Zhuan Pei, Jorn-Steffen Pischke and Hannes Schwandt

Poorly measured confounders are more useful on the left than on the right

Article (Accepted version) (Refereed)

Original citation:
Pei, Zhuan and Pischke, Jorn-Steffen and Schwandt, Hannes (2018) *Poorly measured confounders are more useful on the left than on the right*. Journal of Business and Economic Statistics. ISSN 0735-0015 (In Press)
DOI: 10.1080/07350015.2018.1462710

Poorly Measured Confounders Are More Useful on the Left Than on the Right

Zhuan Pei
Dept. of Policy Analysis and Management, Cornell University, Ithaca, NY 14853-4401, USA (zhuan.pei@cornell.edu)

Jörn-Steffen Pischke
Dept. of Economics, London School of Economics, London WC2A 2AE, UK (s.pischke@lse.ac.uk)

Hannes Schwandt
Dept. of Economics, University of Zürich, Schönberggasse 1, 8001 Zürich, Switzerland (hannes.schwandt@uzh.ch)

March 2018

Abstract

Researchers frequently test identifying assumptions in regression based research designs (which include instrumental variables or difference-in-differences models) by adding additional control variables on the right hand side of the regression. If such additions do not affect the coefficient of interest (much) a study is presumed to be reliable. We caution that such invariance may result from the fact that the observed variables used in such robustness checks are often poor measures of the potential underlying confounders. In this case, a more powerful test of the identifying assumption is to put the variable on the left hand side of the candidate regression. We provide derivations for the estimators and test statistics involved, as well as power calculations, which can help applied researchers interpret their findings. We illustrate these results in the context of estimating the returns to schooling.

JEL classifications: C31, C52

Keywords: Balancing, variable addition, robustness checks, specification testing, Hausman test

1 INTRODUCTION

The identification of causal effects depends on explicit or implicit assumptions which typically form the core of a debate about the quality and credibility of a particular research design. In regression based strategies, this is the claim that variation in the regressor of interest is as good as random after conditioning on a sufficient set of control variables. In instrumental variables models, it involves the assumption that the instrument is as good as randomly assigned. In panel or differences-in-differences designs, it is the parallel trends assumption. The credibility of a design can be enhanced when researchers can show explicitly that these assumptions are supported by the data. This is often done through some form of balancing tests or robustness checks.

The research designs mentioned above are all variants of regression strategies. If the researcher has access to a variable for a potentially remaining confounder, tests of the identifying assumption take two canonical forms. The variable can be added as a control on the right hand side of the regression. The identifying assumption is confirmed if the estimated effect of interest is insensitive to this variable addition—we call this the coefficient comparison test. Alternatively, the variable can be placed on the left hand side of the regression instead of the outcome variable. A zero coefficient on the causal variable of interest then confirms the identifying assumption. This is the balancing test which is typically carried out using baseline characteristics or pre-treatment outcomes in a randomized trial or in a regression discontinuity design.

Researchers often rely on one or the other of these tests. The main point of our paper is to show that the balancing test, using the proxy for the candidate confounder on the left hand side of the regression, is generally more powerful. This is particularly the case when the available variable is a noisy measure of the true underlying confounder. The attenuation due to measurement error often implies that adding the candidate variable on the right hand side as a regressor does little to eliminate any omitted variables bias. The same measurement error does comparatively less damage when putting this variable on the left hand side. Regression strategies work well in finding small but relevant

amounts of variation in noisy dependent variables. We collect basic results for the relevant parameters in the presence of measurement error in Section 3.

These two testing strategies are intimately related through the omitted variables bias formula. The omitted variables bias formula shows that the coefficient comparison test involves two regression parameters, the coefficient from the balancing test and the coefficient from the added regressor in the outcome equation. Both of these parameters have to be non-zero for the coefficient comparison test to fail and actual confounding to take place. The balancing test focuses on a single parameter. The two tests therefore investigate the same hypothesis under the maintained assumption that the added regressor matters in the outcome equation. The ultimate source of the power loss in the coefficient comparison test comes from estimating a nuisance parameter. This is a standard reason for power differences in the econometrics literature but turns out to be relatively unimportant in the numerical examples we present. The nuisance parameter in the coefficient comparison test is more difficult to estimate when there is more measurement error in the added regressor. In the examples we study in Section 5, measurement error is the source of quantitatively meaningful power differences between the two tests.

A second point we are making is that the two strategies, coefficient comparison and balancing, both lead to explicit statistical tests. The balancing test is a simple t-test used routinely by researchers. When adding a covariate on the right hand side, comparing the coefficient of interest across the two regressions can be done using a generalized Hausman test. In practice, we have not seen this test carried out in applied papers, where researchers typically just eye-ball the results (an exception is Gelbach, 2016). We provide the relevant test statistics and discuss how they behave under measurement error in Section 4. We demonstrate the superior power of the balancing test under different scenarios in Section 5.

The principles underlying our analysis are well known but the consequences do not seem to be fully appreciated in applied work. McCallum (1972) and Griliches (1977) are classic references for the issues arising when regression controls are measured with error. Battistin and Chesher (2014) discuss identi-

fication in the presence of a mismeasured covariate in non-linear models based on assumptions about the degree of measurement error in the covariate. We follow McCallum (1972) and Griliches (1977) in framing our discussion around the omitted variables bias arising in linear regressions, the general framework used most widely in empirical studies. The insights we exploit build on Pischke and Schwandt (2012) but we go beyond the analysis in all of these papers in our explicit discussion of testing, which forms the core of our inquiry.

Our focus is on specification testing for a particular research design. The statistical model we discuss below—a baseline regression and an augmented regression with additional covariates—bears a close relationship to models in a large literature, which attempts to use control strategies for point or interval identification. One recent strand of this literature is interested in the selection of control variables in a causal regression and inference for the parameter of interest after such an initial variable selection step (Belloni, Chernozhukov and Hansen, 2014*a*, *b*; Chernozhukov et al., 2017; Chernozhukov et al., forthcoming). A second strand uses the relationship between a treatment variable of interest and observed covariates to model the corresponding relationship with additional unobserved confounders in order to estimate the true causal effect (Altonji, Elder and Taber, 2005; Altonji et al., 2013; and Oster, forthcoming). Although this literature is focused on identification of the causal parameter, the tools can be used for specification checking as well, so in practice the conceptual difference to our approach may not be quite as sharp. Nevertheless, the parameters of interest are different, and our focus is on statistical inference about the credibility of a given baseline design rather than identification of the causal parameter.

Also related is an older literature by Hausman (1978), Hausman and Taylor (1980), and Holly (1982) (see also the summary in MacKinnon, 1992, section II.9), which considers the relative power of the Hausman test compared to alternatives, in particular an F-test for the added covariates in the outcome equation when potentially multiple covariates are added. This comparison effectively maintains that there is a lack of balance, and instead tests whether the added regressors matter for explaining the outcome. While this is a dif-

ferent exercise from ours, this literature highlights the potential power of the Hausman test when it succinctly transforms a test with multiple restrictions (like the F-test for the added covariates) into a test with a single restriction (the coefficient comparison test). We discuss how to extend our framework to multiple added controls in Section 5.3. Our basic findings largely carry over to this setting but we also reach the conclusion that the Hausman test has a role to play when the goal is to summarize a large number of restrictions.

Griliches (1977) uses estimates of the returns to schooling as example for the methodological points he makes. Such estimates have formed a staple of labor economics ever since. We use Griliches' data from the National Longitudinal Survey of Young Men to briefly illustrate our power results in Section 6. It is well suited for our purposes because the data contain various test score measures which can be used as controls in a regression strategy (as in Griliches, 1977), as well as a myriad of other useful variables on individual and family background. The empirical results illustrate and support our theoretical claims.

2 A SIMPLE FRAMEWORK

Consider the following simple framework starting with a population regression equation

$$y_i = \beta^s s_i + e_i^s \tag{1}$$

where y_i is an outcome like log wages, s_i is the causal variable of interest, like years of schooling, and e_i^s is the regression residual. The researcher proposes this short regression model to be causal, i.e. β^s is the parameter of interest. This might be the case because the data come from a randomized experiment, so the simple bivariate regression is all we need. More likely, the researcher has a particular research design applied to observational data. For example, in the case of a regression strategy controlling for confounders, y_i and s_i would be residuals from regressions of the original outcome and treatment variables on the chosen controls. In the case of panel data or differences-in-differences designs the controls are sets of fixed effects. In the case of instrumental variables,

s_i would be the predicted value from a first stage regression. In practice, (1) encompasses a wide variety of empirical approaches, and should be thought of as a short-hand for these. We have this broader interpretation in mind but for presentational clarity we use the simple bivariate regression throughout the discussion in our paper. All subsequent regression equations and results also inherit the structure of the actual underlying research design but we illustrate results in terms of the simple bivariate formulation in (1). We also suppress constants to avoid clutter.

Now consider the possibility that the population regression parameter β^s from (1) may not actually capture a causal effect. There may be a candidate confounder x_i, so that the long regression

$$y_i = \beta^l s_i + \gamma x_i + e_i \tag{2}$$

generates a coefficient β^l which might differ from β^s. To make things concrete, in the returns to schooling context, x_i would be a measure of the remaining part of an individual's earnings capacity which is also related to schooling, like ability or family background.

Researchers who find themselves in a situation where they start with a proposed causal model (1) and a measure for a candidate confounder x_i typically do one of two things: They either regress x_i on s_i and check whether s_i is significant, or they include x_i on the right hand side of the original regression as in (2), and check whether the estimate of β changes materially when x_i is added to the regression of interest. The first strategy constitutes a test for "balance," a standard check for successful randomization in an experiment. The second strategy is a "coefficient comparison test." An appreciable difference between β^l and β^s suggests that the original estimate β^s does not have a causal interpretation. Researchers typically interpret passing either of these tests as strengthening the case for a causal interpretation of the parameter β^s. In case the tests reject, the researcher concludes that neither parameter is likely to be causal, and the research design is a flawed one.

It is tempting to conclude that strategy (2) is preferable because the comparison of β^l and β^s does not just carry information about the validity of re-

gression (1) but also provides a better estimate β^l. It is important to caution against this interpretation. If x_i is an imperfect control or there are multiple omitted variables in (1) then (2) does not necessarily reduce the omitted variables bias (Frost, 1979 or more recently De Luca, Magnus and Peracchi, 2017 and Kassenboehmer and Schurer, 2017). The literatures along the lines of Altonji, Elder and Taber (2005) and Belloni, Chernozhukov and Hansen (2014*b*) all start from the premise that there is a set of regressors x_i so that regression (2) is preferable, at least in principle. Only in the special case where x_i is the only missing confounder and we measure it without error will the parameter β^l from the controlled regression be the causal effect of interest. In practice, there is usually little reason to believe that these two conditions are met, and hence a difference between β^l and β^s only indicates a poor research design.

The relationship between the two testing strategies is easy to see. Write the regression of x_i on s_i, which we will call the balancing regression, as

$$x_i = \delta s_i + u_i. \tag{3}$$

The change in the coefficient on s_i after adding x_i to the regression (1) is given by the omitted variables bias formula

$$\beta^s - \beta^l = \gamma\delta. \tag{4}$$

This change consists of two components, the coefficient γ on x_i in the outcome equation (2) and the coefficient δ from the balancing regression.

Here we consider the relationship between these two approaches: the balancing test, consisting of an investigation of the null hypothesis

$$H_0 : \delta = 0, \tag{5}$$

compared to the inspection of the coefficient movement $\beta^s - \beta^l$. The latter strategy of comparing β^s and β^l is often done informally, but it can be formalized as a statistical test of the null hypothesis

$$H_0 : \beta^s - \beta^l = 0, \tag{6}$$

which we will call the coefficient comparison test. From (4) it is clear that (6) amounts to

$$H_0 : \beta^s - \beta^l = 0 \Leftrightarrow \gamma = 0 \text{ or } \delta = 0. \tag{7}$$

This highlights that the two approaches formally test the same hypothesis under the maintained assumption $\gamma \neq 0$. We may often have a strong sense that $\gamma \neq 0$; i.e. we are dealing with a variable x_i which we believe affects the outcome, but we are unsure whether it is related to the regressor of interest s_i. In this case, both tests would seem equally suitable. Nevertheless, in other cases γ may be zero, or we may be unsure. In this case, the coefficient comparison test seems to dominate because it directly addresses the question we are after, namely whether the coefficient of interest β is affected by the inclusion of x_i in the regression.

Be this as it may, our main point is a practical one, that the coefficient comparison test suffers particularly when the true confounder ($\gamma \neq 0$) is measured with error. In general, confounders like x_i may not be easy to measure. If the available measure for x_i contains classical measurement error, the estimator of γ in (2) will be attenuated, and the comparison $\beta^s - \beta^l$ will be too small (in absolute value) as a result. The estimator of δ from the balancing regression is still consistent in the presence of classical measurement error; this regression simply loses precision because the mismeasured variable is on the left hand side. The measurement error drives a wedge between the asymptotic values of the two test statistics and the balancing test becomes relatively more powerful than the coefficient comparison test. In order to make these statements precise, we start by reviewing results for the relevant population parameters in the case of classical measurement error in the following section, before moving on to inference, power calculations, and simulations.

3 POPULATION PARAMETERS IN THE PRESENCE OF MEASUREMENT ERROR

The candidate variable x_i is not observed. Instead, the researcher works with the mismeasured variable

$$x_i^m = x_i + m_i. \tag{8}$$

We start by assuming the measurement error m_i is classical, i.e. $E(m_i) = 0$, $Cov(x_i, m_i) = 0$, $Cov(s_i, m_i) = 0$. In Section 5 below we also investigate the impact of mean-reverting measurement error. As a result of the measurement error, the researcher compares the regressions

$$\begin{aligned} y_i &= \beta^s s_i + e_i^s \\ y_i &= \beta^m s_i + \gamma^m x_i^m + e_i^m. \end{aligned} \tag{9}$$

Notice that the short regression does not involve the mismeasured x_i, so that $\beta^s = \beta^l + \gamma\delta$ as before. However, the population regression coefficients β^m and γ^m are now different from β^l and γ from equation (2):

$$\begin{aligned} \beta^m &= \beta^l + \gamma\delta\theta \\ \gamma^m &= \gamma(1-\theta). \end{aligned} \tag{10}$$

The amount of measurement error is captured by the parameter θ:

$$\theta = \frac{\sigma_m^2}{\sigma_u^2 + \sigma_m^2},$$

where $\sigma_\bullet^2$ denotes the variance of the random variable in the subscript (McCallum, 1972; Garber and Klepper, 1980). $1-\theta$ is the multivariate attenuation factor, which takes the role of the familiar attenuation factor $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_m^2)$ in a bivariate regression. Recall that u_i is the residual from the balancing regression (3). Notice that θ involves only the variation in x_i^m which is orthogonal to s_i. This is the part of the variation in x_i^m relevant to the estimate of γ^m in regression (9), which also has s_i as a regressor. Approaches along the lines of Battistin and Chesher (2014), Altonji, Elder and Taber (2005),

and Oster (forthcoming), which effectively treat eq. (2) as structural, require assumptions on θ or a function of it for point identification.

The population coefficient β^m differs from β^l but less so than β^s. In fact, with classical measurement error β^m lies between β^s and β^l, as can be seen from (10). The parameter γ^m is attenuated compared to γ; the attenuation is bigger than in the case of a bivariate regression of y_i on x_i^m without the regressor s_i if x_i^m and s_i are correlated because $\sigma_u^2 < \sigma_x^2$.

These results highlight a number of issues. The gap $\beta^s - \beta^m$ is too small compared to the desired $\beta^s - \beta^l$, directly affecting the coefficient comparison test. This is a consequence of the fact that γ^m is biased towards zero. Ceteris paribus, this is making the assessment of the hypothesis $\gamma = 0$ more difficult, which in turn affects the inference for $\beta^s = \beta^l$.

Finally, with the mismeasured x_i^m, the balancing regression becomes

$$\begin{aligned} x_i^m &= \delta^m s_i + u_i^m \\ &= \delta s_i + u_i + m_i. \end{aligned} \tag{11}$$

This regression involves measurement error in the dependent variable, which has no effect on the population parameter $\delta^m = \delta$. Because the variance of the residual in (11) is larger than in (3) the estimator $\widehat{\delta}^m$ is less efficient than $\widehat{\delta}$ in the case with no measurement error.

4 INFERENCE

In this section, we consider how conventional standard errors and test statistics for the quantities of interest are affected in the homoskedastic case (see Appendix A for details on the setup, derivations, and an extension to robust standard errors). The primitive disturbances are s_i, u_i, e_i, and m_i, which we assume to be uncorrelated with each other. Other variables are determined by (2), (3), and (8). We use these results to analyze the power of the two alternative tests in the next section. Starting with theoretical results for the baseline homoskedastic case, we extend these results in simulations. Our basic conclusions are the same in all these different scenarios.

