

Pei, Zhuan; Pischke, Jörn-Steffen; Schwandt, Hannes

Working Paper

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

IZA Discussion Papers, No. 10647

Provided in Cooperation with:

IZA – Institute of Labor Economics

Suggested Citation: Pei, Zhuan; Pischke, Jörn-Steffen; Schwandt, Hannes (2017) : Poorly Measured Confounders Are More Useful on the Left Than on the Right, IZA Discussion Papers, No. 10647, Institute of Labor Economics (IZA), Bonn

This Version is available at:

<https://hdl.handle.net/10419/161270>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

DISCUSSION PAPER SERIES

IZA DP No. 10647

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

Zhuan Pei
Jörn-Steffen Pischke
Hannes Schwandt

MARCH 2017

DISCUSSION PAPER SERIES

IZA DP No. 10647

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

Zhuan Pei

Cornell University and IZA

Jörn-Steffen Pischke

London School of Economics and IZA

Hannes Schwandt

University of Zürich and IZA

MARCH 2017

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ABSTRACT

Poorly Measured Confounders Are More Useful on the Left Than on the Right¹

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 various strategies which have been suggested to identify the returns to schooling.

JEL Classification: C31, C52

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

Corresponding author:

Jörn-Steffen Pischke
Centre for Economic Performance
London School of Economics
Houghton Street
London WC2A 2AE
United Kingdom
E-mail: s.pischke@lse.ac.uk

¹ This paper builds on ideas from and supersedes the paper “A Cautionary Note on Using Industry Affiliation to Predict Income,” by Pischke and Schwandt (NBER WP18384, 2012). We thank Suejin Lee for excellent research assistance and Alberto Abadie, Josh Angrist, Matias Cattaneo, Bernd Fitzenberger, Brigham Frandsen, Daniel Hungerman, Francesca Molinari, Pedro Souza, and participants at various seminars and conferences for helpful comments.

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, possibly after suitable conditioning. The credibility of a design can be enhanced when researchers can show explicitly that potentially remaining sources of selection bias have been eliminated. This is often done through some form of balancing tests or robustness checks.

The research designs mentioned above can all be thought of as variants of regression strategies. If the researcher has access to a variable for a potentially remaining confounder, tests for 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 causal 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 mea-

surement 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.

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. If the researcher has a strong prior that the added regressor ought to matter for the outcome under study, then the balancing test will provide the remaining information necessary to assess the research design. This maintained assumption is the ultimate source of the superior power of the balancing test. However, we show that quantitatively meaningful differences emerge particularly when there is some substantial amount of measurement error in the added regressor. We derive the relevant parameters in the presence of measurement error in Section 3.

Of course, sometimes researchers may be more agnostic about whether the added regressor matters for the outcome. In case it does not matter, rejecting balance for this variable is of no consequence for this particular research design. In this view, only the coefficient comparison test is really relevant while the balancing test provides no additional information. However, this strikes us as a narrow view and not one shared by many in the experimental community, where balancing tests are commonly used. Lack of balance is seen as an indictment of the randomization in an experiment irrespective of whether the variable in question affects the outcome. Lack of balance with respect to one or more observed covariates raises the possibility that there may also be lack of balance for other unobservables, and would lead a prudent researcher to reassess the credibility of their research design. The same should be true for quasi-experimental research based on observational data.

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 re-

gressions 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.² We provide the relevant test statistics and discuss how they behave under measurement error in Section 4. We also show how the coefficient comparison test is simple to implement for varying identification strategies. We demonstrate the superior power of the balancing test under a variety of scenarios in Section 5.

The principles underlying the points we are making are not new but the consequences do not seem to be fully appreciated in much applied work. Griliches (1977) is a classic reference for the issues arising when regression controls are measured with error. A subsequent literature, for example Rosenbaum and Rubin (1983) and Imbens (2003), has considered omitted variables bias in non-linear models without measurement error. More closely related is Battistin and Chesher (2014), as it discusses identification in the presence of a mismeasured covariate in non-linear models. Like in the literature following Rosenbaum and Rubin (1983), they discuss identification given assumptions about a missing parameter, namely the degree of measurement error in the covariate. We follow Griliches (1977) in framing our discussion around the omitted variables bias arising in linear regressions, the general framework used most widely in empirical studies. Unlike this literature, we are less interested in point identification in the presence of missing information. We go beyond the analysis in all of these papers in our explicit discussion of testing, which forms the core of our study.

Altonji, Elder and Taber (2005) discuss an alternative but closely related approach to the problem. As we noted above, applied researchers often argue that relative stability of regression coefficients when adding additional controls provides evidence for credible identification. Implicit in this argument is the idea that other confounders not controlled for are similar to the controls just added to the regression. The paper by Altonji, Elder and Taber (2005) formalizes this argument. In practice, adding controls will typically move the coefficient of interest somewhat even if it is not by much. Altonji et al. (2013)

²An exception is Gelbach (2016), who discusses the Hausman test in this context.

and Oster (forthcoming) extend the original Altonji, Elder and Taber work by providing more precise conditions for bounds and point identification in this case. The approach in these papers relies on an assumption about how the omitted variables bias due to the observed regressor is related to any remaining omitted variables bias due to unobserved confounders.

The remaining unobserved confounders in this previous work can be thought of as the source of measurement error in the covariate which is added to the regression in our analysis. For example, in our empirical example below, we use mother’s education as a measure for family background but this variable may only capture a small part of all the relevant family background information, a lot of which may be orthogonal to mother’s education. In fact, we show that our formulation and Oster’s (forthcoming) are isomorphic. This means that our framework is a useful starting point for researchers who are willing to make the type of assumptions in Altonji, Elder and Taber (2005) and follow-up papers as well.

Another related strand of work is by Belloni, Chernozhukov and Hansen (2014a, b), who tackle the opposite problem from Altonji, Elder and Taber (2005), namely choosing the best controls when the researcher has a potentially bigger set of candidate controls available than is necessary. This large dimensional set may come from nonlinearities and interactions among regressors. Belloni, Chernozhukov and Hansen (2014b) use Lasso to select regressors which are highly correlated with either the treatment or the outcome conditional on other covariates. They then estimate an outcome equation including as controls all the regressors selected in this preliminary step. In a sense, this is more closely related to our setup than the Altonji, Elder and Taber approach as Belloni, Chernozhukov and Hansen (2014b) also postulate that identification can be achieved when using a subset of the available covariates as controls. Their variable selection problem is related to the two testing strategies we discuss in this paper. However, like Altonji et al. (2013) and Oster (forthcoming), their ultimate interest is in point identification and inference for the treatment effects parameter, not in testing whether a particular specification is subject to remaining confounders. Their setup is also not specifically geared towards

dealing with control variables which are subject to error, which is our focus.

An older literature by Hausman (1978), Hausman and Taylor (1980), and Holly (1982) (see also the summary in MacKinnon, 1992, section II.9) 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 different 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 illustrate our power results in Section 6. In addition to Griliches (1977), this data set has been used in a well known study by Card (1995). 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 investigated by Griliches, 1977), a candidate instrument for college attendance (investigated by Card, 1995), as well as a myriad of other useful variables on individual and family background. The empirical results support and illustrate our theoretical claims.

2 A Simple Framework

Consider the following simple framework starting with a population regression equation

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

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. 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.³

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 causal effect of s_i on y_i would only be obtained conditional on x_i , as in the long regression

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

and the researcher would like to probe whether this is a concern. For example, in the returns to schooling context, x_i might be some 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. In principle, the second strategy has the advantage that it goes beyond testing whether (1)

³Of course, all subsequent regression equations and results also inherit the structure of the actual underlying research design.

qualifies as a causal regression. An appreciable change in β suggests that the original estimate β^s is biased. The results obtained with x_i as an additional control should be closer to the causal effect we seek to uncover. In particular, if x_i were the only relevant confounder and if we measure it without error, the β parameter from the controlled regression is 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 and β^s again only indicates a flawed research design.

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

$$x_i = \delta_0 + \delta s_i + u_i. \quad (3)$$

The change in the coefficient β from adding x_i to the regression (1) is given by the omitted variables bias formula

$$\beta^s - \beta = \gamma \delta. \quad (4)$$

The change in the coefficient of interest β from adding x_i 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

$$H_0 : \delta = 0, \quad (5)$$

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

$$H_0 : \beta^s - \beta = 0, \quad (6)$$

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

$$H_0 : \beta^s - \beta = 0 \Leftrightarrow \gamma = 0 \text{ or } \delta = 0. \quad (7)$$

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.⁴ 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.⁵

Here we make the point that the balancing test adds valuable information particularly when the true confounder is measured with error. In general, 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 $\beta^s - \beta$ 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 measurement error; this regression simply loses precision because the mismeasured variable is on the left hand side. Under the maintained assumption that $0 < \gamma < \infty$, the balancing test is more powerful than the coefficient comparison test. In order to make these statements precise, we collect results for the relevant population parameters for the case of classical measurement error in the following section, before moving on to the test statistics.

⁴One might argue that researchers should only carry out the long regression and not the short regression if they know that $\gamma \neq 0$: if $\delta \neq 0$, not including x in the regression will lead to omitted variable bias; if $\delta = 0$, both $\hat{\beta}^s$ and $\hat{\beta}$ are consistent but $\hat{\beta}^s$ is less efficient than $\hat{\beta}$. As we emphasized in the Introduction, however, the focus of this paper is on testing whether the treatment is plausibly randomly assigned in an (quasi-)experimental design. In the analysis of a randomized controlled trial, for example, researchers may include covariates when estimating the treatment effect but that does not come before a formal test of covariate balance.

⁵Equations (4) and (7) highlight that a regressor ought to be included in the long regression when both $\gamma \neq 0$ and $\delta \neq 0$. This differs from the selection rule chosen by Belloni, Chernozhukov and Hansen (2014b), who include a regressor when either $\gamma \neq 0$ or $\delta \neq 0$ is true.

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

$$x_i^m = x_i + m_i. \quad (8)$$

Here we assume the measurement error m_i is classical, i.e. $E(m_i) = 0$, $Cov(x_i, m_i) = 0$. In Section 5 below we also investigate the impact of non-classical errors. As a result of the measurement error, the researcher compares the regressions

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

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

$$\begin{aligned} \beta^m &= \beta + \gamma\delta \frac{1 - \lambda}{1 - R^2} = \beta + \gamma\delta\theta \\ \gamma^m &= \gamma \frac{\lambda - R^2}{1 - R^2} = \gamma(1 - \theta) \end{aligned} \quad (10)$$

where R^2 is the population R^2 of the regression of s_i on x_i^m and

$$\lambda = \frac{Var(x_i)}{Var(x_i^m)}$$

is the reliability of x_i^m .⁶ λ measures the amount of measurement error present as the fraction of the variance in the observed x_i^m , which is due to the signal in the true x_i . It is also the attenuation factor in a simple bivariate regression on x_i^m . In the multivariate model (9), an alternative way to parameterize the amount of measurement error is

$$\theta = \frac{1 - \lambda}{1 - R^2} = \frac{\sigma_m^2}{\sigma_u^2 + \sigma_m^2}.$$

⁶Note R^2 is also the population R^2 of the regression of x_i^m on s_i .

where σ_{\cdot}^2 denotes the variance of the random variable in the subscript. $1 - \theta$ is the multivariate attenuation factor. Recall that u_i is the residual from the balancing regression (3).

