Is selection on observables informative about selection on unobservables? The correct answer is probably "No."

Abstract: This is a thesis on whether selection on observables is informative about selection on unobservables. By creating my own adaption to the paper of Black et al. (2015), I create a method by which one can investigate whether selection on observables is informative about selection on unobservables. Although this is pioneering research, I am able to provide some suggestive evidence that selection on observables is *not* informative about selection on unobservables. As a 'by-product' I found some evidence that it is sometimes better not to control for observables than to control for these variables.

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master thesis Policy Economics
Name student: Tijmen Kars
Student ID number: 374110

Supervisor: Prof. dr. H.D. Webbink
Second assessor: C.M. Oosterveen MSc

Date final version: May 16, 2017

I would like to thank my supervisor, Dinand Webbink, for supervision and for the idea of relating the papers by Altonji, Elder and Taber (2005) and Oster (forthcoming) to the paper by Black, Joo, LaLonde, Smith and Taylor (2015). I also would like to thank Matthijs Oosterveen for being the second assessor of this thesis.

Table of Contents

1. Introduction

One of the fundamental issues in economics - and in the social sciences in general – is how to credibly estimate causal effects. Credible estimates can be obtained by using experiments and quasi-experiments. However, often, these are not available to use. The alternative is to use observational data. When observational data is used, (partial) associations are used to get estimates of causal effects. However, it could be that there is selection bias in these estimates, even if one controls for a lot of covariates. Recently, new techniques have been developed in order to use observational data – and thus (partial) associations – for credibly estimating causal effects. These techniques rely on the assumption that selection on observable factors is informative about selection on unobservable factors. In this thesis I will investigate whether this assumption is plausible. For this, I will make use of existing (quasi-)experimental studies. Within these studies we know what the true causal effect is for one specific group (the *Compliers*).

In many situations, researchers have used methods that indeed deliver estimates that can be interpreted as causal under some assumptions, examples are Instrumental Variables (IV), Regression Discontinuity Designs and Difference-in-Differences (DID) methods.[1] Moreover, experiments are used in order to establish causal effects. However, most of the time, we do not know whether unobserved factors would have driven the results significantly, if we would have calculated associations between an outcome and treatment instead of using IV or RDD.[2] The reason is that IV and RDD provide a local average treatment effect (LATE), while calculated associations provide estimates of average treatment effects (Angrist and Pischke, 2009).

Angrist and Pischke (2010) write about a "credibility revolution" in economics. They describe how empirical research in economics has evolved over the last decades. One of their main messages is that empirical research in microeconomics has experienced a "credibility revolution". What they mean with this is that first, the field was primarily concerned with improving econometric methods, while it was still very common to look at associations between certain variables in order to estimate causal effects. However, after some time the "credibility revolution" kicked in (the exact starting point is difficult to determine, according to Angrist and Pischke [2010]), and this resulted in a shift in focus from econometric methods to research design. As described by Angrist and Pischke (2010): there are essentially three types of credible quasi-experimental designs: 1) IV; 2) RDD; 3) DID. Of course, experiments are also credible.

[1] DID methods only show a credible estimate of a causal effect if the Common Trends Assumption (Angrist and Pischke, 2009) is likely to hold.

[2] This does not hold for DID methods where the Common Trends Assumption is likely to hold.

This thesis is about whether (controlled) associations can be interpreted as causal effects or can be interpreted as causal effects after making some corrections and/or innocuous assumptions. Consequently, this thesis presents research about whether selection on observable factors is informative about selection on unobservable factors. Therefore, this is an important thesis. The research in this thesis can be considered as pioneering research. As such, there is not yet similar empirical evidence on this topic.

If associations between variables or methods that are based on associations (for example, Oster [forthcoming][3]) deliver credible estimates of causal effects, then this could potentially save a lot of research costs. The reason is that researchers do not need to spend a lot of time on finding a sound quasi-experiment. However, if it seems that selection on observables is not informative about selection on unobservables, then it is confirmed that current solutions (like IV and RDD) are indeed very important. Furthermore, it would indicate that we should be very careful when trying to interpret associations as causal effects.

The practical application of the idea that selection on observables contains information about the selection on unobservables began with the study of Altonji, Elder and Taber (2005). These researchers used this idea to interpret their calculated associations in a causal way after making some assumptions. Recently, Oster developed a formal framework in which selection on observables is used in order to establish causal effects from statistical associations between an outcome and treatment variable. Basically, Oster expands and formalizes the methodological and technical ideas that are present in the study of Altonji et al. (2005).

In this thesis, I will present evidence of whether selection on observables says something about selection on unobservables. The paper by Black, Joo, LaLonde, Smith and Taylor (2015) is at the basis of the methodology I will use. The paper by Black et al. presents a method which indicates the amount of selection bias (that is, the amount of selection on unobservables) which exists between different groups within an IV setting. Thus, it makes use of IV setups and infers the degree of selection bias between different groups within this setting. I will make use of this, by using existing (quasi-)experimental studies and applying the method of Black et al. (2015). The reason is that good (quasi-)experimental studies deliver credible estimates of causal effects for the Compliers within a (quasi-)experiment. Consequently, we can infer the degree of selection bias between the Compliers and the Never Takers and between the Compliers and the Always Takers. Moreover, in order to link this to the selection bias story of calculated associations, I will apply the method by Black et al. with and without control variables. This

[3] From now on, if I write "Oster" without mentioning a year, then this refers to Oster (forthcoming). The reason is that writing "Oster (forthcoming)" many times, probably leads to a decrease in the readability of this thesis.

will give the selection on observables and selection on unobservables. Consequently, I am able to compare the direction of the selection on observables with the direction of the selection on unobservables. As far as I know, I am the first to show this. To my knowledge, I am also the first to show estimates of selection on observables.[4]

The results show that it is probable that quite often, selection on observables is not informative about selection on unobservables. Since it is beyond the scope of the thesis to calculate standard errors for the ratio of selection on unobservables to selection on observables, I cannot draw strong conclusions. However, since many ratios are different from each other it is very likely that selection on observables is not informative about selection on unobservables. This implies that using a correlation or methods that are essentially using a correlation is probably a large mistake. A very surprising result is that sometimes an uncontrolled estimate is better – less biased – than a controlled estimate! There are even results that show that an uncontrolled estimate was unbiased while the controlled estimate was biased!

Because I will go more into detail about the papers by Altonji et al. (2005), Oster and Black et al. (2015) in the next section, I briefly discussed them here. Besides, my thesis does not contain something like a literature review which presents previous empirical results which have been presented in past literature, because such literature does not exist. Instead, I discuss the papers by Altonji et al, Oster and Black et al. while I describe the empirical approach.[5] This is the content of the next section. In section 3, I describe the data and mention the publications that are used. Section 4 contains the results per publication and summarizes all results. Section 5 concludes.

2. Empirical approach

2.1 Selection bias in (partial) associations.

As already said, looking at the (partial) correlation between variables does not automatically deliver credible estimates of the causal effect of one variable on another variable. There is a potential endogeneity problem. ‘Potential’, because we do not know whether this problem arises or not. The endogeneity problem, can be illustrated with the help of this regression formula:

$$Y \ = \ \beta_0 + \beta_1 D + \beta_2 X' + \varepsilon \qquad (1)$$

[4] Altonji et al. (2005), Oster and probably others as well, provide estimates of the absolute of the selection on observables. In contrast, I also show the sign/direction of the selection on observables.

[5] The reason is that I want to limit the number of repetitions.

where Y denotes the outcome, β_0 a constant term, D the treatment variable, X a vector of control variables and ε the error term.[6] Suppose, we are interested in the causal effect of D on Y then we could regress Y on D without controls (so equation (1) without X). However, this will not deliver the correct estimate of the causal effect of D on Y if an endogeneity issue arises. The endogeneity problem means that the D is correlated with the error term. Consequently, the β_1 is potentially biased. Then we could estimate equation (1) – so include controls X – in an attempt to solve the endogeneity problem. However, having X included may not be sufficient, because we do not know whether we have controlled for all confounders. That is, there can be unobserved factors or unobservables which drive (part of) the estimation result. If the unobservables drive (part of) the result, then this means that there is an endogeneity problem. The part of the difference between the true causal effect and the estimated (partial) association that is due to unobservable confounders can be labelled as selection bias or selection on unobservables.

2.2 How to use (partial) associations for estimating causal effects as proposed by Altonji et al. (2005).

Consider again equation (1). Sometimes, it is (nearly) impossible to employ a research design which will give us credible estimates of the effect of D on Y (Altonji et al., 2005). Therefore, Altonji et al. invoke the assumption that selection on observables is informative about selection on unobservables in order to use correlation in a way that tries to account for bias due to these unobservables. This idea of using correlation is molded into two mirroring approaches which both lead to the same conclusion. Since Oster expands and formalizes the two approaches of Altonji et al., I only briefly discuss them here; the next subsection contains a description of the study by Oster which is more elaborate. The first approach is to estimate equation (1) – essentially this is calculating the partial correlation between Y and D – and to accompany this by establishing bounds on the coefficient of D. These bounds are located based on: 1) information from the regression of Y on D and control variables; 2) information about the degree of selection on observables;[7] 3) an assumption about the ratio of selection on unobservables to selection on observables (Altonji et al., assume that this is equal to one, an assumption they consider as conservative). This means that the coefficient on D in equation (1)

[6] Subscripts are suppressed for convenience.