Start with the estimator $\widehat{\delta}^m$ and its associated t-statistic. $\widehat{\delta}^m$ is still a consistent estimator for δ but its standard error is inflated compared to the case with no measurement error. Denoting the estimated standard error of a given estimator by $\widehat{se}(\bullet)$, a test based on the t-statistic $t_{\delta^m} = \widehat{\delta}^m / \widehat{se}\left(\widehat{\delta}^m\right)$ remains consistent because m_i is correctly accounted for in the residual of the balancing regression (11). However, the t-statistic is asymptotically smaller in absolute value than in the error free case. As $n \to \infty$, the scaled t-statistic is

$$\text{plim}\left(\frac{1}{\sqrt{n}} t_{\delta^m}\right) = \sqrt{1-\theta}\frac{\delta}{\left(\frac{\sigma_u}{\sigma_s}\right)}.$$

This means the null hypothesis (5) is rejected less often. The test is less powerful than in the error free case ($\theta = 0$); the power loss is captured by the term $\sqrt{1-\theta}$.

We next turn to $\widehat{\gamma}^m$, the estimator for the coefficient on the mismeasured x_i^m in (9). The parameter γ is of interest since it determines the coefficient movement $\beta^s - \beta^l = \gamma\delta$ in conjunction with the result from the balancing regression. For ease of exposition, we impose conditional homoskedasticity of e_i^m given s_i and x_i^m here and leave the more general case to Appendix A.3.2. Denote the asymptotic standard error by $se(\bullet)$, i.e. $se(\bullet) \equiv \frac{1}{\sqrt{n}}\text{plim}\{\sqrt{n}\widehat{se}(\bullet)\}$. The asymptotic standard error for $\widehat{\gamma}^m$ is

$$se\left(\widehat{\gamma}^m\right) = \frac{\sqrt{1-\theta}}{\sqrt{n}}\sqrt{\theta\gamma^2 + \frac{\sigma_e^2}{\sigma_u^2}}.$$

Measurement error enters the standard error in two ways: the first is an attenuation factor compared to the standard error for a correctly measured x_i, while the second is an additive effect which depends on the value of γ. The parameters in the two terms are not directly related, so $se\left(\widehat{\gamma}^m\right) \gtrless se\left(\widehat{\gamma}\right)$. Measurement error does not necessarily inflate the standard error here.

The two terms have a simple, intuitive interpretation. Measurement error attenuates the parameter γ^m towards zero, the attenuation factor is $1-\theta$. The standard error is attenuated in the same direction; this is reflected in the $\sqrt{1-\theta}$ factor, which multiplies the remainder of the standard error calculation. The second influence from measurement error comes from the term $\theta\gamma^2$,

which results from the fact that the residual variance $Var\left(e_i^m\right)$ is larger when there is measurement error. The increase in the variance is related to the true γ, which enters the residual.

The t-statistic for testing whether $\gamma^m = 0$ has a limit

$$\text{plim}\left(\frac{1}{\sqrt{n}}t_{\gamma^m}\right) = \sqrt{1-\theta}\frac{\gamma}{\sqrt{\theta\gamma^2 + \frac{\sigma_e^2}{\sigma_u^2}}}.$$

In addition to the two sources of measurement error in the standard error, the t-statistic involves the attenuation factor $1-\theta$ for the coefficient γ^m. As in the case for the balancing regression, the t-statistic for $\widehat{\gamma}^m$ is smaller than t_γ for the error free case. But in contrast to the balancing test statistic t_{δ^m}, measurement error reduces t_{γ^m} relatively more, due to the fact that measurement error in a regressor both attenuates the relevant coefficient towards zero (captured by $\sqrt{1-\theta}$) and introduces additional variance into the residual (the $\theta\gamma^2$-term) in the denominator. As a result, classical measurement error makes the assessment of whether $\gamma = 0$ more difficult compared to the assessment of whether $\delta = 0$. As we will see, this contributes to the greater power of the balancing test statistic.

Finally, consider the quantity $\beta^s - \beta^m$, which enters the coefficient comparison test. Before proceeding we note that the covariance term in the expression for the asymptotic variance of $\widehat{\beta}^s - \widehat{\beta}^m$

$$Var\left(\widehat{\beta}^s - \widehat{\beta}^m\right) = Var\left(\widehat{\beta}^s\right) + Var\left(\widehat{\beta}^m\right) - 2Cov\left(\widehat{\beta}^s, \widehat{\beta}^m\right) \tag{12}$$

reduces the sampling variance of $\widehat{\beta}^s - \widehat{\beta}^m$. This covariance term is positive and generally sizable compared to $Var\left(\widehat{\beta}^s\right)$ and $Var\left(\widehat{\beta}^m\right)$ since the regression residuals e_i^s and e_i^m are highly correlated. Because $2Cov\left(\widehat{\beta}^s, \widehat{\beta}^m\right)$ gets subtracted, looking at the standard errors of $\widehat{\beta}^s$ and $\widehat{\beta}^m$ alone can potentially mislead the researcher into concluding that the two coefficients are not significantly different from each other when in fact they are.

The coefficient comparison test itself can be formulated as a t-test as well, since we are interested in the movement in a single parameter, i.e.

$$t_{(\beta^s - \beta^m)} = \frac{\widehat{\beta}^s - \widehat{\beta}^m}{\widehat{se}(\widehat{\beta}^s - \widehat{\beta}^m)}$$

where $\widehat{se}(\widehat{\beta}^s - \widehat{\beta}^m)$ is a consistent standard error estimator. Using (4) and (10) we obtain

$$\text{plim}\left(\frac{1}{\sqrt{n}}t_{(\beta^s-\beta^m)}\right) = \sqrt{1-\theta}\frac{\delta\gamma}{\sqrt{\gamma^2\frac{\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \delta^2\frac{\sigma_e^2}{\sigma_u^2}}} \tag{13}$$

Under the alternative hypothesis ($\delta \neq 0$) and the maintained assumption $\gamma \neq 0$, the three test satistics are asymptotically related in the following way:

$$\text{plim}\left(\frac{1}{\frac{1}{\sqrt{n}}t_{(\beta^s-\beta^m)}}\right)^2 = \text{plim}\left(\frac{1}{\frac{1}{\sqrt{n}}t_{\delta^m}}\right)^2 + \text{plim}\left(\frac{1}{\frac{1}{\sqrt{n}}t_{\gamma^m}}\right)^2 \tag{14}$$

This result highlights a number of things. First of all, under the maintained hypothesis $\gamma \neq 0$, the balancing test alone is more powerful. This is not surprising at all, since the balancing test only involves estimating the parameter δ while the coefficient comparison test involves estimating both δ and γ. Imposing $\gamma \neq 0$ in the coefficient comparison test is akin to $t_{\gamma^m} \to \infty$, and this would restore the equivalence of the balancing and coefficient comparison tests. Note that the power advantage from imposing $\gamma \neq 0$ exists regardless of the presence of measurement error.

The second insight is that measurement error affects the coefficient comparison test in two ways. The test statistic is subject to both the attenuation factor $\sqrt{1-\theta}$ and the term $\theta\delta^2\gamma^2$ in the variance, which is inherited from the t-statistic for $\widehat{\gamma}^m$. Importantly, however, all these terms interact in the coefficient comparison test. In our numerical exercises below, it turns out that the way in which measurement error attenuates γ^m compared to γ is a major source of the power disadvantage of the coefficient comparison test. Our simulations demonstrate that the differences in power between the coefficient comparison and balancing tests can be substantial when there is considerable measurement error in x_i^m.

5 POWER COMPARISONS

5.1 Asymptotic and Monte Carlo Results with Classical Measurement Error

The ability of a test to reject when the null hypothesis is false is described by the power function of the test. The power functions here are functions of d, the values the parameter δ might take on under the alternative hypothesis, while we keep $\gamma \neq 0$ fixed. Using our results from the previous section, it is easy to demonstrate that under the alternative hypothesis $\delta \neq 0$,

$$Power_{t_{\delta^m}}(d) > Power_{t_{(\beta^s - \beta^m)}}(d; \gamma). \tag{15}$$

We give a full derivation in Appendix A.

In practice, this result may or may not be important. In addition, when the standard error is estimated, the powers of the two tests may differ from the theoretical results above. Therefore, we carry out a number of Monte Carlo simulations to assess the performance of the two tests.

Table 1: Parameters for Power Calculations and Implied R^2s

$\sigma_s^2 = 1$		$\beta = 1$	
$\sigma_u^2 = 3$		$\gamma = 3$	
$\sigma_e^2 = 30$		$n = 100$	
		R^2	
d	$\theta = 0$	$\theta = 0.7$	$\theta = 0.85$
0	0.48	0.16	0.09
0.5	0.53	0.23	0.16
1.0	0.59	0.33	0.27
1.5	0.66	0.44	0.39
2.0	0.72	0.54	0.50

NOTE: The implied population R^2's do not depend on n, but the subsequent power calculations do.

Table 1 displays the parameter values we use as well as the implied values of the population R^2 of regression (9). The values were chosen so that for

intermediate amounts of measurement error in x_i^m the R^2s are reflective of regressions fairly typical of those in applied microeconomics, for example, a wage regression. Note that the amounts of measurement error we consider are comparatively large. In our empirical application we use variables like mother's education and the presence of a library card in the household as measures of family background. We suspect that these variables pick up at most a minor part of the true variation of family background, even in the presence of other covariates, so that values of $\theta = 0.7$ or $\theta = 0.85$ for the measurement error are not unreasonable.

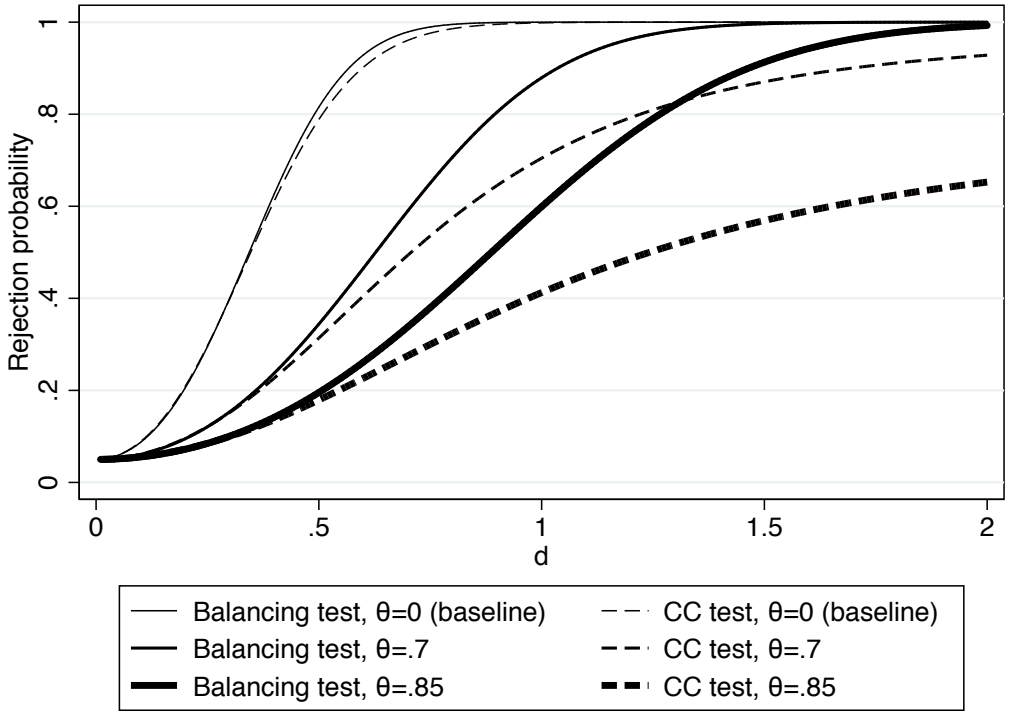

Figure 1: Theoretical Rejection Rates. d is the value the coefficient in the balancing equation takes on under the alternative hypothesis.

In Figure 1, we start by plotting the theoretical power functions for both tests for three different magnitudes of the measurement error. We calculate these power functions using the t-distribution with $n-2$ degrees of freedom, consistent with how Stata 14 performs the balancing test (this distribution

choice makes little difference with our sample size of 100). The thin lines show the power functions with no measurement error. The power functions can be seen to increase quickly with d, and both tests reject with virtual certainty once d exceeds values of 1. The balancing test is slightly more powerful but this difference is small, and only visible in the figure for a small range of d.

The medium thick lines correspond to $\theta = 0.7$, i.e. 70% of the variance of x_i^m is measurement error after partialling out s_i. Measurement error of that magnitude visibly affects the power of both tests. The balancing test still rejects with certainty for $d > 1.5$, while the coefficient comparison test does not reject with certainty for the parameter values considered in the figure. This discrepancy becomes even more pronounced when we set $\theta = 0.85$ (thick lines). The power of the coefficient comparison test does not rise above 0.65 in this case, while the balancing test still rejects with probability 1 when d is around 2.

The results in Figure 1 highlight that there are parameter combinations where the balancing test has substantially more power than the coefficient comparison test. In other regions of the parameter space, the two tests have more similar power, for example, when $d < 0.5$. While we highlight the consequences of measurement error throughout the paper, we should note that formally any particular value of θ can be mimicked by an appropriate combination of values for γ and σ_u^2. This is an immediate consequence of the fact that the classical measurement error model is underidentified by one parameter. In that sense "measurement error" is simply a label for a certain set of parameter values. It is always difficult to choose empirically relevant values for simulations, and we take comfort from the fact that the results emerging from this section are also reflected in the empirical example in Section 6.

Before going on to simulations of more complicated cases, we contrast the theoretical power functions in Figure 1, based on asymptotic approximations, to simulated rejection rates of the same tests in Monte Carlo samples. Figure 2 shows the power functions for the two tests without measurement error ($\theta = 0$) and with a large amount of measurement error ($\theta = 0.85$), as well as their simulated counterparts. We computed 25,000 replications in these

simulations, and each repeated sample contains 100 observations. Without measurement error, the theoretical power functions are closely aligned with the empirical rejection rates (thinner lines). Adding measurement error, this is also true for the balancing test (the solid thicker lines are on top of each other and not distinguishable) but not for the coefficient comparison test (broken thicker lines).

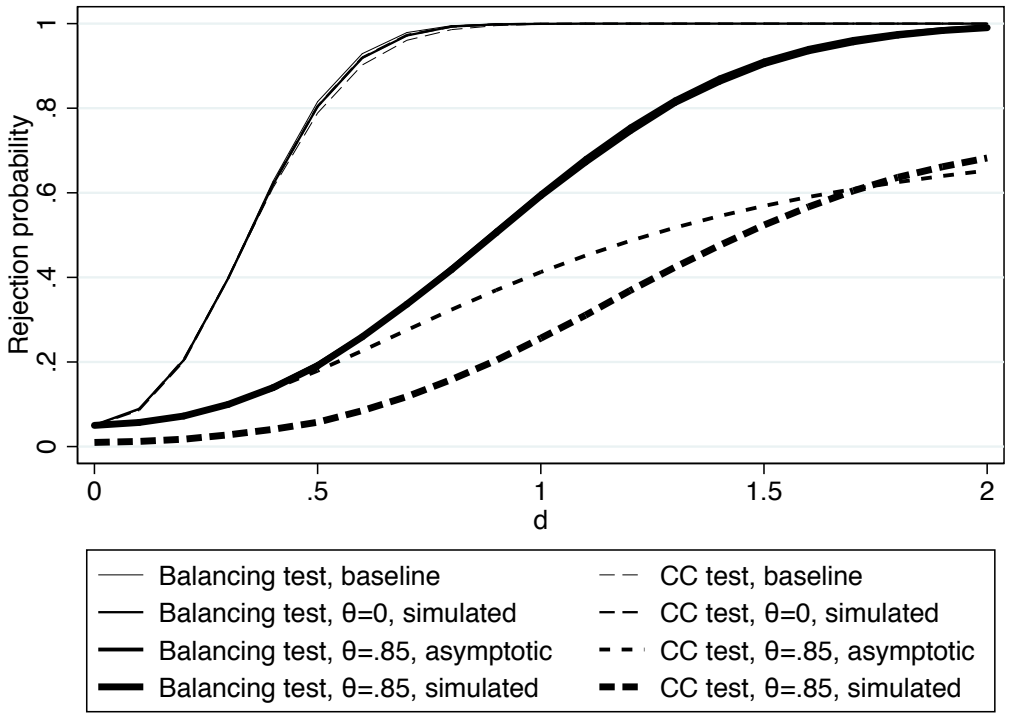

Figure 2: Theoretical and Simulated Rejection Rates. Comparison of asymptotic rejection rates (from Figure 1) with rejection rates based on Monte Carlo simulations. Baseline refers to the theoretical rejection rates without measurement error. d is the value the coefficient in the balancing equation takes on under the alternative hypothesis.

