Contaminated Control Variables in 2SLS Models

Rob Schonlau
Associate Professor of Finance
Colorado State University
schonlau@colostate.edu

Asad Dossani
Assistant Professor of Finance
Colorado State University
asad.dossani@colostate.edu

Jeffrey P. Dotson
Associate Professor of Marketing
The Ohio State University
dotson.83@osu.edu

this version:

November 18, 2024[1]

[1]Acknowledgments: We thank the seminar participants and discussants at the 2024 Western Economic Association International conference, the 2024 Society for Economic Measurement conference, Colorado State University, and at Brigham Young University for valuable comments and suggestions. All errors are our own.

Contaminated Control Variables in 2SLS Models

Abstract

Despite guidance in the theoretical literature that there needs to be as many exogenous instruments as endogenous variables for identification when using 2SLS, many papers in empirical finance instrument only the key variable of interest but then include, as though exogenous, an assortment of control variables that may themselves also be endogenous. We discuss the tradeoff between the omitted variable bias associated with not including these variables versus the bias created by including endogenous control variables. We suggest a new diagnostic test when thinking about this tradeoff in a 2SLS setting and suggest a way to calculate the maximum possible bias in the coefficient of interest coming from the control variables. Using simulated data and an empirical example from the diversification discount literature, we show how the new test and bias calculations can help researchers better understand and troubleshoot their 2SLS models.

Keywords: Endogeneity, 2SLS, IV Methods, Contaminated Controls, Diversification Discount

1 Introduction

In the presence of endogeneity it is challenging to identify the causal effect that a key variable of interest has on a specific outcome. A common empirical approach in this setting uses instrumental variables in two-stage-least-squares (2SLS) models to address the endogeneity. Standard practice often leads researchers to include an assortment of control variables in addition to the key variable of interest on the right-hand-side of the equation to mitigate the potential for omitted variable bias. The instrument(s) are typically well motivated in the various papers' discussions of the relevancy and exclusion conditions insofar as the instrument(s) relate specifically to the key endogenous variable of interest and the error term. But, in almost all of these papers, minimal consideration is given to the possibility that the other control variables might also be endogenous and hence also correlated with the error term. This ignored correlation—what we are calling contaminated controls—can have a direct and strong effect on the researcher's ability to draw inference from the 2SLS results if the endogenous control variables are also correlated with the instrument(s) used with the key variable of interest.

The idea that endogenous control variables create problems for identification is not new. The theoretical literature and various econometric textbooks clearly indicate that there needs to be at least as many excluded instruments as there are endogenous variables in order for the parameters in a system to be identified.[1] But it is clear from a survey of even recent empirical work that there is ongoing disagreement in practice about how best to operationalize this point with many researchers either including multiple control variables in 2SLS models with minimal discussion of their potential endogeneity or other researchers simply dropping the control variables altogether. Indeed, of the approximately 400 papers using 2SLS models we surveyed in the Journal of Finance, Journal of Financial Economics, and the Review of Financial Studies from 2010 - 2023 almost 70 percent of them provide minimal or no discussion of the potential endogeneity of the control

[1] For examples of several textbooks and papers that discuss this point see chapter 5 of Wooldridge (2002), chapter 8 of Davidson and MacKinnon (2004), Murray (2006), or section 3 of Roberts and Whited (2012).

variables or simply assert that the controls are exogenous without any supporting discussion.[2] Of the remaining 30 percent, most of these papers either don't tabulate the first-stage results or appear to simply drop all control variables from both stages. Thus, while the ideal is clearly to have at least as many excluded instruments as there are endogenous variables, the challenge in finding even one good instrument is apparently leading empirical researchers to compromise with a narrow focus on the key variable of interest while ignoring the effects that the other potentially endogenous control variables might have on the inference around the key variable of interest.