[7] Although it is not explicitly mentioned by Altonji et al. (2005), according to me, the sign/direction of the selection on observables is an assumption.

is the upper-bound estimate of the effect and, roughly speaking, the lower-bound is estimated by adjusting the regression by invoking the assumption that selection on unobservables has the same strength as the selection on observables.[8] The second approach is calculating how high the ratio of selection on unobservables to selection on observables should be in order to explain away the estimated effect. That is, how high would this ratio be if the true causal effect were to be equal to zero. If this ratio is assumed to be unrealistic, then the coefficient on D is assumed to be a credible estimate of the causal effect. As noted by Altonji et al. (2005) these approaches "may be helpful" (p. 153) in estimating causal effects, when calculating a correlation is the only method you can use.

2.3 How to use (partial) associations for estimating causal effect: Oster expands and improves Altonji et al. (2005).

Oster notes that after the publication of Altonji et al. (2005), few researchers applied their method (or similar methods) when the only thing they could do was calculating (partial) correlations. Oster shows that – under some assumptions – the core idea from Altonji et al. can be considered as an improvement over only calculating correlations. Oster expands the two approaches from Altonji et al. Therefore, I will focus on Oster's methods and technical applications in this section and not on the study by Altonji et al.

Oster also contains two mirroring methods – which are very similar to the ones in Altonji et al. (2005): one that creates a bounding set for the true causal effect based on calculated correlations and an assumption about the ratio of selection on unobservables to selection on observables and another that calculates the ratio of selection on unobservables to the selection on observables if the true causal effect were to be equal to a certain value (in most contexts a value of zero would likely be the appropriate hypothesis).[9] I will now elaborate on this (one should consult Oster if one wants to know all assumptions and technical aspects).

Oster notes that a common practice within empirical economics is to check for the robustness of estimates to omitted variables bias (OVB) by including controls and looking at the movements of the parameter (or coefficient) of interest. If this parameter does not change dramatically, then it is commonly assumed that the parameter is robust to OVB. Oster argues

[8] In Altonji et al. (2005) effects are positive.

[9] Oster also contains a third method. This method encompasses calculating bounds for the R^2 that would result if one would estimate equation (1) with all relevant observed and unobserved variables included. This "could then be discussed in terms of whether it is plausible that the unobservables explain more of the variance than implied by this value" (Oster, forthcoming, p. 25). I do not discuss this third method in my thesis, because I consider the third method as less appropriate and less transparent than the other two methods from Oster within the context of my thesis.

that looking at the movements of the parameter of interest is not enough in order to conclude that a parameter is robust to OVB. One should also account for the movements in R^2 (Oster, forthcoming). The reason, according to Oster, is that some controls do not explain a lot of the variance in the dependent variable. Thus, if a researcher includes a lot of these controls in the analysis, then it is indeed not unexpected that the parameter of interest does not change dramatically.

The two mirroring methods of Oster are based on one estimator. Oster presents two types of the estimator: the restricted estimator and the unrestricted estimator. The restricted estimator employs somewhat different assumptions than the unrestricted estimator. In most applications, the restricted estimator is inaccurate – it provides a simple robustness calculation and it shows a lot of the intuition behind the unrestricted estimator. When doing research the unrestricted estimator is more appropriate to use, therefore I will focus on the unrestricted estimator. When I only write 'estimator' within the context of Oster , it indicates the unrestricted estimator. Oster has written a STATA program that accompanies her paper. This program is called *psacalc.*[10]

The following regression model is an adaptation from equation (1) in Oster[11], she uses this for explanation:

$$Y = \gamma_0 + \beta D + \Psi\omega^o + W_2 + \epsilon \qquad (2)$$

γ_0 denotes the constant term, D is "(scalar) treatment and ω^o is a vector of the observed controls, ω_1^o, ..., ω_J^o. The index W_2 is not observed. Define $W_1 = \Psi\omega^o$ and assume that all elements of ω^o are orthogonal to W_2, so W_1 and W_2 are orthogonal (Oster, forthcoming, p. 13)." W_2 is similar to W_1, with the only difference that W_1 is observed. I think it may be possible that W_2 does not exist within certain contexts. Suppose it exists and that there are J unobserved variables within W_2, then one could define $W_2 = \Phi\omega^u$, where ω^u is a vector of unobserved variables, ω_1^u, ..., ω_J^u (Oster, 2015).[12] Oster notes that the assumption that W_1 and W_2 are orthogonal might be considered as somewhat strange if one cares about the relation between the two. However, she shows that the orthogonality assumption and correlation between W_1

[10] *Psacalc* can be downloaded via the ssc in STATA

[11] There are two minor differences: 1) Oster's equation used the notation X for denoting (scalar) treatment (I use D instead for avoiding confusion and because it is used often in the treatment evaluation literature); 2) Oster's formula includes a constant term, however, this is not clearly mentioned by her, according to me. Therefore, I explicitly display a constant term.

[12] This is not explicitly noted in Oster, however, it was noted in the working paper (Oster, 2015) and I think this is implicitly present in Oster. I consider it as still noteworthy, because it makes more clear the meaning of W_2.

and W_2 are not contradicting each other (Oster, forthcoming, Appendix A). Notice that Oster tries to improve the common practice for checking the robustness of estimates to OVB, by accounting for R^2 movements. Thus, a crucial assumption in her estimator is the assumption that selection on observables is informative about selection on unobservables. The degree to which selection on observables is informative about selection on unobservables can be expressed in a ratio: "the proportional selection relationship". The proportional selection relationship can be defined as follows: $\delta \frac{Cov(W_1,D)}{Var(W_1)} = \frac{Cov(W_2,D)}{Var(W_2)}$ (Oster, forthcoming, p. 13). δ is called the "coefficient of proportionality" (Oster, forthcoming, p. 13). If $|\delta| = 1$, then this means that selection on observables is of the same importance as the selection on unobservables. If $|\delta| > 1$, then selection on observables has less strength than selection on unobservables and if $|\delta| < 1$, then selection on observables has more strength than selection on unobservables. If $\delta < 0$, then selection on observables is of the opposing direction compared to the selection on observables. For example, suppose a regression which includes the observed controls results in 2 as the estimate for β and without controls it would be equal to 1.5, while $\delta > 0$, then hypothetically including the unobservables as additional controls would lead to an estimate for β which is above 2. If $\delta < 0$, then adding the unobservables would lead to an estimate of β below 2.

The estimator of Oster consists of three steps: 1) run a regression of Y on D without controls (see equation (3) below). The coefficient on D is labelled as $\dot{\beta}$ and the R^2 of this regression is labelled as $\dot{R}$. 2) run a regression of Y on D with controls included (see equation (4) below). The coefficient on D is labelled as $\tilde{\beta}$ and the R^2 of this regression is labelled as $\tilde{R}$. 3) Use the information from steps 1 and 2 to investigate whether the $\tilde{\beta}$ is robust to OVB. Here the two mirroring methods of Oster come into play. The first method is calculating a bounding set for $\tilde{\beta}$ and the second method is calculating the δ – the coefficient of proportionality – that would occur if the true causal effect would be equal to zero, this δ is labelled as $\hat{\delta}$. If $\hat{\delta}$ has a value that is unlikely to occur in reality, the estimated $\tilde{\beta}$ is assumed to be robust to OVB.

For step 1, we can rewrite equation (2) as:

$$Y = \gamma_0 + \dot{\beta} D + \epsilon \tag{3}$$

For step 2, we can rewrite equation (2) as:

$$Y = \gamma_0 + \tilde{\beta} D + \Psi \omega^o + \epsilon \tag{4}$$

The bounding set for the true causal effect, β, is called Δ_s. $\Delta_s = [\tilde{\beta}, \beta^*]$. Thus, $\tilde{\beta}$ is one bound and β^* is the other bound. β^* has multiple solutions if δ is assumed to be unequal to 1 and/or if the direction of the covariance between W_1 and D is changed by the bias due to unobservables, but it has one solution if $\delta = 1$ and if the direction of the covariance between W_1 and D is not changed by the bias due to unobservables (Oster, forthcoming). Oster argues that $\delta = 1$ is a plausible assumption and that for implementation matters the direction of $Cov(W_1, D)$ is assumed to be unaffected by bias from unobservables. Unfortunately, Oster does not provide any formula for β^* in the case of the unrestricted estimator (she only provides it in the context of the restricted estimator), but by employing these two assumptions, calculation of β^* basically boils down to the following: β^* is calculated in such a way that the direction of the movement from $\tilde{\beta}$ to β^* is the same as the direction of the movement from $\dot{\beta}$ to $\tilde{\beta}$. In addition, the calculation of β^* takes into account $\dot{R}$ and $\tilde{R}$.[13] An example: if $\dot{\beta} = -4$ and $\tilde{\beta} = 4$, then $\beta^* > 4$.

