NBER WORKING PAPER SERIES

UNOBSERVABLE SELECTION AND COEFFICIENT STABILITY: THEORY AND VALIDATION

Emily Oster

Working Paper 19054
http://www.nber.org/papers/w19054

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2013

Ling Zhong and Unika Shrestha provided excellent research assistance. I thank David Cesarini, Todd Elder, Matt Gentzkow, Chad Syverson, Azeem Shaikh, Jesse Shapiro, Matt Taddy and participants in seminar at University of Chicago Booth School of Business, Wharton and Yale for helpful comments. I gratefully acknowledge financial support from the Neubauer Family. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Unobservable Selection and Coefficient Stability: Theory and Validation
Emily Oster
NBER Working Paper No. 19054
May 2013
JEL No. C01,I1,I12

ABSTRACT

A common heuristic for evaluating the problem of omitted variable bias in economics is to look at coeffcient movements after inclusion of controls. The theory under which this is informative is one in which the selection on observables is proportional to selection on unobservables, an idea which is formalized in Altonji, Elder and Taber (2005). However, this connection is rarely made explicit and the underlying assumption is rarely tested. In this paper I first show how, under proportional selection, coeffcient movements, along with movements in R-squared values, can be used to calculate a measure of omitted variable bias. I then undertake two validation exercises. First, I relate maternal behavior on child birth weight and IQ. Simple controlled regressions give misleading estimates; estimates adjusted with a proportional selection adjustment fit significantly better. Second, I match observational and randomized trial data for 29 relationships in public health. I show that on average bias-adjusted coeffcients perform much better than simple controlled coeffcients and I suggest that a simple form of this adjustment could dramatically improve inference in many public health contexts.

Emily Oster
University of Chicago
Booth School of Business
5807 South Woodlawn Ave
Chicago, IL 60637
and NBER
eoster@uchicago.edu

1 Introduction

Concerns about omitted variable bias are common to most or all non-experimental empirical work in economics, other social sciences and the natural sciences. And although randomized experiments are common in natural sciences and becoming increasingly common within economics, the majority of empirical work in both settings is still not randomized.[1] Within economics, a common heuristic for evaluating the robustness of a result to omitted variable bias concerns is to look at the sensitivity of the treatment effect to added controls. The heuristic suggests that if a coefficient is stable as controls are added, this is a good sign that there is little remaining bias. In a review of non-structural, non-experimental empirical work published in three top economics journals[2] in 2012, 75% of papers explored the sensitivity of the results to varying control sets, and a number of these papers were quite explicit about the relationship between coefficient stability and omitted variable bias.[3]

Although it is rarely made explicit, this coefficient stability heuristic relies on the idea that the selection on observable covariates is informative about the selection on *unobservable* covariates, an idea which is formalized in Altonji, Elder and Taber (2005) and suggested informally in Murphy and Topel (1990). I will refer to this as the *proportional selection assumption*. In the context of a linear model, these papers show how this assumption can be used to calculate a causal treatment effect. Neither paper formalizes the link with coefficient movements.

The fact that the link between the proportional selection assumption and coefficient movements is not explicit in either the underlying theoretical work or in the common robustness checks creates two problems. First, the use of this as a robustness test is rarely done in the most efficient or informative way. Second, there has been little or no effort to test whether the proportional selection assumption is better than alternatives (for example, than the alternative that the unobservables are related to the treatment but there is no information provided about that relationship by the link between treatment and observables). The informativeness of robustness tests which rely on this proportional selection theory rests crucially on whether it is empirically valid.

In this paper I take up both of these issues. I begin by expanding on the theory laid out in Altonji, Elder and Taber (2005) (hence, AET) and connecting the omitted variable bias directly to coefficient movements. I provide some explicit guidance for performing a bias adjustment based on this theory. I then present two validation exercises, both of which take advantage of settings in which I observe a "true" treatment

[1] For example: in 2012 *JAMA* published 133 major research papers, only 53 of which were randomized. The *American Journal of Public Health* published 128, only 14 of which were randomized. The combination of the *American Economic Review,* the *Quartly Journal of Economics* and the *Journal of Political Economy* published 69 empirical, non-structural papers, only 11 of which were randomized.

[2] *American Economic Review, Journal of Political Economy* and *Quarterly Journal of Economics.*

[3] For example, Chiappori et al (2012) state: "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 et al (2012) state: "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."

effect matched to possibly biased estimates.

I begin in Section 2 with theory. I consider a simple setup: an outcome Y is fully determined by a treatment variable X, a vector of observable controls W, a vector of unobserved controls W' and an error term, ϵ, which is orthogonal to X,W and W'. The true effect of X on Y is β. I introduce the proportional selection assumption: $\delta\frac{Cov(W,X)}{Var(W)} = \frac{Cov(W',X)}{Var(W')}$. Since W' is unobserved, the coefficient on X in a regression of Y on X and W is biased. I demonstrate that, under the proportional selection assumption, a measure of this bias can be calculated from (1) the coefficients on X with and without controls for W; (2) the R-squared values from controlled and uncontrolled regressions; and (3) an assumption about the R-squared of a (hypothetical) regression controlling for X, W and W'.

Denote the uncontrolled coefficient ξ, the coefficient with controls Λ, the two R-squared values as R_1 and R_2 and the maximum R-squared as R_{max}. The bias on Λ is very closely approximated by $\delta\frac{(\xi-\Lambda)(R_{max}-R_2)}{(R_2-R_1)}$. Under an equal selection assumption ($\delta = 1$) this recovers the bias exactly. Under the assumption that $\delta < 1$, suggested by AET, this is a very close upper bound on the bias, allowing us to infer a lower bound on β. The intuition behind this adjustment is straightforward: if the coefficient moves a lot when the controls are added *and* there is a lot of remaining variation in Y which could be explained by related variables, the bias on Λ is large. I will refer to this procedure for recovering β as the *proportional selection adjustment.*[4]

This baseline discussion links residual omitted variable bias to coefficient movements. I explore two practical extensions. First, I consider the case in which there is a second vector of controls, M, which is correlated with X and Y, but orthogonal to W. If M is fully observed – that is, does not have any unobserved components – I show the same bias-calculation results go through when M is included in both the controlled and uncontrolled regressions. Put differently, the relevant coefficient movements are those which occur when one adds the variables which relate to the omitted variables. More generally, I show that even with remaining omitted variables this procedure can recover the effect one would estimate if W' was observed.

Second, I consider the case where the vector W has multiple components which can be included in turn. I ask whether observing that treatment effect converge as better controls are added should lead one to conclude further controls would not alter the coefficient. Although this is a common heuristic, it implies that the controlled effect equals the true treatment effect only if the R-squared simultaneously converges to the maximum R-squared.

Following the theory, I turn to two validation exercises. It is not possible to directly test the proportional selection assumption, but I argue I can test the assumption indirectly – and the methodology more generally – by asking whether estimates generated by the proportional selection adjustment are closer to the true causal effect.[5] Doing this requires an estimate of the true causal effect. Given this, validation could

[4]Bellows and Miguel (2009) derive an adjustment like this using only coefficient movements. Their derivation, however, works only in the case where the variance of W and W' are identical, which is a very strong assumption and not what is laid out in AET. The difference is discussed briefly in Section 2.

[5]Altonji et al (2008) also compare results from their adjustment to randomized results in a single case (catheterization), although

take two forms. First, I can ask whether *some* value of δ would match possibly biased regression coefficients to a true effect. Second, more constrained, I can ask whether a single value of δ might organize a number of findings within a setting or settings.

In Section 3, I consider links between maternal behavior (prenatal and early life), child birth weight and child IQ. Many studies – in economics and elsewhere – have suggested links between maternal behaviors and child outcomes, but most studies are subject to significant concerns about omitted variable bias, notably associated with socioeconomic status. I use data from the National Longitudinal Survey of Youth (NLSY) and US Natality Detail Files to estimate (1) the impact of breastfeeding, drinking in pregnancy and low birth weight/prematurity on child IQ and (2) the impact of maternal drinking and smoking on child birth weight.

I estimate regressions with and without controls for maternal socioeconomic status, and use the coefficients and R-squared values to perform the proportional selection adjustment. I then ask whether there is a value of δ which matches the adjusted β to the true β. Doing this requires an assumption on R_{max} – the amount of variation in Y which could be explained if we observed W' – and also a measure of the true β. In this case, because the the central concern is omitted family background, I draw estimates of R_{max} from published sibling correlations in IQ and birth weight and use sibling fixed effects regressions in the NLSY to estimate the true β.[6] The basic conclusions from these sibling regressions are validated with external evidence.

The proportional selection adjustment performs well. In the simplest (and weakest) validation test, I show that, in all seven relationships estimated, there is a value of δ for which the adjusted coefficient matches the true treatment effect estimate. Further, all of the estimated δ values hover around 1. Following on this, I show a stronger validation test: performing the proportional selection adjustment using a value of $\delta = 1$, with bootstrapped standard errors, would have led to much improved inference across the settings considered here.

The baseline fully controlled regressions suggest that breastfeeding and maternal drinking are associated with higher IQ (note that there is no reason to believe maternal drinking would increase IQ and this must be due to selection) and low birth weight/prematurity is associated with lower IQ. In sibling fixed effects regressions the first two relationships are very close to zero, while the low birth weight effect is still negative and marginally significant. The inferred coefficients after the proportional selection adjustment with a $\delta = 1$ are also very close to zero in the case of breastfeeding and drinking, and negative and marginally significant for prematurity. In the case of birth weight, in both the NLSY and Natality Detail Files, baseline controlled regressions show lower birth weight associated with both maternal smoking and maternal drinking; only the former is supported in sibling fixed effects regressions and, again, the adjusted coefficients reflect this. Further, even in the case of smoking and birth weight, the adjusted coefficients are much closer to the sibling fixed effects estimates, whereas the coefficients with controls overstate the effect.

Section 4 takes this approach a step further and asks whether we might use it to suggest a general form

they consider only the test of the null hypothesis rather than comparing magnitudes.

[6] Note that I will refer to this as the true β for simplicity, while accepting that it may still have bias.

of this adjustment that could improve inference in a particular setting. I consider links between positive health behaviors and health outcomes, an area of much policy interest where many existing studies suffer from omitted variable bias concerns. I combine NHANES data (for observational correlations) with randomized evidence in two settings: the relationship between exercise and a number of health measures and the relationship between vitamin D/calcium (CaD) supplementation and a similar set of measures. I generate a total of 29 treatment-outcome pairs where I can estimate a relationship in the NHANES and match the point estimate to a treatment effect from a randomized trial.

In this section I begin where I ended in Section 3, with the assumption that $\delta = 1$. I then *estimate* a value for R_{max} which would rationalize each point estimate. The first step is to ask whether there is a value of R_{max} which would match the adjusted to true β in each setting. Second, once I have estimated these values for each treatment-outcome pair, I ask whether a single value of R_{max}(or, in fact, a parametrization as a function of R_2 and R_1) could improve inference across all the settings. This exercise is in part validation but goes further: ultimately, it asks whether a procedure using *only* results from observable regressions might be used to improve inference in these settings and parallel settings.

The evidence suggest that in 22 of the 29 settings I could match the treatment point estimate with some value of R_{max}. In all of the remaining I could match a value within the randomized confidence interval (these tend to be cases where none of the coefficients – randomized or observational – are significant). Second, more importantly, I find that a single parametrization of R_{max} would improve the point estimate in 21 of 29 cases and decrease the overall error by 30%. Many of the cases are ones in which the controlled coefficients overstate the benefit of the intervention and the adjusted coefficients match the truth.

I parametrize R_{max} as a function of R_2 and R_1 which means that this adjustment can be done without seeing the R-squared values. The results suggest a value of $\beta_{adj} = \Lambda - 1.018(\xi - \Lambda)$ would, on average, be closer to the true β than Λ in these settings. I perform several out-of-sample tests and show this performs well. I argue this adjustment may be applicable to a wide swath of the public health literature where the outcome is a health outcome, treatment is a good health behavior and we see only imperfect socioeconomic status controls. This adjustment would be easy for researchers (or research consumers) to perform, and could be helpful in evaluating the plausibility of results.

2 Theory

This section outlines the theory. Section 2.1 develops the baseline result and shows the bias calculation under the assumption of proportional selection. Section 2.2 describes an extension to a case where there is an additional orthogonal category of controls. Section 2.3 discusses a closely related heuristic of looking for increasing stability of coefficients as more controls are added. Section 2.4 summarizes the practical guidance

from the theory before I take up validation in Section 3.