The above discussion highlights several questions that empirical researchers using 2SLS confront. For example, if the research focus is on one key variable of interest and there exists both a strong instrument for that specific variable as well as a set of potentially endogenous control variables that might also relate to the outcome of interest, is the researcher better off estimating the overall 2SLS model with or without the other control variables?[3] What effect does the inclusion of the other endogenous control variables have on the 2SLS estimate of the key variable of interest given a strong instrument for that one variable that itself is not correlated with the error term? Is it possible to quantify the potential bias in the estimated marginal effect of interest coming from the inclusion of specific control variables? Is there information to be gained by estimating the key 2SLS result both with and without the other control variables in the system and then comparing the results? If so, then what does the comparison reveal? Is there a statistical test that reveals whether specific control variables are endogenous and might be affecting the inference around the key variable of interest?[4] Is there a cost to using multiple instruments in an overidentified system if some of the instruments are correlated with the control variables? And, if more than one strong

[2]The set of papers using 2SLS in these journals were identified using Google Scholar searches and the following search strings: "two-stage", "two stage", "first-stage", "first stage", "2sls", "tsls", "exclusion", and "instrument(s)". This set of papers was then reviewed in more detail to ensure the paper used 2SLS. Not every paper that used these terms reported first-stage results or tabulated all of the control variables. Thus the summary numbers and percentages reported above are approximate based on the information provided.

[3]Note that this question is not about whether to drop the control variables from the first stage alone. Rather the question is about whether to drop the potentially endogenous control variables from the overall system of equations.

[4]Note that this question is not about whether the key variable of interest is endogenous. Rather it is about whether or not a control variable in the system might be creating bias in the 2SLS estimate for the key variable of interst.

instrument is available, but different instruments lead to different inferences for the key variable of interest, how should one decide which instrument should be used? Given that literally hundreds of papers at top finance journals have used 2SLS methods in recent years combined with (1) the widespread lack of consideration of the potential endogeneity of the control variables, (2) the implicit disagreement in practice evidenced by the existence of many recent papers that either include or exclude the control variables from the analysis, and (3) the common use of multiple instruments in overidentified 2SLS systems, there is obviously a need in the literature for a paper that discusses the exact tradeoffs involved in these decisions and provides clear practical advice for empirical researchers.

Our paper adds to the literature by addressing these questions and makes several contributions. First, we draw attention to a common problem affecting inference with 2SLS that has been largely ignored in recent empirical work. Given the prevalence of this problem, with literally hundreds of recent examples in top journals, a discussion of the issues and consequences of the inclusion or exclusion of contaminated controls for inference with 2SLS seems important. In exploring this issue we provide intuition from both analytical expressions for the bias related to endogenous control variables as well as simulation exercises. Second, we propose a new diagnostic test that allows researchers to directly test whether the contaminated controls problem might exist in their data. Thus, unlike the exclusion condition, which is not directly testable, it is possible to ascertain whether the inclusion of specific endogenous controls might be affecting the key estimate in specific models. As part of this discussion we also provide a formula for the maximum possible bias (MPB) in the main coefficient of interest coming from each endogenous control variable. The combination of the new statistical test together with the new MPB calculation will not only help researchers better understand how robust their 2SLS inferences are for their main variable of interest but will also direct their attention to which specific control variables need further consideration. To our knowledge, we are the first to propose both the test and the MPB calculations when thinking about inference in a 2SLS system. Third, using simulation studies we address the question of whether the

key 2SLS estimate is better estimated with or without the inclusion of the potentially endogeneous control variables. Although not contended in theory, this question is clearly contended in practice given many recent examples of papers that either include or drop the control variables. As part of this discussion we explore practical suggestions for what to do if the control variables are contaminated and show how using multiple instruments can lead to bias if some are correlated with the other control variables. The ability to test for possible contaminated control bias and to estimate the maximum size of the bias as a function of specific instruments and control variables provides a new and detailed way to evaluate which among several instruments in an overidentified system are likely providing the least biased 2SLS estimate for the key variable of interest, provides an explanation for why different instruments that are each "strong" may sometimes point to different 2SLS results, and it highlights a potential issue with using an overidentified model–as is common in practice if the researcher has more than one instrument–if one of the instruments is strongly correlated with the control variables.