Let us now turn to the second method. $\hat{\delta}$ has always a unique solution, so if the direction of $Cov(W_1, D)$ is assumed to be affected by bias due to unobservables and/or one suspects $\delta \neq 1$, then the second method is an easy way to go (Oster, forthcoming). $\hat{\delta}$ provides the coefficient of proportionality that would arise if the true causal effect would be equal to zero by taking into account $\dot{\beta}$, $\tilde{\beta}$, $\dot{R}$ and $\tilde{R}$. Thus, $\hat{\delta}$ is calculated by setting $\beta^* = 0$.[14, 15] To consider the formula for $\hat{\delta}$ for the unrestricted estimator one should consult Oster, because this formula contains a lot of notation.

2.4 The assumption that selection on observables is informative about selection on unobservables is crucial.

It can be concluded that the assumption that selection on observables is informative about selection on unobservables is crucial in Oster and Altonji et al. (2005). Moreover, the same holds for all papers that calculate correlations and perform coefficient stability tests with this (for example, by adding extra controls). The reason is that if one employs coefficient stability

[13] The R^2 that is linked to β^* is $1.3\tilde{R}$. It is called R_{max} by Oster. R_{max} is the R^2 that is assumed to result if equation (2) would be estimated with all relevant observed and unobservable variables included as controls. One could think that it would be conservative to assume $R_{max} = 1$, however, it is noted by Oster that this is *too* conservative in most cases. Based on empirical analyses of experimental data, Oster concludes that $1.3\tilde{R}$ provides a reasonable upper bound for R_{max} if $1.3\tilde{R} \leq 1$. Therefore, $R_{max} = min(1.3\tilde{R}, 1)$.

[14] Note that Oster indicates that $\hat{\delta}$ can be calculated for any hypothesis concerning β. However, Oster writes that $\beta = 0$ (or $\beta^* = 0$) is appropriate in probably most cases.

[15] Here too, the R^2 that is linked to β^* is $1.3\tilde{R}$. See footnote 13 for more information.

tests based on correlations, one implicitly (if not explicitly stated) assumes that selection on observables is informative about selection on unobservables. We can ask ourselves whether this is truly so. By using the paper by Black et al. (2015) and making some adaptions to their paper, I am able to uncover empirical evidence which will shed light on the plausibility of the assumption that selection on observables is informative about selection on unobservables. Note that I will only consider binary treatment indicators and binary instruments.

2.5 Calculation of selection on observables and selection on unobservables.

Black et al. (2015) provide a method which shows the amount of selection bias between three different groups within a LATE framework. These three groups are: Always Takers, Compliers and Never Takers. Thus, it makes use of the IV setup. When I write about "IV setup" I mean that the setup applies to the Instrumental Variables (IV) estimator, fuzzy Regression Discontinuity Design (fuzzy RDD) and experiments without perfect compliance. The IV setup consists of two equations which are simultaneously solved:

$$D_i = \gamma_6 + \tau_0 Z_i + \tau_1 {\omega^o}'_i + \xi_i \tag{5}$$

$$Y_i = \gamma_7 + \theta_0 \widehat{D}_i + \theta_1 {\omega^o}'_i + \eta_i \tag{6}$$

γ_6 and ξ (γ_7 and η) denote the constant term and error term respectively in equation (5) (in equation (6)), D is the treatment variable, $\widehat{D}$ is the fitted value of D obtained by estimating equation (5), Y is the outcome, Z is the instrument and ω^o is a vector of control variables. Subscript i indicates the i^{th} cross-sectional unit of observation.[16] The effect for compliers is θ_0 (Angrist and Pischke, 2009).

The Always Takers are the units for which D always equals 1, the Compliers are the units that show a change in D due to a change in Z and Never Takers are the units for which D always equals 0. This can be formulated in the following way:

- The group $(D = 1|Z = 1)$ can contain Always Takers and Compliers.
- The group $(D = 1|Z = 0)$ only contains Always Takers.
- The group $(D = 0|Z = 1)$ only contains Never Takers.
- The group $(D = 0|Z = 0)$ can contain Never Takers and Compliers.

I illustrate this in Table 1.

[16] If one has to deal with time-series data, i can be replaced by t. If one has to deal with panel data, i can be replaced by it.

Table 1: illustration of groups within the IV setup

	$Z = 0$	$Z = 1$
$D = 0$	Compliers and/or Never Takers	Never Takers
$D = 1$	Always Takers	Compliers and/or Always Takers

Within the IV setup, Black et al. (2015) demonstrate a way of calculating the selection bias (that is, selection on unobservables), between Always Takers and Compliers and between Never Takers and Compliers. The selection bias between Always Takers and Compliers is calculated based on the subsample for which $D = 1$ and the selection bias between Never Takers and Compliers is calculated based on the subsample for which $D = 0$. In Black et al. (2015), this is carried out by regressing Y on Z and a vector of controls for each subsample separately. The regression can be displayed in the following way:

$$Y_i = \gamma_0 + \tilde{\alpha} Z_i + \Psi {\omega^o}'_i + \epsilon_i \quad (7)$$

where Y is the outcome, γ_0 the constant term, Z the instrument, ${\omega^o}'_i$ a vector of controls and ϵ is the error term. $\tilde{\alpha}$ shows (a part of) the selection bias. The reason of why it could show a *part* of selection bias is that some Always Takers could have $Z = 1$, while the compliers always have $Z = 1$ in subsample $D = 1$, similarly, Never Takers could have $Z = 0$, while compliers always have $Z = 0$ in subsample $D = 0$. Therefore, $\tilde{\alpha}$ potentially displays an underestimation of the true degree of selection bias. In principle, if $\tilde{\alpha}$ displays an underestimation of the true degree of selection bias, $\tilde{\alpha}$ could be transformed to the true degree of selection bias by accounting for the proportion of Always Takers, Never Takers and Compliers in the subsamples. I will not do this in this paper, because it is not necessary for the calculation of δ; I will return to this issue later. By calculating $\tilde{\alpha}$ – which is selection on unobservables – we get a crucial ingredient for testing the assumption that selection on observables is informative about selection on unobservables. By adapting the method of Black et al. (2015) we can also calculate the selection on observables. The first step is to remove all controls from equation (7) – except the constant term. This yields equation (8):

$$Y_i = \gamma_0 + \dot{\alpha} Z_i + \epsilon_i \quad (8)$$

where $\dot{\alpha}$ shows (a part of) the selection bias. $\dot{\alpha}$ is almost identical to $\tilde{\alpha}$, the only difference is that $\dot{\alpha}$ is an uncontrolled estimate. Then we follow the same interpretation procedure as for equation (7).

Note that $\dot{\alpha}$ and $\tilde{\alpha}$ both yield the selection on unobservables. However, in the case of $\dot{\alpha}$ this selection bias is calculated without controls and in the case of $\tilde{\alpha}$ it is calculated with controls, so the difference between the two coefficients shows the amount of selection on observables![17] I define: $\breve{\alpha}$ equals the amount of selection observables. Therefore,

$$\breve{\alpha} = \dot{\alpha} - \tilde{\alpha}. \tag{9}$$

If $\breve{\alpha} = 0$, then adding controls ω^o has no effect on changing the amount of selection bias. If $|\dot{\alpha}| > |\tilde{\alpha}|$, then adding controls ω^o leads to less selection bias. If $|\dot{\alpha}| < |\tilde{\alpha}|$, then adding controls ω^o leads to more selection bias.

In order to avoid confusion, I will differentiate between the terms "selection bias on observables" or "selection on observables" and "selection bias on unobservables" or "selection on unobservables" and I will introduce the term "total bias". I define $\dot{\alpha}$ as "Total bias".[18] The reason is that $\dot{\alpha}$ can be considered as the sum of selection on unobservables (when controls are included) and selection on observables. To see this, one can rewrite equation (9) in the following way:

$$\dot{\alpha} = \breve{\alpha} + \tilde{\alpha}.$$

Suppose we are interested in the effect of D on Y, and β is the coefficient that captures this effect. Suppose the true β is 2 and having no controls (except for a constant term) would lead to an estimated β of 2. Suppose that adding controls (the observables) would lead to an estimated β of 3, in this case selection on observables equals -1 (2 minus 3) and selection on unobservables equals 1. Suppose we would be able to – hypothetically – add the unobservables as additional controls, then the estimated β would be 2 again. The regression without any controls leads to an unbiased estimate of 2, so total bias is zero. This means that selection bias on observables and selection bias on unobservables compensate each other. If one would add the observables as controls, then there would be bias in the estimate. The reason that bias emerges as controls are added, is the fact that (part of) the compensating selection is removed. With "compensating selection" I mean that this selection compensates the selection caused by unobservables. As soon as controls are added, a biased estimate of β emerges.

I call the method of Black et al. (2015) combined with my own adaptation of that method the "alpha method". Analyzing data with help of the alpha method sheds light on the plausibility

[17] I would like to thank my supervisor, Dinand Webbink, for making me aware of this.

