standard for robustness relying on this estimator that could be easily implemented by researchers. A key issue is the need to make an assumption about the share of the variance of the dependent variable which is jointly explained by the observed and unobserved variables that are correlated with it. I suggest a standard based on the performance of this estimator in randomized data.

This provides one general approach to developing intuition about R_{max} but it is worth noting that within a given context it may be possible to develop a better intuition. Some examples of this are provided earlier in the article. In the case of education and wages, I develop a value of R_{max} (used for constructing the data) by looking at how much of current year wages are explained by past year wages; the theory is that any ability/motivation/family background confounders are determined prior to the previous year’s wages. In the analysis of maternal behavior and child outcomes I use sibling correlations as a benchmark since sibling share the same family background. In two articles following on their original article (Altonji, Elder, and Taber 2005b; Altonji et al. 2008) Altonji and coauthors suggested two methods for adjusting for idiosyncratic variance, an approach parallel to my use of R_{max} .

The core insight is to recognize that coefficient stability on its own is at best uninformative and at worst very misleading. It must be combined with information about R -squared movements to develop an argument.

The robustness approach in this article addresses concerns related to unobservables that are related to the observables. A key issue that must still be addressed is the appropriate choice of observables (as discussed in Angrist and Pischke 2010). If there are unobservables related to the treatment for which we cannot learn about this relationship using the relationship between treatment and observed controls then this result breaks down. Recognizing this issue may help improve the control sets used in empirical work.

SUPPLEMENTARY MATERIALS

Supplementary materials are three appendices. Appendix A: Theoretical results, including proofs. Appendix B: Additional tables and figures. Appendix C: List of citations for Section 5.

ACKNOWLEDGMENTS

Ling Zhong, Unika Shrestha, Damian Kozbur, Guillaume Pouliot, David Birke and Angela Li provided excellent research assistance. The author thanks David Cesarini, Raj Chetty, Todd Elder, Amy Finkelstein, Guido Imbens, Larry Katz, Jonah Gelbach, Matt Gentzkow, Matt Notowidigdo, Chad Syverson, Manisha Shah, Azeem Shaikh, Jesse Shapiro, Bryce Steinberg, Matt Taddy, Heidi Williams, and participants in seminars at Brown University, University of Chicago Booth School of Business, Wharton and Yale for helpful comments. The author is grateful to a number of authors for providing replication files or rerunning analysis by request. The author gratefully acknowledges financial support from the Neubauer Family. *Stata* code to perform the calculations described in this article is available from the author's website or through ssc under the name *psacalc*.

[Received March 2016. Revised August 2016.]