Figure 2 reveals that the empirical rejection rates of the coefficient comparison test in the presence of measurement error deviate substantially from the power function calculation based on the asymptotic approximation. This discrepancy is almost completely explained by the fact that we use the asymptotic

values of standard errors in the calculations but estimated standard errors in the simulations. The joint distribution between the coefficient and standard error estimators is difficult to characterize, especially in the case of the coefficient comparison test, so we abstract away from the sampling variation in estimating the standard errors in the theoretical derivations of the power functions. Figure 2 shows that the test is severely distorted under the null in the simulations; it barely rejects more than 1% of the time for a nominal size of 5%. While this problem leads to too few rejections under the null, it is important to note that the same issue arises for positive values of d until about $d < 1.5$. For larger values of d the relationship reverses. In other words, for moderate values of d the coefficient comparison test statistic is biased downwards under the alternative, and the test has too little power. This highlights another advantage of the balancing test—a standard t-test where no such problem arises. We note that this is a small sample problem, which goes away when we increase the sample size (in unreported simulations). We suspect that this problem is related to the way in which the coefficient comparison test effectively combines the simple t_{δ^m} and t_{γ^m} test statistics in a non-linear fashion, as can be seen in equation (14), and the fact that t_{γ^m} sometimes is close to 0 in small samples despite the fact that we fix γ substantially above 0.

5.2 Monte Carlo Results with Mean-reverting Measurement Error

The homoskedastic case with classical measurement error is highly stylized and does not correspond well to the situations typically encountered in empirical practice. We explore the case of mean reverting measurement error (Bound et al., 1994) using simulations in this sub-section. Some additional results can be found in Appendix D. We generate measurement error as

$$m_i = \kappa x_i + \mu_i$$

where κ is a parameter and $Cov(x_i, \mu_i) = 0$, so that κx_i captures the error related to x_i and μ_i the unrelated part. When $-1 < \kappa < 0$, the error is mean

reverting, i.e. the κx_i-part of the error reduces the variance in x_i^m compared to x_i.

The case of mean reverting measurement error captures a variety of ideas, including the one that we may observe only part of a particular confounder made up of multiple components. Imagine we would like to include in our regression a variable $x_i = w_{1i} + w_{2i}$, where w_{1i} and w_{2i} are two orthogonal variables. We observe $x_i^m = w_{1i}$. For example, x_i may be family background, w_{1i} is mother's education and other parts of family background correlated with it, and w_{2i} are all relevant parts of family background which are uncorrelated with mother's education. As long as selection bias due to w_{1i} and w_{2i} is the same, this amounts to the mean reverting measurement error formulation above. Note that $\lambda = Var(x_i)/Var(x_i^m) > 1$ in this case, so the mismeasured x_i^m has a lower variance than the true x_i. This scenario is also isomorphic to the model studied by Oster (forthcoming). See Appendix B for details.

The mismeasured x_i^m can now be written as

$$x_i^m = (1+\kappa)\,\delta s_i + (1+\kappa)\,u_i + \mu_i,$$

so mean reversion directly affects the coefficient in the balancing regression, which will be smaller than δ for a negative κ. As a result, the balancing test will reject less often. At the same time, a negative κ offsets and possibly reverses the attenuation bias on γ. This brings the power functions of the balancing and coefficient comparison tests closer together.

For the simulations we set $\kappa = -0.5$, so the error is mean reverting. We also fix σ_μ^2 in the simulations. However, it is important to note that the nature of the measurement error will change as we change the value of d under the alternative hypotheses. x_i depends on δ and the correlated part of the measurement error depends in turn on x_i. We show results for two cases with $\sigma_\mu^2 = 0.75$ and $\sigma_\mu^2 = 2.25$. Under the null, these two parameter values correspond to $\lambda = 2$ and $\lambda = 1$, respectively. The case $\lambda = 2$ corresponds to the Oster (forthcoming) model just described with $Var(w_{1i}) = Var(w_{2i})$. These models exhibit relatively large amounts of mean reversion. Figure 3 demonstrates that the balancing test again dominates for these parameter

values. The gap is small for the $\sigma_\mu^2 = 0.75$ case but grows with σ_μ^2, the classical portion of the measurement error. This finding is not surprising as the mean-reversion part in the measurement error biases the estimate of γ in the opposite direction from the classical part and can in principle flip the sign of the bias around. As a result, the coefficient comparison test could have greater power.

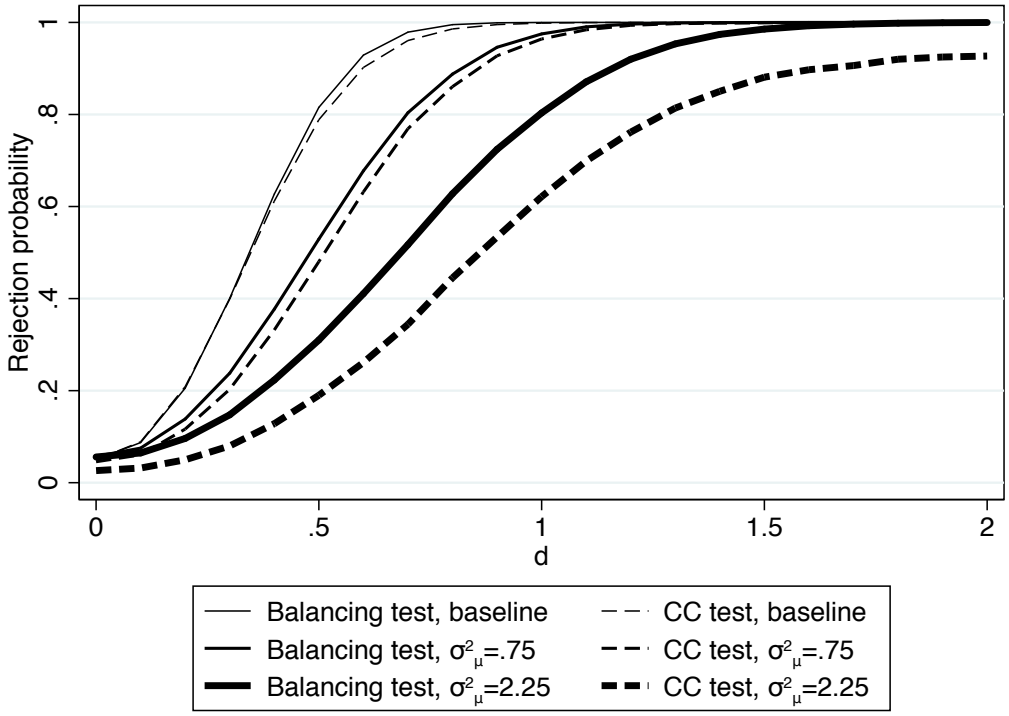

Figure 3: Simulated Rejection Rates with Mean Reverting Measurement Error. Comparison of baseline rejection rates (from Figure 1) with simulated rejection rates based on mean reverting measurement error and robust standard errors. d is the value the coefficient in the balancing equation takes on under the alternative hypothesis.

5.3 Multiple Controls

So far we have concentrated on the case of a single added regressor x_i. Often in empirical practice we may want to add a set of additional covariates at once. It is straightforward to extend our framework to that setting. Some interesting

new issues arise in this analysis.

Suppose there are k added regressors, i.e. $\mathbf{x}_i$ is a $k \times 1$ vector, and

$$\begin{aligned} y_i &= \beta^l s_i + \mathbf{x}_i' \boldsymbol{\gamma} + e_i \\ \mathbf{x}_i &= \boldsymbol{\delta} s_i + \mathbf{u}_i \\ \beta^s - \beta^l &= \boldsymbol{\gamma}' \boldsymbol{\delta} \end{aligned} \tag{16}$$

where $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and $\mathbf{u}_i$ are $k \times 1$ vector analogs of their scalar counterparts in Section 2. The coefficient comparison test compares the βs from eqs. (1) and (16) just as before. Lee and Lemieux (2010) suggest a balancing test for multiple covariates in the context of evaluating regression discontinuity designs. Let $\mathbf{x}_{(j)}$ denote the $n \times 1$ vector of all the observations on the j-th x-variable. Stack all the x-variables on the left-hand-side of the regression to obtain

$$\begin{bmatrix} \mathbf{x}_{(1)} \\ \mathbf{x}_{(2)} \\ ... \\ \mathbf{x}_{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{s} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{s} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{s} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ ... \\ \delta_k \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{(1)} \\ \mathbf{u}_{(2)} \\ ... \\ \mathbf{u}_{(k)} \end{bmatrix},$$

where $\mathbf{s} = [s_1, s_2, ..., s_n]'$ and $\mathbf{u}_{(j)}$ is the vector of residuals corresponding to covariate $\mathbf{x}_{(j)}$. The balancing test is an F-test for the joint significance of the $\boldsymbol{\delta}$ coefficients, the null is $\boldsymbol{\delta} = \mathbf{0}$.

We will call this stacking of equations the left-hand-side (LHS) balancing test. While it is the natural multivariate extension, an alternative would be to regress s on the covariates $\boldsymbol{x}$

$$s_i = \boldsymbol{\pi}' \mathbf{x}_i + v_i$$

(including any other covariates implicit in the regressions in eq. (16)) and test whether the coefficient vector $\boldsymbol{\pi}$ is significantly different from zero. This is a standard F-test. We refer to this test as the right-hand-side (RHS) balancing test. Notice that even though the balancing variables are now on the right, this is conceptually still a balancing test. Applied researchers sometimes use this RHS test; for example, Bruhn and McKenzie (2009) report it being used in some experimental studies in development economics.

While putting the balancing variables on the RHS might at first glance seem unusual, it turns out that the LHS and RHS tests are closely related. This should not be surprising as both tests exploit the joint covariance matrix of the $\mathbf{x}_{(j)}$ and $\mathbf{s}$. This can be seen most clearly in the case of a single covariate x_i (i.e. $k = 1$), where the LHS and the RHS tests using a conventional covariance matrix for homoskedastic residuals are numerically identical.

The intuiton for this is the following: In the single covariate case, the F-test amounts to the overall F-test for the significance of the regression. This, in turn, is a function of the R^2 of the regression. Since only two variables x_i and s_i are involved, this is the square of the correlation coefficient between the two. But the correlation coefficient is not directional, so the forward and reverse regression have to deliver the same F-statistic (in the case when covariates are present in the regression, replace the R^2 and correlation coefficient with their partial equivalents in this argument).

With multiple covariates ($k > 1$), the LHS and RHS tests are no longer equivalent. However, the scaled F-statistics of the two tests have the same probability limit in the special case where the LHS regression has a spherical error structure $var(\boldsymbol{u}_i) = \sigma^2 I_k$ and the RHS regression is homoskedastic, as we show in Appendix C. (See Ludwig, Mullainathan and Spiess, 2017 for a similar result).

How do the balancing tests with multiple covariates perform in practice? Figures 4 and 5 show simulations using a similar design as described in Table 1 for all k balancing equations. We set $k = 4$ and generate normally distributed, spherical errors and impose homoskedasticity and independence when performing the joint test of the δ_j's or the $\pi_j' s$. Our experiments with other moderate values of k for the most part did not reveal different insights. With multiple covariates there are different ways of specifying the alternative hypotheses now. The null hypothesis may fail for one, various, or all of the k covariates. We show rejection rates under two polar versions of the alternative hypothesis. Figure 4 shows simulations for the case where all covariates are unbalanced, i.e. $\delta_1 = \delta_2 = \ldots = \delta_k = d$. Figure 5 analyzes the case where only the first covariate is unbalanced while the others remain balanced, i.e.

$\delta_1 = d,\ \delta_2 = \ldots = \delta_k = 0.$

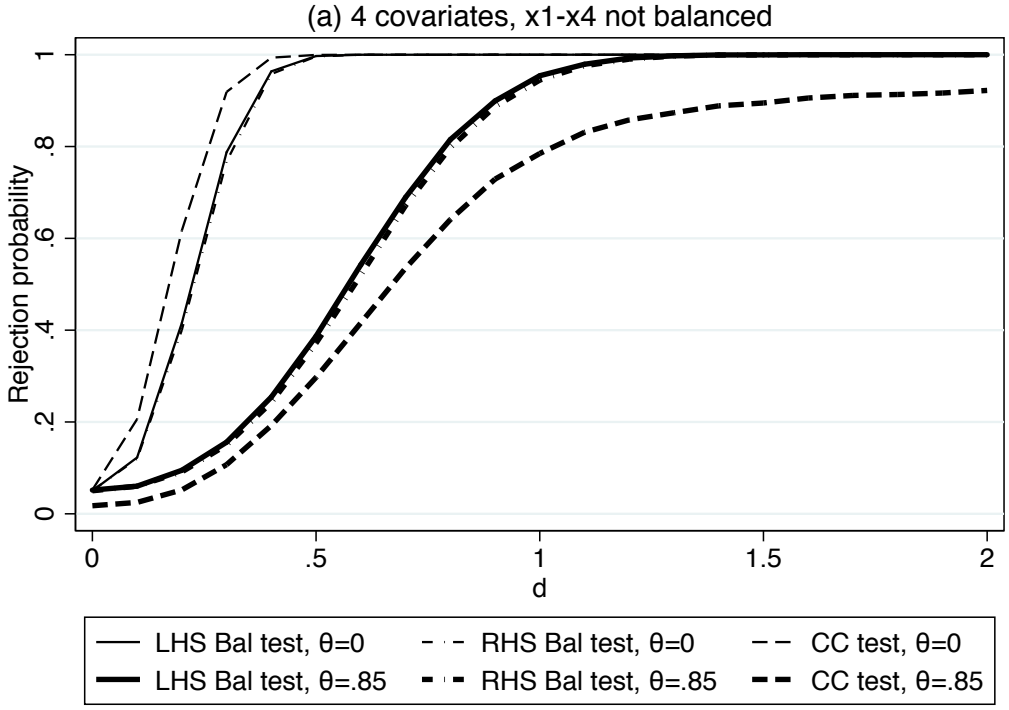

Figure 4: Simulated Rejection Rates with Multiple Controls: All Covariates Unbalanced. Simulated rejection rates for simultaneous tests for adding 4 additional covariates at once. All covariates are unbalanced under the alternative hypothesis; d is the value the coefficient in the balancing equation takes on under the alternative hypothesis for all covariates simultaneously.

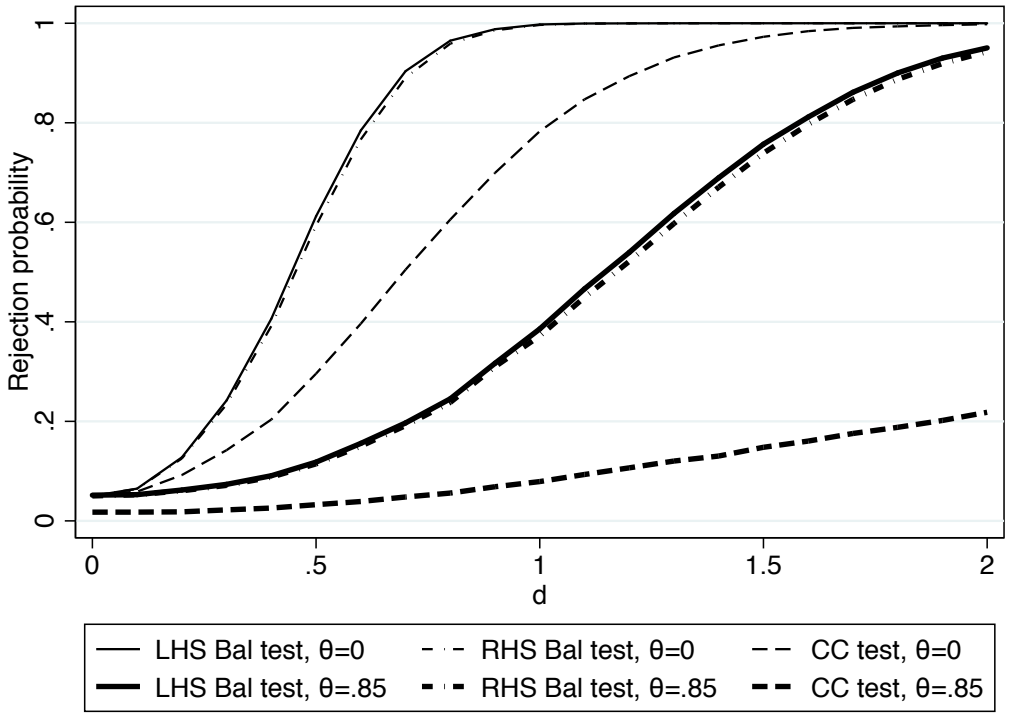

Figure 5: Simulated Rejection Rates with Multiple Controls: One Covariate Unbalanced. Simulated rejection rates for simultaneous tests for adding 4 additional covariates at once. Only one of the 4 covariates is unbalanced under the alternative hypothesis; d is the value the coefficient in the balancing equation takes on under the alternative hypothesis for this covariate.