[18] I would like to thank my supervisor, Dinand Webbink, for the idea for this definition.

of the assumption that selection on observables is informative about selection on unobservables. If $|\dot{\alpha}| < |\tilde{\alpha}|$, then selection on observables is of opposite direction compared to the selection on unobservables. One may consider it as evidence against the assumption that selection on observables is informative about selection on unobservables. However, this is wrong: if all or almost all analyses show that $|\dot{\alpha}| < |\tilde{\alpha}|$, then observables can be considered as informative about selection on unobservables. The reason is that we would know that the direction of change in the parameter of interest probably should be turned around when interpreting the result. For example, in an empirical effect study using correlations, if the uncontrolled estimate is 4 and adding controls leads to an estimate of 5, then we know that the true causal effect should be lower than 5. If differences between the magnitudes of $|\dot{\alpha}|$ and $|\tilde{\alpha}|$ differ among analyses – that is, sometimes $|\dot{\alpha}| = |\tilde{\alpha}|$ (or $|\dot{\alpha}| \approx |\tilde{\alpha}|$), sometimes $|\dot{\alpha}| > |\tilde{\alpha}|$ and sometimes $|\dot{\alpha}| < |\tilde{\alpha}|$ – then assuming that selection on observables is informative about selection on unobservables would be wrong in general.[19] Consequently, in general when carrying out an empirical effect study using correlations, one cannot know whether a changed estimate due to controlling for observables improves or worsens the uncontrolled estimate. However, if the estimated coefficient remains the same we know that controlling for the included observables does not matter. But in this latter case one still cannot know whether the unobservables confound the estimated effect, if differences between the magnitudes of $|\dot{\alpha}|$ and $|\tilde{\alpha}|$ differ among analyses in this thesis (and perhaps, in future research). From now on I will use the notation $Y_{D=1,i} = Y_{1,i}$, $Y_{D=0,i} = Y_{0,i}$, $\delta_{D=1,i} = \delta_{1,i}$, $\delta_{D=0,i} = \delta_{0,i}$, $\alpha_{D=1,i} = \alpha_{1,i}$ and $\alpha_{D=0,i} = \alpha_{0,i}$, where α stands for $\dot{\alpha}$, $\tilde{\alpha}$ or $\breve{\alpha}$.[20]

As already noted, δ is the ratio of selection on unobservables to selection on observables. With the alpha method we can calculate the selection on unobservables and selection on observables. Therefore,

$$\delta = \frac{\tilde{\alpha}}{\breve{\alpha}} \tag{10}$$

Thus, by employing the alpha method and using equation (10) one can show which δ results when using the alpha method. As noted by Oster, a δ in the interval of $[0, 1]$ is considered as plausible, with $\delta = 1$ being a conservative choice. However, if it seems that empirically δ is far outside the $[0, 1]$ interval, this means that the we have to adjust the view of Oster. It does not mean that Oster's estimator is completely useless, I will come back to this issue later. As I

[19] I explicitly mentions that this would be the case "in general", since it might be the case that some research is carried out within a very special context.
[20] Depending on context, subscripts i may be suppressed.

wrote earlier, it is not necessary to adjust $\tilde{\alpha}$ for the proportion of Always Takers, Compliers and Never Takers for the calculation of δ. The reason is that if one would adjust $\tilde{\alpha}$ one would also adjust $\dot{\alpha}$ and consequently adjust $\breve{\alpha}$. This adjustment would then be carried out to the denominator and numerator in equation (9), consequently, the calculated δ would exactly be the same as without the adjustment.

Unfortunately, it is beyond the scope of this thesis to calculate standard errors for $\breve{\alpha}$ and δ. However, since I calculate many δs, I will be able to assess the likelihood that δ is always in the interval $[0, 1]$. Again note that this thesis is pioneering research. Besides, Altonji et al. (2005) do not provide standard errors for δ either. In addition, Oster does not explicitly mention a way of calculating standard errors for $\hat{\delta}$.[21]

Note that if we would – hypothetically - add all relevant observables and unobservables to equation (7) (or equation (8)), we would have a multicollinearity issue. We can think of this as if we were interested in the 'effect' of being an Always Taker or Complier on Y_1 and in the 'effect' of being a Never Taker or Complier on Y_0 in the context of equations (7) and (8), while noting that being an Always Taker, Complier or Never Taker is a function of observed and/or unobserved factors. If we would add all relevant observed and unobserved controls to the regression and retain Z, Z and the controls would show perfect multicollinearity. Notice that it is difficult to calculate for software (or by hand) which variable(s) would be redundant. The observed and unobserved variables would redundant if one prefers to have Z included in the regression. However, it is most straightforward to consider Z as the redundant variable. This implies that the α after controlling for all relevant factors is zero.

3. Empirical studies which will be used

In order to test the assumption that selection on observables is informative about selection on unobservables, I will make use of existing (quasi-)experimental studies. First, it is important that the studies have a binary treatment indicator and a binary instrument. Second, the studies should be replicable (that is, data is available and if necessary original programming code is available as well). The studies I will use are the following: Angrist and Evans (1998a); Angrist et al. (2002a); Abadie, Angrist and Imbens (2002) and Kazianga et al. (2013a). For these studies I will estimate equations (7), (8), (9) and (10) for the treatment and control group separately. The first two studies have been published in the American Economic Review, the third study

[21] Oster only mentions a bootstrap procedure in order to calculate standard errors for β^*.

has been published in Econometrica and the last study has been published in the American Economic Journal: Applied Economics. These are well-respected journals, so the quality of the instruments is expected to be high. For the study of Angrist and Evans (1998a) I downloaded the data from Angrist and Evans (1998b) and the programs from Angrist and Evans (1998c). For the study of Angrist et al. (2002a), I downloaded the data from Angrist et al. (2009a) and the programs from Angrist et al. (2009b). In addition, I consulted Angrist et al. (2002b) for additional information on the use of the data and programs. For the study of Abadie et al. (2002), I downloaded the data from Abadie, Angrist and Imbens (2009). Because the variables in the dataset of Abadie et al. (2009) were unlabeled I had to infer which label belongs to which variable by looking at the means, standard deviations, etc. Unfortunately, I could not identify all relevant variables. I got the remaining labels from the "jtpa.dta" file from Froelich and Blaise (2008). Replication of the relevant analyses of Abadie et al. (2002) showed that I correctly identified all relevant variables. For the study of Kazianga et al. (2013a), I downloaded the data and programs from Kazianga et al. (2013b).

The paper by Angrist and Evans (1998a) is about the effect of having more than two children compared to having two children on the labor supply decisions of American mothers. Since the number of children can be considered as endogenous, Angrist and Evans (1998a) use an instrument for it. Angrist and Evans (1998a) note that mothers whose first two children are of the same sex are more likely to have more than two children. The instrument Angrist and Evans use is whether the first two children are of the same sex or not. The sex of children is thus assumed to be random. The data Angrist and Evans (1998a) use is from the 1980 and 1990 US Censuses. I focus on the 1980 US Census results, because Black et al. (2015) also do that. More specifically, I focus on the results of columns (1) and (2) of Table 7 in Angrist and Evans (1998a), the reason is that Black et al. also do that.[22] The controls which are used are the age of the mother in years, the age of the mother in years at the first birth and different dummy variables for whether the first child was a boy, whether the second child was a boy, whether someone belongs to the black people, whether someone belongs to the Hispanic people and whether someone belongs to another race (so not black, Hispanic or white) (Angrist and Evans, 1998a; 1998b).

The study of Angrist et al. (2002a) is about the effect of schooling vouchers on different educational and social outcomes. This study was carried out using data on the *Programa de Ampliación de Cobertura de la Educación* Secundaria (PACES), which was a Colombian

[22] I do not use "ln (*family income)*" as outcome variable, because Black et al. (2015) also ignore that variable.

policy intervention that distributed schooling vouchers that covered a substantial part of private secondary schooling costs. The vouchers were randomly distributed in some parts of Colombia by means of a lottery. The random distribution of these vouchers was used as an instrument for whether a pupil ever received a scholarship (pupils also acquired scholarships from other sources). The study of Angrist et al. (2002a) employs two different samples: pupils who applied for the lottery in Bogotá in 1995 and a sample consisting of pupils who applied for the lottery in Bogotá in 1995, in Bogotá in 1997 or in Jamundi in 1993. The latter sample is called the 'combined sample'. The quality of the second sample is lower than the quality of the Bogotá 1995 sample, since the distribution of vouchers in the Jamundi 1993 subsample may be nonrandom and since the Bogotá 1997 subsample was "too recent for a good reading on some outcomes" (Angrist et al., 2002a, p. 1540).[23] However, the vast majority of observations in the combined sample is from the Bogotá 1995 cohort. I focus on the results in columns (2)-(5) of Table 7 of Angrist et al. (2002a) for the analyses I am going to carry out. The controls for the Bogotá 1995 sample are the following (analyses with "Test scores (total points)" as outcome employ a different set of controls, see below): a dummy for whether the survey was carried out in person, a dummy for whether "survey was completed using new survey" (Angrist et al., 2009a)[24], age, sex, dummies for the strata of residence and dummies for the month of interview. Beside these controls, additional controls are included when analyzing the combined sample: a dummy for whether the individual has access to a phone[25], a city dummy and dummies which