In addition to the analytical and simulation-based results, we also provide an empirical example of our test based on a paper from the diversification discount literature published in the Journal of Finance. For this example, we first show 2SLS results using our sample that are similar to results published in the earlier paper suggesting that firms with multiple divisions experience a valuation premium rather than the diversification discount commonly reported in this literature. We then show that this unexpected result can be explained by the contaminated controls in the model and that using the diagnostic test and MPB calculations proposed in this paper would have identified the issue. We then use this example as the basis of a discussion for how researchers can explore their 2SLS results if they find possible evidence of contaminated control bias.

The paper is organized as follows. In Section 2 we describe the 2SLS estimator under ideal conditions and then provide a detailed description of how the inclusion of endogenous control variables affect these estimates. In Section 3, we propose a test for contaminated controls, and derive the relevant distribution for the test statistic. As part of this discussion we show the magnitude of

the problem using simulated data. In Section 4, we suggest a way to calculate the maximum possible bias that can occur in the key variable of interest due to the observed correlations. In Section 5, we run a series of simulations to validate the proposed test, and show the effect of contaminated controls on 2SLS estimates of the key variable of interest. In Section, 6 we present an empirical example to illustrate the use of our test and the expression for the bias. In Section 7, we conclude.

2 OLS and 2SLS estimates

To facilitate the discussion of how contaminated control variables affect 2SLS estimation, it is helpful first to briefly review the equations involved. In this section, we start with a general regression model and show the form of the bias created when using OLS to estimate marginal effects in the presence of endogeneity. The setting we consider is general and could be motivated using omitted variables, measurement error, or simulteneity (e.g., see discussion in section 4.1 Wooldridge (2002)). For our purposes we model the endogeneity as coming from an omitted variable in the discussion below. After showing the omitted variable bias in an OLS setting, we show the form of the bias in a 2SLS setting and demonstrate how the 2SLS bias is affected by the inclusion of endogenous controls. Finally, we present the relevant expressions for the general case of endogeneity.[5]

2.1 Bias in OLS estimates

Suppose we are interested in explaining the effect that a particular explanatory variable x_1 has on the outcome of interest y. Assume the data generating process for y is a function of x_1, x_2, and x_m

[5]Throughout this paper, when we use the term bias, we are referring to the asymptotic or large sample bias, computed using the probability limit of the estimator. We note that 2SLS estimators are known to be biased in finite samples but can be consistent in large samples (e.g., see Angrist and Krueger (2001) and chapter 5 of Wooldridge (2002)). Hence the focus in the literature on the large sample properties of 2SLS estimators.

as shown in equation 1 with $\mathrm{E}(u|x_1, x_2, x_m) = 0$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_m x_m + u \tag{1}$$

As is common in empirical research, assume the data generating process is only partially observable with x_m being omitted in this example. Thus the estimable model has an error term, $w = \beta_m x_m + u$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + w \tag{2}$$

The OLS estimate of β_1 is biased and inconsistent if $\mathrm{E}(w|x_1, x_2) \neq 0$. In a multivariate setting, the formula for the OLS estimate of β_1 measures the partial effect that x_1 has on y after netting out x_2. By application of the Frisch-Waugh-Lovell theorem the OLS estimate of β_1 is equivalent to a regression of y^* on x_1^*, which represent the residuals from regressions of y and x_1 on the other explanatory variables from the model (in this case just x_2), respectively. [6] Thus x_1^* is x_1 after partialling out the effects of the other control variables; x_1^* is the portion of x_1 uncorrelated with the other control variables; y^* is the portion of y uncorrelated with the other control variables (not including the key variable of interest x_1). Substituting the full model for y, from equation 1, into the $\hat{\beta}_1$ formula highlights the factors that affect the bias in the OLS estimate as shown in the equations below. We will use this "double residual regression" notation for the coefficients throughout the paper because this approach lends itself to intuitive analytical expressions for the bias that exists