With the mismeasured x_i^m , the balancing regression becomes

$$x_i^m = \delta_0^m + \delta^m s_i + u_i + m_i, \quad (11)$$

which implies that

$$\lambda = 1 - \frac{\sigma_m^2}{\text{Var}(x_i^m)} > 1 - \frac{\sigma_u^2 + \sigma_m^2}{\text{Var}(x_i^m)} = R^2.$$

As a result

$$\begin{aligned} 0 &< \frac{1 - \lambda}{1 - R^2} < 1 \\ 0 &< \frac{\lambda - R^2}{1 - R^2} < \lambda. \end{aligned}$$

θ is an alternative way to parameterize the degree of measurement error in x_i compared to λ and R^2 . The θ parameterization uses 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. θ turns out to be a useful parameter in many of the derivations that follow.

The population coefficient β^m differs from β but less so than β^s . In fact, β^m lies between β^s and β , 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 ($R^2 > 0$).

These results highlight a number of issues. The gap $\beta^s - \beta^m$ is too small compared to the desired $\beta^s - \beta$, 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. Finally, the balancing regression (11) with the mismeasured x_i^m involves measurement error in the dependent variable, which has no effect on the population parameter $\delta^m = \delta$, but the estimator $\hat{\delta}^m$ is less efficient than $\hat{\delta}$.

The results here are also useful for thinking about the identification of β and γ in the presence of measurement error. Rearranging (10) yields

$$\begin{aligned}\gamma &= \gamma^m \frac{1 - R^2}{\lambda - R^2} \\ \beta &= \beta^m - \delta \gamma^m \frac{1 - \lambda}{\lambda - R^2}.\end{aligned}\tag{12}$$

Since R^2 can be estimated from the data, these expressions only involve the unknown parameter λ . If we are willing to make an assumption about the measurement error, we are able to point identify β . Even if λ is not known precisely, (12) can be used to bound β for a range of plausible reliabilities. Alternatively, (10) can be used to derive the value of λ for which $\beta = 0$. These calculations are similar in spirit to the ones suggested by Oster (forthcoming) in a setting that is closely related.

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.⁷ We present the theoretical power functions for the two alternative test statistics; derivations are in Appendix A, which also shows that our results carry over to robust standard errors. We extend the power results to the heteroskedastic case and non-classical measurement error in simulations. Our basic conclusions are the same in all these different scenarios.

Start with the standard error of estimator $\hat{\delta}^m$ from the balancing regression:

$$\sqrt{n} \widehat{se}(\hat{\delta}^m) \xrightarrow{p} \sqrt{\frac{\sigma_u^2 + \sigma_m^2}{\sigma_s^2}} = \frac{1}{\sqrt{1 - \theta}} \frac{\sigma_u}{\sigma_s},$$

where we use $\widehat{se}(\bullet)$ to denote the estimated standard error of a given estimator. Let $se(\bullet)$ denote the *asymptotic* standard error of an estimator, i.e., $se(\bullet) \equiv$

⁷See Appendix A for the precise setup of the model. 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).

$\frac{1}{\sqrt{n}}\text{plim}\{\sqrt{n}\widehat{se}(\bullet)\}$. In the case of $\widehat{\delta}^m$,

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

Comparing the asymptotic standard error of $\widehat{\delta}^m$ to its counterpart in the case with no measurement error,

$$se\left(\widehat{\delta}\right) = \frac{1}{\sqrt{n}} \frac{\sigma_u}{\sigma_s},$$

we have

$$se\left(\widehat{\delta}^m\right) = \frac{se\left(\widehat{\delta}\right)}{\sqrt{1-\theta}}.$$

Since $0 < \theta < 1$, the standard error is inflated compared to the case with no measurement error.

A test based on the t -statistic

$$t_{\delta^m} = \frac{\widehat{\delta}^m}{\widehat{se}\left(\widehat{\delta}^m\right)}$$

remains consistent because m_i is correctly accounted for in the residual of the balancing regression (11), but the t -statistic is asymptotically smaller than in the error free case: As $n \rightarrow \infty$, the comparison of the scaled t -statistics when $\delta > 0$ is (without loss of generality, we are assuming that δ is either zero or positive)

$$\frac{1}{\sqrt{n}}t_{\delta^m} \xrightarrow{p} \sqrt{1-\theta} \frac{\delta}{\left(\frac{\sigma_u}{\sigma_s}\right)} < \frac{\delta}{\left(\frac{\sigma_u}{\sigma_s}\right)} \xleftarrow{p} \frac{1}{\sqrt{n}}t_{\delta}$$

This means the null hypothesis (5) is rejected less often. The test is less powerful than in the error free case; 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 $\beta^s - \beta = \gamma\delta$ in conjunction with the result from the balancing regression. Let \tilde{x}_i^m be the residual from the population regression of x_i^m on s_i . 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.2.3. The standard error for $\hat{\gamma}^m$ in the limit is

$$\begin{aligned} se(\hat{\gamma}^m) &= \frac{1}{\sqrt{n}} \frac{\sqrt{Var(e_i^m)}}{\sqrt{Var(\tilde{x}_i^m)}} \\ &= \frac{1}{\sqrt{n}} \sqrt{\frac{\gamma^2 \theta \sigma_u^2 + \sigma_e^2}{\sigma_u^2 + \sigma_m^2}} \\ &= \frac{1}{\sqrt{n}} \sqrt{1 - \theta} \sqrt{\theta \gamma^2 + \frac{\sigma_e^2}{\sigma_u^2}}, \end{aligned}$$

while

$$se(\hat{\gamma}) = \frac{1}{\sqrt{n}} \sqrt{\frac{\sigma_e^2}{\sigma_u^2}}.$$

$se(\hat{\gamma}^m)$ involves two terms: the first term is an attenuated version of $se(\hat{\gamma})$ from the corresponding regression with the correctly measured x_i , while the second term depends on the value of γ . The parameters in the two terms are not directly related, so $se(\hat{\gamma}^m) \geq se(\hat{\gamma})$. 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 $Var(e_i^m)$ 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 $\gamma^m = 0$ is

$$t_{\gamma^m} = \frac{\hat{\gamma}^m}{se(\hat{\gamma}^m)}$$

and it follows that when $\gamma > 0$

$$\frac{1}{\sqrt{n}} t_{\gamma^m} \xrightarrow{p} \sqrt{1 - \theta} \frac{\gamma}{\sqrt{\theta \gamma^2 + \frac{\sigma_e^2}{\sigma_u^2}}} < \frac{\gamma}{\sqrt{\frac{\sigma_e^2}{\sigma_u^2}}} \xleftarrow{p} \frac{1}{\sqrt{n}} t_{\gamma}.$$

As in the case of $\widehat{\delta}^m$ from 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, namely due to the term $\theta\gamma^2$ in the denominator, in addition to the attenuation factor $\sqrt{1-\theta}$. This is due to the fact that measurement error in a regressor both attenuates the relevant coefficient towards zero and introduces additional variance into the residual. Though interestingly, $\theta\gamma^2$ captures the additional residual variance while the factor $\sqrt{1-\theta}$ now captures the attenuation of γ^m . In the balancing test statistic, $\sqrt{1-\theta}$ accounted for the residual variance. The upshot from this discussion is that 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 is the source of the greater power of the balancing test statistic.

Finally, consider the quantity $\beta^s - \beta^m$, which enters the coefficient comparison test. To form a test statistic for this quantity we need the expression for the asymptotic variance of $\widehat{\beta}^s - \widehat{\beta}^m$, which we derive through an application of the delta method to the omitted variables bias formula

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

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:

$$\begin{aligned} Var(\widehat{\beta}^s - \widehat{\beta}^m) &= \gamma^2 (1 - \theta)^2 Var(\widehat{\delta}^m) + \delta^2 Var(\widehat{\gamma}^m) \\ &\quad + 2\delta\gamma (1 - \theta) Cov(\widehat{\delta}^m, \widehat{\gamma}^m). \end{aligned} \tag{13}$$

Using $Var(\widehat{\delta}^m)$ and $Var(\widehat{\gamma}^m)$, which we derived above, and the fact that $Cov(\widehat{\delta}^m, \widehat{\gamma}^m) = 0$, which we show in Appendix A.2.2, we get

$$Var(\widehat{\beta}^s - \widehat{\beta}^m) = \frac{1}{n} (1 - \theta) \left(\gamma^2 \frac{\sigma_u^2}{\sigma_s^2} + \theta \delta^2 \gamma^2 + \delta^2 \frac{\sigma_e^2}{\sigma_u^2} \right).$$

It is easy to see that, like $Var(\widehat{\gamma}^m)$, $Var(\widehat{\beta}^s - \widehat{\beta}^m)$ has both an attenuation factor as well as an additional positive term compared to the case where $\theta =$

0, i.e. $Var(\hat{\beta}^s - \hat{\beta})$. Measurement error may therefore raise or lower the sampling variance for the coefficient comparison test.

Before we proceed to discuss the power of the coefficient comparison test, we note that the covariance term in

$$Var(\hat{\beta}^s - \hat{\beta}^m) = Var(\hat{\beta}^s) + Var(\hat{\beta}^m) - 2Cov(\hat{\beta}^s, \hat{\beta}^m)$$

reduces the sampling variance of $\hat{\beta}^s - \hat{\beta}^m$. In fact, this covariance term is positive, and it is generally sizable compared to $Var(\hat{\beta}^s)$ and $Var(\hat{\beta}^m)$ since the regression residuals e_i^s and e_i^m are highly correlated. Because $2Cov(\hat{\beta}^s, \hat{\beta}^m)$ gets subtracted, looking at the standard errors of $\hat{\beta}^s$ and $\hat{\beta}^m$ 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. Define

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

where $\widehat{se}(\hat{\beta}^s - \hat{\beta}^m)$ is a consistent standard error estimator. Since

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

we have

$$\begin{aligned} \frac{1}{\sqrt{n}}t_{(\beta^s - \beta^m)} &\xrightarrow{p} \frac{\delta\gamma(1 - \theta)}{\sqrt{(1 - \theta)\left(\gamma^2\frac{\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \delta^2\frac{\sigma_e^2}{\sigma_u^2}\right)}} \\ &= \sqrt{1 - \theta} \frac{\delta\gamma}{\sqrt{\gamma^2\frac{\sigma_u^2}{\sigma_s^2} + \theta\delta^2\gamma^2 + \delta^2\frac{\sigma_e^2}{\sigma_u^2}}}. \end{aligned} \quad (14)$$

Under the alternative hypothesis ($\delta \neq 0$) and the maintained assumption $\gamma \neq 0$, the limits for the other two test statistics can be written as

$$\begin{aligned} \frac{1}{\sqrt{n}}t_{\delta^m} &\xrightarrow{p} \sqrt{1 - \theta} \frac{\delta\gamma}{\sqrt{\gamma^2\frac{\sigma_u^2}{\sigma_s^2}}} \\ \frac{1}{\sqrt{n}}t_{\gamma^m} &\xrightarrow{p} \sqrt{1 - \theta} \frac{\delta\gamma}{\sqrt{\theta\delta^2\gamma^2 + \delta^2\frac{\sigma_e^2}{\sigma_u^2}}}. \end{aligned}$$

Hence, using (14), it is apparent that under these conditions the three tests are asymptotically related in the following way:

$$\text{plim} \left(\frac{1}{\frac{1}{\sqrt{n}}t_{(\beta^s - \beta^m)}} \right)^2 = \text{plim} \left(\frac{1}{\frac{1}{\sqrt{n}}t_{\delta^m}} \right)^2 + \text{plim} \left(\frac{1}{\frac{1}{\sqrt{n}}t_{\gamma^m}} \right)^2 \quad (15)$$

These results highlight 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 $t_{\gamma^m} \rightarrow \infty$, 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 $\hat{\gamma}^m$. 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 . Before we turn to these results, we briefly note how the coefficient comparison test can be implemented in practice.