[23] Just before submitting this thesis, I discovered that there are a few small signs of potential nonrandomness in the distribution of vouchers in the Bogota cohorts by looking at Table 2 of Angrist et al. (2002a). This table shows the difference in characteristics between winners and losers. It seems that the difference in *Father's highest grade* is significant (coefficient is -0.431 and the standard error is 0.199) in the whole Bogota 1995 sample, while the difference in *Father's wage (>2 min wage)* seems significant for the subsample of test takers of the Bogota 1995 cohort. The Bogota 1997 sample also show some signs of nonrandomness in the distribution of vouchers, but this is less important since Angrist et al. (2002a) already note that this sample has a lower quality. Nowhere in Angrist et al. (2002a) one can find that the authors note that there is nonrandomness in the instrument for the Bogota cohorts. Instead, they write this: "There is little evidence of any association between win/loss status and the individual characteristics measured in our data from Bogotá" (Angrist et al., 2002a, pp. 1539-1540). However, Bettinger et al. (2016) note that the difference in *age at time of application* in the Bogota 1995 sample can be considered as significant. In addition, during a presentation Saavedra (2012, from 22:15 till 23:52) said that he suspects that there is a structural difference in *age at time of application* in all cohorts. Later on in this thesis, I indicate which results should be given less weight. I do not consider the existing evidence as sufficiently strong to conclude that the instrument is seriously flawed in the Bogota 1995 sample. In addition, Angrist et al. (2006), Bettinger et al. (2010) and Huber et al. (2017) consider the instrument as random in the Bogota 1995 sample. Therefore, I do not label the results for the Bogota 1995 sample as results which should be given less weight.

[24] It is noted in the datasets provided by Angrist et al. (2009a), however it is not clear to me what it exactly means. The reason is that Angrist et al. (2002a) do not provide (clear) information about it.

[25] Actually, this is also included in the analyses for the Bogotá 1995 sample (by Angrist et al. [2002a]), however it is omitted due to multicollinearity in my analyses as well as in the relevant analyses by Angrist et al. (2002a).

control for the different years (Angrist et al., 2002a; 2002b; 2009a; 2009b).[26] The controls for the analyses with "Test scores (total points)" as outcome are the following: a dummy for whether the survey was carried out in person, a dummy for whether "survey was completed using new survey" (Angrist et al., 2009a)[27], age, sex, dummies for the strata of residence, dummies for the month of interview, dummies for test site, the highest school grade completed by the father, the highest school grade completed by the mother, *dad_miss* and *mom_miss* (Angrist et al., 2002a; 2002b; 2009a; 2009b).[28]

Abadie et al. (2002) is a study about the effect of job training on earnings for different income quantiles in the United States. To be specific, the study uses Quantile regression and Quantile IV regression. However, the study also provides a 'normal' IV. This 'normal' IV is what I will use. Therefore, my analyses will be based on column 1 of Table III in Abadie et al. (2002). Since the effects presented in Table III of Abadie et al. (2002) are estimated for each sex separately, I will present results for men and women separately. The job training was provided by the Job Training Partnership Act (JTPA) and randomly distributed to individuals. However, 60% of the ones that received the right to make use of the training made use of the training (Abadie et al., 2002). The controls which are used are: "dummies for black and Hispanic applicants, a dummy for high-school graduates (including GED holders), dummies for married applicants, 5 age-group dummies, and dummies for AFDC receipt (for women) and whether the applicant worked at least 12 weeks in the 12 months preceding random assignment. Also included are dummies for the original recommended service strategy (classroom, OJT/JSA, other) and a dummy for whether earnings data are from the second follow-up survey" (Abadie et al., 2002, p. 101).

The last study I make use of is the study by Kazianga et al. (2013a). In contrast to the other three studies, this study employs an RDD. The study is about the effect of "girl-friendly"

[26] The city dummy and (some) year dummies are collinear and the number of collinear dummies (city dummy and year dummies) depends on the outcome variable. This is also the case in Angrist et al. (2002a), however, they are not very clear about it. They mention that they control for, among other variables, city and year of application. If the city dummy and a certain year dummy are collinear and, for example, the year dummy is included and the city dummy is omitted, then the year dummy also controls for city.

[27] It is noted in the datasets provided by Angrist et al. (2009a), however it is not clear to me what it exactly means. The reason is that Angrist et al. (2002a) do not provide (clear) information about it.

[28] *Dad_miss* and *mom_miss* do not contain labels in Angrist et al. (2009a) and no (clear) information is provided by Angrist et al. (2002a). A careful look at the data suggests that *dad_miss* and *mom_miss* are probably assigned a value of one if the schooling data for that particular parent are missing and a value of zero if the schooling data were available. Looking at all observations (*N = 1,212*) in the dataset "tab7test" (Angrist et al., 2009a) reveals that 97% of the individuals with a father whose highest grade completed is above zero also show a value of zero for the variable *dad_miss*. For the data concerning mothers this percentage is 98. This also suggests that if schooling data was unavailable for a parent, this parent was assigned a value of zero for the highest school grade completed. Again, no clear information is provided by Angrist et al. (2002a, 2002b; 2009a; 2009b).

primary schools on different schooling outcomes in Burkina Faso. Villages in Burkina Faso were assigned “a score based largely on the estimated number of children to be served from the proposed and neighboring villages, giving additional weight to girls” (Kazianga et al., 2013a, p. 44). Within each department, the top 50% villages (that is, villages with the highest scores) were eligible for a girl-friendly school, whereas the lower scoring villages were not eligible for such a school. Based on this, Kazianga et al. (2013a) inferred the cut-off score for each department. Compliance with the cut-off score was not perfect: not all eligible villages received a girl-friendly school and some ineligible villages eventually got a girl-friendly school. The unit of observation in Kazianga et al. (2013a) is individuals. Since Kazianga et al. (2013a) work mostly under the assumption that compliance was strict, there are not many Two-Stage Least Squares estimations presented in their paper. Consequently, I only focus on Column (7) of Table 7 for my analyses.[29] The fuzzy RDD presents a dilemma for me: the researchers calculated the relative scores for each village (village score minus the cut-off score) and use this and a squared term of this as control variables. It is of course true that villages below the cut-off score have a different score compared to the villages above the cut-off. I therefore choose to include the relative score and the squared term of the relative score in the regressions ‘without’ control variables. The controls variables which are used are “Head is male”, “Head’s age”, “Head years of schooling”, “Number of members”, “Number of children”, “Child is female”, “Head’s child”, “Head’s grandchild”, “Head’s niece/nephew”, “Muslim”, “Animist”, “Christian”, “Fulfulde language”, “Gulmachema language”, “Moore language”, “Gourmanch ethnicity”, “Mossi ethnicity”, “Peul ethnicity”, “Basic flooring”, “Basic roofing”, “Number of radios”, “Number of phones”, “Number of watches”, “Number of bikes”, “Number of cows”, “Number of motorbikes”, “Number of carts” (Kazianga et al., 2013a, p. 50), dummies for the age of the child and department fixed effects (Kazianga et al., 2013a).[30]

Over the last years, economics has evolved. The same holds for developments in estimating standard errors. One example is that it is common practice to use different standard errors than the ‘normal’ standard errors. This creates a dilemma since I use somewhat older studies. I have decided to use the same standard errors as are used in the studies I make use of. The reason for this is that I remain close to the original study.

[29] Column (7) of Table 7 from Kazianga et al. (2013a) shows the effect of a child’s enrollment in a girl-friendly school on the child’s total test score.

[30] With the exception of the child’s sex and age, all these variables are measured at the household level.

4. Results

4.1 Main results

First, I show the results for equations (7), (8), (9) and (10) from the data used in Angrist and Evans (1998a). These results can be found in Table 2.[31] Here the treatment indicator equals 1 if mothers have more than 2 children and zero if mothers have only 2 children. The sample is limited to mothers with at least 2 children (Angrist and Evans, 1998a). The instrument equals 1 if the sex of the first 2 children is equal and 0 if the sex of the first 2 children is not the same. In Table 3 I show a cross table for the treatment variable and the instrument. From this one can infer the number of Always Takers, Compliers and Never Takers.[32] The results from the data from Angrist and Evans (1998b) show a striking result for the subsample $D = 0$: they show that adding controls leads to (more) bias! In the case of the first three outcome variables, adding controls leads to significant selection bias, while for the outcome variable "income mom", adding controls leads to more pronounced selection bias (it is more significant and the coefficient is higher). Therefore, all δs are negative in the subsample $D = 0$. On the side of $D = 1$, the first three outcome variables show a negative δ, but adding controls does not lead to significant selection bias. Adding controls leads to insignificant selection bias for the last two outcome variables. Basically, only the two last outcome variables show evidence in favor of Oster's methods.