These figures highlight a number of results. The LHS and RHS balancing tests are indeed very similar as their power functions virtually lie on top of each other in both figures. When all covariates are unbalanced as in Figure 4 and when measurement error is absent, the Hausman test turns out to be an efficient test in combining the k separate hypotheses into one single test-statistic, which is generated from the estimates of only two parameters, the long and short β's. The balancing tests, on the other hand, have to rely on the estimation of k parameters. In this case, the rejection rates for the coefficient comparison test (thin broken lines) therefore lie above the ones for both the balancing tests (thin solid and dash-dot lines). In the presence of measurement error, however, the balancing tests are again more powerful than the coefficient

comparison test as can be seen from the juxtaposition of the thicker lines.

This power advantage of the balancing tests is greater when only one covariate is unbalanced as can be seen in Figure 5. Both tests are less powerful in this case, but the power loss for the coefficient comparison test is now much more pronounced. This is particularly noticeable in the case with measurement error in the covariates (thick lines) but the balancing tests outperform the coefficient comparison test even without measurement error in this case. Empirically relevant cases may often lie in between these extremes. Researchers may be faced with a set of potential controls to investigate, some of which may be unbalanced with the treatment while others are not. Figures 4 and 5 demonstrate that the balancing test will frequently be the most powerful tool in such a situation, but the coefficient comparison test also has a role to play in the multivariate case.

The simulations reveal some further insights. With measurement error, the small sample issue of the coefficient comparison test, which we highlighted in Figure 2, arises again. On top of this, we found in unreported simulations that both the LHS and RHS balancing tests with robust standard errors (clustered standard errors across equations for the LHS test and heteroskedasticity-robust standard errors for the RHS test) have a size distortion under the null hypothesis and reject too often. This is due the standard small sample distortion of these covariance matrices discussed in the literature (MacKinnon and White, 1985; Chesher and Jewitt, 1987; Angrist and Pischke, 2009, chap. 8). We find that this bias tends to get worse when more covariates are added. Applied researcher may be most interested in the testing strategies discussed here when k is large (so that a series of single variable balancing tests is unattractive), and will want to rely on a robust covariance matrix. An upward size distortion may be less of an issue for a conservative researcher in a balancing test (where it means the researcher will falsely decide not to go ahead with a research design where the covariates are actually balanced) than in a test for the presence of non-zero treatment effects (where the same bias leads to false discoveries). Nevertheless, we suspect that most applied researchers would prefer a test with a correct size under the null and a steep power function. As

a result, research on improvements for the bias problem in multivariate tests is therefore particularly important (we discuss some current approaches in our working paper Pei, Pischke and Schwandt, 2017).

The upshot is that it is in principle straightforward to extend the balancing test to multiple covariates. An interesting finding is that a RHS test offers a computationally simple alternative that closely mimics the performance of the more standard LHS balancing test. Yet, at this point implementation issues related to the small sample bias of robust covariance estimators also hamper our ability to confidently carry out balancing tests for multiple covariates. Moreover, sometimes we are interested in the robustness of the original results when the number of added regressors is very large. An example would be a differences-in-differences analysis in a state-year panel, where the researcher is interested in checking whether the results are robust to the inclusion of state specific trends. The balancing test does not seem to be the right framework to deal with this situation. The coefficient comparison test has a role to play in this scenario.

6 EMPIRICAL ANALYSIS

We illustrate the theoretical results in the context of estimating the returns to schooling using data from the National Longitudinal Survey of Young Men (NLS). This is a panel study of about 5,000 male respondents interviewed from 1966 to 1981. The data set has featured in many prominent analyses of the returns to education, including Griliches (1977) and Card (1995). We use the NLS extract posted by David Card and augment it with the variable on body height measured in the 1973 survey. We estimate regressions similar to equation (2). The variable y_i is the log hourly wage in 1976 and s_i is the number of years of schooling reported by the respondent in 1976. Our samples are restricted to observations without missing values in any of the variables used.

Table 2 presents OLS regressions for the return to schooling controlling for the respondent's score on the Knowledge of the World of Work test (KWW), a

Table 2: Regressions for Returns to Schooling and Specification Checks Controlling for the KWW Score

	Dependent Variable							
	Log hourly earnings					Mother's years of education	Library card at age 14	Body height in inches
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.0609	0.0596	0.0608	0.0603	0.0591	0.2500	0.0133	0.0731
	(0.0059)	(0.0060)	(0.0059)	(0.0059)	(0.0060)	(0.0422)	(0.0059)	(0.0416)
KWW score	0.0070	0.0068	0.0069	0.0069	0.0067	0.0410	0.0076	0.0145
	(0.0015)	(0.0016)	(0.0016)	(0.0015)	(0.0016)	(0.0107)	(0.0016)	(0.0117)
Mother's years of education		0.0053			0.0048			
		(0.0037)			(0.0037)			
Library card at age 14			0.0097		0.0045			
			(0.0215)		(0.0216)			
Body height in inches				0.0078	0.0075			
				(0.0034)	(0.0034)			
p-values								
Coefficient comparison test		0.161	0.651	0.156	0.084			
LHS balancing test: individual						0.000	0.025	0.079
LHS balancing test: joint							0.000	
RHS balancing test: joint							0.000	

NOTE: The number of observations is 1,773 in all regressions. Heteroskedasticity robust standard errors in parentheses. The joint LHS balancing test is conducted via the `suest` Stata command. All regressions control for experience, experience-squared, indicators for black, for southern residence and residence in an SMSA in 1976, indicators for region in 1966 and living in an SMSA in 1966.

variable used by Griliches (1977) as a proxy for ability. Additional covariates are experience, race, and past and present residence. The estimated return to schooling is 0.061.

In columns (2) to (4) we include variables which might proxy for the respondent's family background, mother's education (column 2), whether the household had a library card when the respondent was 14 (column 3), and body height measured in inches (column 4). Mother's education captures an important component of a respondent's family background. The library card measure has been used by researchers to proxy for parental attitudes (e.g. Farber and Gibbons, 1996). Body height is determined by parents' genes and by nutrition and disease environment during childhood. It is unlikely a particularly powerful control variable but it is predetermined and correlated with family background, self-esteem, and ability (e.g. Persico, Postlewaite and Silverman, 2004; Case and Paxson, 2008).

Conditional on the KWW score, these three variables are only weakly correlated with earnings and only the coefficient for body height is marginally significant. The estimated return to education moves very little when these additional controls are included; the differences to column (1) are in the order of 0.001. In column (5) we enter all three variables simultaneously. The coefficients on the controls are slightly attenuated, and the return to education falls slightly further to 0.059. Below the estimates in columns (2) to (5), we display the p-values comparing each of the estimated returns to education to the one from column (1). None of the tests reject at the 5% level. These results from the coefficient comparison test seem to confirm the impression that the coefficient movements are not statistically significant.

It might be tempting to conclude from this evidence that the return to schooling estimated in column (1) should be given a causal interpretation but this conclusion is premature. A first caution actually comes from the coefficient comparison test in column (5), which is significant at the 10% level. The coefficient movement of 0.002 is not large, and the individual standard errors in columns (1) and (5) of 0.006 do not suggest that this movement might be significant. Eq. (12) warns that relying on the individual standard errors can

be rather misleading. Nevertheless, most researchers would probably not find the evidence in columns (1) to (5) worrisome enough to abandon their research project.

More potent warnings emerge from the balancing regressions in columns (6) to (8), where we regress maternal education, the library card, and body height on education while controlling for the KWW score. The education coefficient is positive and strongly significant for mother's education and the library card, and more marginally so for body height. Moreover, both the LHS and RHS joint balancing tests reject the hypothesis that all three controls are balanced with a p-value of virtually zero. The magnitudes of the coefficients, particularly mother's education, are substantively important. These estimates reflect selection bias: individuals with more education have significantly better educated mothers, were more likely to grow up in a household with a library card, and experienced more body growth when young. Our interpretation of these results is that education levels are related to family background in these regressions but the available background measures are fairly useless as controls when put on the right-hand side. These measurement problems matter less for the estimates in columns (6) to (8), and these specifications are therefore informative about the role of selection. Comparing the p-values at the bottom of the table to the corresponding ones for the coefficient comparison test in columns (2) to (4) demonstrates the superior power of the balancing test and illustrates the message of our paper in a forceful fashion.

7 CONCLUSION

Using predetermined characteristics as dependent variables offers a useful specification check for a variety of identification strategies popular in empirical economics. We argue that this is the case even for variables which might be poorly measured and are of little value as control variables. Such variables are available in many data sets. We encourage researchers to be more inventive in finding such measures and perform balancing tests with them more frequently. We show that this is generally a more powerful strategy than adding the same

variables on the right hand side of the regression as controls and looking for movement in the coefficient of interest.

We have illustrated our theoretical results with an application to the returns to education. We find the balancing test indeed to be useful for gauging selection bias due to confounders, even when they are potentially measured poorly. It is important to point out that the balancing test does not address any other issues which may also haunt a successful empirical investigation of causal effects. One possible issue is measurement error in the variable of interest. This is exacerbated as more potent controls are added to a regression. Griliches (1977) shows that a modest amount of measurement error in schooling may explain patterns of returns in controlled and uncontrolled regressions. Another issue, also discussed by Griliches, is that controls like test scores might themselves be influenced by schooling, which would make them bad controls.

ACKNOWLEDGEMENTS

We thank Suejin Lee for excellent research assistance and the editor Rajeev Dehejia, two referees, an associate editor, Alberto Abadie, Josh Angrist, Panle Jia Barwick, Matias Cattaneo, Bernd Fitzenberger, Brigham Frandsen, Daniel Hungerman, Rachael Meager, Doug Miller, Francesca Molinari, Franco Peracchi, Pedro Souza, and participants at various seminars and conferences for helpful comments. This research has been supported by a grant from the Economic and Social Research Council [ES/H02123X/1] to the Centre for Economic Performance at the LSE.

References

Altonji, Joseph G., Timothy Conley, Todd E. Elder, and Christopher R. Taber. 2013. "Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables." mimeographed.

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

Angrist, Joshua, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton Univeristy Press.

Battistin, Erich, and Andrew Chesher. 2014. "Treatment Effect Estimation with Covariate Measurement Error." *Journal of Econometrics*, 178(2): 707–715.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014*a*. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives*, 28(2): 29–50.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014*b*. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies*, 81(2): 608–650.

Bound, John, Charles Brown, Greg J Duncan, and Willard L Rodgers. 1994. "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data." *Journal of Labor Economics*, 12(3): 345–368.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(2): 200–232.

Card, David. 1995. "Using Geographic Variations in College Proximity to Estimate the Returns to Schooling." In *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp.* , ed. Loizos Nicolaou Christofides,

John Vanderkamp, E. Kenneth Grant and Robert Swidinsky. Toronto: University of Toronto Press.

Case, Anne, and Christina Paxson. 2008. "Stature and Status: Height, Ability, and Labor Market Outcomes." *Journal of Political Economy*, 116(3): 499–532.

Chernozhukov, Victor, Denis Chetverikov, Mert Demire, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review Papers and Proceedings*, 107(5): 261–265.

Chernozhukov, Victor, Denis Chetverikov, Mert Demire, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. forthcoming. "Double Machine Learning for Treatment and Structural Parameters." *Econometrics Journal*.

Chesher, Andrew, and Ian Jewitt. 1987. "The Bias of a Heteroskedasticity Consistent Covariance Matrix Estimator." *Econometrica*, 55(5): 1217–1222.

De Luca, Guiseppe, Jan R. Magnus, and Franco Peracchi. 2017. "Balanced Variable Addition in Linear Models." mimeographed.

Farber, Henry S., and Robert Gibbons. 1996. "Learning and Wage Dynamics." *The Quarterly Journal of Economics*, 111(4): 1007–1047.

Frost, Peter A. 1979. "Proxy Variables and Specification Bias." *Review of Economics and Statistics*, 61(2): 323–325.

Garber, Steven, and Steven Klepper. 1980. "Extending the Classical Normal Errors-in-Variables Model." *Econometrica*, 48(6): 1541–1546.

Gelbach, Jonah B. 2016. "When Do Covariates Matter? And Which Ones, and How Much?" *Journal of Labor Economics*, 34(2): 509–543.

Griliches, Zvi. 1977. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica*, 45(1): 1–22.

Hansen, Bruce E. 2017. "The Exact Distribution of the t-Ratio with Robust and Clustered Standard Errors." mimeographed.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica*, 46(6): 1251–71.

Hausman, Jerry A., and William E. Taylor. 1980. "Comparing Specification Tests and Classical Tests." MIT Dept. of Economics Working Paper no. 266.

Hausman, Jerry A., Jason Abrevaya, and F.M. Scott-Morton. 1998. "Misclassification of the Dependent Variable in a Discrete-response Setting." *Journal of Econometrics*, 87(2): 239–269.

Holly, Alberto. 1982. "A Remark on Hausman's Specification Test." *Econometrica*, 50(3): 749–759.

Kassenboehmer, Sonja C., and Stefanie Schurer. 2017. "Survey Item-response Behavior as a Proxy for Unobserved Ability: Theory and Application." mimeographed.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281–355.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess. 2017. "Machine Learning Tests for Effects on Multiple Outcomes." mimeographed.

MacKinnon, James G. 1992. "Model Specification Tests and Artificial Regressions." *Journal of Economic Literature*, 30(1): 102–146.

MacKinnon, James G., and Halbert White. 1985. "Some Heteroskedasticity-consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics*, 29(3): 305–325.

McCallum, Bennett T. 1972. "Relative Asymptotic Bias from Errors of Omission and Measurement." *Econometrica*, 40(4): 757–758.

Miller, Kenneth S. 1981. "On the Inverse of the Sum of Matrices." *Mathematics Magazine*, 54(2): 67–72.

Oster, Emily. forthcoming. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*.

Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt. 2017. "Poorly Measured Confounders Are More Useful on the Left Than on the Right." NBER Working Papper 23232.

Persico, Nicola, Andrew Postlewaite, and Dan Silverman. 2004. "The Effect of Adolescent Experience on Labor Market Outcomes: The Case of Height." *Journal of Political Economy*, 112(5): 1019–1053.

Pischke, Jörn-Steffen, and Hannes Schwandt. 2012. "A Cautionary Note on Using Industry Affiliation to Predict Income." NBER Working Papper 18384.

Zellner, Arnold. 1963. "Estimators for Seemlingly Unrelated Regression Equations: Some Exact Finite Sample Results." *Journal of the American Statistical Association*, 58(304): 977–992.

Appendix (For Online Publication Only)

A Derivations of Key Results

A.1 Statistical Framework and Population Regression Parameters

In this subsection, we summarize the baseline statistical framework along with the balancing and coefficient comparison tests introduced in Section 2. Under the assumptions we propose as part of the statistical framework, we find the expressions of various population regression parameters. These population parameters will be used in deriving the power functions of the two statistical tests in subsequent subsections.

Let (s_i, x_i, y_i, m_i) for $i = 1, ...n$ be i.i.d. variables. The five relevant regressions are:[1]

$$x_i = \delta s_i + u_i \tag{A1}$$

$$y_i = \beta^s s_i + e_i^s \tag{A2}$$

$$y_i = \beta^l s_i + \gamma x_i + e_i \tag{A3}$$

$$x_i^m = \delta^m s_i + u_i^m \tag{A4}$$

$$y_i = \beta^m s_i + \gamma^m x_i^m + e_i^m \tag{A5}$$

with $x_i^m = x_i + m_i$ being a noisy measure of x_i. In this subsection, we simply think of these equations as projections: the coefficients are population regression parameters, and the residuals u_i, e_i^s, e_i, u_i^m, e_i^m are orthogonal to the respective regressors by construction.

As mentioned in Section 2, a researcher may propose to give β^s a causal interpretation. This is because the researcher has carried out a randomized experiment or applied a particular research design to observational data. In the case of a regression strategy with controls, for example, y_i and s_i are residuals from regressions of the original outcome and treatment variables on the chosen

[1] To simplify exposition, we omit the constant in the regressions throughout the paper with the understanding that s, x, m, and y are normalized to have mean zero.

controls. The researcher is interested in testing her design using additional confounders. If she were to directly observe x_i, the balancing test would entail testing $\delta = 0$ in regression (A1), and the coefficient comparison test would entail testing $\beta^s - \beta^l = 0$ from regressions (A2) and (A3). Unfortunately, she only has access to x_i^m, so the actual balancing test entails testing $\delta^m = 0$ in regression (A4), and the actual coefficient comparison test entails testing $\beta^s - \beta^m = 0$ from regressions (A2) and (A5).