4.1 Implementing the Coefficient Comparison Test

The balancing test is a straightforward t -test, which regression software calculates routinely. We noted that the coefficient comparison test is a generalized Hausman test. Regression software will typically calculate this as well if it allows for seemingly unrelated regression estimation (SURE). SURE takes $\text{Cov}(e_i^s, e_i^m)$ into account and therefore facilitates the test. In Stata, this is

implemented via the `suest` command. Generically, the test would take the following form:

```
reg y s
est store reg1
reg y s x
est store reg2
suest reg1 reg2
test[reg1_mean]s=[reg2_mean]s
```

The test easily accommodates covariates or can be carried out with the variables `y`, `s`, and `x` being residuals from a previous regression (hence facilitating large numbers of fixed effects though degrees of freedom may have to be adjusted in this case).

As far as we can tell, the Stata `suest` or `3reg` commands don't work for the type of IV regressions we might be interested in here. An alternative, which also works for IV, is to take the regressions (1) and (2) and stack them:

$$\begin{bmatrix} y_i \\ y_i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha^s \\ \alpha \end{bmatrix} + \begin{bmatrix} s_i & 0 \\ 0 & s_i \end{bmatrix} \begin{bmatrix} \beta^s \\ \beta \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & x_i \end{bmatrix} \begin{bmatrix} 0 \\ \gamma \end{bmatrix} + \begin{bmatrix} e_i^s \\ e_i \end{bmatrix}.$$

Testing $\beta^s - \beta = 0$ is akin to a Chow test across the two specifications (1) and (2). Of course, the data here are not two subsamples but rather duplicates of the original data set. To take account of this and allow for the correlation in the residuals across duplicates, it is crucial to cluster standard errors on the observation identifier i .

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. Because the joint distribution between the coefficient and standard error estimators is

difficult to characterize, especially in the case of the coefficient comparison test, we abstract away from the sampling variation in estimating the standard errors in the theoretical derivations of this section. The resulting t -statistic for the null hypothesis that the coefficient δ is zero in the balancing test is

$$t_{\delta^m} = \frac{\hat{\delta}^m}{se(\hat{\delta}^m)} = \frac{\sqrt{n} \cdot \hat{\delta}^m}{\frac{\sqrt{\sigma_u^2 + \sigma_m^2}}{\sigma_s}} = \frac{\sqrt{n} \cdot \hat{\delta}^m}{\frac{\sigma_u}{\sigma_s \sqrt{1-\theta}}}.$$

Similarly, we use

$$t_{(\beta^s - \beta^m)}(d; \gamma) = \frac{\hat{\beta}^s - \hat{\beta}^m}{se(\hat{\beta}^s - \hat{\beta}^m)} = \frac{\sqrt{n}(\hat{\beta}^s - \hat{\beta}^m)}{\sqrt{V_\beta(d; \gamma)}}$$

where

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

in the derivation of the power function for the coefficient comparison test.

As shown in Appendix A, the power function for a 5% critical value of the balancing test is

$$\begin{aligned} Power_{t_{\delta^m}}(d) &= 1 - \Phi \left(1.96 - d \frac{\sqrt{n} \sigma_s \sqrt{1-\theta}}{\sigma_u} \right) \\ &\quad + \Phi \left(-1.96 - d \frac{\sqrt{n} \sigma_s \sqrt{1-\theta}}{\sigma_u} \right), \end{aligned} \quad (16)$$

where $\Phi(\bullet)$ is the standard normal cumulative distribution function. The power function for the coefficient comparison test is

$$\begin{aligned} Power_{t_{(\beta^s - \beta^m)}}(d; \gamma) &= 1 - \Phi \left(1.96 - d \frac{\sqrt{n} \gamma (1-\theta)}{\sqrt{V_\beta(d; \gamma)}} \right) \\ &\quad + \Phi \left(-1.96 - d \frac{\sqrt{n} \gamma (1-\theta)}{\sqrt{V_\beta(d; \gamma)}} \right). \end{aligned} \quad (17)$$

Note that the power function for the balancing test does not involve the

parameter γ . Using our results above, for $0 < \gamma < \infty$ it can be written as

$$\begin{aligned} Power_{t_{\delta m}}(d) = 1 - \Phi \left(1.96 - d \frac{\sqrt{n}\gamma(1-\theta)}{\sqrt{V_{\delta}(d;\gamma)}} \right) \\ + \Phi \left(-1.96 - d \frac{\sqrt{n}\gamma(1-\theta)}{\sqrt{V_{\delta}(d;\gamma)}} \right), \end{aligned} \quad (18)$$

where

$$V_{\delta}(d;\gamma) = (1-\theta) \frac{\gamma^2 \sigma_u^2}{\sigma_s^2}.$$

It is hence apparent that $V_{\beta}(d;\gamma) > V_{\delta}(d;\gamma)$, i.e. the coefficient comparison test has a larger variance. As a result, when $d \neq 0^8$

$$Power_{t_{\delta m}}(d) > Power_{t_{(\beta s - \beta m)}}(d;\gamma). \quad (19)$$

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 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^2 s 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 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.

⁸To see this, define $f(t) = 1 - \Phi(1.96 - t) + \Phi(-1.96 - t)$ and denote the probability density function of a standard normal distribution by ϕ . The f notation allows us to rewrite the expressions for the power functions $Power_{t_{\delta m}}(d)$ and $Power_{t_{(\beta s - \beta m)}}(d;\gamma)$ in equations (17) and (18) simply as $f(t_1)$ and $f(t_2)$. When $d \neq 0$, $V_{\beta}(d;\gamma) > V_{\delta}(d;\gamma)$ implies that $|t_1| > |t_2| > 0$. Since $f'(t) = \phi(1.96 - t) - \phi(1.96 + t)$ is positive for all $t > 0$ and negative for all $t < 0$, $f(t_1) > f(t_2)$ given $|t_1| > |t_2| > 0$, and equation (19) follows.

In Figure 1, we start by plotting the theoretical power functions for both tests for three different magnitudes of the measurement error.⁹ The black/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 blue/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 $\theta = 0.85$ (red/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 $d < 0.5$.¹⁰

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

⁹The power function for the balancing test in equation (16) is written using the normal distribution, but we actually calculate it using the t -distribution with $n - 2$ degrees of freedom. This is consistent with how Stata version 14 performs the balancing test following the command `reg x s` or `reg x s, r`, even though this distribution choice makes little difference given our sample size ($n = 100$).

¹⁰While 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.

2 shows the power functions for the two tests without measurement error ($\theta = 0$) and with ($\theta = 0.85$), as well as their simulated counterparts.¹¹ Without measurement error, the theoretical power functions are closely aligned with the empirical rejection rates (black lines). Adding measurement error, this is also true for the balancing test (solid red and blue/thick lines) but not for the coefficient comparison test (broken red and blue/thick lines).

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 empirical test is severely distorted under the null; 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 (15), 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 beyond the Benchmark Model

The homoskedastic case with classical measurement error might be highly stylized and not correspond well to the situations typically encountered in empirical practice. We therefore explore some other scenarios using simulations

¹¹We did 25,000 replications in these simulations, and each repeated sample contains 100 observations.

in this section. Figure 3 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 $\bar{\sigma}_u^2 = 3$ and $\bar{\sigma}_e^2 = 30$ 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 3, 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 (black/thin lines). Otherwise, the main findings look very similar to those in Figure 1. Heteroskedasticity does not seem to alter the basic conclusions appreciatively.

Next, we explore mean reverting measurement error (Bound et al., 1994). 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 $\lambda = \text{Var}(x_i) / \text{Var}(x_i^m) > 1$ 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.

Notice that x_i^m can now be written as

$$x_i^m = (1 + \kappa) \delta_0 + (1 + \kappa) \delta s_i + (1 + \kappa) u_i + \mu_i,$$

so this parameterization 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 should bring 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 $\text{Var}(w_{1i}) = \text{Var}(w_{2i})$. These models exhibit relatively large amounts of mean reversion. Figure 4 demonstrates that the balancing test again dominates. 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 mean-reverting measurement error does less damage in terms of biasing the estimate of γ .

A particular case of mean reverting measurement error is the one where x_i is a dummy variable, so we provide some simulation results for this case. 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:

$$\Pr(x_i = 1) = \Phi(\delta s_i), \quad (20)$$

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

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

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 (20). One is the fact that $\text{plim}(\hat{\delta})$ from estimating (11) linearly is not going to equal the δ we generated in the probit equation (20) to generate x . The relationship between $\text{plim}(\hat{\delta})$ and δ is concave. In Figure 5, we plot rejection rates against values of δ , although the quantity $\text{plim}(\hat{\delta})$ 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 (20). 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 the corollary of our result that mean reverting measurement error also reduces the power of the balancing test. Of course, we know from the relationship (15) between the test statistics that the coefficient comparison test will also suffer from the same power loss.

The black/thin lines in Figure 5 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 (20) 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 (20) 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 (15), the power of the coefficient comparison test will equal the power of t_γ when $t_\delta \rightarrow \infty$. 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 .¹²

Adding measurement error to the binary regressor x_i makes things worse as is visible from the red/thick lines in Figure 5. 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.

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.

¹²The 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 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.