[6] For a more detailed discussion of the multivariate OLS coefficient formula and the Frisch-Waugh-Lovell theorem see Wooldridge (2003) pages 78-79, Davidson and MacKinnon (2004) section 2.4, Greene (2003) page 27, and Lovell (1963). Using an astericks to identify the residual is similar to the notation used by Greene but written without the matrix notation. See Greene pg 27 for a matrix version of this formula. This idea is sometimes discussed as "the double residual regression" (e.g., see section 17.3 in Goldberger (1991) for example).

in the key 2SLS coefficient of interest. [7]

$$\begin{aligned}
\text{plim}\,\hat{\beta}_{1,OLS} &= \frac{cov(x_1^*, y^*)}{var(x_1^*)} = \frac{cov(x_1^*, y)}{var(x_1^*)} \\
&= \frac{\beta_1 cov(x_1^*, x_1) + \beta_2 cov(x_1^*, x_2) + \beta_3 cov(x_1^*, x_m) + cov(x_1^*, u)}{var(x_1^*)} \\
&= \beta_1 + \underbrace{\beta_m \frac{cov(x_1^*, x_m)}{var(x_1^*)}}_{\text{bias}} \qquad (3)
\end{aligned}$$

The OLS omitted variable bias is a function of 3 factors: (1) the marginal effect of the omitted variable β_m, (2) the covariance of the omitted variable, x_m, with the portion of the key variable of interest that is uncorrelated with the other controls, and (3) the variance in x_1^*. Intuitively, the size of the bias is increasing in the magnitude of the omitted marginal effect and in the covariance of the partial effect of the variable of interest with the omitted variable. The bias can be either positive or negative depending on the sign of the omitted marginal effect and the covariance between the variable of interest and the omitted variable.

2.2 Bias in 2SLS estimates with one control and one instrument

Empirical researchers often rely on instruments in a 2SLS framework to address the omitted variable bias. The first stage in the 2SLS model is a regression of the endogeneous variable of interest x_1 on the instrument z and control variable x_2. From this, we compute the fitted values $\hat{x}_1$.

$$\begin{aligned}
x_1 &= \gamma_0 + \gamma_1 z + \gamma_2 x_2 + e \\
\hat{x}_1 &= \hat{\gamma}_0 + \hat{\gamma}_1 z + \hat{\gamma}_2 x_2 \qquad (4)
\end{aligned}$$

[7] In simplifying the $\hat{\beta}_1$ expression, $\text{cov}(x_1^*, x_2) = 0$ given that the residuals x_1^* are orthogonal to x_2, $cov(x_1^*, x_1)/var(x_1^*) = 1$, and $\text{cov}(x_1^*, u)$ is assumed to be 0 given that $\text{E}(u|x_1, x_2, x_m) = 0$.

The second stage in 2SLS is a regression of y on $\hat{x}_1$ and x_2.

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 x_2 + v \tag{5}$$

The theoretical literature is clear on the conditions required for the 2SLS estimate of β_1 to be consistent. These conditions are discussed in econometric textbooks (e.g., see Wooldridge (2002), Angrist and Pischke (2009)) as well as in various well-known papers (e.g., see Bound et al. (1995), Angrist and Krueger (2001), Murray (2006), and Roberts and Whited (2012)) and typically focus on the relevancy and exclusion conditions. The relevancy condition requires that the instrument (strongly) correlate with the endogenous variable after controlling for the effects of the other variables, i.e $\hat{\gamma}_1 \neq 0$ in the first stage equation. The exclusion condition requires that the first stage regressors (the instrument and control variable) not be correlated with the error term, that is, $cov(z, v) = 0$ and $cov(x_2, v) = 0$. The exclusion condition is not directly testable and hence is motivated based on logic and theory. The relevancy and exclusion conditions together ensure that the $\mathrm{E}(v|\hat{x}_1, x_2) = 0$ in the second stage model.