In the baseline framework, we assume that m is classical and hence uncorrelated with s and u, and therefore x:

Assumption 1. $Cov(m_i, s_i) = Cov(m_i, u_i) = 0$.

As mentioned in Section 5.1, we also impose conditional homoskedasticity of u_i and m_i given s_i in the theoretical derivations, and we abstract away from the sampling variation in estimating the standard errors:

Assumption 2. *a) The variances of u_i, m_i and s_i exist, which we denote by σ_u^2, σ_m^2 and σ_s^2 respectively. $Var(u_i|s_i) = \sigma_u^2$ and $Var(m_i|s_i) = \sigma_m^2$; b) σ_u, σ_m and σ_s are known constants.*

Define $\theta \equiv \sigma_m^2 / (\sigma_u^2 + \sigma_m^2)$. We collect expressions for the population regression coefficients δ^m, β^m, and γ^m in terms of the other model parameters, as discussed in Section 3, in Lemma 1.

Lemma 1. *Under Assumptions 1 and 2,*

a) $\delta^m = \delta$ and $u_i^m = u_i + m_i$,

b) $\gamma^m = \gamma(1 - \theta)$,

c) $\beta^m = \beta^l + \gamma\delta\theta$,

d) $\theta = \frac{1-\lambda}{1-R^2}$, where $\lambda = \frac{Var(x_i)}{Var(x_i^m)}$ is the reliability of x_i^m, and R^2 is the population R^2 of the regression of x_i^m on s_i.

Proof. For part a), under the two Assumptions,

$$\delta^m = \frac{Cov(x_i^m, s_i)}{Var(s_i)} = \frac{Cov(x_i, s_i)}{Var(s_i)} = \delta$$

and $u_i^m = u_i + m_i$ directly follows.

For part b), performing an anatomy to the multiple regression (A5), we have

$$\gamma^m = \frac{Cov(y_i, u_i + m_i)}{Var(u_i + m_i)} = \gamma \frac{\sigma_u^2}{\sigma_u^2 + \sigma_m^2}, \tag{A6}$$

where $u_i + m_i$ is the residual from the population regression of x_i^m on s_i under Assumptions 1 and 2. Equation (A6) becomes

$$\gamma^m = \gamma(1 - \theta). \tag{A7}$$

For part c), the omitted variable bias formula implies

$$\begin{aligned} \beta^s &= \beta^l + \gamma\delta \\ \beta^s &= \beta^m + \gamma^m\delta, \end{aligned}$$

and therefore

$$\beta^m = \beta^l + \gamma\delta\theta. \tag{A8}$$

To see why

$$\theta = \frac{1 - \lambda}{1 - R^2}, \tag{A9}$$

in part d) holds, notice that

$$\begin{aligned} Var(x_i) &= \delta^2\sigma_s^2 + \sigma_u^2 \\ Var(x_i^m) &= \delta^2\sigma_s^2 + \sigma_u^2 + \sigma_m^2 \\ R^2 &= 1 - \frac{\sigma_u^2 + \sigma_m^2}{\delta^2\sigma_s^2 + \sigma_u^2 + \sigma_m^2}, \end{aligned}$$

from which equation (A9) mechanically follows. □

A.2 Balancing Test

In this subsection, we derive the power function of the balancing test.

Proposition 1. *Under Assumptions 1 and 2, the large-sample power function of the balancing test at the five-percent level when $\delta = d$ is*[2]

$$1 - \Phi\left(1.96 - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right).$$

[2]It is also possible to specify the value of δ under the alternative hypothesis as a function of n, $\delta = h/\sqrt{n}$. This local alternative framework prevents t_{δ^m} from exploding. The two formulations lead to the same power function.

Proof. Under Assumptions 1 and 2, the asymptotic variance of $\widehat{\delta}^m$ can be directly calculated using part a) of Lemma 1, and the resulting test statistic for the null hypothesis that the balancing coefficient δ is zero is

$$t_{\delta^m} = \frac{\widehat{\delta}^m}{se\left(\widehat{\delta}^m\right)} = \frac{\widehat{\delta}^m}{\frac{1}{\sqrt{n}}\frac{\sqrt{\sigma_u^2+\sigma_m^2}}{\sigma_s}}.$$

Note that

$$\begin{aligned} \theta &= \frac{\sigma_m^2}{\sigma_u^2+\sigma_m^2} \\ &\Rightarrow \quad \sigma_u^2+\sigma_m^2 = \frac{\sigma_u^2}{1-\theta} \end{aligned}$$

Hence

$$t_{\delta^m} = \widehat{\delta}^m \frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}.$$

The rejection probability when $\delta = d$ and when using critical value C is

$$\begin{aligned} \Pr\left(|t_{\delta^m}| > C\right) &= \Pr\left(t_{\delta^m} > C\right) + \Pr\left(t_{\delta^m} < -C\right) \\ &= \Pr\left(\frac{\widehat{\delta}^m}{se\left(\widehat{\delta}^m\right)} > C\right) + \Pr\left(\frac{\widehat{\delta}^m}{se\left(\widehat{\delta}^m\right)} < -C\right) \\ &= \Pr\left(\frac{\widehat{\delta}^m - d}{se\left(\widehat{\delta}^m\right)} > C - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) \\ &\quad + \Pr\left(\frac{\widehat{\delta}^m - d}{se\left(\widehat{\delta}^m\right)} < -C - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) \\ &\approx 1 - \Phi\left(C - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) + \Phi\left(-C - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) \end{aligned}$$

when n is large.[3] Therefore, the large-sample power function of the balancing

[3] Assuming σ_u, σ_m and σ_s to be constants in Assumption 2 conveniently allows us to apply large sample normal approximation. In the case of the balancing test with normal u and m, however, we can also proceed with the estimated $se\left(\widehat{\delta}^m\right)$ based on the homoskedasticity-only estimator. It is a standard result that the t-statistic is then the ratio of a normal random variable and an independent χ^2 random variable, which will allow us to characterize the power function. The characterization becomes much more difficult, however, a) when we employ the heteroskedasticity-robust standard error estimator or b) for the coefficient comparison test. For a) see ongoing work in Hansen (2017), and for b) see Zellner (1963) for a related derivation.

test at the five percent level is

$$Power_{t_{\delta^m}}(d) = 1 - \Phi\left(1.96 - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\sigma_s\sqrt{1-\theta}}{\sigma_u}\right).$$

□

A.3 The Coefficient Comparison Test

A.3.1 Coefficient Comparison Test under Homoskedasticity of e_i^m

For the coefficient comparison test $\beta^s - \beta^m = 0$, the test statistic is

$$t_{(\beta^s-\beta^m)} = \frac{\widehat{\beta}^s - \widehat{\beta}^m}{\sqrt{Var(\widehat{\beta}^s - \widehat{\beta}^m)}},$$

which is asymptotically standard normal. We apply the delta method to the omitted variables bias formula

$$\widehat{\beta}^s - \widehat{\beta}^m = \widehat{\delta}^m\widehat{\gamma}^m$$

to derive the variance. Specifically, we can relate $Var(\widehat{\beta}^s - \widehat{\beta}^m)$ to the asymptotic variances of $\widehat{\delta}^m$ and $\widehat{\gamma}^m$ and their asymptotic covariance:

$$\begin{aligned} Var\left(\widehat{\beta}^s - \widehat{\beta}^m\right) = {} & \gamma^2(1-\theta)^2 Var\left(\widehat{\delta}^m\right) + \delta^2 Var\left(\widehat{\gamma}^m\right) \\ & + 2\delta\gamma(1-\theta)\, Cov\left(\widehat{\delta}^m, \widehat{\gamma}^m\right). \end{aligned} \tag{A10}$$

We have already shown in Proposition 1 that

$$Var(\widehat{\delta}^m) = \frac{1}{n}\frac{\sigma_u^2}{(1-\theta)\sigma_s^2}, \tag{A11}$$

and we derive $Var\left(\widehat{\gamma}^m\right)$ and $Cov\left(\widehat{\delta}^m, \widehat{\gamma}^m\right)$ in the remainder of this subsection.

To simplify the derivation, we make the following three assumptions:

Assumption 3. *s_i, u_i, e_i and m_i are mutually independent.*

Assumption 4. *$E[u_i^3] = 0$.*

Assumption 5. *$Var(e_i^m | s_i, x_i^m)$ is constant.*

Assumption 3 strengthens Assumption 1, and is needed in the proofs of Propositions 2 and 3 below. Assumptions 4 and 5 are imposed to simplify the expressions in Proposition 2, although we relax Assumption 5 in the next subsection and provide the a more general result regarding the power function of the coefficient comparison test.

Note that all of the Assumptions 1-5 are satisfied in the DGP's we adopt for the Monte Carlo simulations underlying Figure 2, that is, when s_i, u_i, e_i, m_i follow a joint normal distribution with a diagonal variance covariance matrix (see subsection D.1 for details). In subsection A.3.2, we also derive the general expression of $Var(\widehat{\beta}^s - \widehat{\beta}^m)$ when Assumption 5 is relaxed.

Proposition 2. *Under Assumptions 2-5, the large-sample power function of the coefficient comparison test at the five-percent level when $\delta = d$ is*

$$1 - \Phi\left(1.96 - d\frac{\sqrt{n}\gamma\left(1-\theta\right)}{\sqrt{V_\beta\left(d;\gamma\right)}}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\gamma\left(1-\theta\right)}{\sqrt{V_\beta\left(d;\gamma\right)}}\right) \tag{A12}$$

where

$$V_\beta\left(d;\gamma\right) \equiv \left(1-\theta\right)\left(\frac{\gamma^2\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \frac{\delta^2\sigma_e^2}{\sigma_u^2}\right).$$

Proof. As mentioned above, we apply the delta method to $\widehat{\beta}^s - \widehat{\beta}^m = \widehat{\delta}^m\widehat{\gamma}^m$. With $Var(\widehat{\delta}^m)$ already derived in Proposition 1, we need the expressions of $Var(\widehat{\gamma}^m)$ and $Cov\left(\widehat{\delta}^m, \widehat{\gamma}^m\right)$. In order to derive $Var(\widehat{\gamma}^m)$, first note that under Assumptions 2, 3, and 5

$$Var\left(\widehat{\gamma}^m\right) = \frac{1}{n}\frac{Var\left(e_i^m\right)}{Var\left(u_i + m_i\right)}, \tag{A13}$$

where $u_i + m_i$ is the residual from the population regression of x_i^m on s_i. Since $Var\left(u_i + m_i\right) = \sigma_u^2 + \sigma_m^2$, the missing piece in equation (A13) is $Var\left(e_i^m\right)$. Plugging the results from Lemma 1 into (9), we get

$$\begin{aligned} y_i &= \beta^m s_i + \gamma^m x_i^m + e_i^m \\ &= \left(\beta^l + \gamma\delta\theta\right) s_i + \gamma\left(1-\theta\right) x_i^m + e_i^m \\ &= \left(\beta^l + \gamma\delta\right) s_i + \gamma\left(1-\theta\right)\left(u_i + m_i\right) + e_i^m \end{aligned}$$

Since

$$\begin{aligned} y_i &= \beta^l s_i + \gamma\left(\delta s_i + u_i\right) + e_i \\ &= \left(\beta^l + \gamma\delta\right) s_i + \gamma u_i + e_i, \end{aligned}$$

matching residuals yields

$$\begin{aligned} \gamma u_i + e_i &= \gamma\left(1-\theta\right)\left(u_i + m_i\right) + e_i^m \\ e_i^m &= \gamma\theta u_i - \gamma\left(1-\theta\right) m_i + e_i \\ Var\left(e_i^m\right) &= \gamma^2\theta^2\sigma_u^2 + \gamma^2\left(1-\theta\right)^2\sigma_m^2 + \sigma_e^2 \\ &= \gamma^2\left(\left(\frac{\sigma_m^2}{\sigma_u^2+\sigma_m^2}\right)^2\sigma_u^2 + \left(\frac{\sigma_u^2}{\sigma_u^2+\sigma_m^2}\right)^2\sigma_m^2\right) + \sigma_e^2 \\ &= \gamma^2\theta\sigma_u^2 + \sigma_e^2. \end{aligned}$$

So

$$\begin{aligned} Var\left(\widehat{\gamma}^m\right) &= \frac{1}{n}\frac{\gamma^2\theta\sigma_u^2 + \sigma_e^2}{\sigma_u^2+\sigma_m^2} \\ &= \frac{1-\theta}{n}\left(\gamma^2\theta + \frac{\sigma_e^2}{\sigma_u^2}\right). \end{aligned} \tag{A14}$$

As for $Cov(\widehat{\delta}^m, \widehat{\gamma}^m)$, first note that

$$\widehat{\delta}^m - \delta = \frac{\sum_i (u_i + m_i) s_i}{\sum_i s_i^2} \tag{A15}$$

$$\widehat{\gamma}^m - \gamma^m = \frac{\sum_i e_i^m \widetilde{x}_i^m}{\sum_i (\widetilde{x}_i^m)^2} \tag{A16}$$

where $\widetilde{x}_i^m = x_i^m - \widehat{\delta}^m s_i$ is the residual from regressing x_i^m on s_i. By Assumption 3 along with the fact that and $\hat{\delta}^m \xrightarrow{p} \delta$, the asymptotic joint distribution of the numerators in equations (A15) and (A16) is

$$\begin{aligned} &\frac{1}{\sqrt{n}}\begin{bmatrix} \sum_i (u_i + m_i) s_i \\ \sum_i e_i^m \widetilde{x}_i^m \end{bmatrix} \\ &\xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \left(\sigma_u^2 + \sigma_m^2\right)\sigma_s^2 & E[s_i(u_i+m_i)^2 e_i^m] \\ E[s_i(u_i+m_i)^2 e_i^m] & E[(u_i+m_i)^2 (e_i^m)^2] \end{bmatrix}\right). \end{aligned}$$

By Assumptions 3 and 4,

$$\begin{aligned} E[s_i(u_i + m_i)^2 e_i^m] &= E[s_i(u_i + m_i)^2(\gamma\theta u_i - \gamma(1-\theta)m_i + e_i)] \\ &= 0. \end{aligned}$$

Since the denominators of equations (A15) and (A16) converge in probability to positive constants,

$$Cov(\widehat{\delta}^m, \widehat{\gamma}^m) = 0. \tag{A17}$$

Plugging equations (A11), (A14) and (A17) into (A10) yields

$$\begin{aligned} Var(\widehat{\beta}^s - \widehat{\beta}^m) &\equiv \frac{1}{n} V_\beta(d;\gamma) \\ &= \frac{1}{n}(1-\theta)\left(\frac{\gamma^2\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \frac{\delta^2\sigma_e^2}{\sigma_u^2}\right). \end{aligned} \tag{A18}$$

Recall from the proof of Lemma 1 that

$$\beta^s - \beta^m = \delta\gamma^m = \delta\gamma(1-\theta),$$

so the power function of the coefficient comparison test is

$$Power_{t_{(\beta^s-\beta^m)}}(d;\gamma) = 1 - \Phi\left(1.96 - d\frac{\sqrt{n}\gamma(1-\theta)}{\sqrt{V_\beta(d;\gamma)}}\right) + \Phi\left(-1.96 - d\frac{\sqrt{n}\gamma(1-\theta)}{\sqrt{V_\beta(d;\gamma)}}\right).$$

□

A.3.2 Relaxing Homoskedasticity of e_i^m

In this subsection, we provide the expression for $Var(\widehat{\beta}^s - \widehat{\beta}^m)$ while relaxing the conditional homoskedasticity of e_i^m, i.e. Assumption 5. Our derivation of this asymptotic variance expression still relies on equation (A10). Since equations (A11) and (A17) are not affected by Assumption 5, we will only need the general expression for $Var(\widehat{\gamma}^m)$.

Proposition 3. *Under Assumptions 2-4, the large-sample power function of the coefficient comparison test at the five-percent level when $\delta = d$ is given by*

(A12), where

$$V_\beta(d;\gamma) = (1-\theta)\left(\frac{\gamma^2\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \frac{\delta^2\sigma_e^2}{\sigma_u^2}\right) + \gamma^2\delta^2\left[\frac{(\kappa_u - 3\sigma_u^4)\theta^2}{(\sigma_m^2+\sigma_u^2)^2} + \frac{(\kappa_m - 3\sigma_m^4)(1-\theta)^2}{(\sigma_m^2+\sigma_u^2)^2}\right]$$

with $\kappa_m \equiv E[m_i^4]$ *and* $\kappa_u \equiv E[u_i^4]$.