It is straightforward to extend our framework to that setting. In this section, we describe this multivariate extension, and provide some simulation results. 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

$$\begin{aligned} y_i &= \alpha + \beta s_i + \mathbf{x}_i' \boldsymbol{\gamma} + e_i \\ \mathbf{x}_i &= \boldsymbol{\delta}_0 + \boldsymbol{\delta} s_i + \mathbf{u}_i \\ \beta^s - \beta &= \boldsymbol{\gamma}' \boldsymbol{\delta} \end{aligned} \tag{21}$$

where $\boldsymbol{\gamma}$, $\boldsymbol{\delta}_0$, $\boldsymbol{\delta}$ and \mathbf{u}_i are $k \times 1$ vector analogs of their scalar counterparts in Section 2. 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. We can stack all the x -variables on the left-hand-side of the regression to obtain

$$\begin{bmatrix} \mathbf{x}_{(1)} \\ \mathbf{x}_{(2)} \\ \dots \\ \mathbf{x}_{(k)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\iota} \delta_{01} \\ \boldsymbol{\iota} \delta_{02} \\ \dots \\ \boldsymbol{\iota} \delta_{0k} \end{bmatrix} + \begin{bmatrix} \mathbf{s} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{s} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{s} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \dots \\ \delta_k \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{(1)} \\ \mathbf{u}_{(2)} \\ \dots \\ \mathbf{u}_{(k)} \end{bmatrix},$$

where $\boldsymbol{\iota}$ is an $n \times 1$ vector of ones, $\mathbf{s} = [s_1, s_2, \dots, s_n]'$, and $\mathbf{u}_{(j)}$ the vector of residuals corresponding to covariate $\mathbf{x}_{(j)}$. We can then perform an F -test for the joint significance of the $\boldsymbol{\delta}$ coefficients. This left-hand-side (LHS) balancing test is similar to the way we implemented the coefficient comparison test above in Section 4.1.

The drawback of the LHS test is that stacking equations is non-standard, and requires some extra programming to carry it out. It therefore seems appealing to consider the alternative of regressing s on the covariates \mathbf{x}

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

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. Applied researchers sometimes use this RHS balancing 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 deliver very similar results. In the case of a single covariate x_i (i.e. $k = 1$) the LHS and the RHS tests using a conventional covariance matrix for homoskedastic residuals are numerically identical.¹³ This is no longer true with multiple covariates ($k > 1$). 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 $\text{var}(\mathbf{u}_i) = \sigma^2 I_k$ and the RHS regression is homoskedastic, as we show in Appendix C.

How do the balancing tests with multiple covariates perform in practice? Figure 6 shows simulations using a similar design as described in Table 1 for all k balancing equations. However, 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: first, for the case where all covariates are unbalanced, i.e. $\delta_1 = \delta_2 = \dots = \delta_k = d$, and then for the case where only the first covariate is unbalanced while the others remain balanced, i.e. $\delta_1 = d, \delta_2 = \dots = \delta_k = 0$. We generate normally distributed, spherical errors and impose homoskedasticity and independence when performing the joint test of the δ_j 's or the π_j 's. There are four panels in Figure 6: the top row has 4 added covariates, and the bottom row 8; the left hand column shows the case where all covariates are unbalanced while the right hand column displays the case where only the first covariate is unbalanced.

Figure 6 highlights 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 all four panels. When all covariates are unbalanced and when measurement error is absent, the Hausman test turns out to be an efficient test

¹³The F -test in this case 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 there other covariates present in the regression, replace the R^2 and correlation coefficient with their partial equivalents in this argument).

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.¹⁴ In this case, the rejection rates for the coefficient comparison test (black/thin broken lines) therefore lie above the ones for both the balancing tests (black/thin solid and dash-dot lines), as can be seen in the left-hand panels. 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 red lines.

This power advantage of the balancing tests is greater when only one covariate is unbalanced. 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 (red/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. Figure 6 demonstrates that the balancing tests will frequently be the most powerful tools in such a situation, but the coefficient comparison test also has a role to play in the multivariate case.

The simulations reveal a number of 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 the standard small sample distortion of these covariance matrices discussed in the literature (MacKinnon and White, 1985; Chesher and Jewitt, 1987; Angrist and Pis-

¹⁴The analyses in Hausman (1978), Hausman and Taylor (1980), Holly (1982), and MacKinnon (1992) section II.9, which compare the power of the coefficient comparison test to the F -test for $\gamma = \mathbf{0}$, highlight a similar result.

chke, 2009, chapter 8). We find that the 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 remedies for the bias problem in multivariate tests is therefore particularly important.¹⁵

While we find few differences between the power of the LHS and RHS tests in our simulations, we know from the theoretical analysis in Appendix C that the test statistics will differ asymptotically when the third and fourth moments of the underlying data deviate from the normally distributed case. It is therefore interesting to probe how the two tests perform in an example with real data.

We therefore pooled data from the 2010 - 2014 American Community Surveys (ACS). Our data set consists of white and African American individuals

¹⁵We find in unreported simulations that the classic small sample corrections HC2 and HC3 by MacKinnon and White (1985) still have size distortions under the null. There is currently an active literature on how to better deal with this small sample bias of the robust or clustered covariance estimator. For example, Young (2016) suggests an adjustment of the degrees of freedom of hypothesis tests but this adjustment is only implemented for one coefficient at a time, so does not work for testing multiple linear restrictions at once. Cattaneo, Jansson and Newey (2017) present an adjustment of the entire covariance matrix but only consider the case of heteroskedasticity and do not allow for clustering. As a result, neither of these can currently be applied to our LHS balancing test.

Another alternative is to rely on a series of single coefficient tests and adjust the resulting test statistics for multiple testing. Akin to the size distortion of robust test statistics, without adjustment such multiple testing will reject too often under the null as first noted by Bonferroni (1935). There is a sizable literature in statistics and theoretical econometrics on this topic with modern approaches based either on the influential work by Westfall and Young (1993) or by Benjamini and Hochberg (1995). Examples of empirical applications in economics are Kling, Liebman and Katz (2007), Anderson (2008), and Duflo, Dupas and Kremer (2017). But these examples remain rare, and no clear choices among the multitude of theoretical alternatives have yet emerged among applied researchers.

aged 21 to 64 with non-missing annual earnings. This data set has 5,644,865 observations. We generated a binary treatment s_i according to

$$\Pr(s_i = 1) = \omega F(\text{educ}_i) + (1 - \omega) U.$$

educ_i is the years of schooling of individual i , $F(\text{educ}_i)$ is its cumulative distribution function, U is a uniform random variable, and ω is a weight akin to the parameter d in our earlier Monte Carlo experiments. Under the null hypothesis, $\omega = 0$, and the treatment s_i consists solely of the generated noise U . For values of $\omega > 0$, the treatment s_i is related to the education level of the individual, which in turn is correlated with other individual covariates \mathbf{x}_i . Our vector of covariates (\mathbf{x}_i) contains the six variables: female, black, age, age squared, log family size, and log income. These variables take on very different distributions from simple binary for female and black to skewed distributions for family size and income. They are also the types of variables researchers will likely use to check for balance when working with individual household data. The bigger the ω , the more likely the balancing test relating s_i and \mathbf{x}_i should reject. In our simulations, we draw samples of size 1,000 with replacement from the original 5,644,865 observations in the ACS dataset. We perform 10,000 replications and carry out the LHS and RHS tests for various values of ω . Figure 7 shows the results for the two balancing tests. The rejection rates are virtually indistinguishable. We find no evidence that the performance of the two tests differs in this setting.¹⁶ This does not mean that the LHS and RHS test statistics are identical in any given sample. Particularly under the null we sometimes find sizable disparities in p -values.

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.

¹⁶We have also experimented with basing selection into treatment s_i on income and including education among the added covariates instead. The results are very similar.

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 an important 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 in a particular table or set of tables.

We start in Table 2 by presenting simple OLS regressions controlling for experience, race, and past and present residence. The estimated return to schooling is 0.075. This estimate may not reflect the causal effect of education on income because important confounders, such as ability or family background, are not controlled for.

In columns (2) to (5) we include variables which might proxy for the respondent's family background. In column (2) we include mother's education, in column (3) whether the household had a library card when the respondent was 14, and in column (4) we add body height measured in inches. Each of these variables is correlated with earnings, and the coefficient on education moves moderately when these controls are included. 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 important parental attitudes (e.g. Farber and Gibbons, 1996). Body height is a variable 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). The return to education falls by 0.1 to 0.2 log points when these controls are added. In column (5) we enter all three variables simultaneously. The coefficients on the controls are somewhat attenuated, and the return to education falls slightly further to 0.071.

It might be tempting to conclude from the relatively small change in the estimated returns to schooling that this estimate should be given a causal interpretation. We provide a variety of evidence that this conclusion is unlikely to be a sound one. Below the estimates in columns (2) to (5), we display the p -values from the coefficient comparison test, comparing each of the estimated returns to education to the one from column (1). Although the coefficient movements are small, the tests all reject at the 5% level, and in columns (4) and (5) they reject at the 1% level. These results might not be expected from the size of the coefficient movements and the individual standard errors on the years of education coefficients alone, highlighting the importance for the formal coefficient comparison test.

The results in columns (6) to (8), where we regress maternal education, the library card, and body height on education, further magnify the concern. The education coefficient is positive and strongly significant in all three regressions, with t -values ranging from 4.4 to 13.1, and 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 are substantively important. It is difficult to think of these results as causal effects: the respondent’s education should not affect predetermined proxies of family background. Instead, 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. Measurement error leads to attenuation bias when these variables are used on the right-hand side which renders them fairly useless as controls. The measurement error matters less for the estimates in columns (6) to (8), and these are 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.

The following tables have the same general layout. In Table 3 we change the baseline specification by including the respondent's score on the Knowledge of the World of Work test (KWW), a variable used by Griliches (1977) as a proxy for ability. The sample size is reduced due to the exclusion of missing IQ values in the test score for consistency with a subsequent table. Estimated returns without the KWW score in this restricted sample (unreported) are very similar to those in Table 2. Adding the KWW score reduces the coefficient on education by almost 20%, from 0.075 to 0.061. Adding our additional controls maternal education, the library card, and body height to this new specification does very little to the estimated returns to education. The coefficient comparison test indicates that none of the small changes in the returns to education are statistically significant. Controlling for the KWW scores has largely knocked out the library card effect but done little to the coefficients on maternal education and body height. The relatively small and insignificant coefficient movements in columns (2) to (5) suggest that the specification controlling for the KWW score might solve the ability bias problem.

Columns (6) to (8), however, show that the three covariates are still mostly unbalanced with respect to education even when the KWW score is in the regression. This raises the possibility that the estimated returns in columns (1) to (5) might remain biased by selection. The estimated coefficients on education for the three controls are on the order of half their value from Table 2, and the body height measure is now only significant at the 10% level. But the relationship between mother's and own education is still sizable, so that this measure continues to indicate the possibility of important selection. Balance

in library card ownership is rejected despite the fact that a comparison of the coefficients in columns (1) and (3) indicates no role for this variable at all. A joint balancing test with all three controls strongly rejects the hypothesis that they are balanced. The results in this table illustrate the message of our paper in a powerful fashion.

While the KWW score might be a potent control, it is likely also measured with substantial error, for example, due to testing noise. Griliches (1977) proposes to instrument this measure with an independent IQ test score variable, which is also contained in the NLS data, to eliminate at least some of the consequences of this measurement error. In Table 4, we take the specification one step further by instrumenting the KWW score with IQ. The coefficient on the KWW score almost triples, in line with the idea that an individual test score is a very noisy measure of ability. The education coefficient now falls to only about half its previous value from 0.061 to 0.034. This might be due to positive omitted variable bias present in the previous regressions which is eliminated by IQ-instrumented KWW (although there may be other possible explanations for the change as well, like measurement error in schooling). Both the coefficient comparison tests and the balancing tests (individual and joint) indicate no evidence of selection any more. This is due to a combination of lower point estimates and larger standard errors. We note that the joint LHS and RHS balancing tests produce somewhat different test statistics in this case, although both p -values are well above conventional rejection levels. The contrast between Tables 3 and 4 highlights the usefulness of the balancing test: it warns about the Table 3 results, while the coefficient comparison test delivers insignificant differences in either case.