REFERENCES

- Almond, D., and Currie, J. (2011), "Killing Me Softly: The Fetal Origins Hypothesis," *Journal of Economic Perspectives*, 25, 153–172. [198]
- Almond, D., and Mazumder, B. (2011), "Health Capital and the Prenatal Environment: The Effect of Ramadan Observance During Pregnancy," *American Economic Journal: Applied Economics*, 3, 56–85. [198]
- Altonji, J., Conley, T., Elder, T., and Taber, C. (2011), "Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables," Mimeo, Yale University. [187,189]
- Altonji, J. G., Elder, T. E., and Taber, C. R. (2002), "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schools," NBER Working Paper No. 9358. [188,189,193]
- (2005a), "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *Journal of Human Resources*, 40, 791–821. [187,188,189]
- (2005b), "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184. [203]
- (2008), "Using Selection on Observed Variables to Assess Bias From Unobservables When Evaluating Swan-Ganz Catheterization," *American Economic Review*, 98, 345–350. [198,203]
- Angrist, J. D., and Pischke, J.-S. (2010), "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24, 3–30. [187,189,191,196,203]
- Bellows, J., and Miguel, E. (2009), "War and Local Collective Action in Sierra Leone," *Journal of Public Economics*, 93, 1144–1157. [191,201]
- Chiappori, P.-A., Oreffice, S., and Quintana-Domeque, C. (2012), "Fatter Attraction: Anthropometric and Socioeconomic Matching on the Marriage Market," *Journal of Political Economy*, 120, 659–695. [187]
- Dehejia, R., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062. [189]
- de Kieviet, J. F., Zoetebier, L., van Elburg, R. M., Vermeulen, R. J., and Oosterlaan, J. (2012), "Brain Development of Very Preterm and Very Low-Birthweight Children in Childhood and Adolescence: A Meta-Analysis," *Developmental Medicine & Child Neurology*, 54, 313–323. [198]
- Falgreen-Eriksen, H. L., Mortensen, E. L., Kilburn, T., Underbjerg, M., Bertrand, J., Stavring, H., Wimberley, T., Grove, J., and Kesmodel, U. S. (2012), "The Effects of Low to Moderate Prenatal Alcohol Exposure in Early Pregnancy on IQ in 5-Year-Old Children," *BJOG*, 119, 1191–1200. [198]
- Gelbach, J. B. (2016), "When do Covariates Matter? And Which Ones, and How Much?" *Journal of Labor Economics*, 34, 509–543. [189]
- Heckman, J., and Hotz, J. (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–874. [189]
- Henderson, J., Gray, R., and Brocklehurst, P. (2007), "Systematic Review of Effects of Low-Moderate Prenatal Alcohol Exposure on Pregnancy Outcome," *BJOG*, 114, 243–252. [198]
- Imbens, G. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132. [188,189]
- Kramer, M. S., Aboud, F., Mironova, E., et al. (2008), "Breastfeeding and Child Cognitive Development: New Evidence From a Large Randomized Trial," *Archives of General Psychiatry*, 65, 578–584. [198]
- Lacetera, N., Pope, D. G., and Sydnor, J. R. (2012), "Heuristic Thinking and Limited Attention in the Car Market," *American Economic Review*, 102, 2206–2236. [187]
- LaLonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620. [189]
- Leamer, E. (1978), *Specification Searches: Ad Hoc Inference With Nonexperimental Data*, New York: Wiley. [189,191]
- Lumley, J., Chamberlain, C., Dowswell, T., Oliver, S., Oakley, L., and Watson, L. (2009), "Interventions for Promoting Smoking Cessation During Pregnancy," *Cochrane Database of Systematic Reviews*, CD001055. [198]
- Manski, C. (2003), *Partial Identification of Probability Distributions* (Springer Series in Statistics), New York: Springer. [196]
- Mazumder, B. (2011), "Family and Community Influences on Health and Socioeconomic Status: Sibling Correlations Over the Life Course," *The B.E. Journal of Economic Analysis & Policy*, 11, 2876. [198]
- Murphy, K., and Topel, R. (1990), "Efficiency Wages Reconsidered: Theory and Evidence," in *Advances in the Theory and Measurement of Unemployment*, eds. Y. Weiss, and G. Fishelson, London: Palgrave MacMillan, pp. 204–240. [187,189]
- Nunn, N., and Wantchekon, L. (2011), "The Slave Trade and the Origins of Mistrust in Africa," *American Economic Review*, 101, 3221–3252. [188,191,201]
- O'Callaghan, F. V., O'Callaghan, M., Najman, J. M., Williams, G. M., and Bor, W. (2007), "Prenatal Alcohol Exposure and Attention, Learning and Intellectual Ability at 14 Years: A Prospective Longitudinal Study," *Early Human Development*, 83, 115–123. [198]
- Olken, B. A., and Barron, P. (2009), "The Simple Economics of Extortion: Evidence from Trucking in Aceh," *Journal of Political Economy*, 117, 417–452. [190]
- Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, Cambridge: Cambridge University Press. [189]
- Pischke, J.-S., and Schwandt, H. (2015), "Poorly Measured Confounders are Useful on the Left But Not on the Right," London School of Economics Working Paper. [189]
- Rosenbaum, P., and Rubin, D. (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society, Series B*, 45, 212–218. [189]
- Salt, A., and Redshaw, M. (2006), "Neurodevelopmental Follow-up After Preterm Birth: Follow up After Two Years," *Early Human Development*, 82, 185–197. [198]
- Tamer, E. (2010), "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195. [196]