Most of the empirical papers we surveyed in top finance journals in recent decades tend to discuss the exclusion condition solely in terms of whether the instrument for the key variable of interest is correlated with the error term, and not whether the other controls may also be correlated with the error. We will refer to this as the "narrow exclusion restriction" (i.e. $\mathrm{cov}(z, v) = 0$) to distinguish it from the complete set of exclusion conditions noted in the econometric textbooks and methodology papers. A researcher's empirical focus is often on a single key variable of interest, and most of the papers we surveyed discussed endogeneity and instrument(s) in terms of the focus variable only while the rest of the variables are carried along as "controls" without careful consideration of their potential correlation with either the error term or with the instruments. Indeed, of the almost 400 papers that use instrumental variables with 2SLS in the Journal of Finance, the Journal of Financial Economics, and the Review of Financial Studies between 2010 and 2023,

a large majority of them include various control variables with minimal or no discussion of the potential endogeneity of the control variables.

The potential bias due to endogeneity in the 2SLS estimate of β_1 is of similar form as in Equation 3 but with both $\hat{x}_1$ and $\hat{x}_1^*$ used in place of x_1 and x_1^*. Consistent with the astericks notation used above, $\hat{x}_1^*$ represents the portion of $\hat{x}_1$ uncorrelated with the other controls.

$$\begin{aligned}
\text{plim}\,\hat{\beta}_{1,2SLS} &= \frac{cov(\hat{x}_1^*, y)}{var(\hat{x}_1^*)} \\
&= \frac{\beta_1 cov(\hat{x}_1^*, x_1) + \beta_2 cov(\hat{x}_1^*, x_2) + \beta_m cov(\hat{x}_1^*, x_m) + cov(\hat{x}_1^*, u)}{var(\hat{x}_1^*)} \\
&= \beta_1 + \underbrace{\beta_m \frac{cov(\hat{x}_1^*, x_m)}{var(\hat{x}_1^*)}}_{\text{bias}}
\end{aligned} \tag{6}$$

Given the widespread inclusion of other control variables in 2SLS models in the literature without corresponding discussion of the control variables' potential correlation with the error term, a common implicit assumption in the literature must be that if an instrument z for the key variable of interest x_1 satisfies the narrow exclusion condition, i.e. if $cov(z,v) = 0$ or $cov(z,x_m) = 0$, then $cov(\hat{x}_1^*, x_m) = 0$. But this is not necessarily true. Indeed, in the discussion below we show that even if the narrow exclusion condition is satisfied with the $cov(z,v) = 0$ the bias in the main variable of interest can be non-zero if the controls are endogenous. In the above expressions, $\hat{x}_1^*$ is the residuals from the regression of $\hat{x}_1$ on the control variables (i.e., x_2 in this example), and hence is orthogonal to whatever control variables are included in the model. The relation between $\hat{x}_1$ and $\hat{x}_1^*$ is shown below for a model with a single control variable x_2.

$$\begin{aligned}
\hat{x}_1 &= \lambda_1 + \lambda_2 x_2 + \xi \\
&= \hat{\lambda}_1 + \hat{\lambda}_2 x_2 + \hat{x}_1^*
\end{aligned} \tag{7}$$

To facilitate understanding for how the 2SLS bias in Equation 6 is directly affected by endoge-

nous controls we rewrite $\hat{x}_1^*$ as a function of the control variable x_2 and instrument z. To do this we set the two expressions for $\hat{x}_1$ from Equations 4 and 7 equal and solve for $\hat{x}_1^*$.

$$\begin{aligned} \hat{\gamma}_0 + \hat{\gamma}_1 z + \hat{\gamma}_2 x_2 &= \hat{\lambda}_1 + \hat{\lambda}_2 x_2 + \hat{x}_1^* \\ \hat{x}_1^* &= (\hat{\gamma}_0 - \hat{\lambda}_1) + \hat{\gamma}_1 z + (\hat{\gamma}_2 - \hat{\lambda}_2) x_2 \end{aligned} \tag{8}$$