Finding an instrumental variable for education is an alternative to control strategies, such as using test scores. In Table 5 we follow Card's (1995) analysis and instrument education using distance to the nearest college, while dropping the KWW score. We use the same sample as in Table 2, which differs from Card's sample.¹⁷ Our IV estimates of the return to education are slightly higher

¹⁷Unlike Card, who uses two dummies for proximity to a two- and a four-year college, we use a single dummy variable for whether there is a four-year college in the county as

than in Table 2 but lower than in Card (1995) at around 8%. The IV returns estimates are relatively noisy, with a t -statistic of about 2. Columns 1-5 of Table 5 show that the IV estimate on education, while bouncing around a bit, does not change significantly when maternal education, the library card, or body height is included. In particular, if these three controls are included at the same time in column (5), the point estimate is indistinguishable from the unconditional estimate in column (1) both substantively and statistically.

IV regressions with pre-determined variables on the left hand side can be thought of as a test for random assignment of the instruments. In this case the selection regressions in columns (6)-(8) are imprecise, just like the IV returns estimates, and as a result less informative. The coefficients in the regressions for mother's education and body height have the wrong sign but confidence intervals cover anything ranging from zero selection to large positive amounts. Only the library card measure is large, positive, with a p -value of around 0.06, possibly indicative of some remaining potential for selection even in the IV regressions. However, with p -values of around 0.29, both the LHS and RHS joint balancing tests fail to reject the null hypothesis that all three controls are balanced. In other words, the college distance instrument passes the balancing test, but the data do not speak clearly in this particular case.

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 should be available in many data sets, and we encourage researchers to perform such balancing tests 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

instrument, and we instrument experience and experience squared by age and age squared. We restrict Card's sample to non-missing values in maternal education, the library card, and body height.

interest.

We have illustrated our theoretical results with an application to the returns to education. Taking our assessment from this exercise at face value, a reader might conclude that the results in Table 4, returns around 3.5%, can safely be regarded as causal estimates. Of course, this is not the conclusion reached in the literature, where much higher IV estimates like those in Table 5 are generally preferred (see e.g. Card, 2001 or Angrist and Pischke, 2015, chapter 6). This serves as a reminder that the discussion here is focused on sharpening one particular tool in the kit of applied economists. Successfully passing the balancing test should be a necessary condition for a successful research design but it is not sufficient.

The balancing test and other statistics we discuss here are useful for gauging selection bias due to observed confounders, even when they are potentially measured poorly. It 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, which is also exacerbated as more potent controls are added. Griliches (1977) shows that a modest amount of measurement error in schooling may be responsible for the patterns of returns we have displayed in Tables 2 to 4. Another issue, also discussed by Griliches, is that controls like test scores might themselves be influenced by schooling, which would make them bad controls. For all these reasons, other approaches like IV estimates of the returns may be preferable.

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.
- Anderson, Michael L.** 2008. “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects.” *Journal of the American Statistical Association*, 103(484): 1481–1495.
- Angrist, Joshua, and Jörn-Steffen Pischke.** 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.
- Angrist, Joshua, and Jörn-Steffen Pischke.** 2015. *Mastering Metrics: The Path from Cause to Effect*. Princeton University 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.** 2014a. “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.** 2014b. “Inference on Treatment Effects after Selection among High-Dimensional Controls.” *The Review of Economic Studies*, 81(2): 608–650.
- Benjamini, Yoav, and Yosef Hochberg.** 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300.

- Bonferroni, Carlo Emilio.** 1935. *Il Calcolo delle Assicurazioni su Gruppi di Teste*. Tipografia del Senato.
- 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.
- Card, David.** 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica*, 69(5): 1127–1160.
- Case, Anne, and Christina Paxson.** 2008. "Stature and Status: Height, Ability, and Labor Market Outcomes." *Journal of Political Economy*, 116(3): 499–532.
- Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey.** 2017. "Inference in Linear Regression Models with Many Covariates and Heteroskedasticity." mimeographed, Michigan, Berkeley, and MIT.
- Chesher, Andrew, and Ian Jewitt.** 1987. "The Bias of a Heteroskedasticity Consistent Covariance Matrix Estimator." *Econometrica*, 55(5): 1217–1222.
- Duflo, Esther, Pascaline Dupas, and Michael Kremer.** 2017. "The Impact of Free Secondary Education: Experimental Evidence from Ghana." mimeographed, MIT, Stanford, and Harvard.
- Farber, Henry S., and Robert Gibbons.** 1996. "Learning and Wage Dynamics." *The Quarterly Journal of Economics*, 111(4): 1007–1047.

- 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.
- 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.
- Imbens, Guido W.** 2003. “Sensitivity to Exogeneity Assumptions in Program Evaluation.” *American Economic Review*, 93(2): 126–132.
- Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz.** 2007. “Experimental Analysis of Neighborhood Effects.” *Econometrica*, 75(1): 83–119.
- Lee, David S., and Thomas Lemieux.** 2010. “Regression Discontinuity Designs in Economics.” *Journal of Economic Literature*, 48(2): 281–355.
- 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.

- 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*.
- 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.
- Rosenbaum, Paul R., and Donald B. Rubin.** 1983. “Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2): 212–218.
- Westfall, Peter H., and S. Stanley Young.** 1993. *Resampling Based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley and Sons.
- Young, Alwyn.** 2016. “Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections.” mimeographed, LSE.

Figure 1: Theoretical Rejection Rates

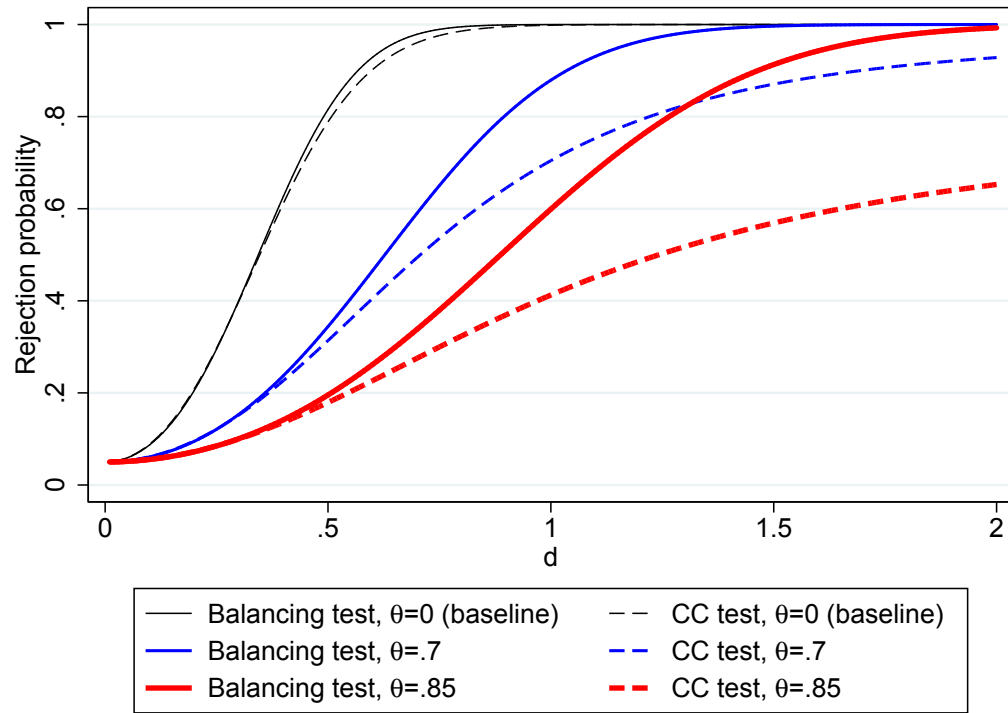
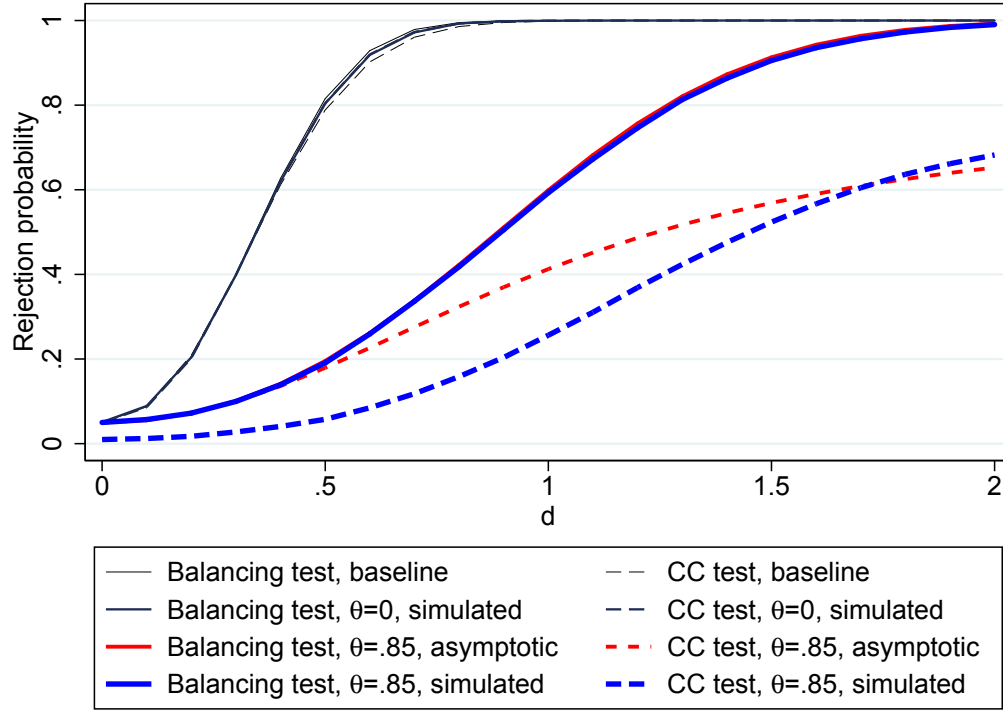
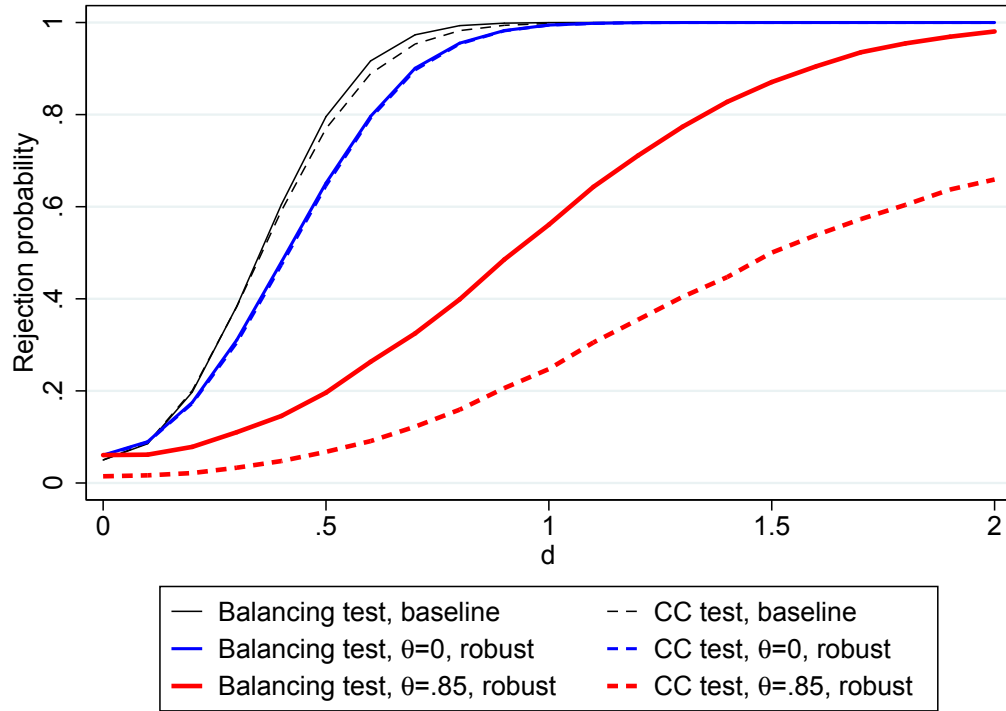