2.1 Baseline Result: Bias Calculation Under Proportional Selection

Consider a linear model relating an outcome Y to treatment X.

$$Y = \alpha + \beta X + W + W' + \epsilon \tag{1}$$

W and W' are indexes of control variables which are related to both X and Y. W and W' are orthogonal to each other; the researcher observes W but not W'. The final term, ϵ, is an *iid* noise term. Without loss of generality I assume the variance of X and W are equal to 1, and the variance of W' is $V_{w'}$.

The key assumption is of proportional selection: the relationship between W and X is informative about the relationship between W' and X. Formally, denote the covariance between W and X as C_{wx} and between W' and X as $C_{w'x}$. Proportional selection assumes the following equality holds for some δ :

$$\delta C_{wx} = \frac{C_{w'x}}{V_{w'}}$$

AET provide the formal theory underlying this assumption. Among other things, they point out that if W is chosen randomly from the set $\{W, W'\}$ then the assumption that $\delta = 1$ will hold. If not, then δ may be larger or smaller than 1, although they suggest that in practice $\delta < 1$ may be an appropriate assumption if the controls are chosen to be among the most important variables.

Were both W and W' observed, it would be possible to recover β from a standard linear regression model. With W' unobserved, the researcher is able to estimate two equations:

$$\begin{aligned} Y &= \hat{\alpha} + \xi X + \wr + \epsilon & (2) \\ Y &= \tilde{\alpha} + \Lambda X + \Psi W + \tau + \epsilon & (3) \end{aligned}$$

ξ is the coefficient on X with no controls, and Λ is the coefficient on X when including all the observed controls. Both ξ and Λ are subject to omitted variable bias. Since the models are linear, the relationship between these coefficients and the true β is straightforward:

$$\begin{aligned} \xi &= \beta + C_{wx} + C_{w'x} \\ \Lambda &= \beta + \frac{C_{w'x}}{V_{\tilde{x}}} \end{aligned}$$

where $\tilde{X}$ is the residual from a bi-variate regression of X on the observed controls W. AET note that under

the maintained assumption that W and W' are orthogonal and the proportional selection assumption, $C_{w'\tilde{x}} = C_{w'x} = \delta C_{wx} V_{w'}$. Further, $V_{\tilde{x}} = 1 - C_{xw}^2$. They show, therefore, the treatment effect β can be recovered as $\beta = \Lambda - \delta \frac{C_{wx} V_{w'}}{1 - C_{wx}^2}$. This latter term is the bias from omission of the W' vector. The central question is how this bias relates to the movement in coefficients from ξ (no controls) to Λ(with observed controls). The result is summarized in Proposition 1.

Proposition 1. *Denote the R-squared in equation (2) as R_1 and the R-squared in equation (3) as R_2. Further, denote the full R-squared from Equation (1) as R_{max}. Under the proportional selection assumption, $\delta \frac{(\xi - \Lambda)(R_{max} - R_2)}{(R_2 - R_1)}$ is a very close upper bound on the bias if $\delta < 1$, a close lower bound on the bias if $\delta > 1$ and exactly equal to the bias if $\delta = 1$.*

Proof. The bias is given above: $\frac{\delta C_{wx} V_{w'}}{1 - C_{wx}^2}$. The difference between ξ and Λ is $\xi - \Lambda = C_{wx} + \delta C_{wx} V_{w'} - \frac{\delta C_{wx} V_{w'}}{1 - C_{wx}^2}$. Dividing, I can express the relationship between this coefficient difference and the bias:

$$\xi - \Lambda = \left(\frac{C_{wx}(1 - C_{wx}^2 - \delta C_{wx}^2 V_{w'})}{\delta C_{wx} V_{w'}} \right) \frac{\delta C_{wx} V_{w'}}{1 - C_{xw}^2}$$

Now consider the variances from equations (2) and (3):

$$\begin{aligned} V_\wr &= 1 + V_{w'} - C_{xw}^2 [1 + \delta V_{w'}]^2 \\ V_\tau &= V_{w'} - \frac{[\delta C_{wx} V_{w'}]^2}{1 - C_{wx}^2} \end{aligned}$$

Straightforward simplification yields:

$$\frac{V_\tau}{V_\wr - V_\tau} = \frac{V_{w'}(1 - C_{xw}^2 - \delta^2 C_{xw}^2 V_{w'})}{(1 - C_{xw}^2 - \delta C_{xw}^2 V_{w'})^2}$$

Note that the definition of the R-squared in a linear model means that $\frac{V_\tau}{V_\wr - V_\tau} = \frac{(R_{max} - R_2)}{(R_2 - R_1)}$. Together, this implies that : $\delta \frac{(\xi - \Lambda)(R_{max} - R_2)}{(R_2 - R_1)} = \left[\frac{\delta C_{wx} V_{w'}}{1 - C_{xw}^2} \right] \frac{(1 - C_{xw}^2 - \delta^2 C_{xw}^2 V_{w'})}{(1 - C_{xw}^2 - \delta C_{xw}^2 V_{w'})}$. The first term on the right hand side is the bias. The second term is very close to 1 for values of δ close to 1. If $\delta = 1$ it is exact. If $\delta < 1$ the expression is an upper bound on the bias and if $\delta > 1$ it is a lower bound. □

The result directly relates coefficient movements to the bias, as well as giving a way to calculate the bias (or at least a close bound on it). This calculation requires observing both coefficients and R-squared values from these regressions *and* making an assumption about the maximum R-squared. One assumption is that this value is 1: that if all of the unobservables were observed, they would explain all variation in Y. This assumption may be too strong in many cases where there is some either random component of Y (measurement error, for example) or some variables which predict Y but are orthogonal to X.

Incorporating the R-squared values is, of course, crucial here. The movement in coefficients must be

scaled by the amount of variation in Y explained by observed and unobserved components. Bellows and Miguel (2009) derive derive and apply a version of this adjustment (in an Appendix) but do not incorporate the R-squared values. Their adjustment, based only on the coefficient movements, is valid only in a case where the amount of variation in Y explained by the observables is the same as the amount explained by the unobservables. This assumption is unlikely to hold in general.

It is important to recall that the innovation here is simply to connect the formula for the bias – $\frac{\delta C_{wx} V_{w\prime}}{1-C_{xw}^2}$ – directly to coefficient movements. The argument that this represents the bias under proportional selection is made in more technical detail in AET and their related work. It is also possible to view this as an alternative to the calculation methodology suggested in AET, which uses the data directly to calculate this object. Their calculation is described in more detail in Appendix A and works out to the identical formula – again, exact when $\delta = 1$.

2.2 Extension: Additional Observed Controls

I now consider the extension to a case where there is another index of observed controls – call these M – which are fully observed, do not have a related unobserved component and are orthogonal to W and W'. In a health context these could be, for example, age or sex – baseline variables which explain some of the variation in Y and related to X but do not generate omitted variable concerns.[7] In this case, Equation (4) below is the full equation, and the two estimable equations are (5) and (6):

$$Y = \alpha + \beta X + W + W' + M + \epsilon \tag{4}$$

$$Y = \hat{\alpha} + \xi X + \Gamma M + \wr + \epsilon \tag{5}$$

$$Y = \tilde{\alpha} + \Lambda X + \Psi W + \Gamma' M + \tau + \epsilon \tag{6}$$

Proposition 2 summarizes the bias calculation in this case.

Proposition 2. *Denote the R-squared in equation (4) as R_{max}, the R-squared in equation (5) as R_1 and the R-squared in equation (6) as R_2. Under the assumption of proportional selection, $\delta \frac{(\xi-\Lambda)(R_{max}-R_2)}{(R_2-R_1)}$ is a very close upper bound on the bias if $\delta < 1$, a close lower bound on the bias if $\delta > 1$ and exactly equal to the bias if $\delta = 1$. .*

Proof. The bias on Λ is now $\frac{\delta C_{wx} V_{w\prime}}{1-C_{mx}^2-C_{wx}^2}$. The same algebra as above yields $\xi - \Lambda = \frac{C_{wx}(1-C_{mx}^2-C_{wx}^2(1+\delta V_{w\prime}))}{(1-C_{mx}^2)(1-C_{mx}^2-C_{wx}^2)}$ and $\frac{(R_{max}-R_2)}{(R_2-R_1)} = \frac{V_z(1-C_{mx}^2-C_{wx}^2(1+\delta^2 V_{w\prime}))(1-C_{mx}^2)}{(1-C_{mx}^2-C_{wx}^2(1-\delta V_{w\prime}))^2}$. Combining, I find that $\delta \frac{(\xi-\Lambda)(R_{max}-R_2)}{(R_2-R_1)} = \left[\frac{\delta C_{wx} V_{w\prime}}{1-C_{mx}^2-C_{wx}^2}\right]\left[\frac{(1-C_{mx}^2-C_{wx}^2(1+\delta^2 V_{w\prime}))}{(1-C_{mx}^2-C_{wx}^2(1-\delta V_{w\prime}))}\right]$ which is exactly as above. □

[7] In practice, obvious elements of M like age or sex are often correlated with possible omitted variables. This is fine, I simply define the W and W' category as the parts of those variables which are orthogonal to M.

In practice, this indicates that if there are a set of important, fully observed controls (could be age dummies, or sex, year fixed effects, etc) those should be included in both the "controlled" and "uncontrolled" regressions. The relevant coefficient movements are those which occur between the regression with X and M and the regression with X, M and W.

As a corollary, consider the case where M also has a related unobserved vector M', so the true model is:

$$Y = \alpha + \beta X + W + W' + M + M' + \epsilon$$

Consider the modified model:

$$Y = \alpha + \beta' X + W + W' + \Delta M + \tilde{\epsilon}$$

where $\beta \neq \beta'$ because M' is omitted. In this case, by the same theorem and proof as in Proposition 2, the coefficients and R-squared values from equations (5) and (6) above can be used to recover β'.

This latter observation indicates that it is possible to use the coefficient movements to recover the effect we would estimate if we could observe the unobservables related to W, even if we accept that may not be the true treatment effect. This may be useful in a case where, for example, we wish to assume that the bias associated with the unobserved portion of the M category is small.

A related question is whether we can recover the true β in the case above when we see W and M and make proportional selection assumptions about W and W' and about M and M'. This case is worked out in Appendix A. The bias does relate to the coefficient movements between uncontrolled and fully controlled regression in that case but the calculation is not straightforward.

2.3 Extension: Bias Results with Added Precision

The proposition in Section 2.1 gives a method for calculating bias using information on the movement of coefficients from the fully uncontrolled to the fully controlled regression. It follows simply from that result that if the coefficient on X doesn't change much from the fully uncontrolled to the fully controlled regression, this suggests limited bias. Effectively, this will only occur if C_{wx} is small, which then means the remaining bias is also small.

A common related heuristic is to look for stabilizing coefficients as the number of controls increases. Even if there is a large change in the coefficient when some controls are added, if further controls do not change the coefficient very much, the conclusion is that the result is approaching the causal coefficient.

I capture this setup with the assumption that the true model is as follows:

$$Y = \alpha + \beta X + W_1 + W_2 + W' + \epsilon$$

In this case, I imagine both W_1 and W_2 are observed, while W' is unobserved. I retain the assumption of proportional selection, and assume that the variance of W_1 is 1. I allow the degree of proportionality to vary. In particular, I assume:

$$C_{w_1 x} = \frac{C_{w_2 x}}{\delta V_{w_2}} = \frac{C_{w' x}}{\phi V_{w'}}$$

A common procedure in this case is to run the three regressions below in order, and compare the coefficients ξ, Λ_1 to Λ_2.

$$Y = \alpha + \xi X + \wr + \epsilon \tag{7}$$

$$Y = \tilde{\alpha} + \Lambda_1 X + \Omega_1 W_1 + \tau + \epsilon \tag{8}$$

$$Y = \hat{\alpha} + \Lambda_2 X + \Psi_1 W_1 + \Psi_2 W_2 + \kappa + \epsilon \tag{9}$$

All three coefficients are biased, with the exact formulas given below.

$$\begin{aligned} \xi &= \beta + C_{w_1 x}(1 + \delta V_{w_2} + \phi V_{w'}) \\ \Lambda_1 &= \beta + \frac{C_{w_1 x}(\delta V_{w_2} + \phi V_{w'})}{1 - C^2_{w_1 x}} \\ \Lambda_2 &= \beta + \frac{C_{w_1 x} \phi V_{w'}}{(1 - C^2_{w_1 x}(1 + \delta^2 V_{w_2}))} \end{aligned}$$

I note first that if ξ and Λ_2 are close to each other then this suggests limited remaining bias, as described in the main results above. The stability heuristic is (possibly) useful here in a case where ξ and Λ_1 are far apart. In that case, the common heuristic would be that if Λ_1 and Λ_2 are close together, then the remaining bias on Λ_2 is small. The proposition below summarizes the condition for this to be the case.