We now substitute Equation 8 into Equation 6 to show how the 2SLS bias is affected by endogenous controls – even in the case that the $cov(z, x_m) = 0$.

$$\begin{aligned} \text{plim}\, \hat{\beta}_{1,2SLS} &= \beta_1 + \beta_m \frac{cov(\hat{x}_1^*, x_m)}{var(\hat{x}_1^*)} \\ &= \beta_1 + \underbrace{\beta_m \frac{cov(\hat{\gamma}_1 z, x_m)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related} \\ \text{to the narrow} \\ \text{exclusion condition}}} + \underbrace{\beta_m (\hat{\gamma}_2 - \hat{\lambda}_2) \frac{cov(x_2, x_m)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related to} \\ \text{endogenous controls}}} \end{aligned} \tag{9}$$

The bias in the 2SLS estimate is a function of several factors: Focusing on the narrow exclusion condition related term, the bias is increasing in the magnitude of the covariance of the instrument z, with the the omitted variable x_m. Focusing on the relevancy condition, the size of the denominator in the bias expression is increasing in the strength of the instrument. To see this note that the $var(\hat{x}_1^*) = cov(\hat{x}_1^*, x_1) = cov(f(z), x_1)$. Thus the size of the denominator in the bias expression is increasing in $|(cov(z, x_1)|$. The second term on the right side of Equation 9 shows mechanically how both the exclusion and relevancy conditions affect the bias with weak instruments causing the denominator to be close to 0, and exclusion condition violations causing the numerator to be far different from zero. The last term in Equation 9 shows how the bias is also related to the covariance of the control variables with the omitted variable and hence highlights the cost of including endogenous control variables in a 2SLS model. Like the bias that comes from violations of the narrow exclusion condition, the bias in the key coefficient of interest that comes from the inclusion of endogenous control variables is also exacerbated by weak instruments.

There are two situations where the bias in $\hat{\beta}_{1,2SLS}$ from the control variables will be zero. The first situation occurs if the control variables are exogenous and hence $cov(x_2, x_m) = 0$. This outcome is not testable for the same reason that the exclusion condition is not testable: x_m is not observed. In contrast, the second situation is empirically testable and occurs when $cov(z, x_2) = 0$; when the instrument is not correlated with the control variable, $\hat{\gamma}_2 = \hat{\lambda}_2$ in Equation 9, causing the last term in the bias expression to be zero. Researchers can check whether their key estimate is possibly affected by endogenous control variable bias by checking whether $(\hat{\gamma}_2 - \hat{\lambda}_2)$ is close to 0. The new diagnostic test we propose in this paper is based on this intuition. If this difference is close to zero then the bias from the contaminated controls is small. In Section 3 we discuss the details of how to use this difference as a diagnostic test for contaminated control bias.

2.3 Bias in 2SLS estimates with multiple controls and instruments

We now generalize the results to the case of multiple control variables and instruments. In the case of multiple instruments, we assume again that x_1 is the key variable of interest and is the only variable being instrumented in a first stage equation. Suppose we have a vector of J controls $\boldsymbol{x_2} \equiv (x_{21}, \ldots, x_{2J})'$ and K instruments $\boldsymbol{z} \equiv (z_1, \ldots, z_K)'$. Let $\boldsymbol{\beta_2} \equiv (\beta_{21}, \ldots, \beta_{2J})'$. Generalizing Equations 1 and 2, the data generating process and estimable model are given by:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \boldsymbol{\beta_2'} \boldsymbol{x_2} + \beta_m x_m + u \\ y &= \beta_0 + \beta_1 x_1 + \boldsymbol{\beta_2'} \boldsymbol{x_2} + w \end{aligned} \tag{10}$$