Figure 2: Theoretical and Simulated Rejection Rates



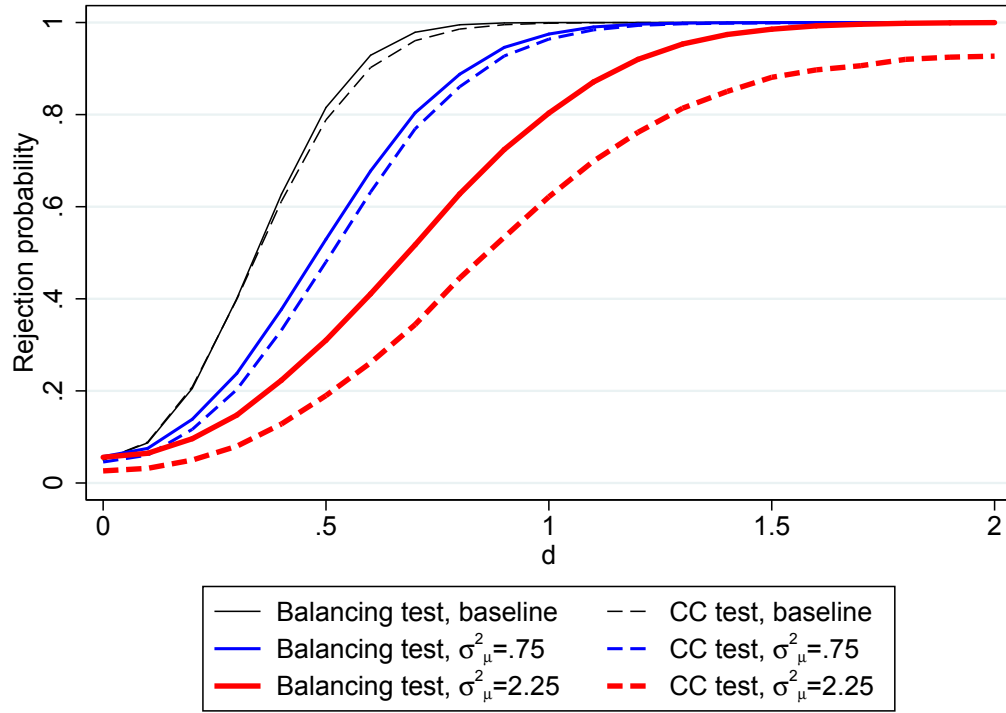
Note: Comparison of asymptotic rejection rates with rejection rates based on Monte Carlo simulations. Baseline refers to the theoretical rejection rates without measurement error.

Figure 3: Simulated Rejection Rates with Heteroskedasticity



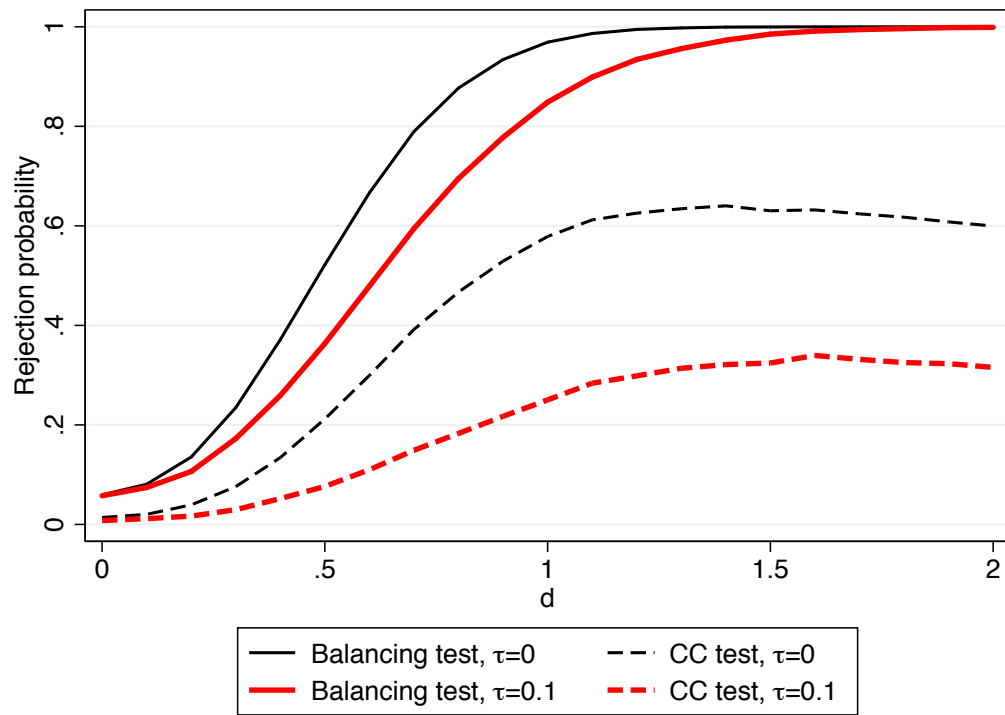
Note: Comparison of baseline rejection rates (from Figure 1) with simulated rejection rates based on heteroskedastic errors and robust standard errors.

Figure 4: Simulated Rejection Rates with Mean Reverting Measurement Error



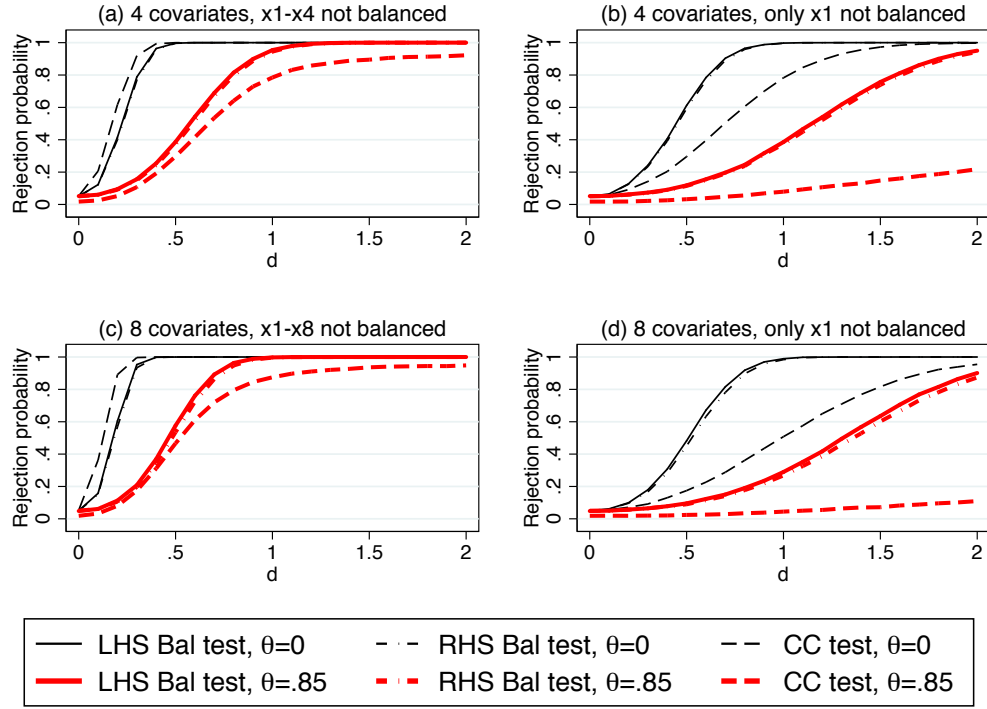
Note: Comparison of baseline rejection rates (from Figure 1) with simulated rejection rates based on mean reverting measurement error and robust standard errors.

Figure 5: Simulated Rejection Rates with Binary Control and Misclassification



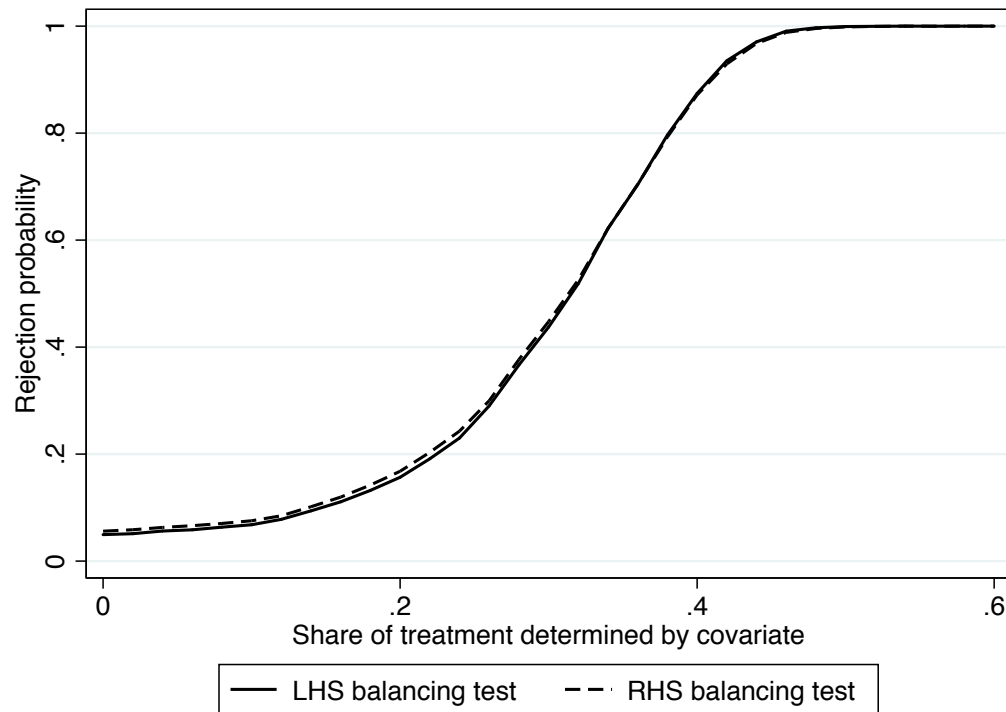
Note: Rejection rates for a binary control variable that is misclassified (i.e. its binary value is flipped) with probability τ .