Proposition 3. *Stabilization of coefficients implies a small remaining bias if and only if a small δV_{w_2} implies that $\phi V_{w'}$ is small.*

Proof. The stabilization heuristic implies two things. First, $\Lambda_1 - \Lambda_2$ is small and, second, that $\frac{\Lambda_1 - \Lambda_2}{\xi - \Lambda_1}$ is small. Note that:

$$\frac{\Lambda_1 - \Lambda_2}{\xi - \Lambda_1} = \frac{\delta V_{w_2}(1 - C^2_{w_1 x}(1 + \delta^2 V_{w_2} + \delta \phi V_{w'}))}{(1 - C^2_{w_1 x}(1 + \delta^2 V_{w_2}))(1 - C^2_{w_1 x}(1 + \delta V_{w_2} + \phi V_{w'}))}$$

For a value of δ close to 1, as I have generally been considering, $\frac{(1 - C^2_{w_1 x}(1 + \delta^2 V_{w_2} + \delta \phi V_{w'}))}{(1 - C^2_{w_1 x}(1 + \delta V_{w_2} + \phi V_{w'}))} \approx 1$ and this collapses to $\frac{\Lambda_1 - \Lambda_2}{\xi - \Lambda_1} \approx \frac{\delta V_{w_2}}{(1 - C^2_{w_1 x}(1 + \delta^2 V_{w_2}))}$. The question is whether finding that this is small implies that the remaining bias, $\frac{C_{w_1 x} \phi V_{w'}}{(1 - C^2_{w_1 x}(1 + \delta^2 V_{w_2}))}$ is also small. It is clear that $C_{w_1 x}$ is not small (because ξ and Λ_1 are not close) which means the bias will be small only if $\frac{\phi V_{w'}}{(1 - C^2_{w_1 x}(1 + \delta^2 V_{w_2}))}$ is small. The ratio above has the same denominator. In all, I conclude that a small value of $\frac{\Lambda_1 - \Lambda_2}{\xi - \Lambda_1}$ implies a small remaining bias only if small δV_{w_2} implies a small $\phi V_{w'}$. □

This proposition argues that using this stability heuristic requires a further assumption along with the proportional selection – namely, that the fact that the second set of observables included are less important than the first set implies that the unobservables are *also* less important. This may be a palatable assumption but it is certainly not obvious. It may, however, be partially informed by taking a view on the maximum R-squared.

Consider V_κ (from Equation 9 above) which can be recovered from the difference between R_{max} and the fully controlled R-squared.

$$V_\kappa = V_{w'}(1 - \phi C_{w_1 x} \frac{C_{w_1 x} \phi V_{w'}}{(1 - C^2_{w_1 x}(1 + \delta^2 V_{w_2}))})$$

If V_κ is large, this implies that $V_{w'}$ must be large. If $V_\kappa$is small, this implies either a small $V_{w'}$ or a large ϕ.

What this suggests is that if the fully controlled R-squared is very far from the hypothesized R_{max}, $V_{w'}$ is not small and, therefore, we cannot infer from the stabilizing coefficients that Λ_2 is close to β. If the fully controlled R-squared is close to R_{max} it may still be the case that the bias is large, since ϕ may be large. In other words, this is a necessary but not sufficient condition.

2.4 Summary

The results in this section formalize some commonly used heuristics. Together, they suggest several things.

First, movement in the coefficient of interest when controls are added is informative about remaining bias under the proportional selection assumption, but must be used along with some assumption about the maximum amount of variance explained by the observables and unobservables together.

Second, the relevant movement in the coefficient is that which occurs after inclusion of the set of controls for which one is concerned about omitted components. For example, if a coefficient moves a lot after inclusion of a precise measure of individual age, this is not informative about how much further movement would be observed with controls for socioeconomic status. Controls of this type should be included in all regressions.

Third, stability in the coefficient of interest as controls are added is reassuring *only* if the R-squared stabilizes at or close to the maximum R-squared.

Together, this provides guidance in how these heuristics might be better used in practice. But it does not provide evidence on whether this adjustment is effective in identifying causal impacts. To learn that, it is necessary to perform some validation, which I turn to below.

3 Validation: Maternal Behavior, Birth Weight and Child IQ

The results above provide a way to recover an estimate of "causal" treatment effects under the assumption that selection on observables and unobservables is proportional. This assumption is fairly strong

and not directly testable. Indirectly, I can test the assumption – and the methodology more generally – by asking whether estimates generated by this procedure are closer to the true causal effect. Discussing that requires a setting in which I can match (possibly) biased estimates to some "true" estimate of a treatment effect.

Given such a setting, validation could take two forms. Most generally, I could ask whether there is a value of δ which would match the adjusted effect from the observational regressions to the true treatment effect. The other elements of the adjustment are observable (in the case of the coefficients and R-squared values from regressions) or knowable in principle (in the case of the maximum R-squared). The value of δ is a free parameter, so one simple validation exercise is to ask whether some value of δ would work. This is equivalent to asking whether the coefficient on the treatment moves closer to the truth when controls are added.

A more constrained test is to ask whether a single value of δ might organize a number of findings within a particular setting (or settings). If yes, this would suggest at a minimum that this technique works well in comparing the robustness of multiple findings within a given setting and, more tentatively, could suggest a value of δ that might be used in other settings.

In this section I undertake both the less constrained and more constrained versions of this validation test in the context of the relationship 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, however, are subject to significant omitted variable concerns. The behaviors which are linked to good child outcomes tend to also be closely linked to maternal socioeconomic status.

I consider five relationships in all: the relationship between child IQ and breastfeeding, drinking during pregnancy, low birth weight/prematurity and the relationship between birth weight (as the outcome) and maternal drinking and smoking in pregnancy. Section 3.1 below describes the data, Section 3.2 the empirical strategy and Section 3.3 the results.

3.1 Data

I use data from the National Longitudinal Survey of Youth Children and Young Adult Survey (NLSY) and data from the US Natality Detail Files (from 2001 and 2002).

NLSY

The NLSY is a longitudinal survey of women, and the Children and Young Adult module collects information on the children of NLSY participants. These data contain information on both IQ and birth weight. In the case of IQ, the outcome of interest is PIAT test scores for children aged 4 to 8. The treatments of interest are: months of breastfeeding, any drinking of alcohol in pregnancy and an indicator for being low birth weight and premature (<2500 grams and <37 weeks of gestation). These variables are summarized in the first rows of Panel A of Table 1.

For birth weight, the outcome is simply birth weight in grams. Here, I use all children. The treatments are whether the mother smokes in pregnancy and maternal drinking intensity during pregnancy. These variables are summarized in Panel B of Table 1.

The NLSY data also contain demographic controls. These are summarized in the remainder of Panel A and Panel B of Table 1 (I summarize these twice since the sample differs for the IQ and birth weight analyses). They include: child age and sex, race, maternal age, maternal education, maternal income and maternal marital status.

Natality Detail Files

The US Natality detail files contain data on all births in the US. I use data from 2001 and 2002 and focus on birth weight as the outcome. The treatments are, again, whether the mother smokes during pregnancy and maternal drinking intensity. I recode drinking data to match the NLSY definitions. The natality detail files also include demographics: child sex, maternal race, age, education and marital status. These data do not report income.

Panel C of Table 1 reports summary statistics.

3.2 Empirical Strategy

The baseline empirical strategy is straightforward. Assume that I observe a measure of the true β, denoted β_{true}. In addition, I can calculate a measure of the bias-adjusted β as described in Section 2, and denoted β_{adj} :

$$\beta_{adj} = \Lambda - \delta \frac{(\xi - \Lambda)(R_{max} - R_2)}{(R_2 - R_1)}$$

The first stage of validation here involves asking whether there is a value of δ for which $\beta_{true} = \beta_{adj}$. Note that for this exercise I assume that the bias calculation is exact even with a value of $\delta \neq 1$. As I note in Section 2, this is close to the truth. In a second stage I ask whether a single value of δ can provide improved inference across all five settings considered.

Behind this empirical strategy there are a number of open questions. First, what controls are in W and which, if any, are in M? Second, what is the value of R_{max}? The other parameters are observed in the regressions but R_{max} cannot be observed since we do not see the unobservables. Third, what is the value of β_{true}?

I begin with the choice of W and M. I argue that the primary omitted variables relate to socioeconomic status; although there may be others, if we were able to eliminate the bias from this source, it would be a significant step forward. Therefore the elements of W will be the observed components of socioeconomic status: maternal education, income, race, marital status and age. Recall that M should contain any elements which may impact the coefficient but do not have omitted counterparts. In this case, it seems appropriate to include child sex and, in the case of IQ, child age.

Second, consider the choice of R_{max}. In theory, this should reflect how much of the variation in child IQ or birth weight could be explained if we had full controls for family background. This is a figure for which we need to go outside the data. Neither IQ nor birth weight seem likely to have an R_{max} of 1. Even identical twins raised together do not have the same IQ scores or identical birth weight. I suggest that the appropriate figure in either case is the correlation between siblings raised together, which will capture the full effect of family background. For IQ, I use a value of 0.385, based on the average correlations from two studies reported in Scarr and Weinberg (1983).[8] For birth weight, I use a value of 0.5, drawn from Mazumder (2011).

Finally, the estimation requires values of $\beta_{true.}$ One natural approach would be to match the observational analysis with evidence from randomized controlled trials which estimate similar parameters. Indeed, this is the approach I will take in the next section. This is not feasible here. Even in the two cases (breastfeeding and smoking) where I do have some randomized or quasi-random estimates on which to rely, the magnitudes are not comparable.

Instead, I take advantage of the family structure in the NLSY to estimate sibling fixed effects models. Although of course these may also be subject to concerns about causality, they should address most or all of the concerns about omitted family background, *per se*, which I argued was the primary omitted variable concern. I can therefore take them as the appropriate estimate of the impact of the treatment if I could adjust fully for family background – the sibling fixed effects estimates are β_{true}.

As a check on this, I can use outside results not to generate magnitudes but as tests of the null of a treatment effect or not. I can then ask whether the relationships which are more robust to the adjustment suggested here are also those which are confirmed in better data. Among the relationships I consider, randomized evidence suggest 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, for example: Falgreen-Eriksen et al, 2012; O'Callaghan et al, 2007). In contrast, low birth weight and prematurity do seem

[8]This is consistent with other overview studies which suggest values in the range of 0.35 to 0.4 – see, for example, Bouchard and McGue, 2003.

to be consistently linked to low IQ (Salt and Redshaw, 2006), a link which also has a biological underpinning (de Kieviet et al, 2012). On the birth weight side, 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).

3.3 Results

Table 2 reports the results: Panel A shows data on child IQ from the NSLY, Panel B data on birth weight from the NLSY and Panel C data on birth weight from the Natality Detail Files.

The first two columns in each panel show estimated treatment effects and R-squared values with only sex (or age and sex in the case of IQ) as controls. Columns 3 and 4 show similar treatment effects with the full control set. 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. There is no biological reason to think this is the case: it *must* be due to selection. Both samples show smoking and drinking are associated with lower birth weight. All seven analyses reported here 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 5 shows the sibling fixed effects estimates; in Panel C, I report the estimates drawn from the NLSY for these outcomes, since the Natality files do not link mothers across births. The positive impacts of breastfeeding and maternal drinking are eliminated. The impact of low birth-weight and prematurity on IQ remains fairly large – about 0.10 standard deviations – but has a p-value of 0.11. In the case of birth weight, the impact of smoking on child birth weight remains strongly significant in these regressions, but there is no measured impact of maternal drinking. These results – the lack of an impact for breastfeeding and maternal drinking, the possible impact of low birth weight on child IQ and the strong impact of smoking on birth weight – line up well with the conclusions on null hypotheses from the literature described above.