The results for equations (7), (8), (9) and (10) for the data from Angrist et al. (2002a) can be found in Tables 4 and 7. Here the instrument equals 1 if a student won the scholarship voucher lottery and 0 if a student did not win. The treatment indicator equals 1 if a student ever made use of a scholarship and 0 if a student never made use of a scholarship. So if a student did not win the scholarship voucher lottery, but obtained a scholarship via another way, the student is considered as being treated. Table 4 shows the results for the sample Bogota 1995 from Angrist et al. (2002a) and Table 7 shows the results for the combined sample (which is Bogotá 1995 plus Jamundi 1993 and Bogotá 1997) from Angrist et al. (2002a). Tables 5, 6 and 8 show the cross tables for the treatment variable and instrument. Tables 4 and 7 display 22 δs and only 2 of them are in the interval $[0, 1]$ (and one δ is very close: 1.06). However, in most cases both the uncontrolled as well as the controlled estimate of α show insignificant selection bias. The overall picture of Tables 4 and 7 is that adding controls does not substantially change selection

[31] Note that the results for $\tilde{\alpha}$ are virtually a replication of Black et al. (2015).

[32] For example, the number of Always Takers is: $72{,}643 + \frac{72{,}643}{195{,}292} \cdot 199{,}548 \approx 146{,}869$.

bias; in most of the cases there is no significant total bias and no significant selection bias on unobservables. However, there are some exceptions!

Table 2: results for Angrist and Evans (1998a)

	Results for subsample $D = 0$				*Results for subsample $D = 1$*			
Outcome Variable	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ
Mom worked last year	0.000782 (0.00201) [0.000]	0.00465** (0.00197) [0.040]	-0.003871	-1.202	-0.00280 (0.00252) [0.000]	0.00191 (0.00247) [0.040]	-0.00471	-0.405
Weeks worked	0.0759 (0.0930) [0.000]	0.298*** (0.0902) [0.062]	-0.2223	-1.342	-0.169 (0.108) [0.000]	0.0652 (0.104) [0.061]	-0.2346	-0.278
Hours worked	0.0398 (0.0774) [0.000]	0.205*** (0.0753) [0.057]	-0.1651	-1.241	-0.218** (0.0949) [0.000]	0.00293 (0.0924) [0.053]	-0.22050	-0.013
Income mom	102.3** (46.59) [0.000]	188.3*** (45.37) [0.054]	-86.0	-2.190	-91.63* (49.38) [0.000]	-1.630 (48.24) [0.047]	-90.001	0.018
N	236,089				158,751			

Notes: $\dot{\alpha}$ stands for "Total bias", $\tilde{\alpha}$ means "Selection bias on unobservables", $\breve{\alpha}$ stands for "Selection bias on observables" and δ means "Ratio of selection on unobservables to selection on observables". Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. R^2 of the whole regression between brackets.

Table 3: cross table – treatment & instrument

		Same sex		Total
		0	1	
More than 2 children	0	122,649	113,440	236,089
	1	72,643	86,108	158,751
Total		195,292	199,548	394,840

Table 4: results for the Bogotá 1995 sample of Angrist et al. (2002a)

	Results for subsample $D = 0$					*Results for subsample* $D = 1$				
Outcome Variable	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N
Highest grade completed	-0.201 (0.177) [0.004]	-0.231 (0.176) [0.114]	0.030	-7.698	*474*	0.137 (0.0881) [0.005]	0.137* (0.0833) [0.118]	-0.00003	-4,515.064	*673*
In school	-0.155** (0.0696) [0.015]	-0.120** (0.0601) [0.196]	-0.035	3.470	*474*	0.0263 (0.0347) [0.001]	0.0223 (0.0321) [0.191]	0.0040	5.574	*673*
Total repetitions since lottery	-0.0304 (0.0728) [0.000]	-0.0627 (0.0754) [0.060]	0.0323	-1.942	*474*	-0.0144 (0.0397) [0.000]	-0.0185 (0.0394) [0.031]	0.0041	-4.494	*673*
Finished 8th grade	-0.111 (0.0753) [0.005]	-0.0885 (0.0780) [0.086]	-0.0225	3.929	*474*	0.101** (0.0444) [0.009]	0.0968** (0.0439) [0.071]	0.0039	25.116	*673*
Test scores (total points)	-0.490 (0.298) [0.020]	-0.197 (0.328) [0.314]	-0.292	0.675	*105*	-0.0797 (0.210) [0.001]	-0.157 (0.196) [0.281]	0.0769	-2.036	*175*
Married or living with companion	0.00394 (0.0212) [0.000]	0.00861 (0.0186) [0.068]	-0.00467	-1.843	*474*	-0.00899 (0.0108) [0.002]	-0.00965 (0.0101) [0.079]	0.00066	-14.604	*672*

Notes: $\dot{\alpha}$ stands for "Total bias", $\tilde{\alpha}$ means "Selection bias on unobservables", $\breve{\alpha}$ stands for "Selection bias on observables" and δ means "Ratio of selection on unobservables to selection on observables". Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. R^2 of the whole regression between brackets.

Table 5: cross table – treatment & instrument (Bogotá 1995 sample)

		Student won voucher		Total
		0	1	
Student used a scholarship	0	425	49	474
	1	137	536	673
Total		562	585	1,147

Table 6: cross table – treatment & instrument (test scores subsample)

		Student won voucher		Total
		0	1	
Student used a scholarship	0	95	10	105
	1	28	147	175
Total		123	157	280

Table 7: results for the combined sample of Angrist et al. (2002a)

	Results for subsample $D = 0$					*Results for subsample* $D = 1$				
Outcome Variable	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N
Highest grade completed	-0.427*** (0.146) [0.016]	-0.220* (0.128) [0.379]	-0.207	1.060	*651*	-0.000127 (0.0922) [0.000]	0.102 (0.0678) [0.535]	-0.102559	-0.999	*926*
In school	-0.141*** (0.0523) [0.015]	-0.139*** (0.0460) [0.200]	-0.002	84.815	*651*	0.0203 (0.0276) [0.001]	0.00978 (0.0260) [0.171]	0.01054	0.928	*926*
Total repetitions since lottery	0.0101 (0.0594) [0.000]	0.0135 (0.0603) [0.056]	-0.0034	-4.008	*651*	-0.0364 (0.0328) [0.001]	-0.0397 (0.0327) [0.046]	0.0033	-11.986	*926*
Finished 8th grade	-0.150** (0.0700) [0.009]	-0.125* (0.0722) [0.101]	-0.024	5.164	*522*	0.0718* (0.0389) [0.005]	0.0821** (0.0383) [0.090]	-0.0103	-7.945	*782*
Married or living with companion	0.00794 (0.0175) [0.000]	0.0154 (0.0170) [0.066]	-0.00746	-2.064	*651*	-0.0151 (0.0106) [0.004]	-0.0143 (0.0106) [0.066]	-0.0008	17.375	*925*

Notes: $\dot{\alpha}$ stands for "Total bias", $\tilde{\alpha}$ means "Selection bias on unobservables", $\breve{\alpha}$ stands for "Selection bias on observables" and δ means "Ratio of selection on unobservables to selection on observables". Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. R^2 of the whole regression between brackets. The results for "finished 8th grade" do not include the Bogotá 1997 cohort (see Angrist et al. 2002a).

Table 8: cross table – treatment & instrument (Combined sample)

		Student won voucher		Total
		0	1	
Student used a scholarship	0	567	84	651
	1	194	732	926
Total		761	816	1,577

The results for equations (7), (8), (9) and (10) from the data used in Abadie et al. (2002) can be found in Table 9.[33] The instrument equals 1 if one got assigned to the job training and 0 otherwise. The treatment indicator equals 1 if an individual received job training and the indicator equals 0 if the individual did not receive job training. Tables 10 and 11 display the cross tables of the treatment variable and the instrument for men and women respectively. As can be seen in Tables 10 and 11, there are almost no Always Takers.[34] This is potentially a very serious issue.[35] Therefore, less weight should be given to all results from the treated subsample. When we consider the results from the untreated subsample, we can see that control variables do not seem to have a large influence on selection bias. For men, the $\dot{\alpha}_0$ as well as the $\tilde{\alpha}_0$ show significant selection bias. Thus, this can be considered as strong evidence against Oster's approach and other studies which rely on coefficient stability tests. This evidence can be considered as even stronger if one compares the size of the selection bias – which is in absolute terms higher than 2,000 – with the estimated effect for men in column 1 of Table III in Abadie et al. (2002) which is 1,593. For women the $\dot{\alpha}_0$ as well as the $\tilde{\alpha}_0$ do not show significant selection bias. However, for both men and women δ_0 is far outside the interval $[0, 1]$. The results for subsample $D = 1$ show mixed evidence concerning Oster's approach: the subsample of women shows a δ which is in the interval $[0, 1]$, while the subsample of men shows a δ which is substantially above 1.