Figure 6: Simulated Rejection Rates with Multiple Controls



Note: Rejection rates for simultaneous tests for adding 4 or 8 additional covariates at once under different specifications for the alternative hypotheses.

Figure 7: Rejection Rates with Multiple Controls in Actual Data from the ACS



Note: Rejection rates based on drawing random samples of size 1,000 from the American Community Surveys. See text for details.

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

$\sigma_s^2 = 1$	$\beta = 1$		
$\sigma_u^2 = 3$	$\gamma = 3$		
$\sigma_e^2 = 30$	$n = 100$		
	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 2: Baseline Regressions for Returns to Schooling and Specification Checks

	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	0.0751 (0.0040)	0.0728 (0.0042)	0.0735 (0.0040)	0.0740 (0.0040)	0.0710 (0.0042)	0.3946 (0.0300)	0.0371 (0.0040)	0.1204 (0.0273)
Mother's years of education		0.0059 (0.0029)			0.0044 (0.0030)			
Library card at age 14			0.0428 (0.0183)		0.0361 (0.0184)			
Body height in inches				0.0090 (0.0027)	0.0084 (0.0027)			
<i>p</i> -values								
Coefficient comparison test		0.044	0.022	0.009	0.002			
LHS balancing test: individual						0.000	0.000	0.000
LHS balancing test: joint							0.000	
RHS balancing test: joint							0.000	

Note: The number of observations is 2,500 in all regressions. Heteroskedasticity robust standard errors in parentheses, and 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 a standard metropolitan statistical area (SMSA) in 1976, indicators for region in 1966 and living in an SMSA in 1966.

Table 3: 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 (0.0059)	0.0596 (0.0060)	0.0608 (0.0059)	0.0603 (0.0059)	0.0591 (0.0060)	0.2500 (0.0422)	0.0133 (0.0059)	0.0731 (0.0416)
KWW score	0.0070 (0.0015)	0.0068 (0.0016)	0.0069 (0.0016)	0.0069 (0.0015)	0.0067 (0.0016)	0.0410 (0.0107)	0.0076 (0.0016)	0.0145 (0.0117)
Mother's years of education		0.0053 (0.0037)			0.0048 (0.0037)			
Library card at age 14			0.0097 (0.0215)		0.0045 (0.0216)			
Body height in inches				0.0078 (0.0034)	0.0075 (0.0034)			
<i>p</i> -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, due to missing values in IQ. Heteroskedasticity robust standard errors in parentheses, and 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.

Table 4: Regressions for Returns to Schooling and Specification Checks Instrumenting 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	0.0340 (0.0139)	0.0339 (0.0139)	0.0342 (0.0138)	0.0343 (0.0139)	0.0345 (0.0138)	0.0234 (0.0952)	0.0168 (0.0134)	-0.0486 (0.0998)
KWW score instrumented by IQ	0.0199 (0.0062)	0.0195 (0.0063)	0.0200 (0.0063)	0.0194 (0.0062)	0.0191 (0.0064)	0.1496 (0.0422)	0.0060 (0.0060)	0.0728 (0.0449)
Mother's years of education		0.0028 (0.0039)			0.0026 (0.0039)			
Library card at age 14			-0.0130 (0.0245)		-0.0154 (0.0243)			
Body height in inches				0.0070 (0.0034)	0.0069 (0.0034)			
<i>p</i> -values								
Coefficient comparison test		0.818	0.634	0.636	0.552			
Balancing test: individual						0.806	0.212	0.626
LHS balancing test: joint							0.593	
RHS balancing test: joint							0.137	

Note: The number of observations is 1,773 in all regressions, due to missing values in IQ. Heteroskedasticity robust standard errors in parentheses, and the joint LHS balancing test is conducted by a stacked IV regression with standard errors clustered across the three additional covariates. 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.

Table 5: Regressions for Returns to Schooling and Specification Checks Instrumenting Schooling by Proximity to College

	Dependent Variable							
	Log hourly earnings				Mother's years of education		Library card at age 14	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education instrumented by college proximity	0.0816 (0.0431)	0.0818 (0.0417)	0.0778 (0.0518)	0.0845 (0.0418)	0.0822 (0.0466)	-0.0952 (0.3594)	0.1015 (0.0542)	-0.3658 (0.3681)
Mother's years of education		0.0030 (0.0143)			0.0012 (0.0140)			
Library card at age 14			0.0367 (0.0886)		0.0237 (0.0581)			
Body height in inches				0.0081 (0.0044)	0.0079 (0.0032)			
<i>p</i> -values								
Coefficient comparison test		0.873	0.686	0.380	0.908			
Balancing test: individual						0.791	0.061	0.321
LHS balancing test: joint							0.290	
RHS balancing test: joint							0.291	

Note: The number of observations is 2,500 in all regressions. Heteroskedasticity robust standard errors in parentheses, and the joint LHS balancing test is conducted by a stacked IV regression with standard errors clustered across the three additional covariates. 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.

Appendix (For Online Publication Only)

A Power Functions

A.1 The Balancing Test

The desired balancing regression is

$$x_i = \delta_0 + \delta s_i + u_i,$$

but x_i is measured with error

$$x_i^m = x_i + m_i.$$

Effectively, we run the balancing regression

$$x_i^m = \delta_0^m + \delta^m s_i + u_i + m_i.$$

As mentioned in Section 5.1, in the theoretical derivation of the power functions we abstract away from the sampling variation in estimating the standard errors by treating σ_u , σ_m and σ_s as known constants. In this case, the asymptotic variance of $\hat{\delta}^m$ can be directly calculated, and the resulting test statistic for the null hypothesis that the balancing coefficient δ is zero is

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

Define

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

Hence

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

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

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

when n is large. This is the power function of the balancing test

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

A.2 The Coefficient Comparison Test

The short and long regressions are

$$\begin{aligned}
 y_i &= \alpha^s + \beta^s s_i + e_i^s \\
 y_i &= \alpha + \beta s_i + \gamma x_i + e_i,
 \end{aligned}$$

and

$$x_i = \delta_0 + \delta s_i + u_i.$$

Adding measurement error in x_i :

$$x_i^m = x_i + m_i,$$

we have

$$\begin{aligned}
 y_i &= \alpha^s + \beta^s s_i + e_i^s \\
 y_i &= \alpha^m + \beta^m s_i + \gamma^m x_i^m + e_i^m \\
 x_i^m &= \delta_0 + \delta s_i + u_i + m_i.
 \end{aligned}$$

Treat s_i , u_i , e_i , and m_i as the underlying random variables which determine x_i , y_i , e_i^s and e_i^m . We normalize s_i to a mean zero variable. For the derivations in the remainder of this section, we make the following assumptions:

Assumption A1: s_i , u_i , e_i and m_i are mutually independent;

Assumption A2: $E[u_i^3] = 0$.

Note that Assumptions A1 and A2 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 the first two moments specified according to

$$\begin{pmatrix} s_i \\ u_i \\ e_i \\ m_i \end{pmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_s^2 & 0 & 0 & 0 \\ 0 & \sigma_u^2 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & \sigma_m^2 \end{bmatrix} \right). \quad (\text{A1})$$

A.2.1 Population Parameters

In this subsection, we derive the expressions of population regression coefficients β^m and γ^m in terms of the model parameters, as discussed in Section 3. Performing an anatomy to the multiple regression (9), we have

$$\gamma^m = \frac{\text{Cov}(y_i, u_i + m_i)}{\text{Var}(u_i + m_i)} = \gamma \frac{\sigma_u^2}{\sigma_u^2 + \sigma_m^2}, \quad (\text{A2})$$

where $u_i + m_i$ is the residual from the population regression of x_i^m on s_i . Using θ as defined above, equation (A2) becomes

$$\gamma^m = \gamma(1 - \theta). \quad (\text{A3})$$

By the omitted variable bias formula, we have

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

and therefore

$$\beta^m = \beta + \gamma\delta\theta. \quad (\text{A4})$$

As mentioned in the main text, an alternative representation of θ is

$$\theta = \frac{1 - \lambda}{1 - R^2}, \quad (\text{A5})$$

where

$$\lambda = \frac{\text{Var}(x_i)}{\text{Var}(x_i^m)}$$

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 . To see why (A5) holds, notice that

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

from which equation (A5) mechanically follows.

A.2.2 Asymptotic Variance in the Coefficient Comparison Test under Homoskedasticity

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

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

which is asymptotically standard normal. As mentioned in Section 4, we rely on the delta method equation (13) to derive $\text{Var}(\hat{\beta}^s - \hat{\beta}^m)$. We have already shown in the previous subsection that

$$\text{Var}(\hat{\delta}^m) = \frac{1}{n} \frac{\sigma_u^2}{(1 - \theta) \sigma_s^2}, \quad (\text{A6})$$

and we derive $\text{Var}(\hat{\gamma}^m)$ and $\text{Cov}(\hat{\delta}^m, \hat{\gamma}^m)$ in the remainder of this subsection. For simplicity of exposition, we make an additional assumption:

Assumption A3: $\text{Var}(e_i^m | s_i, x_i^m)$ is constant.

Like Assumptions A1 and A2, Assumption A3 is also satisfied in the DGP's underlying Figure 2. In the subsection below, we also derive the general expression of $\text{Var}(\hat{\beta}^s - \hat{\beta}^m)$ when Assumption A3 is relaxed.