Proof. Representing model (9) in matrix form,

$$y_i = \mathbf{W}_i'\mathbf{\Gamma} + e_i^m,$$

where $\mathbf{W}_i = (s_i, x_i^m)'$ and $\mathbf{\Gamma} = (\beta^m, \gamma^m)'$. The asymptotic variance-covariance matrix of the regression estimator $\widehat{\mathbf{\Gamma}}$ is

$$\frac{1}{n}E[\mathbf{W}_i\mathbf{W}_i']^{-1}E[\mathbf{W}_i\mathbf{W}_i'(e_i^m)^2]E[\mathbf{W}_i\mathbf{W}_i']^{-1}.$$

Expressing $E[\mathbf{W}_i\mathbf{W}_i']$ in terms of the fundamental model parameters is straightforward:

$$\begin{aligned} E[\mathbf{W}_i\mathbf{W}_i'] &= E\begin{bmatrix} s_i^2 & s_i x_i^m \\ s_i x_i^m & (x_i^m)^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_s^2 & \delta\sigma_s^2 \\ \delta\sigma_s^2 & \delta^2\sigma_s^2 + \sigma_u^2 + \sigma_m^2 \end{bmatrix}. \end{aligned}$$

Writing out the entries in the matrix $E[\mathbf{W}_i\mathbf{W}_i'(e_i^m)^2]$:

$$\begin{aligned} & E[\mathbf{W}_i\mathbf{W}_i'(e_i^m)^2] \\ = \; & E\begin{bmatrix} \underbrace{s_i^2(e_i^m)^2}_{\text{(i)}} & \underbrace{s_i x_i^m(e_i^m)^2}_{\text{(ii)}} \\ s_i x_i^m(e_i^m)^2 & \underbrace{(x_i^m)^2(e_i^m)^2}_{\text{(iii)}} \end{bmatrix}. \end{aligned}$$

Below we express quantities (i) to (iii) in terms of the model parameters. Utilizing Assumptions 2-4, we have the expressions for (i) to (ii):

$$\begin{aligned} E[s_i^2(e_i^m)^2] &= E[s_i^2(\gamma\theta u_i - \gamma(1-\theta)m_i + e_i)^2] \\ &= \sigma_s^2(\gamma^2\theta^2\sigma_u^2 + \gamma^2(1-\theta)^2\sigma_m^2 + \sigma_e^2), \end{aligned} \tag{i}$$

$$\begin{aligned}
E[s_i x_i^m (e_i^m)^2] &= E[s_i(\delta_0 + \delta s_i + u_i + m_i) \cdot (e_i^m)^2] \\
&= \delta_0 E[s_i (e_i^m)^2] + \delta E[s_i^2 (e_i^m)^2] \\
&+ E[s_i u_i (\gamma\theta u_i - \gamma(1-\theta) m_i + e_i)^2] \\
&+ E[s_i m_i (\gamma\theta u_i - \gamma(1-\theta) m_i + e_i)^2] \\
&= \delta \sigma_s^2 (\gamma^2 \theta^2 \sigma_u^2 + \gamma^2 (1-\theta)^2 \sigma_m^2 + \sigma_e^2), \qquad \text{(ii)}
\end{aligned}$$

For the expression of (iii)

$$\begin{aligned}
E[(x_i^m)^2 (e_i^m)^2] &= E[(\delta_0 + \delta s_i + u_i + m_i)^2 (e_i^m)^2] \\
&= \delta_0^2 E[(e_i^m)^2] + \delta^2 E[s_i^2 (e_i^m)^2] \\
&\quad + E[u_i^2 (\gamma\theta u_i - \gamma(1-\theta) m_i + e_i)^2] \\
&\quad + E[m_i^2 (\gamma\theta u_i - \gamma(1-\theta) m_i + e_i)^2] \\
&\quad + 2\delta_0 \delta E[s_i (e_i^m)^2] + 2\delta_0 E[u_i (e_i^m)^2] \\
&\quad + 2\delta_0 E[m_i (e_i^m)^2] + 2\delta E[s_i u_i (e_i^m)^2] \\
&\quad + 2\delta E[s_i m_i (e_i^m)^2] + 2E[u_i m_i (e_i^m)^2].
\end{aligned}$$

Note that

$$\begin{aligned}
E[s_i (e_i^m)^2] &= 0 \\
E[u_i (e_i^m)^2] &= E[m_i (e_i^m)^2] = 0 \\
E[s_i u_i (e_i^m)^2] &= E[s_i m_i (e_i^m)^2] = 0,
\end{aligned}$$

and we only need to find the expressions for

$$\begin{aligned}
& E[u_i^2 (\gamma\theta u_i - \gamma(1-\theta) m_i + e_i)^2] \\
= \; & E[u_i^2 \{\gamma^2 \theta^2 u_i^2 + \gamma^2 (1-\theta)^2 m_i^2 + e_i^2 \\
& - 2\gamma^2 \theta (1-\theta) u_i m_i + 2\gamma\theta u_i e_i - 2\gamma(1-\theta) m_i e_i\}] \\
= \; & \gamma^2 \theta^2 E[u_i^4] + \gamma^2 (1-\theta)^2 \sigma_u^2 \sigma_m^2 + \sigma_u^2 \sigma_e^2 \\
= \; & \gamma^2 \theta^2 \kappa_u + \gamma^2 (1-\theta)^2 \sigma_u^2 \sigma_m^2 + \sigma_u^2 \sigma_e^2,
\end{aligned}$$

$$\begin{aligned}
& E[m_i^2 (\gamma\theta u_i - \gamma(1-\theta) m_i + e_i)^2] \\
= \; & E[m_i^2 \{\gamma^2 \theta^2 u_i^2 + \gamma^2 (1-\theta)^2 m_i^2 + e_i^2 \\
& - 2\gamma^2 \theta (1-\theta) u_i m_i + 2\gamma\theta u_i e_i - 2\gamma(1-\theta) m_i e_i\}] \\
= \; & \gamma^2 \theta^2 \sigma_u^2 \sigma_m^2 + \gamma^2 (1-\theta)^2 \kappa_m + \sigma_m^2 \sigma_e^2,
\end{aligned}$$

and

$$\begin{aligned} E[u_i m_i (e_i^m)^2] &= E[u_i m_i (\gamma\theta u_i - \gamma(1-\theta)m_i + e_i)^2] \\ &= E[u_i m_i \{\gamma^2\theta^2 u_i^2 + \gamma^2(1-\theta)^2 m_i^2 + e_i^2 \\ &\quad -2\gamma^2\theta(1-\theta)u_i m_i + 2\gamma\theta u_i e_i - 2\gamma(1-\theta)m_i e_i\}] \\ &= -2\gamma^2\theta(1-\theta)\sigma_u^2\sigma_m^2. \end{aligned}$$

Putting these terms together,

$$\begin{aligned} E[(x_i^m)^2(e_i^m)^2] &= \delta_0^2 E[(e_i^m)^2] + \delta^2 E[s_i^2(e_i^m)^2] \\ &\quad + E[u_i^2(\gamma\theta u_i - \gamma(1-\theta)m_i + e_i)^2] \\ &\quad + E[m_i^2(\gamma\theta u_i - \gamma(1-\theta)m_i + e_i)^2] \\ &\quad + 2E[u_i m_i (e_i^m)^2] \\ &= \delta_0^2\{\gamma^2\theta^2\sigma_u^2 + \gamma^2(1-\theta)^2\sigma_m^2 + \sigma_e^2\} \\ &\quad + \delta^2\sigma_s^2(\gamma^2\theta^2\sigma_u^2 + \gamma^2(1-\theta)^2\sigma_m^2 + \sigma_e^2) \\ &\quad + \{\gamma^2\theta^2\kappa_u + \gamma^2(1-\theta)^2\sigma_u^2\sigma_m^2 + \sigma_u^2\sigma_e^2\} \\ &\quad + \{\gamma^2\theta^2\sigma_u^2\sigma_m^2 + \gamma^2(1-\theta)^2\kappa_m + \sigma_m^2\sigma_e^2\} \\ &\quad - \{4\gamma^2\theta(1-\theta)\sigma_u^2\sigma_m^2\}. \end{aligned} \tag{iii}$$

Now that we have the expression for both $E[\mathbf{W}_i\mathbf{W}_i']$ and $E[\mathbf{W}_i\mathbf{W}_i'(e_i^m)^2]$, we can compute the asymptotic variance of $\widehat{\gamma}^m$

$$\begin{aligned} Var\left(\widehat{\gamma}^m\right) = \frac{1}{n}\Bigg\{ &(1-\theta)\left(\gamma^2\theta + \frac{\sigma_e^2}{\sigma_u^2}\right) \\ &+ \underbrace{\gamma^2\left[\frac{(\kappa_u - 3\sigma_u^4)\theta^2}{(\sigma_m^2+\sigma_u^2)^2} + \frac{(\kappa_m - 3\sigma_m^4)(1-\theta)^2}{(\sigma_m^2+\sigma_u^2)^2}\right]}_{(a)} \Bigg\}. \end{aligned}$$

Compared to its expression under homoskedasticity (A14), we have an extra term (a) that accounts for the excess kurtosis of the u and m distributions. It

follows that

$$\begin{aligned} V_\beta(d;\gamma) &= n \cdot Var\left(\widehat{\beta}^s - \widehat{\beta}^m\right) \\ &= (1-\theta)\left(\frac{\gamma^2\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \frac{\delta^2\sigma_e^2}{\sigma_u^2}\right) \\ &\quad + \gamma^2\delta^2\left[\frac{(\kappa_u - 3\sigma_u^4)\theta^2}{(\sigma_m^2+\sigma_u^2)^2} + \frac{(\kappa_m - 3\sigma_m^4)(1-\theta)^2}{(\sigma_m^2+\sigma_u^2)^2}\right]. \end{aligned}$$

Note that when u_i and m_i are normal, $\kappa_u - 3\sigma_u^4 = 0$ and $\kappa_m - 3\sigma_m^4 = 0$, and the variance expression above simplifies to that of equation (A18). Since $Var\left(\widehat{\beta}^s - \widehat{\beta}^m\right)$ increases in κ_u and κ_m while the balancing test is unaffected by the heteroskedasticity of e^m, the power advantage of the balancing test is larger when u_i and m_i have thicker tails than a normal distribution. □

B Comparison with Oster (forthcoming)

The Oster (forthcoming) formulation of the causal regression takes the form

$$y_i = \beta s_i + \gamma w_{1i} + w_{2i} + e_i,$$

where w_{1i} is an observed covariate and w_{2i} is an unobserved covariate, uncorrelated with w_{1i}. To map this into our setup, think of the true x_i as capturing both w_{1i} and w_{2i}, i.e. $x_i = w_{1i} + \frac{1}{\gamma}w_{2i}$. Furthermore, there is equal selection, i.e.

$$\frac{Cov(s_i, \gamma w_{1i})}{\gamma^2\sigma_1^2} = \frac{Cov(s_i, w_{2i})}{\sigma_2^2},$$

where σ_1^2 and σ_2^2 are the variances of w_{1i} and w_{2i}, respectively. Then, Oster's (forthcoming) regression can be written as

$$y_i = \beta s_i + \gamma x_i + e_i,$$

which is our regression (2).

Our observed $x_i^r = w_{1i}$, so measurement error is $m_i = -\frac{w_{2i}}{\gamma}$. Measurement error here is mean reverting (hence we use the r superscript on x to distinguish it from the classical measurement error case), i.e.

$$m_i = \kappa x_i + \mu_i \tag{A19}$$

with $\kappa < 0$. Notice that

$$Cov\left(m_i, x_i\right) = -\frac{\sigma_2^2}{\gamma^2},$$

and hence

$$\kappa = \frac{-\sigma_2^2/\gamma^2}{\sigma_1^2 + \sigma_2^2/\gamma^2} \tag{A20}$$

and

$$\begin{aligned}
\mu_i &= -\frac{w_{2i}}{\gamma} - \kappa\left(w_{1i} + \frac{w_{2i}}{\gamma}\right) \\
&= -\kappa w_{1i} - \frac{(1+\kappa)}{\gamma} w_{2i} \\
&= \frac{\sigma_2^2/\gamma^2}{\sigma_1^2 + \sigma_2^2/\gamma^2} w_{1i} - \frac{\sigma_1^2/\gamma}{\sigma_1^2 + \sigma_2^2/\gamma^2} w_{2i}.
\end{aligned}$$

It turns out that μ_i implicitly defined in (A19) and κ given by (A20) imply $Cov(x_i, \mu_i) = 0$ and $Cov(s_i, \mu_i) = 0$. Hence, these two equations represent mean reverting measurement error as defined in the body of the manuscript.[4] In the subsections that follow, we will provide expressions of our two tests as well as those in Oster (forthcoming) using the seven model parameters, β, γ, σ_s^2, σ_e^2, σ_1^2, σ_2^2, and $\sigma_{1s} \equiv Cov(w_{1i}, s_i)$.[5]

B.1 The Balancing and Coefficient Comparison Tests with Mean Reverting Measurement Error

In this subsection, we denote the relevant quantities with the r superscript to signify that we are working with the mean reverting measurement error. The balancing regression equation is

$$x_i^r = \delta^r s_i + u_i^r,$$

[4]Note that $Cov(s_i, \mu_i) = 0$ depends on the equal selection assumption. With proportional selection, i.e.

$$\phi \frac{Cov(s_i, \gamma w_{1i})}{\gamma^2 \sigma_1^2} = \frac{Cov(s_i, w_{2i})}{\sigma_2^2},$$

and $\phi \neq 1$ we would have $Cov(s_i, \mu_i) \neq 0$.

[5]Note that $\sigma_{2s} \equiv Cov(w_{2i}, s_i)$ is a function of σ_1^2, σ_2^2, σ_{1s}, and γ following the equal selection condition.

with the population parameter δ^r being

$$\delta^r = \frac{Cov(x_i^r, s_i)}{Var(s_i)} = \frac{Cov(w_{1i}, s_i)}{\sigma_s^2} = \frac{\sigma_{1s}}{\sigma_s^2}.$$

The asymptotic standard deviation of $\widehat{\delta^r}$ is

$$se(\widehat{\delta^r}) = \frac{1}{\sqrt{n}} \frac{\sqrt{Var(u_i^r)}}{\sigma_s}.$$

Since

$$\begin{aligned} Var(u_i^r) &= Var(x_i^r) - (\delta^r)^2 \sigma_s^2 \\ &= \sigma_1^2 - \frac{\sigma_{1s}^2}{\sigma_s^2}, \end{aligned}$$

$$\begin{aligned} se(\widehat{\delta^r}) &= \frac{1}{\sqrt{n}} \frac{\sqrt{Var(u_i^r)}}{\sigma_s} \\ &= \frac{1}{\sqrt{n}} \frac{\sqrt{\sigma_1^2 \sigma_s^2 - \sigma_{1s}^2}}{\sigma_s^2}. \end{aligned}$$

For the coefficient comparison test, we run the two regressions

$$\begin{aligned} y_i &= \beta^s s_i + e_i^s \\ y_i &= \beta^r s_i + \gamma^r w_{1i} + e_i^r. \end{aligned}$$

The omitted variable bias formula gives us

$$\begin{aligned} \beta^s &= \beta + \frac{\gamma \sigma_{1s}}{\sigma_s^2} + \frac{\sigma_{2s}}{\sigma_s^2} \\ &= \beta + \gamma \delta^r + \frac{\sigma_{2s}}{\sigma_s^2} \end{aligned}$$

and

$$\begin{aligned} \begin{pmatrix} \beta^r \\ \gamma^r \end{pmatrix} &= \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \frac{1}{\sigma_1^2 \sigma_s^2 - \sigma_{1s}^2} \begin{pmatrix} \sigma_1^2 & -\sigma_{1s} \\ -\sigma_{1s} & \sigma_s^2 \end{pmatrix} \begin{pmatrix} \sigma_{2s} \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \frac{\sigma_1^2 \sigma_{2s}}{\sigma_1^2 \sigma_s^2 - \sigma_{1s}^2} \\ -\frac{\sigma_{1s} \sigma_{2s}}{\sigma_1^2 \sigma_s^2 - \sigma_{1s}^2} \end{pmatrix}. \end{aligned}$$

Therefore,

$$\begin{aligned}\beta^r - \beta^s &= \frac{\sigma_1^2\sigma_{2s}}{\sigma_1^2\sigma_s^2 - \sigma_{1s}^2} - \gamma\frac{\sigma_{1s}}{\sigma_s^2} - \frac{\sigma_{2s}}{\sigma_s^2} \\ &= \frac{\sigma_1^2\sigma_{2s}}{\sigma_1^2\sigma_s^2 - \sigma_{1s}^2} - \frac{\gamma\sigma_{1s} + \sigma_{2s}}{\sigma_s^2}.\end{aligned}$$