Column 6 combines these figures with the estimates of R_{max} (0.385 in the case of IQ, 0.5 in the case of birth weight) and calculates the δ which would match β_{adj} to β_{true}. In all seven rows this δ is defined and is positive. That is, these all pass the most basic validation test: the coefficients move toward the truth when the controls are added and there is therefore some value of δ which would match. The values of δ range between about 0.5 and 1.5.

In Column 7 I report the value of δ which would return an adjusted effect of 0 (rather than the value which would match to the sibling fixed effects). This is done for two reasons. First, for those outcomes (maternal drinking, breastfeeding) where other literature suggests we reject the null, a value of 0 is another estimate of the true effect. Second, this allows me to ask in general whether the less robust results are those where a smaller value of δ would eliminate the effect. In the case of the first objective, this column suggests a

smaller range of δ – around 1 – would match each of these effects to zero. In the case of the second I find support: the "true" effects would require, on average, a larger value of δ to produce an adjusted effect of zero.

Finally, Column 8 asks the second form of validation I describe above: could a single value of δ generate better inference across all these settings. I use a value of $\delta = 1$. This is done for two reasons. First, it seems a natural focal point. Second, looking at the values in Columns 6 and 7, this would appear to fit well. Standard errors in this column are calculated with a bootstrap over individuals, although it is worth keeping in mind that these are very sensitive to sample size. This adjustment appears to perform well. The coefficient moves closer to the true treatment effect in all cases. After the adjustment only the impacts of smoking remain significant and sizable.

Coefficient Stability

The above analysis suggests that performing the proportional selection adjustment improves the conclusions. It seems useful to consider whether a similar conclusion could have been reached from using the "coefficient stability" heuristic. To do this, for each treatment I run regressions progressively including controls. I choose the order of controls by ranking the demographics based on the amount of variation in child IQ or birth weight that they explain in the data. I include these controls in the same order for each analysis within outcome (the order differs for IQ and birth weight). Figures 1a-1g show coefficients and R-squared values for the seven analyses.

These figures suggest coefficient stability is not useful distinguishing among these analyses. All show a very similar pattern of stabilizing coefficients. Based on these alone it would be quite difficult to identify some of the relationships as more robust than the others. In line with the discussion in Section 2.3, the issue is clear: the R-squared in the fully controlled regressions here is around 0.25 for IQ and less than 0.1 for birth weight, far below the figures of 0.385 or 0.5 that were drawn from existing data. Given this, the fact that the coefficient has stabilized is not fully informative.

Summary

The results in this section – in particular, in Table 2 – are quite supportive of this approach. It passes the most basic validation test by showing that the coefficient movements are informative and larger movements point to less robust results. Perhaps more surprisingly, the results show that a single value of δ ($\delta = 1$) performs well across all the settings. Returning to the question of applications in economics, this suggests support for the coefficient movement robustness test (although not the coefficient stability test) and may even suggest an adjustment based on $\delta = 1$ might be a helpful statistic to report (similarly one could report a value of δ which generates $\beta_{adj} = 0$ and use $\delta = 1$ as a benchmark).

In the next section I move to a somewhat more aggressive application and ask whether this approach might be used in the context of a large public health literature to improve inference.

4 Application: Health Behaviors and Health Outcomes

A large literature in epidemiology and public health looks to estimate the relationship between positive heath behaviors and health outcomes. Do individuals who exercise live longer? Does taking a vitamin supplement lower your blood pressure? Observational studies in this literature suffer from clear omitted variable bias problems, largely stemming from correlations between high socioeconomic status and both positive health behaviors and good health outcomes. Likely due to this issue, when randomized studies are run to look at similar questions the results are often at odds with what was seen in observational data. A classic example is the exploration of the link between diet and health. For years the medical profession recommended a low-fat, high carbohydrate diet as a key to better health. It turned out this was based on biased estimates. When randomized data from a large study was released in 2006, this result was seriously weakened (Prentice et al, 2006; Beresford et al, 2006; Howard et al, 2006).

Given that many of the central issues facing this literature can be boiled down to omitted socioeconomic status variables, it seems natural to ask whether the selection-on-observables adjustment procedure could improve inference.

In this section I combine observational data on a number of relationships estimated in the public health literature with randomized evidence on those relationships. In some cases, randomized trials have confirmed observational links and in others they have not. I use a comparable observational population and match the magnitude of the randomized impacts to the observational ones. As in the analysis above, I ask whether a version of the selection-bias adjustment procedure could match the observational estimates to the randomized effects.

I go further in this section than in Section 3 in two ways. First, motivated in part by the above, I assume a value $\delta = 1$ and then *estimate* a value for R_{max} for each match. This effectively asks whether it is possible to perform a successful version of this adjustment without thinking carefully about the $R_{max.}$ Second, once I have estimated values for R_{max} across a number of settings (this section considers 29 outcome-treatment pairs), I ask whether a single value of R_{max} (or, in fact, a single parametrization as a function of R_2 and R_1) could improve inference across all the settings.

In the end, I can ask whether a simple version of this procedure – using only the results from the observable regressions – might be used to improve inference across these settings and parallel settings, which I argue would encompass much of the public health literature. The last part of this section provides some simple out of sample tests.

It is worth noting that this section also has a role as validation, just as Section 3 above, in asking whether a version of this adjustment can match the true impacts.

Section 4.1 below describes data, Section 4.2 the empirical strategy, and Section 4.3 the results, including out-of-sample tests.

4.1 Data

This section considers two treatments: exercise and vitamin D+Calcium supplementation. In each case I consider the relationship between the treatment and a range of outcomes. This analysis requires two pieces of data: randomized trial results and observational data.

Randomized Trials

Randomized trial results are drawn from existing work.

Exercise Evidence on the impact of exercise is drawn from several papers which are summarized in a Cochrane Review meta-analysis (Shaw et al, 2006). I consider only studies which compared exercise to no exercise (this excluded studies which also used diet). Outcomes considered include weight, blood pressure, cholesterol, blood glucose and triglycerides.

Vitamin D and Calcium Evidence on the impact of vitamin D and calcium supplementation comes from the Women's Health Initiative, a large scale study of post-menopausal women which has run a number of important interventions. One trial within the study involved randomizing women into receiving vitamin D and calcium supplements (treatment) or not (control). Outcomes include bone density, lipids, blood pressure, exercise, and weight.

In Appendix Table A.1 I list the citation for each outcome-treatment pair, the treatment and any restrictions on age or gender in the study recruitment.

Observational Data

Exercise Exercise data are drawn from the National Health and Nutrition Examination Survey (NHANES), Wave III. Individuals are asked detailed questions about exercise. I use this to create a treatment measure as close as possible to the treatment in each study. In most cases the study includes some kind of jogging three times a week. Exact populations used are listed in Column 3 of Appendix Table A.1 for each paper, but in general these tend to focus on middle-aged individuals. Exercise data and the outcomes variables considered are summarized in Panel A of Table 3.

Vitamin D and Calcium Data on vitamin D and calcium supplementation also comes from the NHANES-III. Individuals are asked about vitamin and mineral supplements, which allows me to create an indicator for taking vitamin D and calcium supplementation. To match the Women's Health Initiative data I use women aged 55 to 85 (recruitment in this study is women 50 to 80, but evaluation is several years later).

Summary statistics on share of women using supplements and outcomes variables are in Panel B of Table 3.

4.2 Empirical Strategy

In this section, I employ the assumption that $\delta = 1$. The bias-adjusted coefficient, β_{adj} is therefore calculated: $\beta_{adj} = \Lambda - \frac{(\xi - \Lambda)(R_{max} - R_2)}{R_2 - R_1}$. In this case, β_{true} is drawn from randomized trial results. As in the analysis above there is an important choice of what is in W and what will be in M. As above, I include in W the standard socioeconomic status measures: education, income, marital status and race. This reflects the observation that the bulk of the omitted variable issue are likely to be socioeconomic status. In M I include age dummies and sex and, in cases where the outcome is weight in kilograms, measures of height.

The first step in the empirical strategy is to simply estimate value of R_{max} for which $\beta_{adj} = \beta_{true.}$ This will be defined as long as Λ is closer to β than ξ is (that is, as long as the coefficients move in the correct direction), although here it is possible that the estimated value of R_{max} could be larger than 1, which would be a failure from a validation standpoint.

The second step is to ask whether there is a single value of R_{max} which would improve inference across all these settings. In fact, this is unlikely to be the case since the settings I consider here differ widely in their predictability, so a single value of R_{max} is not likely to work well. Instead, I parametrize R_{max} as a function of R_1 and $R_2 : R_{max} = R_2 + \psi(R_2 - R_1)$. I then estimate ψ.

Effectively, this assumes that the amount of Y which is explained by the observables is a guide to how much would be explained by the unobservables. A value of $\psi = 1$ would imply that the unobservables explain as much of the variation in Y as the observables.

In addition to having some intuitive appeal, this is a convenient assumption when the goal is to use the conclusions to evaluate existing work. With this assumption, the calculation of the bias-adjusted coefficient collapses to $\beta_{adj} = \Lambda - \psi(\xi - \Lambda)$ and it is not necessary to observe the R-squared values. Since published papers in public health and epidemiology only very rarely report these values, this makes this procedure significantly more useful.

Given this assumption, the full estimation is straightforward. For outcome-treatment pair i denote the adjusted coefficient $\beta^i_{adj}(\psi)$ and the true effect β^i_{true}. The trial also produces a standard error, denoted σ^i. I calculate the difference between the bias-adjusted and true coefficient, scaled by the standard error. I sum these over the outcome-treatment pairs and minimize the sum over the choice of ψ. Formally, I solve:

$$\hat{\psi} = argmin_{\psi} \sum_i \left(\frac{\beta^i_{adj}(\psi) - \beta^i_{true}}{\sigma^i} \right)^2$$

Given this value it is then possible to explore the performance of this adjustment in several ways. First, I can compare the magnitude of the error under the maximum likelihood value of ψ relative to the assumption

that $\psi = 0$ (which is the benchmark controlled regression coefficient). Second, I can compare the performance on each outcome-treatment pair, using bootstrapped standard errors, and ask whether I would have drawn more accurate conclusions about the null hypothesis from the adjusted analysis. Finally, there are a few outcomes for which the trials suggest a conclusion about the null hypothesis but where matching magnitudes is difficult. It is possible to perform an out-of-sample test using these outcomes and exploring whether the same adjustment would lead to more accurate conclusion in these cases.

4.3 Results

The first five columns of Table 4 show the first step of the results. Column 1 lists the outcome and, in the case of exercise where there are typically multiple studies per outcome, information on the citation. The second and third columns list the uncontrolled and controlled effects, their standard errors and the R-squared values. The effects are significant in many but certainly not all cases, and generally in the expected direction, with exercise and vitamin supplementation linked to improved health outcomes.

Column 4 reports the impact from the randomized trials. These impacts are less often significant than the controlled effects and typically much, much smaller. If I consider the relationship between vitamin supplementation and weight, for example, both the randomized and controlled effects are significant, but the randomized effect is tiny relative to the impact with controls. The controlled effect indicates a difference of about one and a half pounds, and the randomized effect a difference of just about one-tenth of a pound. This is an economically significant difference.

Column 5 reports the value of R_{max} which would match β_{adj} and $\beta_{true.}$ In the case where this is not defined (is either less than R_2 or greater than 1) I simply report N.V. ("not valid"). In 22 of 29 outcomes the adjustment works – there is a value of R_{max} for which the adjusted coefficient matches the truth. The cases in which there is no match are all ones where the observational effect is not significant and neither is the randomized effect. These are inherently somewhat noisy, which makes it perhaps less surprising that the coefficient movements are not informative. In fact, if I ask the broader question of whether some value of R_{max} could generate estimates inside the randomized confidence interval, the answer is yes in all 29 of the cases. This finding supports the first validation hurdle: some version of this adjustment with a value of $\delta = 1$ works in most cases.

Turning to the second step, the full estimation procedure described above yields a value of $\psi = 1.018$. This suggests that the omitted characteristics explain approximately as much of the variation in outcome as the included characteristics. I can illustrate the overall impact of this bias adjustment. To do so I re-scale each outcome so the 95% confidence interval from the randomized trial ranges from 0 to 1 (and thus the randomized point estimate is close to 0.5); this is necessary for visualization since the scale of the effects varies widely across outcomes. I then convert first the standard controlled coefficient and then the bias-adjusted coefficient

onto this scale. Figure 2a shows the interval for the randomized trial (open circles) and the controlled coefficient (filled in circle). Although the controlled and true coefficient are similar in some cases, especially when they are both close to zero, in others the controlled coefficient is wildly outside the confidence interval.