The results for equations (7), (8), (9) and (10) from the data used in Kazianga et al. (2013a) can be found in Table 12. The instrument equals 1 if the child lives in a village which was eligible for a girl-friendly school and 0 otherwise. The treatment equals 1 if the child enrolled in a girl-friendly school and 0 otherwise. Table 13 shows the cross table of the treatment and the instrument. As can be seen in Table 12, the δ is outside of the interval $[0, 1]$ in both subsamples. In addition, results from subsample $D = 0$ show that there is no selection bias when there are no controls included (besides the relative score and the relative score squared) and that inclusion of control variables does not cause significant total selection bias. The results from subsample $D = 1$ show that there is significant selection bias without controls and that significant selection bias remains after the inclusion of control variables. The δ_1 is far

[33] It is not entirely clear to me which robust standard errors are reported in Column 1 of Table III of Abadie et al. (2002). I replicated that column and used the "vce(robust)" STATA command to calculate standard errors. Deviations from the standard errors reported in Column 1 of Table III of Abadie et al. (2002) are small. Besides, I replicated the OLS regressions corresponding to the results of that column and the standard errors seemed to be exactly the same. Therefore, I assume that "vce(robust)" provides the correct standard errors.

[34] This is also noted by Abadie et al. (2002).

[35] For example, the 'normal' standard errors for $\dot{\alpha}_1$ and $\tilde{\alpha}_1$ for the sample of men are much larger than the robust standard errors.

above the value of one. This can be considered as evidence against the assumption that δ equals one. Table A1 in the Appendix shows the results when one does not include relative score and the relative score squared in the regression for $\dot{\alpha}$. As can be seen there, the δ is radically different in the subsample $D = 0$. Since it is not crystal clear to me that the relative score variables should be included in the regressions for $\dot{\alpha}$, I give less weight to the results from the subsample $D = 0$. The δ from subsample $D = 1$ is almost identical in Table 12 and Table A1, so it does not really matter which δ_1 one prefers.

Table 9: results for Abadie et al. (2002)

	Results for treatment subsample $D = 0$					*Results for treatment subsample $D = 1$*				
	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N
	Men					Men				
30-month Earnings	-2235.2*** (711.5) [0.003]	-2126.1*** (686.2) [0.085]	-109.1	19.485	2,966	7357.1*** (2708.6) [0.001]	4624.2* (2415.9) [0.088]	2732.9	1.692	2,136
	Women					Women				
30-month Earnings	-246.1 (466.1) [0.000]	-345.4 (452.6) [0.101]	99.3	-3.478	3,380	1061.2 (2393.0) [0.000]	120.9 (2255.4) [0.096]	940.3	0.129	2,722

Notes: $\dot{\alpha}$ stands for "Total bias", $\tilde{\alpha}$ means "Selection bias on unobservables", $\breve{\alpha}$ stands for "Selection bias on observables" and δ means "Ratio of selection on unobservables to selection on observables". Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. R^2 of the whole regression between brackets.

Table 10: cross table – treatment & instrument (for men)

		Assignment		Total
		0	1	
Training	0	1,684	1,282	2,966
	1	19	2,117	2,136
Total		1,703	3,399	5,102

Table 11: cross table – treatment & instrument (for women)

		Assignment		Total
		0	1	
Training	0	1,979	1,401	3,380
	1	35	2,687	2,722
Total		2,014	4,088	6,102

Table 12: results for Kazianga et al. (2013a)

	Results for subsample $D = 0$					*Results for subsample* $D = 1$				
	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N
Total scores	0.0464 (0.0316) [0.005]	0.0513 (0.0345) [0.100]	-0.0048	-10.580	9,719	0.270*** (0.0919) [0.017]	0.221*** (0.0583) [0.367]	0.050	4.458	8,251

Notes: $\dot{\alpha}$ stands for "Total bias", $\tilde{\alpha}$ means "Selection bias on unobservables", $\breve{\alpha}$ stands for "Selection bias on observables" and δ means "Ratio of selection on unobservables to selection on observables". Standard errors clustered at village level in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. R^2 of the whole regression between brackets. The regressions for $\dot{\alpha}$ include relative score and the squared term of relative score as explanatory variables.

Table 13: cross table – treatment & instrument

		Village selected for a "girl-friendly" school		Total
		0	1	
Self-reported attendance	0	5,962	3,757	9,719
	1	3,161	5,090	8,251
Total		9,123	8,847	17,970

4.2 Summary of results

In this sub-section I summarize the results for the αs and δs. First, I count the number of situations in which $\dot{\alpha}$ was not significantly different from zero, but where the $\tilde{\alpha}$ was significantly different from zero. Second, I count the number of situations in which the $\dot{\alpha}$ was significantly different from zero, but where the $\tilde{\alpha}$ was not significantly different from zero. Third, I count the number of situations where both αs were significantly different from zero. Finally, I count the number of situations where both αs were not significantly different from zero.

For this practice I regard something as significantly different from zero if the p-value is lower than 5%. One could argue that I should rather look to changes in p-values, while others could argue that I also should look at whether the $\dot{\alpha}$ and the $\tilde{\alpha}$ are significantly different from each other. However, suppose we would actually be carrying out an effect study in which we use OLS, then my approach is more appropriate. Suppose an uncontrolled estimate of an effect contains selection bias which significant at the 40% level, while the controlled estimate contains selection bias which is significant at the 38%. In both cases one could ignore selection bias. So looking at p-values does not offer much information. Suppose now that an uncontrolled estimate of an effect contains selection bias which significant at the 6% level, while the controlled estimate contains selection bias which is significant at the 4%. In the first case one could say that selection bias can be ignored, whereas in the latter case one should take selection bias seriously. However, in most situations the uncontrolled estimate and controlled estimate will not be statistically different. My approach takes into account that controls can push selection bias to be statistically significant or to be statistically insignificant. Suppose that the p-value of $\dot{\alpha}$ is 0.06 and that the p-value of $\tilde{\alpha}$ is 0.04, while $\dot{\alpha}$ and $\tilde{\alpha}$ are not statistically significant different from each other. One could then say that this is not evidence that controls cause selection bias. However, one could also say – and this the approach I take – that controls caused selection bias to be significant.

Table 14 shows the summary of the αs. Column (1) of Table 14 shows the summary of all αs, while Column (2) shows a summary of αs which excludes αs which should be given less weight due to issues with estimation.[36] These are the results which are excluded in Column (2): The results for the combined sample from Angrist et al. (2002a) and the results for the treated

[36] Note that I exclude the α results from the subsample of treated men from Abadie et al. (2002) from Column (1) as well, because there are virtually no male Always Takers and it seems that as a consequence of this 'normal' standard errors for $\dot{\alpha}_1$ and $\tilde{\alpha}_1$ of the sample of treated men are much larger than the robust standard errors. These facts point to a really serious issue in the estimation results for this particular subsample.

subsamples (men and women) from Abadie et al. (2002).[37] As can be seen in Table 14, in 11.4% (Column (1)) or 12.5% (Column (2)) of cases $\dot{\alpha}$ is not significant, while controls result in a significant $\tilde{\alpha}$. This percentage is quite high. In 8.6% (Column (1)) or 4.2% (Column (2)) of cases, controls solve significant selection bias. However, it seems that in the majority of cases controls do not affect the conclusion that a certain estimate is significant or insignificant. Note that the αs show an underestimation of the selection bias between Always Takers and Compliers and between Compliers and Never Takers. I could transform them as described in Section 3, but it would be beyond the scope of the thesis to calculate standard errors for the transformed αs. The reason is that Black et al. (2015) do not provide a method for calculating standard errors. Thus, the underestimation issue is a limitation. However, we could reinterpret the αs as indicating the amount of selection bias between some Always Takers and the group of the other Always Takers and Compliers and between some of the Never Takers and the group of the other Never Takers and Compliers.

Table 14: Summary of αs

	Column (1)	Column (2)
$\dot{\alpha}$ is not significant & $\tilde{\alpha}$ is significant	4	3
$\dot{\alpha}$ is significant & $\tilde{\alpha}$ is not significant	3	1
$\dot{\alpha}$ is significant & $\tilde{\alpha}$ is significant	6	5
$\dot{\alpha}$ is not significant & $\tilde{\alpha}$ is not significant	22	15
Total number of results	35	24

Notes: This table shows how often a certain situation occurs concerning the significance of the αs in Tables 2, 4, 7, 9 and 12. An α is considered as significant if the p-value is below 5%. Column (1) shows the summary for all α results, except for the results from the subsample of treated men from Abadie et al. (2002). Column (2) shows a summary of αs which should be given the most weight, so this column excludes results for the combined sample from Angrist et al. (2002a) and the results for the treated subsamples (both men and women) from Abadie et al. (2002).

Summarizing the results for δ is less complicated. I simply count the number of δs which are below 0, the number of δ which are in the interval $[0, 1]$ and the number of δs which are above 1. Table 15 shows the summary of the δs. Column (1) of Table 15 shows the δs for all results, while Column (2) shows a summary of δs which excludes δs which should be given less weight due to issues with estimation.[38] These are the results which are excluded in Column (2): The δs

[37] Note that I include the results for the αs from the subsample $D = 0$ from Kazianga et al. (2013a) in both Columns. The reason is that $\dot{\alpha}_0$ is insignificant with and without the 'relative score' variables.