In order to derive $Var(\hat{\gamma}^m)$, first note that

$$Var(\hat{\gamma}^m) = \frac{1}{n} \frac{Var(e_i^m)}{Var(u_i + m_i)}, \quad (A7)$$

where, as mentioned above, $u_i + m_i$ is the residual from the population regression of x_i^m on s_i . Since $Var(u_i + m_i) = \sigma_u^2 + \sigma_m^2$, the missing piece in equation (A7) is $Var(e_i^m)$. Plugging (A3) and (A4) into (9), we get

$$\begin{aligned} y_i &= \alpha^m + \beta^m s_i + \gamma^m x_i^m + e_i^m \\ &= \alpha^m + (\beta + \gamma\delta\theta) s_i + \gamma(1 - \theta) x_i^m + e_i^m \\ &= (\alpha^m + \gamma(1 - \theta)\delta_0) + (\beta + \gamma\delta) s_i + \gamma(1 - \theta)(u_i + m_i) + e_i^m \end{aligned}$$

Since

$$\begin{aligned} y_i &= \alpha + \beta s_i + \gamma(\delta_0 + \delta s_i + u_i) + e_i \\ &= (\alpha + \gamma\delta_0) + (\beta + \gamma\delta) s_i + \gamma u_i + e_i, \end{aligned}$$

matching residuals yields

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

So

$$\begin{aligned} Var(\hat{\gamma}^m) &= \frac{1}{n} \frac{\gamma^2\theta\sigma_u^2 + \sigma_e^2}{\sigma_u^2 + \sigma_m^2} \\ &= \frac{1 - \theta}{n} \left(\gamma^2\theta + \frac{\sigma_e^2}{\sigma_u^2} \right). \end{aligned} \quad (A8)$$

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

$$\hat{\delta}^m - \delta = \frac{\sum_i (u_i + m_i)(s_i - \bar{s})}{\sum_i (s_i - \bar{s})^2} \quad (A9)$$

$$\hat{\gamma}^m - \gamma^m = \frac{\sum_i e_i^m (\tilde{x}_i^m - \bar{\tilde{x}}^m)}{\sum_i (\tilde{x}_i^m - \bar{\tilde{x}}^m)^2} \quad (A10)$$

where \bar{s} and \tilde{x}^m are the sample averages of s_i and \tilde{x}_i^m with $\tilde{x}_i^m = x_i^m - \hat{\delta}_0 - \hat{\delta}^m s_i$ being the residual from regressing x_i^m on s_i . By Assumption A1 along with the fact that $\hat{\delta}_0 \xrightarrow{p} \delta_0$ and $\hat{\delta}^m \xrightarrow{p} \delta$, the asymptotic joint distribution of the numerators in equations (A9) and (A10) is

$$\frac{1}{\sqrt{n}} \left[\begin{array}{c} \sum_i (u_i + m_i)(s_i - \bar{s}) \\ \sum_i e_i^m (\tilde{x}_i^m - \tilde{x}^m) \end{array} \right] \xrightarrow{d} N \left(0, \left[\begin{array}{cc} (\sigma_u^2 + \sigma_m^2) \sigma_s^2 & E[s_i(u_i + m_i)^2 e_i^m] \\ E[s_i(u_i + m_i)^2 e_i^m] & E[(u_i + m_i)^2 (e_i^m)^2] \end{array} \right] \right).$$

By Assumptions A1 and A2,

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

Since the denominators of equations (A9) and (A10) converge in probability to positive constants,

$$Cov(\hat{\delta}^m, \hat{\gamma}^m) = 0. \quad (\text{A11})$$

Plugging equations (A6), (A8) and (A11) into (13) yields

$$\begin{aligned} Var(\hat{\beta}^s - \hat{\beta}^m) &\equiv \frac{1}{n} V_\beta(d; \gamma) \\ &= \frac{1}{n} (1 - \theta) \left(\frac{\gamma^2 \sigma_u^2}{\sigma_s^2} + \theta \delta^2 \gamma^2 + \frac{\delta^2 \sigma_e^2}{\sigma_u^2} \right). \end{aligned} \quad (\text{A12})$$

Recall 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.2.3 Relaxing Assumption A3

In this subsection, we provide the expression for $Var(\hat{\beta}^s - \hat{\beta}^m)$ while relaxing the conditional homoskedasticity of e_i^m , i.e. Assumption A3. Our derivation

of this asymptotic variance expression still relies on equation (13). Since equations (A6) and (A11) are not affected by Assumption A3, we will only need the general expression for $Var(\hat{\gamma}^m)$.

Representing model (9) in matrix form,

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

where $\mathbf{W}_i = (1, s_i, x_i^m)'$ and $\boldsymbol{\Gamma} = (\alpha^m, \beta^m, \gamma^m)'$. The asymptotic variance-covariance matrix of the regression estimator $\hat{\boldsymbol{\Gamma}}$ 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:

$$\begin{aligned} E[\mathbf{W}_i \mathbf{W}_i'] &= E \begin{bmatrix} 1 & s_i & x_i^m \\ s_i & s_i^2 & s_i x_i^m \\ x_i^m & s_i x_i^m & (x_i^m)^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & \delta_0 \\ 0 & \sigma_s^2 & \delta \sigma_s^2 \\ \delta_0 & \delta \sigma_s^2 & \delta_0^2 + \delta^2 \sigma_s^2 + \sigma_u^2 + \sigma_m^2 \end{bmatrix}. \end{aligned}$$

As before, we set $E[s_i] = 0$, which sacrifices no generality since the mean does not enter the variance calculation in any case.

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

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

Below we express quantities (i) to (vi) in terms of the fundamental model parameters. Letting $\kappa_m = E[m_i^4]$ and $\kappa_u = E[u_i^4]$ and utilizing Assumptions 1 and 2, we have the expressions for (i) to (vi):

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

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

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

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

and

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

Finally, for the expression of (vi)

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

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,

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

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 $\hat{\gamma}^m$

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

Compared to its expression under homoskedasticity (A8), we have an extra term (a) that accounts for the excess kurtosis of the u and m distributions. It follows that

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

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 (A12). Since $Var(\hat{\beta}^s - \hat{\beta}^m)$ increases in κ_u and κ_m and that 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

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

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. $x_i = \rho w_{1i} + w_{2i}$. Furthermore, there is equal selection, i.e.

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

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

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

which is our regression with $\gamma = 1$ (the scaling of x_i is arbitrary of course; it could be $x_i = w_{1i} + w_{2i}/\rho$ instead and $\gamma = \rho$ or anything else).

Our observed $x_i^m = \rho w_{1i}$, so measurement error $m_i = -w_{2i}$. Measurement error here is mean reverting, i.e.

$$m_i = \kappa x_i + \mu_i \tag{A13}$$

with $\kappa < 0$. Notice that

$$Cov(m_i, x_i) = -\sigma_2^2,$$

and hence

$$\kappa = \frac{-\sigma_2^2}{\rho^2 \sigma_1^2 + \sigma_2^2} \tag{A14}$$

and

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

It turns out that μ_i implicitly defined in (A13) and κ given by (A14) 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. However, note that $Cov(s_i, \mu_i) = 0$ depends on the equal selection assumption. With proportional selection, i.e.

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

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

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

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

and the variance estimator $\widehat{\text{var}}(\hat{\boldsymbol{\delta}})$ is

$$\widehat{\text{var}}(\hat{\boldsymbol{\delta}}) = \begin{pmatrix} \sum_i s_i^2 & 0 & \cdots & 0 \\ 0 & \sum_i s_i^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_i s_i^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_i s_i^2 \hat{u}_{(1)i}^2 & \sum_i s_i^2 \hat{u}_{(1)i} \hat{u}_{(2)i} & \cdots & \sum_i s_i^2 \hat{u}_{(1)i} \hat{u}_{(k)i} \\ \sum_i s_i^2 \hat{u}_{(2)i} \hat{u}_{(1)i} & \sum_i s_i^2 \hat{u}_{(2)i}^2 & \cdots & \sum_i s_i^2 \hat{u}_{(2)i} \hat{u}_{(k)i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i s_i^2 \hat{u}_{(k)i} \hat{u}_{(1)i} & \sum_i s_i^2 \hat{u}_{(k)i} \hat{u}_{(2)i} & \cdots & \sum_i s_i^2 \hat{u}_{(k)i}^2 \end{pmatrix} \cdot \begin{pmatrix} \sum_i s_i^2 & 0 & \cdots & 0 \\ 0 & \sum_i s_i^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_i s_i^2 \end{pmatrix}^{-1},$$

which allows for correlations in the error terms across covariates. Under the multivariate analog of Assumption A1,

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

Hence,

$$\frac{k}{n} F_{LHS} \xrightarrow{p} \sigma_s^2 \boldsymbol{\delta}' (E[\mathbf{u}_i \mathbf{u}_i'])^{-1} \boldsymbol{\delta}. \quad (\text{A15})$$

On the other hand, the F -statistic for the RHS balancing test following the regression

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

is

$$F_{RHS} = \frac{1}{k} \widehat{\boldsymbol{\pi}' var(\hat{\boldsymbol{\pi}})}^{-1} \hat{\boldsymbol{\pi}} \quad (\text{A16})$$

The probability limit of $\hat{\boldsymbol{\pi}}$ is

$$\boldsymbol{\pi} = \boldsymbol{\Omega}_x^{-1} \boldsymbol{\varsigma} \quad (\text{A17})$$

where $\boldsymbol{\Omega}_x = var(\mathbf{x}_i)$ and $\boldsymbol{\varsigma} = cov(\mathbf{x}_i, s_i)$. The probability limit of the variance estimator is

$$n \widehat{var}(\hat{\boldsymbol{\pi}}) \xrightarrow{p} \boldsymbol{\Omega}_x^{-1} E[(\mathbf{x}_i \mathbf{x}_i')(s_i - \boldsymbol{\pi}' \mathbf{x}_i)^2] \boldsymbol{\Omega}_x^{-1}. \quad (\text{A18})$$

Plugging (A17) and (A18) into (A16), the probability of the scaled F -stat of the RHS balancing test is

$$\begin{aligned} \frac{k}{n} F_{RHS} &\xrightarrow{p} \boldsymbol{\varsigma}' E[(\mathbf{x}_i \mathbf{x}_i')(s_i - \boldsymbol{\pi}' \mathbf{x}_i)^2]^{-1} \boldsymbol{\varsigma} \\ &= \sigma_s^4 \boldsymbol{\delta}' E[(\mathbf{x}_i \mathbf{x}_i')(s_i - \boldsymbol{\pi}' \mathbf{x}_i)^2]^{-1} \boldsymbol{\delta} \end{aligned} \quad (\text{A19})$$

The probability limits (A15) and (A19) are in general different. An analytical comparison between the two is complicated, as it depends on the higher moments of s and \mathbf{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. As mentioned in Section 5.3, we conduct additional investigations of the relative powers of the two tests via simulation using ACS data.

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, i.e.

$$var(\mathbf{u}_i) = \sigma_u^2 \mathbf{I}_k,$$

which is satisfied if s and \mathbf{u} are both normally distributed. Substituting this into (A15), 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 allows us to write

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

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

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

so

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

with

$$\begin{aligned} \text{var}(\boldsymbol{\pi}' \mathbf{x}_i) &= \boldsymbol{\pi}' \boldsymbol{\Omega}_x \boldsymbol{\pi} \\ &= \boldsymbol{\varsigma}' \boldsymbol{\Omega}_x^{-1} \boldsymbol{\varsigma} \\ &= \sigma_s^4 \boldsymbol{\delta}' \boldsymbol{\Omega}_x^{-1} \boldsymbol{\delta}. \end{aligned} \tag{A20}$$

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

$$\begin{aligned} \boldsymbol{\Omega}_x^{-1} &= \frac{1}{\sigma_u^2} \mathbf{I} - \frac{1}{1 + \frac{\sigma_s^2}{\sigma_u^2} \boldsymbol{\delta}' \boldsymbol{\delta}} \frac{1}{(\sigma_u^2)^2} \sigma_s^2 \boldsymbol{\delta} \boldsymbol{\delta}' \\ &= \frac{1}{\sigma_u^2} \mathbf{I} - \frac{\sigma_s^2}{(\sigma_u^2)^2 + \sigma_u^2 \sigma_s^2 \boldsymbol{\delta}' \boldsymbol{\delta}} \boldsymbol{\delta} \boldsymbol{\delta}'. \end{aligned} \tag{A21}$$

Plugging (A21) into (A20):

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

It follows that

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

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

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

Therefore,

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