Figure 2b shows the coefficients after the bias adjustment is done with the value of $\hat{\psi} = 1.018$. The fit is significantly better; note the large decrease in scale (the bias-adjusted coefficients on the same scale as the controlled coefficients can be seen in Appendix Figure 1). In a number of cases where the controlled coefficient showed significant errors – for example, the impact of vitamin supplementation on weight and exercise – the adjusted coefficients are within or very close to the confidence interval. The overall error is significantly smaller in the bias adjustment case – a reduction of 30% on average.

The final column of Table 4 describes this result numerically: I perform the bias adjustment with $\psi = 1.018$, and generate standard errors using a bootstrap over individuals. Again, it's worth taking the standard errors with caution since the observational studies here are, in some cases, significantly underpowered to pick up impacts of the size seen in randomized trials. The bias-adjusted impacts are much closer to the estimates from the randomized data, on average.

This table makes clear much of the value in the adjustment comes in cases where the controlled coefficients lead to false positive conclusions, or at least to an overstatement of the magnitude of the impact. For example, the controlled coefficients suggest a large and significant impact of vitamin supplementation on exercise[9], whereas the bias-adjusted coefficient is very close to the small and insignificant impact estimated in randomized trials. At the same time, the bias-adjustment retains significant effects in many of the cases where there are large and significant effects estimated in randomized trials – for example, the impact of exercise on weight, blood pressure and some measures of heart health. This bias-adjustment does a good job of identifying true from false associations among those which simple controlled regressions show are significant.

Out of Sample Tests

Within sample, the adjusted coefficients above are closer to the true treatment effect than the controlled coefficients. An important related test is to ask how these perform out of sample. I consider two out-of-sample tests.

In the case of exercise and Vitamin D there are also several outcomes for which randomized experiments have reached a conclusion about the null but where magnitude comparisons are difficult. This may be due to differences in the timing of follow-up, the fact that randomized effects are reported as odds ratios or because generating an exactly parallel analysis is difficult. However, given the adjustment value estimated above it is possible to return to these outcomes and explore whether the adjustment procedure used here leads to correct

[9]The theory under which this might matter is that calcium and vitamin D increase bone health, which improves ability to exercise.

conclusions in these cases.

This is done in Panels A and B of Table 5. This table is structured similarly to Table 4 except that in the third column I simply report the hypothesized direction and significance (or not) of the effect in the randomized trial. In general, the bias-adjustment also performs well here. In the case of exercise, the controlled coefficients show significant impacts on both diabetes and mortality (among individuals with heart disease), and the bias-adjusted coefficients correctly identify only the mortality evidence as robust. In the case of vitamin D the controlled coefficients incorrectly suggest supplementation matters for mortality, a result which is corrected by the bias-adjustment.

A second out-of-sample tests relies on another study, the Physician Health Study (PHS). This work evaluated the impact of beta-carotene, vitamin E and vitamin C on heart disease mortality among men.[10] Published results from the study reject links between mortality and any of these vitamins (Hennekens et al, 1996; Sesso et al, 2008). Because the outcome is mortality and magnitudes are therefore difficult to link, I could not use this study in the estimation, but it is possible to use as an out of sample test. The NHANES-III provides the data, as above.

Panel C of Table 5 shows this evidence. Vitamin E and Vitamin C are both linked to lower mortality in the controlled regressions but not (at least not significantly) in the bias-adjusted coefficients. This provides further out-of-sample support.

Summary

The evidence here provides further validation support for the suggestion that the robustness of results to coefficient movements is informative about their validity. Going further, however, it suggests that across a range of settings in public health, inference might be improved on average with a single, very simple adjustment. Given a controlled coefficient Λ and uncontrolled coefficient ξ, a value of $\beta_{adj} = \Lambda - 1.018(\xi - \Lambda)$ would, at least based on the settings above, be closer on average to $\beta_{true.}$

5 Conclusion

The goal of this paper is two-fold. First, I connect the popular robustness heuristic of exploring coefficient sensitivity to controls to the proportional selection assumption formalized in Altonji, Elder and Taber (2005). I provide some guidance to discipline the use of this coefficient movement heuristic and give a simple form of the adjustment using only information on coefficient and R-squared values. In particular, I argue that under the

[10]This study also evaluated (and supported) the importance of aspirin in preventing heart disease mortality. However, the observational evidence on aspirin is marred by both the omitted variable issue but much more so by a problem of reverse causality. It has long been thought that aspirin was good for heart disease so the kind of people who take it tend to be those with heart disease. This problem crops up in most of the settings I consider but to a much, much lesser extent. When facing this problem a bias adjustment of this type will not address the issue. I therefore do not use this as a test.

proportional selection assumption, with proportionality δ, the causal coefficient β can be recovered from the uncontrolled coefficient, ξ, the coefficient with controls, Λ, the R-squared from the uncontrolled and controlled regressions (R_1 and R_2) and an assumption about the maximum R-squared (R_{max}). The exact calculation is:

$$\beta = \Lambda - \delta \frac{(\xi - \Lambda)(R_{max} - R_2)}{(R_2 - R_1)}$$

Second, I describe two validation exercises. I argue that, regardless of the intuitive appeal of this approach, it has value only if it is effective in drawing better causal conclusions. In both validation exercises I consider cases where there exists both observational data which may be biased alongside either randomized data or better observational studies which are more likely to reflect a "true" relationship.

In the case of the relationship between maternal behavior and child birth weight and IQ, I show that a carefully applied version of this approach produces adjusted coefficients which are a much closer match to the truth.

The second validation exercise takes a number of settings and asks whether I can estimate a general version of the adjustment which would lead to better conclusions. I consider settings where (a) the outcome is a health outcome and the treatment is a health behavior and (b) the primary omitted variable bias comes from socioeconomic status, broadly construed. I argue that this applies to many relationships of interest and is not limited to the ones I consider here. I approach this as an estimation with the assumption that $R_{max} = R_2 + \psi(R_2 - R_1)$, and ψ as a parameter to be estimated. I find that a value of $\psi = 1.018$ provides a much better fit to randomized results than the simple controlled coefficients. With bootstrapped standard errors it rejects a number of false-positive associations with limited cost in terms of rejecting true-positive ones.

To the extent that one is comfortable applying these results in other contexts, this suggests a simple way for researchers in parallel settings to evaluate the plausibility of their results, and for readers of published work to do so, as well, by simply calculating: $\beta_{adj} = \Lambda - 1.018(\xi - \Lambda)$.

Returning to the coefficient heuristic, both empirical exercises here provide support for its use. In related settings, it appears that robustness to this adjustment does suggest causal impacts. The evidence may suggest that $\delta = 1$ – that is, an assumption of equal selection – could be a reasonable benchmark. Researchers looking at coefficient movements may consider reporting the adjusted coefficient with a value of $\delta = 1$ as a way to summarize the degree of stability.

References

Almond, Douglas and Bhashkar Mazumder, "Health Capital and the Prenatal Environment: The Effect of Ramadan Observance during Pregnancy," *American Economic Journal: Applied Economics*, October 2011, *3* (4), 56–85.

___ **and Janet Currie**, "Killing Me Softly: The Fetal Origins Hypothesis," *Journal of Economic Perspectives*, Summer 2011, *25* (3), 153–72.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

___ **, Todd Elder, and Christopher R. Taber**, "Using Selection on Observed Variables to Assess Bias from Unobservables When Evaluating Swan-Ganz Catheterization," *American Economic Review*, 2008, *98* (2), 345–50.

Anderssen, S. A., I. Hjermann, P. Urdal, P. A. Torjesen, and I. Holme, "Improved carbohydrate metabolism after physical training and dietary intervention in individuals with the "atherothrombogenic syndrome'. Oslo Diet and Exercise Study (ODES). A randomized trial," *J. Intern. Med.*, Oct 1996, *240* (4), 203–209.

Bellows, John and Edward Miguel, "War and local collective action in Sierra Leone," *Journal of Public Economics*, December 2009, *93* (11-12), 1144–1157.

Beresford, Shirley et al., "Low-Fat Dietary Pattern and Risk of Colorectal Cancer," *JAMA*, 2006, *295* (6), 643–654.

Bouchard, T. J. and M. McGue, "Genetic and environmental influences on human psychological differences," *J. Neurobiol.*, Jan 2003, *54* (1), 4–45.

Brunner, R. L., B. Cochrane, R. D. Jackson et al., "Calcium, vitamin D supplementation, and physical function in the Women's Health Initiative," *J Am Diet Assoc*, Sep 2008, *108* (9), 1472–1479.

___ **, J. Wactawski-Wende, B. J. Caan et al.**, "The effect of calcium plus vitamin D on risk for invasive cancer: results of the Women's Health Initiative (WHI) calcium plus vitamin D randomized clinical trial," *Nutr Cancer*, 2011, *63* (6), 827–841.

Caan, B., M. Neuhouser, A. Aragaki et al., "Calcium plus vitamin D supplementation and the risk of postmenopausal weight gain," *Arch. Intern. Med.*, May 2007, *167* (9), 893–902.

Chiappori, Pierre-Andrei, Sonia Oreffice, and Climent Quintana-Domeque, "Fatter Attraction: Anthropometric and Socioeconomic Matching on the Marriage Market," *Journal of Political Economy*, 2012, *120* (4), 659 – 695.

de Boer, I. H., L. F. Tinker, S. Connelly et al., "Calcium plus vitamin D supplementation and the risk of incident diabetes in the Women's Health Initiative," *Diabetes Care*, Apr 2008, *31* (4), 701–707.

de Kieviet, J. F., L. Zoetebier, R. M. van Elburg, R. J. Vermeulen, and J. Oosterlaan, "Brain development of very preterm and very low-birthweight children in childhood and adolescence: a meta-analysis," *Dev Med Child Neurol*, Apr 2012, *54* (4), 313–323.

Eriksen, H. L. Falgreen, E. L. Mortensen, T. Kilburn, M. Underbjerg, J. Bertrand, H. Stavring, T. Wimberley, J. Grove, and U. S. Kesmodel, "The effects of low to moderate prenatal alcohol exposure in early pregnancy on IQ in 5-year-old children," *BJOG*, Sep 2012, *119* (10), 1191–1200.

Hellenius, M. L., U. de Faire, B. Berglund, A. Hamsten, and I. Krakau, "Diet and exercise are equally effective in reducing risk for cardiovascular disease. Results of a randomized controlled study in men with slightly to moderately raised cardiovascular risk factors," *Atherosclerosis*, Oct 1993, *103* (1), 81–91.

Henderson, J., R. Gray, and P. Brocklehurst, "Systematic review of effects of low-moderate prenatal alcohol exposure on pregnancy outcome," *BJOG*, Mar 2007, *114* (3), 243–252.

Hennekens, C. H., J. E. Buring, J. E. Manson, M. Stampfer, B. Rosner, N. R. Cook, C. Belanger, F. LaMotte, J. M. Gaziano, P. M. Ridker, W. Willett, and R. Peto, "Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease," *N. Engl. J. Med.*, May 1996, *334* (18), 1145–1149.

Heran, B. S., J. M. Chen, S. Ebrahim, T. Moxham, N. Oldridge, K. Rees, D. R. Thompson, and R. S. Taylor, "Exercise-based cardiac rehabilitation for coronary heart disease," *Cochrane Database Syst Rev*, 2011, (7), CD001800.

Howard, Barbara et al., "Low-Fat Dietary Pattern and Risk of Cardiovascular Disease," *JAMA*, 2006, *295* (6), 655–666.

Howe, T. E., B. Shea, L. J. Dawson et al., "Exercise for preventing and treating osteoporosis in postmenopausal women," *Cochrane Database Syst Rev*, 2011, (7), CD000333.

Jackson, R. D., N. C. Wright, T. J. Beck et al., "Calcium plus vitamin D supplementation has limited effects on femoral geometric strength in older postmenopausal women: the Women's Health Initiative," *Calcif. Tissue Int.*, Mar 2011, *88* (3), 198–208.

Kramer, M. S., F. Aboud, E. Mironova et al., "Breastfeeding and child cognitive development: new evidence from a large randomized trial," *Arch. Gen. Psychiatry*, May 2008, *65* (5), 578–584.