For $Var(\widehat{\beta}^s - \widehat{\beta}^r)$ needed for the coefficient comparison test, we again rely on the omitted variable formula

$$\widehat{\beta}^s - \widehat{\beta}^r = \widehat{\delta}^r\widehat{\gamma}^r$$

and the delta method

$$\begin{aligned}Var(\widehat{\beta}^s - \widehat{\beta}^r) = (\gamma^r)^2 Var(\widehat{\delta}^r) + (\delta^r)^2 Var(\widehat{\gamma}^r) \\ + 2\delta^r\gamma^r Cov(\widehat{\delta}^r, \widehat{\gamma}^r).\end{aligned}$$

We already derived the expression for $Var(\widehat{\delta}^r)$ above and will now derive $Var(\widehat{\gamma}^r)$ and $Cov(\widehat{\delta}^r, \widehat{\gamma}^r)$. For simplicity, we are going to assume homoskedasticity of e_i^r conditional on s_i and w_{1i}, which is true, for example, when s_i, w_{1i}, w_{2i} and e_i are joint normal. It follows that

$$Var(\widehat{\gamma}^r) = \frac{1}{n}\frac{Var(e_i^r)}{Var(u_i^r)},$$

where

$$Var(e_i^r) = Var(y_i) - \begin{pmatrix} \beta + \frac{\sigma_1^2\sigma_{2s}}{\sigma_1^2\sigma_s^2 - \sigma_{1s}^2} \\ \gamma - \frac{\sigma_{1s}\sigma_{2s}}{\sigma_1^2\sigma_s^2 - \sigma_{1s}^2} \end{pmatrix}' \begin{pmatrix} \sigma_s^2 & \sigma_{1s} \\ \sigma_{1s} & \sigma_1^2 \end{pmatrix} \begin{pmatrix} \beta + \frac{\sigma_1^2\sigma_{2s}}{\sigma_1^2\sigma_s^2 - \sigma_{1s}^2} \\ \gamma - \frac{\sigma_{1s}\sigma_{2s}}{\sigma_1^2\sigma_s^2 - \sigma_{1s}^2} \end{pmatrix}$$

with

$$Var(y_i) = \beta^2\sigma_s^2 + \gamma^2\sigma_1^2 + \sigma_2^2 + 2\beta\gamma\sigma_{1s} + 2\beta\sigma_{2s} + \sigma_e^2.$$

For $Cov(\widehat{\delta}^r, \widehat{\gamma}^r)$, we can follow the same reasoning as in subsection A.3.1 and show that it is equal to zero. Plugging in the expressions for γ^r, δ^r, $Var(\widehat{\delta}^r)$, and $Var(\widehat{\gamma}^r)$, we obtain

$$\begin{aligned}&Var(\widehat{\beta}^s - \widehat{\beta}^r) \\ =&\frac{1}{n}\frac{\gamma^2\sigma_1^4\sigma_{1s}^2\sigma_s^2(\sigma_{1s}^2 - \sigma_1^2\sigma_s^2)\left(3\gamma^2\sigma_1^2 + \sigma_2^2 - \sigma_e^2\right) + \gamma^4\sigma_1^{10}\sigma_s^6 + 2\gamma^2\sigma_1^4\sigma_2^2\sigma_s^2\sigma_{1s}^4 - \sigma_{1s}^6\left(\gamma^2\sigma_1^2 + \sigma_2^2\right)^2}{\gamma^2\sigma_1^4\sigma_s^4\left(\sigma_{1s}^2 - \sigma_1^2\sigma_s^2\right)^2}.\end{aligned} \tag{A21}$$

In the simple case where the mean reversion coefficient $\kappa = 0$, we have $\sigma_2 = 0$ from (A20), which implies that $x_i = w_{1i}$ and that $w_{2i} = 0$ (almost surely). This is the same as in the classical measurement case with $\theta = 0$. The reader can easily check that (A21) with $\sigma_2 = 0$ is the same as (A18) with $\theta = 0$.

Unlike in the classical measurement error case, it is not always true that $Var(\widehat{\beta}^s - \widehat{\beta}^r) > Var(\widehat{\delta}^r)$. For example, when σ_e is small, the coefficient comparison test may dominate – this is the result of $\widehat{\delta}^r$ being attenuated. But as σ_e increases, the balancing test regains its advantage. As we will see below, this dependence on σ_e of the coefficient comparison test contrasts with the asymptotic bias in Oster (forthcoming).

B.2 Relevant Quantities in Oster (forthcoming)

We now express the quantities in Oster's Proposition 1, which we restate using population parameters

$$\beta = \tilde{\beta} - [\mathring{\beta} - \tilde{\beta}]\frac{R_{max} - \tilde{R}}{\tilde{R} - \mathring{R}}.$$

Her $\tilde{\beta}$ and $\mathring{\beta}$ correspond to our β^s and β^r respectively, and the various R^2's are:

$$\mathring{R} = 1 - \frac{Var(e_i^s)}{Var(y_i)}$$

$$\tilde{R} = 1 - \frac{Var(e_i^r)}{Var(y_i)}$$

$$R_{max} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

Therefore,

$$\frac{R_{max} - \tilde{R}}{\tilde{R} - \mathring{R}} = \frac{Var(e_i^r) - Var(e_i)}{Var(e_i^s) - Var(e_i^r)}.$$

We have derived above the expression for $Var(e_i^r)$, and analogously the expression for$Var(e_i^s)$ is

$$\begin{aligned} Var(e_i^s) &= Var(y_i) - \beta_s^2\sigma_s^2 \\ &= \beta^2\sigma_s^2 + \gamma^2\sigma_1^2 + \sigma_2^2 + 2\beta\gamma\sigma_{1s} + 2\beta\sigma_{2s} + \sigma_e^2 \\ &\quad - (\beta + \gamma\frac{\gamma\sigma_{1s}}{\sigma_s^2} + \frac{\sigma_{2s}}{\sigma_s^2})^2\sigma_s^2. \end{aligned}$$

It follows that

$$\frac{R^2_{max} - R^2_r}{R^2_r - R^2_s} = \frac{\sigma_1^2\sigma_2^2\sigma_s^2}{\gamma^2\sigma_1^4\sigma_s^2 - \gamma^2\sigma_{1s}^2\sigma_1^2 - \sigma_{1s}^2\sigma_2^2},$$

and that the bias

$$\begin{aligned}\Pi &\equiv [\mathring{\beta} - \tilde{\beta}]\frac{R_{max} - \tilde{R}}{\tilde{R} - \mathring{R}} \\ &= \left[\frac{\sigma_1^2\sigma_{2s}}{\sigma_1^2\sigma_s^2 - \sigma_{1s}^2} - \gamma\delta^r - \frac{\sigma_{2s}}{\sigma_s^2}\right]\frac{\sigma_1^2\sigma_2^2\sigma_s^2}{\gamma^2\sigma_1^4\sigma_s^2 - \gamma^2\sigma_{1s}^2\sigma_1^2 - \sigma_{1s}^2\sigma_2^2}\end{aligned}$$

Note that Π does not depend on σ_e, which is not surprising given that Oster's Proposition 1 focuses on identification. In comparison, our coefficient comparison test depends on σ_e, because the variation in e is important for inference. Given this difference in foci, there is no one-to-one correspondence between Oster's Π and our $t_{(\widehat{\beta}^s - \widehat{\beta}^r)}$: for a given value of Π in Oster, there are different values of $t_{(\widehat{\beta}^s - \widehat{\beta}^r)}$ for the coefficient comparison test, and vice versa.

C Comparison of the LHS and RHS Balancing Tests

We compare the LHS and RHS balancing tests introduced in Section 5.3. The F-statistic of the LHS balancing test is

$$F_{LHS} = \frac{1}{k}\hat{\boldsymbol{\delta}}'\widehat{Var}(\hat{\boldsymbol{\delta}})^{-1}\hat{\boldsymbol{\delta}},$$

where we use a consistent variance estimator $\widehat{Var}(\hat{\boldsymbol{\delta}})$ for $Var(\hat{\boldsymbol{\delta}})$. On the other hand, the F-statistic for the RHS balancing test following the regression

$$s_i = \boldsymbol{\pi}'\boldsymbol{x}_i + v_i$$

is

$$F_{RHS} = \frac{1}{k}\hat{\boldsymbol{\pi}}'\widehat{Var}(\hat{\boldsymbol{\pi}})^{-1}\hat{\boldsymbol{\pi}} \tag{A22}$$

with $\widehat{Var}(\hat{\boldsymbol{\pi}})$ being a consistent estimator for $Var(\hat{\boldsymbol{\pi}})$.

We introduce the multivariate analog of Assumption 2 regarding $\boldsymbol{u}_i$ and s_i in equation (16) and the existence of the relevant moments:

Assumption 6. *The variances of* $\boldsymbol{u}_i$ *and* s_i *exist, as does* $E[(\boldsymbol{x}_i\boldsymbol{x}_i')(s_i-\boldsymbol{\pi}'\boldsymbol{x}_i)^2]$. $Var(\boldsymbol{u}_i|s_i) = E[\boldsymbol{u}_i\boldsymbol{u}_i']$.

Lemma 2. *Under Assumption 6,*

a) $\frac{k}{n}F_{LHS} \xrightarrow{p} \sigma_s^2\boldsymbol{\delta}'(E[\boldsymbol{u}_i\boldsymbol{u}_i'])^{-1}\boldsymbol{\delta}$.

b) $\frac{k}{n}F_{RHS} \xrightarrow{p} \sigma_s^4\boldsymbol{\delta}'E[(\boldsymbol{x}_i\boldsymbol{x}_i')(s_i-\boldsymbol{\pi}'\boldsymbol{x}_i)^2]^{-1}\boldsymbol{\delta}$

Proof. For part a), Assumption 6 implies that

$$n\widehat{Var}(\hat{\boldsymbol{\delta}}) \xrightarrow{p} \frac{1}{\sigma_s^4}E[s_i^2\boldsymbol{u}_i\boldsymbol{u}_i'] = \frac{1}{\sigma_s^2}E[\boldsymbol{u}_i\boldsymbol{u}_i'].$$

Hence,

$$\frac{k}{n}F_{LHS} \xrightarrow{p} \sigma_s^2\boldsymbol{\delta}'(E[\boldsymbol{u}_i\boldsymbol{u}_i'])^{-1}\boldsymbol{\delta}. \tag{A23}$$

For b),

$$\boldsymbol{\pi} = {\boldsymbol{\Omega}_{\boldsymbol{x}}}^{-1}\boldsymbol{\varsigma} \tag{A24}$$

where $\boldsymbol{\Omega}_{\boldsymbol{x}} = Var(\boldsymbol{x}_i)$ and $\boldsymbol{\varsigma} = Cov(\boldsymbol{x}_i, s_i)$. The probability limit of the variance estimator is

$$n\widehat{Var}(\hat{\boldsymbol{\pi}}) \xrightarrow{p} {\boldsymbol{\Omega}_{\boldsymbol{x}}}^{-1}E[(\boldsymbol{x}_i\boldsymbol{x}_i')(s_i-\boldsymbol{\pi}'\boldsymbol{x}_i)^2]{\boldsymbol{\Omega}_{\boldsymbol{x}}}^{-1}. \tag{A25}$$

Plugging (A24) and (A25) into (A22), the probability of the scaled F-stat of the RHS balancing test is

$$\begin{aligned}\frac{k}{n}F_{RHS} &\xrightarrow{p} \boldsymbol{\varsigma}'E[(\boldsymbol{x}_i\boldsymbol{x}_i')(s_i-\boldsymbol{\pi}'\boldsymbol{x}_i)^2]^{-1}\boldsymbol{\varsigma} \\ &= \sigma_s^4\boldsymbol{\delta}'E[(\boldsymbol{x}_i\boldsymbol{x}_i')(s_i-\boldsymbol{\pi}'\boldsymbol{x}_i)^2]^{-1}\boldsymbol{\delta}\end{aligned} \tag{A26}$$

□

The probability limits (A23) and (A26) are in general different. An analytical comparison between the two is complicated, as it depends on the higher moments of s and $\boldsymbol{u}$. However, we show below that the two scaled F-statistics have the same probability limits, in the special case where the LHS balancing regression has a spherical error structure and the RHS balancing regression is homoskedastic.

C.1 Special Case: Spherical LHS Error Structure and Homoskedastic RHS Regression

We consider the special case where the RHS regression is homoskedastic and the LHS balancing regression has a spherical error structure:

Assumption 7. *$Var(v_i|\boldsymbol{x}_i)$ is constant and $Var(\boldsymbol{u}_i) = \sigma_u^2 \boldsymbol{I}_k$.*

Note that Assumption 7 is satisfied if s and $\boldsymbol{u}$ are both normally distributed.

Proposition 4. *Under Assumptions 6 and 7, $plim \frac{k}{n} F_{LHS} = plim \frac{k}{n} F_{RHS}$.*

Proof. Combining Assumption 7 with (A23) in the proof of Lemma 1, the LHS F-statistic simplifies to

$$\frac{k}{n} F_{LHS} \xrightarrow{p} \frac{\sigma_s^2 \boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma_u^2}$$

For the RHS F-statistic, homoskedasticity in Assumption 7 allows us to write

$$E[(\boldsymbol{x}_i \boldsymbol{x}_i')(s_i - \boldsymbol{\pi}'\boldsymbol{x}_i)^2] = E[\boldsymbol{x}_i \boldsymbol{x}_i'] E[(s_i - \boldsymbol{\pi}'\boldsymbol{x}_i)^2]$$

To find the expression of $E[\boldsymbol{x}_i \boldsymbol{x}_i'] E[(s_i - \boldsymbol{\pi}'\boldsymbol{x}_i)^2]$, first note that

$$\sigma_s^2 = Var(\boldsymbol{\pi}'\boldsymbol{x}_i) + E[(s_i - \boldsymbol{\pi}'\boldsymbol{x}_i)^2]$$

so

$$E[(s_i - \boldsymbol{\pi}'\boldsymbol{x}_i)^2] = \sigma_s^2 - Var(\boldsymbol{\pi}'\boldsymbol{x}_i)$$

with

$$\begin{aligned} Var(\boldsymbol{\pi}'\boldsymbol{x}_i) &= \boldsymbol{\pi}' \boldsymbol{\Omega}_{\boldsymbol{x}} \boldsymbol{\pi} \\ &= \boldsymbol{\varsigma}' {\boldsymbol{\Omega}_{\boldsymbol{x}}}^{-1} \boldsymbol{\varsigma} \\ &= \sigma_s^4 \boldsymbol{\delta}' {\boldsymbol{\Omega}_{\boldsymbol{x}}}^{-1} \boldsymbol{\delta}. \end{aligned} \tag{A27}$$

Since $\text{rank}(\boldsymbol{\delta}\boldsymbol{\delta}') = 1$ and $\text{tr}[(\sigma_s^2 \boldsymbol{\delta}\boldsymbol{\delta}')(\frac{1}{\sigma_u^2}\boldsymbol{I}_k)^{-1}] = \frac{\sigma_s^2}{\sigma_u^2}\boldsymbol{\delta}'\boldsymbol{\delta}$, by Miller (1981) we have

$$\begin{aligned} {\boldsymbol{\Omega}_{\boldsymbol{x}}}^{-1} &= \frac{1}{\sigma_u^2}\boldsymbol{I} - \frac{1}{1 + \frac{\sigma_s^2}{\sigma_u^2}\boldsymbol{\delta}'\boldsymbol{\delta}} \frac{1}{(\sigma_u^2)^2} \sigma_s^2 \boldsymbol{\delta}\boldsymbol{\delta}' \\ &= \frac{1}{\sigma_u^2}\boldsymbol{I} - \frac{\sigma_s^2}{(\sigma_u^2)^2 + \sigma_u^2 \sigma_s^2 \boldsymbol{\delta}'\boldsymbol{\delta}} \boldsymbol{\delta}\boldsymbol{\delta}'. \end{aligned} \tag{A28}$$

Plugging (A28) into (A27):

$$\begin{aligned}
Var(\boldsymbol{\pi}'\boldsymbol{x}_i) &= \frac{\sigma_s^4\boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma_u^2} - \frac{\sigma_s^6(\boldsymbol{\delta}'\boldsymbol{\delta})^2}{(\sigma_u^2)^2+\sigma_u^2\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}} \\
&= \frac{\sigma_s^4\boldsymbol{\delta}'\boldsymbol{\delta}[(\sigma_u^2)^2+\sigma_u^2\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}] - \sigma_s^6(\boldsymbol{\delta}'\boldsymbol{\delta})^2\sigma_u^2}{(\sigma_u^2)^2[\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}]} \\
&= \frac{\sigma_s^4\boldsymbol{\delta}'\boldsymbol{\delta}(\sigma_u^2)^2}{(\sigma_u^2)^2[\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}]} \\
&= \frac{\sigma_s^4\boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}}
\end{aligned}$$

It follows that

$$\begin{aligned}
E[(s_i-\boldsymbol{\pi}'\boldsymbol{x}_i)^2] &= \sigma_s^2 - Var(\boldsymbol{\pi}'\boldsymbol{x}_i) \\
&= \sigma_s^2 - \frac{\sigma_s^4\boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}} \\
&= \frac{\sigma_s^2[\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}] - \sigma_s^4\boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}} \\
&= \frac{\sigma_s^2\sigma_u^2}{\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}}
\end{aligned}$$

As a result, the probability limit of $\frac{k}{n}F_{RHS}$ is

$$\begin{aligned}
&\sigma_s^4\boldsymbol{\delta}'E[(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}E[(s_i-\boldsymbol{\pi}'\boldsymbol{x}_i)^2]^{-1}\boldsymbol{\delta} \\
=&\sigma_s^4\boldsymbol{\delta}'\boldsymbol{\Omega}_{\boldsymbol{x}}^{-1}\frac{\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma_s^2\sigma_u^2}\boldsymbol{\delta} \\
=&\sigma_s^4\boldsymbol{\delta}'\left(\frac{1}{\sigma_u^2}\boldsymbol{I} - \frac{\sigma_s^2}{(\sigma_u^2)^2+\sigma_u^2\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}}\boldsymbol{\delta}\boldsymbol{\delta}'\right)\frac{\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma_s^2\sigma_u^2}\boldsymbol{\delta} \\
=&\sigma_s^4\left(\frac{\boldsymbol{\delta}'\boldsymbol{\delta}(\sigma_u^2+\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta})}{\sigma_s^2\sigma_u^4} - \frac{(\boldsymbol{\delta}'\boldsymbol{\delta})^2\sigma_s^2}{\sigma_s^2\sigma_u^4}\right) \\
=&\sigma_s^4\left(\frac{\boldsymbol{\delta}'\boldsymbol{\delta}\sigma_u^2}{\sigma_s^2\sigma_u^4}\right) \\
=&\frac{\sigma_s^2\boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma_u^2}.
\end{aligned}$$

Therefore,

$$\text{plim}\frac{k}{n}F_{LHS} = \text{plim}\frac{k}{n}F_{RHS}.$$

□

D Simulations: Summary of Data Generating Processes and Additional Results

D.1 Summary of Data Generating Processes

In this subsection, we present a succinct summary of the data generating processes (DGP's) used in the Monte Carlo simulations.