[38] Note that I exclude the result for δ from the subsample of treated men from Abadie et al. (2002) from Column (1) as well, because there are virtually no male Always Taker and it seems that as a consequence of this 'normal' standard errors for $\dot{\alpha}_1$ and $\tilde{\alpha}_1$ of the sample of treated men are much larger than the robust standard errors. These facts point to a really serious issue in the estimation results for this particular subsample.

from the combined sample from Angrist et al. (2002a), the δs from the treated subsamples (men and women) from Abadie et al. (2002) and the δ from the subsample of untreated children from Kazianga et al. (2013a). As can be noted in Table 15, it seems that in the majority of cases δ is lower than zero. Only a few δs are in the interval $[0, 1]$, while quite some δs are above 1. Although I cannot calculate standard errors for the δ, it seems quite safe to assert that δ is at least sometimes substantially below zero. The reason is that I have found so many δs below zero, that it seems unlikely that all δs would in reality be in the interval $[0, 1]$. In a similar fashion, it seems quite safe to assert that δ is at least sometimes substantially above one. This result strongly suggests that selection on observables is not informative about selection on unobservables.

Table 15: Summary of δs

	Column (1)	Column (2)
$\delta < 0$	21	15
$0 \leq \delta \leq 1$	4	2
$\delta > 1$	10	6
Total number of results	35	23

Notes: This table shows how often a certain situation occurs concerning δ in Tables 2, 4, 7, 9 and 12. Column (1) shows the summary for all δ results, except for the result from the subsample of treated men from Abadie et al. (2002). Column (2) shows a summary of δs which should be given the most weight, so this column excludes results for the combined sample from Angrist et al. (2002a), the results for the treated subsamples (both men and women) from Abadie et al. (2002) and the δ from the subsample of untreated children from Kazianga et al. (2013a).

5. Conclusions

As I wrote earlier, a fundamental problem in economics – and in the social sciences in general – is whether selection on observables is informative about selection on unobservables. Sometimes researches calculate partial correlations between an outcome and a variable which possibly affects the outcome and look at the stability of the partial correlations when including additional controls. Some time ago Altonji et al. (2005) and recently Oster tried to improve this practice. Both studies assume that the movement of the estimated effect due to inclusion of additional controls is indicative of the movement that would occur if all relevant unobservable confounders could be added to the analysis. In all these studies – studies looking at the stability of the partial correlation, Altonji et al (2005) and Oster – a crucial assumption is that selection on observables is informative about selection on unobservables. My results provide suggestive evidence that this assumption does not always hold. Potentially, it could be that in the majority of situations this assumption does not hold.

When looking only to the αs one can conclude that a controlled estimate is sometimes even worse than an uncontrolled estimate in the sense that it is (more) biased. This is an astonishing and dramatic result for the social sciences! The αs show that sometimes the characteristics of the unobservables are completely different from the characteristics of the observables.

Although I cannot calculate standard errors for the δ, it seems quite safe to assert that the δ is at least sometimes substantially below zero. The reason is that I have found so many δs below zero, that it seems unlikely that all δs would in reality be in the interval $[0, 1]$. In a similar fashion, it seems quite safe to assert that the δ is at least sometimes substantially above one. So this implies that employing Oster's methods with the assumption that $\delta = 1$ is a shaky practice.

It is important to note that the approach of Oster might however be useful when broader ranges of δ are taken into account. It could very well be that future research establishes that most δs are between a certain interval around 0, for example, $[-2, 2]$. Then a researcher can safely implement Oster's method: one has just to take into account that a δ in the interval $[-2, 2]$ is plausible.

My results show that it is not safe to assume that selection on observables is of the same direction as the selection on unobservables. In other words, it is not safe to assume that controlling for observables is better than not controlling for observables. Suppose a researcher would now calculate partial correlation and look at the influence of adding controls or suppose the researcher would employ the Oster estimator, one could ask then whether the controlled estimate is better than the uncontrolled estimate, because my results show that probably sometimes an uncontrolled estimate is better than a controlled estimate. This conclusion shows that we should be glad that a "credibility revolution" (Angrist and Pischke, 2010) in economics has occurred. The fundamental problems with partial correlations remain, but, luckily, due to the "credibility revolution" we can use the quasi-experimental toolbox in some cases.

Appendix

Table A1: results for Kazianga et al. (2013a) without relative score and relative score squared as explanatory variables in the regression for $\dot{\alpha}$

	Results for subsample $D = 0$					*Results for subsample* $D = 1$				
	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N	$\dot{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	δ	N
Total scores	0.0533* (0.0285) [0.005]	0.0513 (0.0345) [0.100]	0.0020	25.971	9,719	0.269*** (0.0797) [0.014]	0.221*** (0.0583) [0.367]	0.048	4.562	8,251

Notes: $\dot{\alpha}$ stands for "Total bias", $\tilde{\alpha}$ means "Selection bias on unobservables", $\breve{\alpha}$ stands for "Selection bias on observables" and δ means "Ratio of selection on unobservables to selection on observables". Standard errors clustered at village level in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. R^2 of the whole regression between brackets.

References

Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica, 70*(1), 91-117.

Abadie, A., Angrist, J., & Imbens, G. (2009). *Replication data for: Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings.* Retrieved January 26, 2017, from hdl:1902.1/11300, Harvard Dataverse, V1, UNF:3:TvYbGDy6UT1468e5vkISYA==

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy, 113*(1), 151-184.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002a). Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. *American Economic Review, 92*(5), 1535-1558.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002b). *MIT Economics: Angrist Data Archive*. Retrieved December 8, 2016, from http://economics.mit.edu/faculty/angrist/data1/data/angetal02

Angrist, J. D., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2009a). *Replication data for: Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment.* Retrieved December 11, 2016, from hdl:1902.1/11298, Harvard Dataverse, V1

Angrist, J. D., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2009b). *Replication data for: Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment.* Retrieved December 8, 2016, from hdl:1902.1/11298, Harvard Dataverse, V1

Angrist, J., Bettinger, E., & Kremer, M. (2006). Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia. *American Economic Review, 96*(3), 847-862.

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

Angrist, J. D., & Evans, W. N. (1998b). *MIT Economics: Angrist Data Archive.* Retrieved May 12, 2016, from http://economics.mit.edu/faculty/angrist/data1/data/angev98

Angrist, J. D., & Evans, W. N. (1998c). *MIT Economics: Angrist Data Archive.* Retrieved May 10, 2016, from http://economics.mit.edu/faculty/angrist/data1/data/angev98

Angrist, J. D., & Pischke, J. -S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J. -S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives, 24*(2), 3–30.

Bettinger, E., Kremer, M., Kugler, M., Medina, C., Posso, C., & Saavedra, J. E. (2016). *Can Educational Voucher Programs Pay for Themselves?* Retrieved May 9, 2017, from http://mit-neudc.scripts.mit.edu/2016/wp-content/uploads/2016/03/paper_208.pdf

Bettinger, E., Kremer, M., & Saavedra, J. E. (2010). Are Educational Vouchers Only Redistributive? *The Economic Journal, 120*, F204-F228.

Black, D. A., Joo, J., LaLonde, R., Smith, J. A., & Taylor, E. J. (2015). *Simple Tests for Selection Bias: Learning More from Instrumental Variables (IZA Discussion Paper No. 9346).* Bonn, Germany: Institute for the Study of Labor.

Froelich, M., & Melly, B. (2008). *Estimation of quantile treatment effects with STATA - Alexandria (file: jtpa.dta).* Retrieved December 15, 2016, from https://www.alexandria.unisg.ch/46580/

Huber, M., Laffers, L., & Mellace, G. (2017). Sharp IV bounds on average treatment effects on the treated and other populations under endogeneity and noncompliance. *Journal of Applied Econometrics, 32*, 56-79.

Kazianga, H., Levy, D., Linden, L. L., & Sloan, M. (2013a). The Effects of “Girl-Friendly” Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. *American Economic Journal: Applied Economics, 5*(3), 41–62.

Kazianga, H., Levy, D., Linden, L. L., & Sloan, M. (2013b). *The Effects of "Girl-Friendly" Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso.* Retrieved February 27, 2017, from https://www.aeaweb.org/articles?id=10.1257/app.5.3.41

Oster, E. (2015). *Unobservable Selection and Coefficient Stability: Theory and Evidence.* Unpublished manuscript (version: 26 January 2015). Brown University and NBER. Retrieved 24 February, 2017, from https://www.brown.edu/research/projects/oster/sites/brown.edu.research.projects.oster/files/uploads/Unobservable_Selection_and_Coefficient_Stability.pdf

Oster, E. (forthcoming). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal Of Business & Economic Statistics*.

Saavedra, J. E. (2012). Collegiate and Labour Market Effects: Effects of Vouchers for Private Schooling in Colombia. USC Rossier School of Education, Los Angeles. Retrieved May 1, 2017, from https://www.youtube.com/watch?v=eGGbmjlalFA