Lacetera, Nicola, Devin G. Pope, and Justin R. Sydnor, "Heuristic Thinking and Limited Attention in the Car Market," *American Economic Review*, August 2012, *102* (5), 2206–36.

LaCroix, A. Z., J. Kotchen, G. Anderson et al., "Calcium plus vitamin D supplementation and mortality in postmenopausal women: the Women's Health Initiative calcium-vitamin D randomized controlled trial," *J. Gerontol. A Biol. Sci. Med. Sci.*, May 2009, *64* (5), 559–567.

Lumley, J., C. Chamberlain, T. Dowswell, S. Oliver, L. Oakley, and L. Watson, "Interventions for promoting smoking cessation during pregnancy," *Cochrane Database Syst Rev*, 2009, (3), CD001055.

Margolis, K. L., R. M. Ray, L. Van Horn et al., "Effect of calcium and vitamin D supplementation on blood pressure: the Women's Health Initiative Randomized Trial," *Hypertension*, Nov 2008, *52* (5), 847–855.

Mazumder, Bhashkar, "Family and Community Influences on Health and Socioeconomic Status: Sibling Correlations Over the Life Course," *The B.E. Journal of Economic Analysis & Policy*, 2011, *11* (3), 1.

Murphy, Kevin and Robert Topel, "Efficiency Wages Reconsidered: Theory and Evidence," in "Advances in the Theory and Measurement of Unemployment" 1990, pp. 204–240.

O'Callaghan, F. V., M. O'Callaghan, J. M. Najman, G. M. Williams, and W. Bor, "Prenatal alcohol exposure and attention, learning and intellectual ability at 14 years: a prospective longitudinal study," *Early Hum. Dev.*, Feb 2007, *83* (2), 115–123.

Orozco, L. J., A. M. Buchleitner, G. Gimenez-Perez, M. Roque I Figuls, B. Richter, and D. Mauricio, "Exercise or exercise and diet for preventing type 2 diabetes mellitus," *Cochrane Database Syst Rev*, 2008, (3), CD003054.

Prentice, Ross et al., "Low-Fat Dietary Pattern and Risk of Invasive Breast Cancer," *JAMA*, 2006, *295* (6), 639–642.

Rajpathak, S. N., X. Xue, S. Wassertheil-Smoller et al., "Effect of 5 y of calcium plus vitamin D supplementation on change in circulating lipids: results from the Women's Health Initiative," *Am. J. Clin. Nutr.*, Apr 2010, *91* (4), 894–899.

Rossom, R. C., M. A. Espeland, J. E. Manson et al., "Calcium and vitamin D supplementation and cognitive impairment in the women's health initiative," *J Am Geriatr Soc*, Dec 2012, *60* (12), 2197–2205.

Salt, A. and M. Redshaw, "Neurodevelopmental follow-up after preterm birth: follow up after two years," *Early Hum. Dev.*, Mar 2006, *82* (3), 185–197.

Scarr, Sandra and Richard Weinberg, "The Minnesota Adoption Studies: Genteic Differences and Malleability," *Child Development*, 1983, *54* (2), 260–267.

Sesso, H. D., J. E. Buring, W. G. Christen, T. Kurth, C. Belanger, J. MacFadyen, V. Bubes, J. E. Manson, R. J. Glynn, and J. M. Gaziano, "Vitamins E and C in the prevention of cardiovascular disease in men: the Physicians' Health Study II randomized controlled trial," *JAMA*, Nov 2008, *300* (18), 2123–2133.

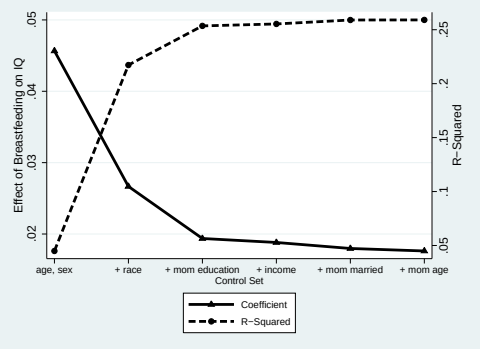
Shaw, Kelly, Hanni Gennat, Peter ORourke, and Chris Del Mar, "Exercise for overweight or obesity," *Cochrane Database of Systematic Reviews*, 2006, (4).

Stefanick, M. L., S. Mackey, M. Sheehan, N. Ellsworth, W. L. Haskell, and P. D. Wood, "Effects of diet and exercise in men and postmenopausal women with low levels of HDL cholesterol and high levels of LDL cholesterol," *N. Engl. J. Med.*, Jul 1998, *339* (1), 12–20.

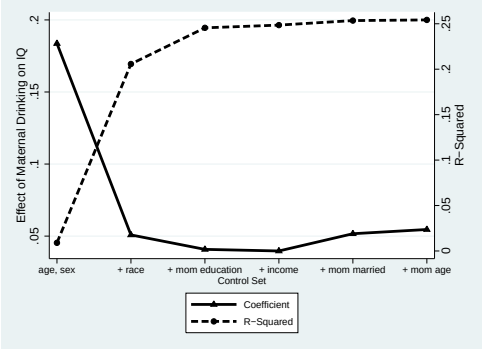
Wood, P. D., M. L. Stefanick, D. M. Dreon et al., "Changes in plasma lipids and lipoproteins in overweight men during weight loss through dieting as compared with exercise," *N. Engl. J. Med.*, Nov 1988, *319* (18), 1173–1179.

Figure 1: **Coefficient Stability, Maternal Behavior, Child Birth Weight and IQ**

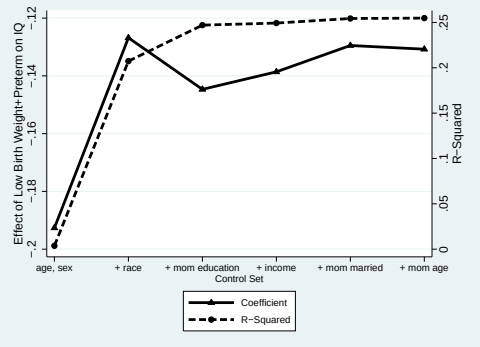
(a) Months of Breastfeeding, IQ



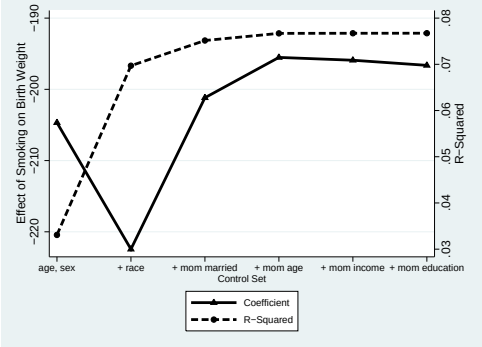
(b) Maternal Drinking Pregnancy, IQ



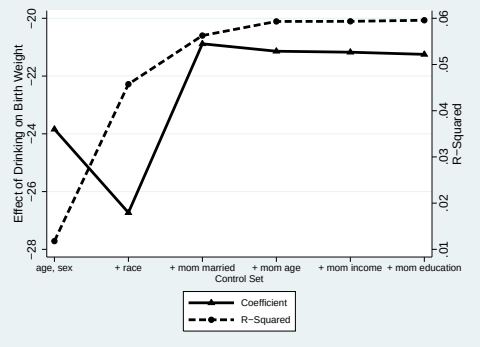
(c) Low Birth Weight+Preterm, IQ



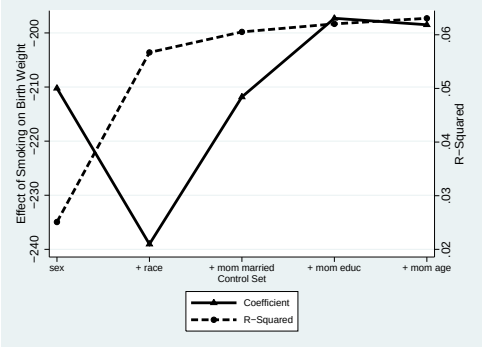
(d) Maternal Smoking, Birth Weight in NLSY



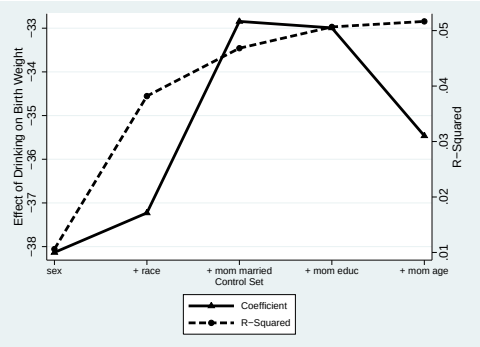
(e) Maternal Drinking, Birth Weight in NLSY



(f) Maternal Smoking, Birth Weight in Natality Files



(g) Maternal Drinking, Birth Weight in Natality Files



Notes: These graphs show the evolution of the estimated relationship between each treatment and child IQ or birth weight as controls are added. Controls are added in the same order within an outcome-cross-dataset. The order is chosen based on ordering the controls by how much of IQ or birth weight they explain and including the most important first.

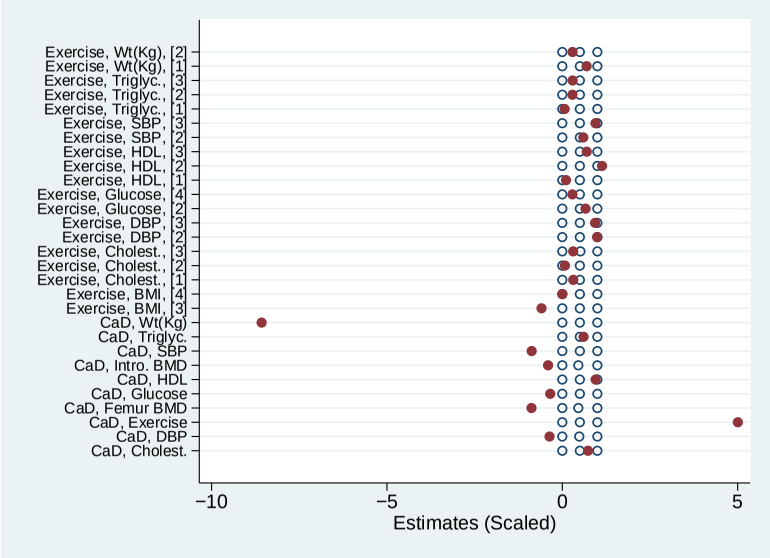
Table 1: **Summary Statistics: Early Life and Child IQ**

Panel A: NLSY Data, IQ Analysis			
	Mean	*Standard Deviation*	*Sample Size*
IQ (PIAT Score, Standardized)	0.026	0.991	6613
Breastfeeding Months	2.32	4.51	6184
LBW + Preterm	0.049	0.217	5896
Mom Drink at all in Pregnancy	0.322	0.467	6225
Age	5.57	1.37	6613
Child Female	0.494	0.500	6613
Mother Black	0.284	0.451	6613
Mother Age	25.1	5.42	6613
Mother Education (years)	12.4	3.1	6613
Mother Income	$39,980	$79,069	6613
Mother Married	0.649	0.477	6613
Panel B: NLSY Data, Birth Weight Analysis			
Birth Weight (grams)	3292.8	604.9	7418
Mom Smoke in Pregnancy	0.290	0.453	7418
Drinking Intensity (0-7)	0.634	1.15	7174
Child Female	0.486	0.499	7418
Mother Black	0.277	0.447	7418
Mother Age	24.2	5.42	7418
Mother Education (years)	12.1	3.1	7418
Mother Income	$31,097	$62,975	7418
Mother Married	0.665	0.471	7418
Panel C: Natality Detail Files			
Birth Weight (grams)	3333.8	575.1	5,886,822
Mom Smoke in Pregnancy	0.123	0.328	5,886,822
Drinking Intensity (0-7)	0.023	0.316	5,886,822
Child Female	0.488	0.499	5,886,822
Mother Black	0.167	0.373	5,886,822
Mother Age	27.2	6.13	5,886,822
Mother Education (1-5)	3.51	1.16	5,886,822
Mother Married	0.658	0.474	5,886,822

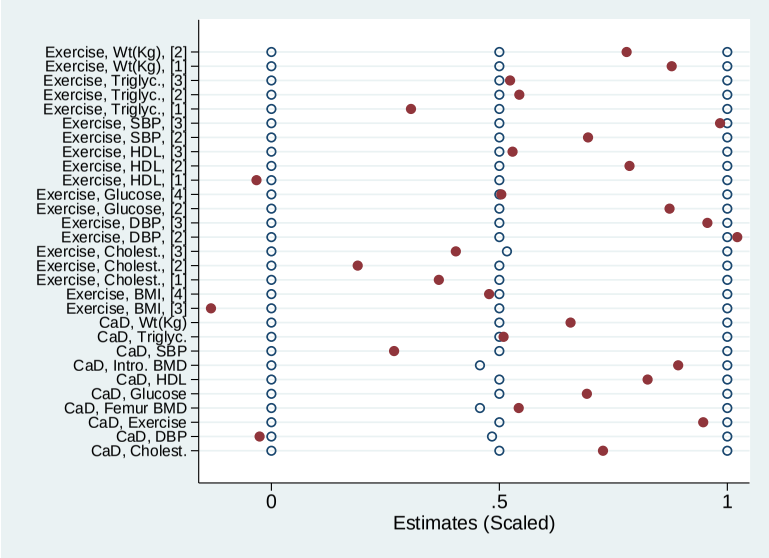
Notes: This table shows summary statistics for the data used in the analysis in Section 3. Drinking intensity is coded from 0 (never) to 7 (every day). Natality detail files are from 2001 and 2002. NLSY data is from the NLSY Children and Young Adults panel.