- Baseline DGP (Figure 2):

$$\begin{pmatrix} s_i \\ u_i \\ e_i \\ m_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 30 & 0 \\ 0 & 0 & 0 & \sigma_m^2 \end{bmatrix} \right)$$

$$y_i = s_i + 3x_i + e_i$$

$$x_i = d \cdot s_i + u_i$$

$$x_i^m = x_i + m_i \tag{A29}$$

  when $\theta = 0$, $\sigma_m^2 = 0$; when $\theta = 0.85$, $\sigma_m^2 = 17$.

- Mean reverting measurement error DGP (Figure 3): Replace equation (A29) in the baseline DGP with

$$x_i^m = -0.5\left(d \cdot s_i + u_i\right) + \mu_i$$

  where $\sigma_\mu^2 = 0.75$ or $\sigma_\mu^2 = 2.25$.

- Multiple controls DGP (Figure 4): in the baseline DGP,
  - Replace u_i with $\boldsymbol{u}_i \sim N(\mathbf{0}, \sigma_u^2 \boldsymbol{I})$ and m_i with $\boldsymbol{m}_i \sim N(\mathbf{0}, \sigma_m^2 \boldsymbol{I})$;
  - Replace $x_i = d \cdot s_i + u_i$ with $\boldsymbol{x}_i = \boldsymbol{d} \cdot s_i + \boldsymbol{u}_i$, where
    - $\boldsymbol{d} = (d, d, ..., d)'$ in panels (a) and (c)
    - $\boldsymbol{d} = (d, 0, ..., 0)'$ in panels (b) and (d);
  - Replace $x_i^m = x_i + m_i$ with $\boldsymbol{x}_i^m = \boldsymbol{x}_i + \boldsymbol{m}_i$.

- Heteroskedastic DGP (Figure D.2):

– Everything is the same as in the baseline DGP, except generate u_i and e_i conditional on s_i:

$$\begin{pmatrix} u_i \\ e_i \end{pmatrix} | s_i \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \left(\frac{e^{|s_i|}}{1+e^{|s_i|}}\right)^2 \sigma_{0u}^2 & 0 \\ 0 & \left(\frac{e^{|s_i|}}{1+e^{|s_i|}}\right)^2 \sigma_{0e}^2 \end{bmatrix} \right)$$

with σ_{0u}^2 and σ_{0e}^2 set so that the unconditional variances $\overline{\sigma}_u^2 = 3$ and $\overline{\sigma}_e^2 = 30$ match those in the baseline DGP.

- Binary x DGP (Figure D.2): in the baseline DGP,

  – Replace $\sigma_s^2 = 1$ with $\sigma_s^2 = 0.25$;

  – Replace $x_i = d \cdot s_i + u_i$ with

$$\Pr(x_i = 1) = \Phi(d \cdot s_i);$$

  – Replace $x_i^m = x_i + m_i$ with

$$\Pr(x_i^m = 1 | x_i = 0) = \Pr(x_i^m = 0 | x_i = 1) = \tau$$

  with $\tau = 0$ or $\tau = 0.1$.

D.2 Additional Results

We explore additional scenarios using simulations in this section beyond those included in the main text. Figure D.2 shows the original theoretical power functions for the case with no measurement error from Figure 1. It adds empirical rejection rates from simulations with heteroskedastic errors u_i and e_i of the form

$$\begin{aligned} \sigma_{u,i}^2 &= \left(\frac{e^{|s_i|}}{1+e^{|s_i|}}\right)^2 \sigma_{0u}^2 \\ \sigma_{e,i}^2 &= \left(\frac{e^{|s_i|}}{1+e^{|s_i|}}\right)^2 \sigma_{0e}^2. \end{aligned}$$

We set the baseline variances σ_{0u}^2 and σ_{0e}^2 so that the unconditional variances $\overline{\sigma}_u^2 = 3$ and $\overline{\sigma}_e^2 = 30$ match the variances in Figure 1. The test statistics used in

the simulations employ robust standard errors. We plot the rejection rates for data with no measurement error and for the more severe measurement error scenario given by $\theta = 0.85$. As can be seen in Figure D.2, both the balancing and the coefficient comparison tests lose some power with heteroskedastic residuals and a robust covariance matrix compared to the conventional, homoskedastic baseline (thin lines). Otherwise, the main findings look very similar to those in Figure 1. Heteroskedasticity does not seem to alter the basic conclusions appreciatively.

We presented results in the main text on mean reverting measurement error when x and m are continuous. Another prominent case of mean reverting measurement error is the one where x_i is a dummy variable. In this case, the balancing equation is a binary choice model, and hence inherently non-linear. While we assume that the researcher continues to estimate (3) as a linear probability model, we generate x_i as follows:

$$\Pr\left(x_i = 1\right) = \Phi\left(\delta s_i\right), \tag{A30}$$

where $\Phi\left(\bullet\right)$ is the normal distribution function as before. Measurement error takes the form of misclassification, and we assume the misclassification rate to be symmetric:

$$\Pr\left(x_i^m = 1 | x_i = 0\right) = \Pr\left(x_i^m = 0 | x_i = 1\right) = \tau.$$

Compared to the baseline parameters in Table 1, we set $\sigma_s^2 = 0.25$, and $\tau = 0.1$ in our simulations. The model remains the same in all other respects. We use robust standard errors in estimating (9) and (11).

Various issues arise from the nonlinear nature of (A30). One is the fact that plim $\left(\hat{\delta}\right)$ from estimating (11) linearly is not going to equal the δ we used in the probit equation (A30) to generate x. The relationship between plim $\left(\hat{\delta}\right)$ and δ is concave. In Figure D.2, we plot rejection rates against values of δ, although the quantity plim $\left(\hat{\delta}\right)$ is probably more comparable to what we put on the x-axis in the previous figures that summarize the simulation results from linear models. We note that results look qualitatively very similar when

we plot rejection rates against the empirical averages of $\hat{\delta}$ from estimating (11) as a linear probability model.

Another issue is that measurement error in x_i will now lead to a biased estimate of δ in estimating (11). This is true even if we were to use a probit and estimated a model like (A30). The bias takes the form of attenuation, just as in the case of a binary regressor with measurement error (see Hausman, Abrevaya and Scott-Morton, 1998). This is a corollary of our result that mean reverting measurement error also reduces the power of the balancing test. Of course, we know from the relationship (14) between the test statistics that the coefficient comparison test will also suffer from the same power loss.

The thin lines in Figure D.2 reveal a sizable power advantage for the balancing test even without any misclassification. This result is in stark contrast to the linear models we have analyzed, where a large power loss for the coefficient comparison test only resulted once we introduced measurement error. In fact, it is possible to think of the binary nature of x_i itself as a form of mismeasurement. Equation (A30) defines $\Pr(x_i = 1)$ as a latent index, but the outcome regression (2) uses a coarse version of this variable in the form of the binary x_i.

In our parameterization, the coefficient comparison test never reaches a rejection rate of 1, and the power function levels off at a far lower level. As d increases, the power of the balancing test goes to 1. In the linear model, the rejection rate of t_γ is independent of d. Because of the nonlinear nature of (A30) this is no longer true here, and the average value of t_γ across repeated samples actually falls for higher values of d. Drawing on (14), the power of the coefficient comparison test will equal the power of t_γ when $t_\delta \to \infty$. This is not a specific feature of the binary case but is generally true for the relationship between the three test statistics. However, in the binary case this implies that the power of the coefficient comparison test may decline with d.[6]

[6]The reason for the decline of t_γ with d in our parameterization is as follows: the standard error of $\hat{\gamma}$ depends on the residual variance of the long regression, which is independent of d, and on the variance of the residual from regressing x_i on s_i (because s_i is partialled out in the long regression). When $d = 0$, this latter residual is just equal to x_i itself, which is binary. But s_i is continuous, so as d increases, partialling out s_i transforms the binary x_i

Adding measurement error to the binary regressor x_i makes things worse as is visible from the thick lines in Figure D.2. The power loss of the balancing test is comparatively minor for the relatively low misclassification rate of $\tau = 0.1$ we are using. Much of the loss for the balancing test results from the binary nature of the x_i variable in the first place. The coefficient comparison test is affected by misclassification error to a much higher degree because t_γ is affected, the Hausman, Abrevaya and Scott-Morton (1998) result notwithstanding.

into a continuous variable, which has less variance than in the $d = 0$ case. As the effective variance in this regressor falls, the standard error of $\hat{\gamma}$ goes up and t_γ goes down.

References

Hansen, Bruce E. 2017. "The Exact Distribution of the t-Ratio with Robust and Clustered Standard Errors." mimeographed.

Hausman, Jerry A., Jason Abrevaya, and F.M. Scott-Morton. 1998. "Misclassification of the Dependent Variable in a Discrete-response Setting." *Journal of Econometrics*, 87(2): 239–269.

Miller, Kenneth S. 1981. "On the Inverse of the Sum of Matrices." *Mathematics Magazine*, 54(2): 67–72.

Oster, Emily. forthcoming. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*.

Zellner, Arnold. 1963. "Estimators for Seemlingly Unrelated Regression Equations: Some Exact Finite Sample Results." *Journal of the American Statistical Association*, 58(304): 977–992.

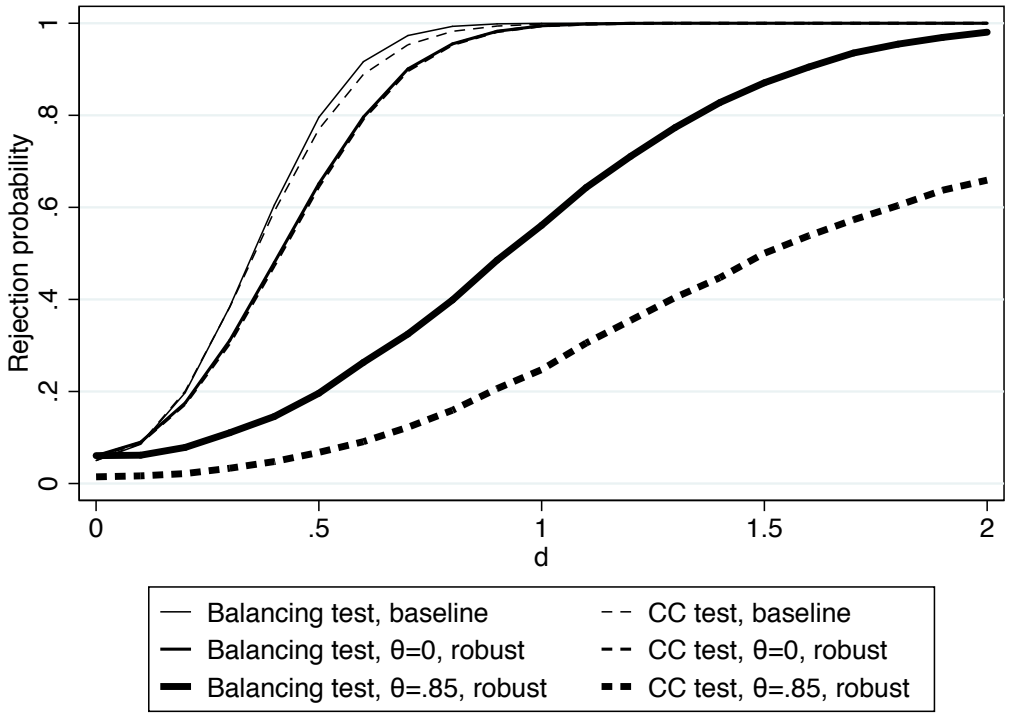

Figure A.1: Simulated Rejection Rates with Heteroskedasticity. Comparison of baseline rejection rates (from Figure 1) with simulated rejection rates based on heteroskedastic errors and robust standard errors. d is the value the coefficient in the balancing equation takes on under the alternative hypothesis.

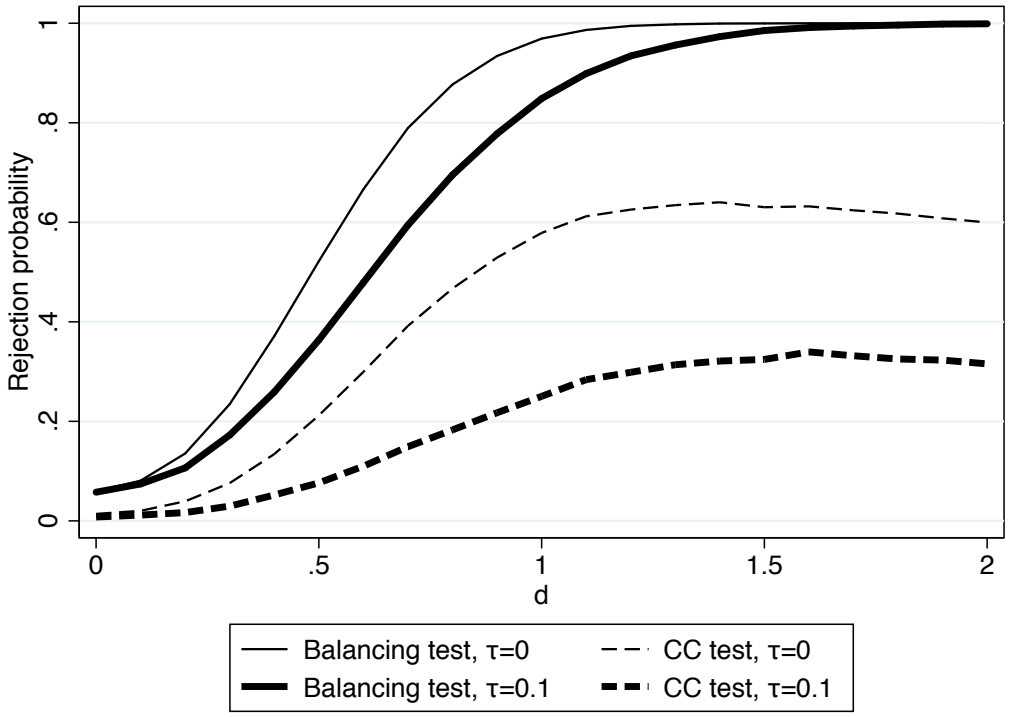

Figure A.2: Simulated Rejection Rates with Binary Control and Misclassification. Rejection rates for a binary control variable that is misclassified (i.e. its binary value is flipped) with probability τ. d is the value the coefficient in the balancing equation takes on under the alternative hypothesis.