Figure 2: **Model Fit With And Without Bias Adjustment**

(a) Controlled Regression



(b) With Bias Adjustment



Notes: These graphs show the randomized effect sizes along with (in Sub-Figure a) the effects estimated in controlled regressions and (in Sub-Figure b) the bias-adjusted coefficients using the best-fit adjustment value of $\psi = 1.018$. Every outcome is scaled so the top and bottom of the 95% confidence interval in the randomized trial take values of 0 and 1 respectively. The mean randomized trial value is typically 0.5, although in some cases it is slightly more or less when the confidence intervals are not symmetric.

Table 2: **Maternal Behavior, Child IQ and Birth Weight**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Child IQ, Standardized (NLSY) ($R_{max} = .385$)								
Treatment Variable	*Baseline Effect*	*Baseline* R^2	*Effect with Full Controls*	*Controls* R^2	*Sibling FE Estimate*	δ to match *Sibling*	δ to match $\beta = 0$	*Bias-Adjusted Coefficient,* $\delta = 1$
Breastfeeding (Months)	0.045*** (.003)	.045	0.017***(.002)	.259	-0.008*(.005)	1.60	1.06	0.001 (.003)
Drinking in Pregnancy (Any)	0.183***(.026)	.009	0.054**(.023)	.254	0.024 (.036)	0.43	0.79	-0.014 (.025)
LBW + Preterm	-0.192***(.058)	.003	-0.131***(.050)	.254	-0.107 (.071)	0.72	4.07	-0.098 (.062)
Panel B: Birth Weight in Grams (NLSY) $R_{max} = .5$								
Treatment Variable	*Baseline Effect*	*Baseline* R^2	*Effect with Full Controls*	*Controls* R^2	*Sibling FE Estimate*	δ to match *Sibling*	δ to match $\beta = 0$	*Bias-Adjusted Coefficient,* $\delta = 1$
Smoking in Pregnancy	-204.7*** (15.2)	.033	-196.63***(15.6)	.076	-74.7***(29.4)	1.56	2.52	-118.66*(66.3)
Drinking in Pregnancy (Intensity)	-23.83***(6.14)	.011	-21.24*** (6.05)	.059	-3.41 (8.74)	0.75	0.89	2.62 (17.23)
Panel C: Birth Weight in Grams (Natality Detail Files) $R_{max} = .5$								
Treatment Variable	*Baseline Effect*	*Baseline* R^2	*Effect with Full Controls*	*Controls* R^2	*Sibling FE Estimate*	δ to match *Sibling*	δ to match $\beta = 0$	*Bias-Adjusted Coefficient,* $\delta = 1$
Smoking in Pregnancy	-210.2***(.699)	.025	-198.9***(.735)	.064	-74.7***(29.4)	0.99	1.58	-73.3***(3.70)
Drinking in Pregnancy (Intensity)	-38.15***(.73)	.010	-34.95***(.72)	.053	-3.41 (8.74)	0.94	1.04	-1.45 (1.73)

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 age dummies in the case of IQ. Full control effects in the NLSY: race, age, education, income, marital status. Full control effects in Natality Detail Files: race, education, marital status and age. Sibling fixed effects estimates come from NLSY in all panels. The value of δ is calculated to match the adjusted β to the sibling fixed effect (Column 6) or to 0 (Column 7). The bias-adjusted effect in Column 8 is generated using the assumption that $\delta = 1$. Standard errors are estimated using a bootstrap over individuals. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

Table 3: **Summary Statistics: Exercise and Vitamins**

Panel A: Exercise [NHANES-III]			
	Mean	*Standard Deviation*	*Sample Size*
Jogging 3+ Times/Wk	.033	.179	9268
BMI	28.0	6.08	9251
Weight (kg)	78.2	18.4	9252
Diastolic Blood Pressure	76.8	10.3	9197
Systolic Blood Pressure	123.9	17.5	9198
Serum Glucose (mmol/l)	5.61	2.17	8712
Triglycerides (mmol/l)	1.71	1.44	8791
Cholesterol (mmol/l)	5.39	1.13	8811
HDL (mmol/l)	1.31	.41	8740
Panel B: Vitamin D and Calcium Supplements [NHANES-III]			
Took VitD+Calcium	.211	.408	3200
Weight (kg)	69.5	16.3	3180
Diastolic Blood Pressure	73.5	10.1	3003
Systolic Blood Pressure	140.2	20.9	3004
Serum Glucose (mg/dl)	111.9	50.5	2937
Triglycerides (mg/dl)	166.4	111.8	2983
Cholesterol (mg/dl)	232.3	45.6	2988
HDL (mg/dl)	55.7	16.9	2972
Exercise Intensity (METS/wk)	14.3	20.4	3196
Femur BMD	.68	.13	2689
Introchanter BMD	.94	.19	2689

Notes: This table shows summary statistics for the data used in Section 4. NHANES-III : National Longitudinal Health and Nutrition Survey, Wave III. For Exercise, the sample restrictions in the analysis differ slightly depending on which paper I am comparing to. For the summary statistics I consider the most inclusive definition.

Table 4: **Selection Adjustments and Randomized Results**

(1)	(2)	(3)	(4)	(5)	(6)
Outcome [Citation]	*Uncontrolled Effect (Std. Error), [R^2]*	*Controlled Effect (Std. Error), [R^2]*	*Randomized Effect*	*R_{max} to match Randomized*	*Bias-Adjusted Effect ($\delta = 1, \phi = 1.018$) (Std. Error)*
Panel A: Exercise					
BMI, [3]	-1.58** (0.44) [0.023]	-1.30** (0.44) [0.078]	-0.60**	0.211	-1.00** (.47)
BMI, [4]	-1.93** (0.41) [0.020]	-1.49** (0.41) [0.048]	-1.01**	0.079	-1.03** (.36)
Weight in Kg, [1]	-4.56** (1.20) [0.177]	-3.98** (1.21) [0.201]	-4.60**	*N.V.*	-3.40** (1.18)
Weight in Kg, [2]	-2.41** (1.09) [0.210]	-1.53 (1.10) [0.230]	-1.15**	0.238	-0.63 (1.11)
Diastolic BP, [2]	-0.090 (0.67) [0.054]	0.004 (0.67) [0.071]	-1.80	*N.V.*	0.099 (1.02)
Diastolic BP, [3]	-0.330 (0.88) [0.017]	-0.197 (0.89) [0.050]	-3.00	*N.V.*	-0.061 (1.08)
Systolic BP, [2]	0.223 (1.00) [0.094]	0.655 (1.01) [0.119]	0.20	*N.V.*	1.09 (1.02)
Systolic BP, [3]	-0.071 (1.23) [0.069]	0.285 (1.24) [0.110]	-4.00	*N.V.*	0.64 (1.40)
Serum Glucose, [2]	-0.205 (0.14) [0.025]	-0.131 (0.14) [0.051]	-0.19**	*N.V.*	-0.055 (.095)
Serum Glucose, [4]	-0.313** (0.12) [0.027]	-0.237** (0.12) [0.048]	-0.16*	0.069	-0.158** (.067)
Cholesterol, [1]	-0.151* (0.09) [0.033]	-0.121 (0.09) [0.062]	-0.02	0.163	-0.091 (.071)
Cholesterol, [2]	-0.123 (0.08) [0.068]	-0.086 (0.08) [0.095]	0.05	0.198	-0.049 (.087)
Cholesterol, [3]	-0.159 (0.11) [0.035]	-0.103 (0.11) [0.084]	0.02	0.191	-0.047 (.123)
Triglycerides, [1]	-0.468** (0.13) [0.034]	-0.359** (0.13) [0.066]	-0.16	0.125	-0.25** (.093)
Triglycerides,[2]	-0.367** (0.10) [0.030]	-0.276** (0.10) [0.062]	-0.20**	0.089	-0.184** (.082)
Triglycerides,[3]	-0.393** (0.15) [0.027]	-0.271* (0.15) [0.072]	-0.16	0.113	-0.147 (.136)
HDL, [1]	0.104** (0.03) [0.016]	0.091** (0.03) [0.104]	0.13**	*N.V.*	0.076** (.031)
HDL, [2]	0.108** (0.03) [0.092]	0.081** (0.03) [0.132]	0.03	0.207	0.052* (.028)
HDL, [3]	0.095** (0.04) [0.022]	0.065* (0.04) [0.116]	0.03	0.228	0.035 (.040)
Weight in Kg	-3.03** (0.77) [0.139]	-1.58** (0.78) [0.184]	-0.13**	0.228	-0.105 (.80)
Diastolic BP	-0.256 (0.46) [0.047]	-0.153 (0.48) [0.065]	0.11	0.109	-0.048 (.52)
Systolic BP	-1.12 (0.93) [0.102]	-0.521 (0.96) [0.129]	0.22	0.162	0.095 (1.05)
Serum Glucose	-6.92** (2.39) [0.021]	-3.59* (2.44) [0.056]	-0.82	0.085	-0.19 (2.59)
Triglycerides	4.47 (5.39) [0.010]	3.03 (5.45) [0.064]	1.43	0.124	1.57 (6.00)
Cholesterol	0.199 (2.17) [0.012]	0.156 (2.23) [0.030]	-1.67	0.815	0.112 (2.52)
HDL	1.28* (0.80) [0.015]	1.02 (0.82) [0.053]	0.05	0.193	0.75 (.91)
Exercise (METS/wk)	5.27** (0.94) [0.028]	2.88** (0.94) [0.092]	0.18	0.165	0.44 (1.20)
Femur BMD	-0.019** (0.01) [0.175]	-0.006 (0.01) [0.260]	0.007**	0.341	0.008 (.007)
Introchanter BMD	-0.020** (0.01) [0.163]	-0.008 (0.01) [0.216]	0.0003	0.253	0.004 (.010)

Notes: This table displays the match between the results from observational data and randomized results. Citation Key: [1] Wood et al, 1988; [2] Stefanick et al, 1998; [3] Hellenius et al, 1993; [4] Anderssen et al, 1996. Full citations for randomized data and observational sample restrictions are in Appendix Table A.1. Controls in Panels A and B include : dummies for age and sex (controlled and uncontrolled regressions), dummies for income, dummies for education category, dummies for race, dummies for detailed marital status (controlled regressions only). The bias-adjustment in Column 4 is performed using a value of $\psi = 1.018$. Standard errors are bootstrapped over individuals. *significant at the 10% level, ** significant at the 5% level. *N.V.* R_{max} value to match to randomized is less than R_2.

Table 5: **Selection Adjustments, Out-of-Sample Outcomes**

Panel A: Exercise				
Outcome	*Uncontrolled Effect (Std. Error)*	*Controlled Effect (Std. Error)*	*Randomized Effect [Possible Direction, Sig.]*	*Bias-Adjusted Effect (Std. Error)*
Ever Diabetes	-0.035**(.009)	-0.019** (.009)	Negative, Not Significant	-0.003 (.010)
Mortality, with heart disease, Men	-0.132**(.041)	-0.115**(.041)	Negative, Significant	-0.098**(.05)
Overall Bone Density, Women	-0.013 (.012)	-0.0003 (.012)	Positive, Not Significant	0.013 (.014)
Panel B: Vitamin D and Calcium Supplementation				
Outcome	*Uncontrolled Effect (Std. Error)*	*Controlled Effect (Std. Error)*	*Randomized Effect [Possible Direction, Sig.]*	*Bias-Adjusted Effect (Std. Error)*
Ever Diabetes	-0.049**(.015)	-0.023 (.016)	Negative, Not Significant	0.002 (.018)
Mortality	-0.058**(.019)	-0.034*(.020)	Negative, Not Significant	-0.010 (.023)
Panel C: Vitamins and Mortality in Physician Health Study				
Outcome	*Uncontrolled Effect (Std. Error)*	*Controlled Effect (Std. Error)*	*Randomized Effect [Possible Direction, Sig.]*	*Bias-Adjusted Effect (Std. Error)*
Beta-Carotene Supplements	-0.035*(.019)	-0.022 (.019)	Negative, Not Significant	-0.009 (.021)
Vitamin E Supplements	-0.033***(.012)	-0.026**(.012)	Negative, Not Significant	-0.017 (.013)
Vitamin C Supplements	-0.029** (.011)	-0.021* (.012)	Negative, Not Significant	-0.013 (.013)

Notes: Exercise treatment: total exercise times per month (in units of 100). All adjustments are done using a value of $\delta = 1$ and $\psi = 1.018$. Citation List: Exercise and (a) diabetes (Orozco et al, 2008); (b) mortality (Heran et al, 2011); (c) bone density (Howe et al, 2011). Vitamin Supplementation and: (a) diabetes (de Boer et al, 2008); (b) mortality (LaCroix et al, 2009); (c) cognitive (Rossom et al, 2012); (d) cancer (Brunner et al, 2011). Physican Health Study: (a) Beta-carotene (Hennekens et al, 1996); vitamins E and C (Sesso et al, 2008).

Appendix A: Further Theoretical Results

This appendix discusses two additional issues related to the theory. Subsection A.1 below briefly contrasts the calculation of bias based on the coefficients to the calculation directly from the data suggested by Altonji, Elder and Taber (2005). Subsection A.2 discusses the case with two partially observed categories.

A.1. Altonji, Elder and Taber (2005) Calculation

Recall the true model:

$$Y = \alpha + \beta X + W + W' + \epsilon$$

and the fully controlled model

$$Y = \tilde{\alpha} + \Lambda X + \Psi W + \tau + \epsilon$$

For simplicity, assume $\epsilon = 0$ so there is no *iid* noise; in this case $R_{max} = 1$. Under the proportional selection assumption, I show that the exact formula for the bias on Λ is $\frac{\delta C_{wx} V_{w'}}{1-C_{wx}^2}$. Further, I show that $\delta \frac{(\xi-\Lambda)(1-R_2)}{(R_2-R_1)} = \left[\frac{\delta C_{wx} V_{w'}}{1-C_{xw}^2}\right] \frac{(1-C_{xw}^2-\delta^2 C_{xw}^2 V_{w'})}{(1-C_{xw}^2-\delta C_{xw}^2 V_{w'})}$ which is a very close approximation to the bias and is exact when $\delta = 1$.

Altonji, Elder and Taber (2005) suggest an alternative way to calculate this object using the raw data. In particular, they do the following:

1. Run the fully controlled regression and calculate V_τ and ΨW.
2. Regress X on ΨW , a regression we will write as:

$$X = \hat{\alpha} + \Gamma(\Psi W) + \tilde{X}$$

3. Calculate $V_{\tilde{x}}$ and extract Γ.

The calculation is then $\frac{\delta \Gamma V_\tau}{V_{\tilde{x}}}$.

Recall that $V_{\tilde{x}} = 1 - C_{xw}^2$ so the denominator is exactly equal to the denominator of the bias. The numerator differs slightly. In particular:

$$\begin{aligned} \Gamma &= \frac{C_{wx}}{1 - \frac{\delta C_{wx}^2 V_{w'}}{1-C_{wx}^2}} \\ V_\tau &= V_{w'}(1 - \delta \frac{\delta C_{wx}^2 V_{w'}}{1 - C_{wx}^2}) \end{aligned}$$

Combining, we find that:

$$\delta \Gamma V_\tau = \delta C_{xw} V_{w'} \frac{(1 - \delta \frac{\delta C_{wx}^2 V_{w'}}{1-C_{wx}^2})}{1 - \frac{\delta C_{wx}^2 V_{w'}}{1-C_{wx}^2}}$$

Simplifying, we find that:

$$\frac{\delta \Gamma V_\tau}{V_{\tilde{x}}} = \left[\frac{\delta C_{wx} V_{w'}}{1 - C_{xw}^2}\right] \frac{(1 - C_{xw}^2 - \delta^2 C_{xw}^2 V_{w'})}{(1 - C_{xw}^2 - \delta C_{xw}^2 V_{w'})}$$

which is exactly the same formula we get from the coefficient movement analysis and, as there, is exactly equal to the bias in the case where $\delta = 1$.

A.2. Extension: Multiple Categories with Proportional Selection

I consider now an extension to a case where there are two omitted categories, both of which have observed and unobserved components and both of which can be described with a proportional selection assumption. For

example, consider the relationship between child health and pollution. One partially observed category is family socioeconomic status. A second is area geography. We may want to ask the question of whether we can use a similar coefficient movement logic to infer true treatment effects in this case, perhaps by looking at the movement of coefficients between the regression with only the treatment as a control and that with the treatment plus observed socioeconomic status and geography controls.

Section 2.3 demonstrated that it is possible to infer the effect we would estimate if we saw *either* category fully observed, but does not discuss inferring the true causal effect.

We consider now the case where the true model is

$$Y = \alpha + \beta X + W + W' + M + M' + \epsilon$$

and X, W and M are observed. The variances of X, M and W are all equal to 1. We will assume equal selection on M and M' and on W and W' and will assume that both M elements are orthogonal to the W elements. Adopting proportional rather than equal selection makes the algebra more confusing without improving intuition. In particular, I assume

$$\begin{aligned} C_{w'x} &= C_{wx}V_{w'} \\ C_{m'x} &= C_{mx}V_{m'} \end{aligned}$$

I consider the following two equations – one with no controls, the second fully controlled:

$$\begin{aligned} Y &= \alpha + \xi X + \wr + \epsilon \\ Y &= \alpha + \Lambda X + \Theta_1 W + \Theta_2 M + \tau + \epsilon \end{aligned}$$

As before, the coefficient on Λ is biased. In this case: $\Lambda = \beta + \frac{C_{wx}V_{w'}+C_{mx}V_{m'}}{1-C_{mx}^2-C_{wx}^2}$.

We can ask directly how the coefficient difference relates to the bias and we find:

$$(\xi - \Lambda)\frac{[C_{wx}V_{w'} + C_{mx}V_{m'}]}{(1 - C_{mx}^2 - C_{wx}^2)(C_{mx} + C_{wx}) - (C_{mx}^2 + C_{wx}^2)[C_{wx}V_{w'} + C_{mx}V_{m'}]} = \frac{C_{wx}V_{w'} + C_{mx}V_{m'}}{1 - C_{mx}^2 - C_{wx}^2}$$

Using the same variance calculation we used before, we find that:

$$\frac{V_\tau}{V_\wr - V_\tau} = \frac{[C_{wx}V_{w'} + C_{mx}V_{m'}] - \frac{(V_{w'}+V_{m'})(1-C_{mx}^2-C_{wx}^2)}{[C_{wx}V_{w'}+C_{mx}V_{m'}]}}{[(1 - C_{mx}^2 - C_{wx}^2)(C_{mx} + C_{wx}) - (C_{mx}^2 + C_{wx}^2)[C_{wx}V_{w'} + C_{mx}V_{m'}]] - \frac{(1-C_{mx}^2-C_{wx}^2)(2-(C_{mx}+C_{wx})(C_{mx}(1+V_{m'})+C_{wx}(}{[C_{wx}V_{w'}+C_{mx}V_{m'}]}}$$

Comparing the two equations, we can see that the exact calculation we had before does not go through here. In particular, there is an added piece in the numerator and the denominator of the variance calculation. To the extent these are small or roughly cancel each other, the formula given in Proposition 1 in the paper will be approximately correct. More generally reassuring, this suggests that in this case the bias does still scale with coefficient movements, and with the r-squared movements.

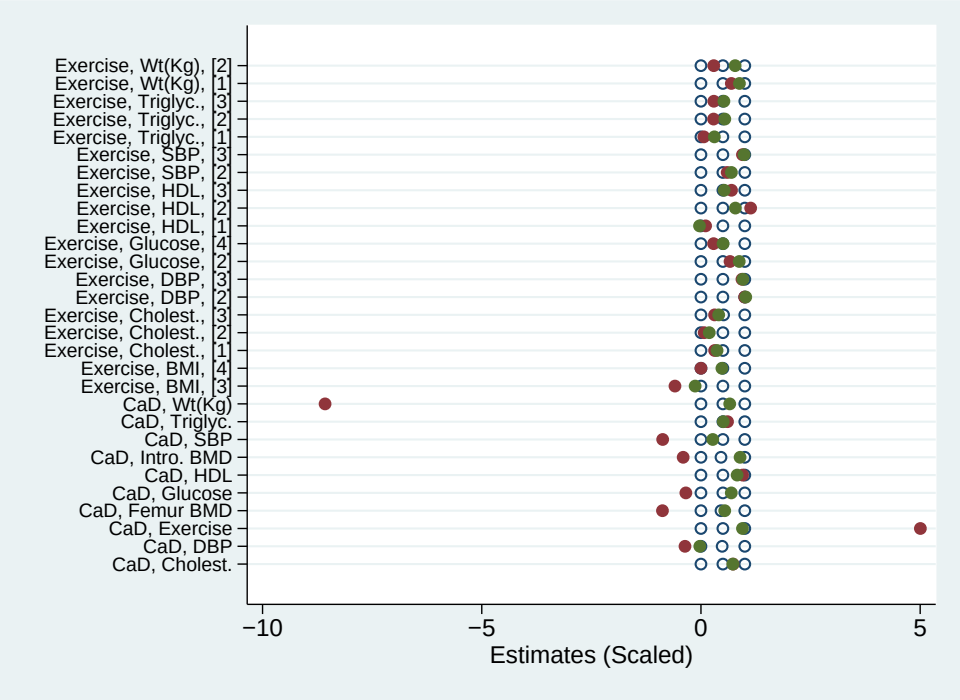
Appendix Tables and Figures

Table A1: Citation for Randomized Outcomes

Outcome	*Citation*	*Sample Restrictions (if any)*
Exercise, BMI, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
Exercise, BMI, [4]	Anderssen et al, 1996	Age 30-50
Exercise, Wt(Kg), [1]	Wood et al, 1988	Female, 30-59
Exercise, Wt(Kg), [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, DBP, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, DBP, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
Exercise, SBP, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, SBP, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
Exercise, Glucose, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, Glucose, [4]	Anderssen et al, 1996	Age 30-50
Exercise, Triglyc, [1]	Wood et al, 1988	Female, 30-59
Exercise, Triglyc, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, Triglyc, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
Exercise, Cholest, [1]	Wood et al, 1988	Female, 30-59
Exercise, Cholest, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, Cholest, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
Exercise, HDL, [1]	Wood et al, 1988	Female, 30-59
Exercise, HDL, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, HDL, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
CaD, Wt(Kg)	Caan et al, 2007	Women, 55-85
CaD, DBP	Margolis et al, 2008	Women, 55-85
CaD, SBP	Margolis et al, 2008	Women, 55-85
CaD, Glucose	de Boer et al, 2008	Women, 55-85
CaD, Triglyc	Rajpathak et al, 2010	Women, 55-85
CaD, Cholest	Rajpathak et al, 2010	Women, 55-85
CaD, HDL	Rajpathak et al, 2010	Women, 55-85
CaD, Exercise	Brunner et al, 2008	Women, 55-85
CaD, Femur BMD	Jackson et al, 2011	Women, 55-85
CaD, Intro. BMD	Jackson et al, 2011	Women, 55-85

Notes: This table shows the source of the randomized estimates. The text of the outcome matches the form of citation in Figure 2.

Appendix Figure 1: Adjusted Coefficients on Controlled Coefficient Scale



Notes: This table shows Figure 2b graphed on the same scale as Figure 2a.