Let $\boldsymbol{\lambda_2} \equiv (\lambda_{21}, \ldots, \lambda_{2J})'$, $\boldsymbol{\gamma_2} \equiv (\gamma_{21}, \ldots, \gamma_{2J})'$, and $\boldsymbol{\gamma_1} \equiv (\gamma_{11}, \ldots, \gamma_{1K})'$. Generalizing equations

4 and 5, the first and second stage estimates are given by:

$$\begin{aligned} x_1 &= \gamma_0 + \boldsymbol{\gamma_1'} \boldsymbol{z} + \boldsymbol{\gamma_2'} \boldsymbol{x_2} + e \\ \hat{x}_1 &= \hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}_1' \boldsymbol{z} + \hat{\boldsymbol{\gamma}}_2' \boldsymbol{x_2} \\ y &= \beta_0 + \beta_1 \hat{x}_1 + \boldsymbol{\beta_2'} \boldsymbol{x_2} + v \end{aligned} \tag{11}$$

Generalizing Equations 7 and 8, we solve for $\hat{x}_1^*$. Note that $\hat{x}_1^*$ now partials out the effects of all control variables $\boldsymbol{x_2}$.

$$\begin{aligned} \hat{x}_1 &= \lambda_1 + \boldsymbol{\lambda_2'} \boldsymbol{x_2} + \xi \\ &= \hat{\lambda}_1 + \hat{\boldsymbol{\lambda}}_2' \boldsymbol{x_2} + \hat{x}_1^* \\ \hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}_1' \boldsymbol{z} + \hat{\boldsymbol{\gamma}}_2' \boldsymbol{x_2} &= \hat{\lambda}_1 + \hat{\boldsymbol{\lambda}}_2' \boldsymbol{x_2} + \hat{x}_1^* \\ \hat{x}_1^* &= (\hat{\gamma}_0 - \hat{\lambda}_1) + \hat{\boldsymbol{\gamma}}_1' \boldsymbol{z} + (\hat{\boldsymbol{\gamma}}_2 - \hat{\boldsymbol{\lambda}}_2)' \boldsymbol{x_2} \end{aligned} \tag{12}$$

Generalizing Equation 9, the expression for the bias is given by:

$$\begin{aligned} \hat{\beta}_{1,2SLS} &= \beta_1 + \underbrace{\beta_m \frac{cov(\hat{x}_1^*, x_m)}{var(\hat{x}_1^*)}}_{\text{bias}} \\ &= \beta_1 + \underbrace{\beta_m \frac{cov(\hat{\boldsymbol{\gamma}}_1' \boldsymbol{z}, x_3)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related} \\ \text{to the narrow} \\ \text{exclusion condition}}} + \underbrace{\beta_m (\hat{\boldsymbol{\gamma}}_2 - \hat{\boldsymbol{\lambda}}_2)' \frac{cov(\boldsymbol{x_2}, x_m)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related to} \\ \text{endogenous controls}}} \end{aligned} \tag{13}$$

Thus the overall bias in the β_1 estimate coming from endogenous controls can come from as many channels as there are control variables, with some channels potentially increasing whereas others potentially decreasing the overall bias. Being able to test whether bias in the key variable of interest might be coming from each of the control variables could be useful in understanding the model. Alternatively it may be useful for applied researchers to perform a single test of the

net effect of all the control variables together. Either can be accomplished by testing whether the difference $(\hat{\boldsymbol{\gamma}}_{\mathbf{2}} - \hat{\boldsymbol{\lambda}}_{\mathbf{2}})$ is close to zero using a Wald test, with varying restrictions depending on the set of control variables to be tested. We note that this test is for a necessary condition for bias from contaminated controls and not for a sufficient condition; i.e, showing the difference is statistically different from zero signals that there may be contaminated control bias in the 2SLS estimate of interest whereas showing that the difference is not statistically different than zero indicates that there is negligible bias from the control variables even if they are also endogenous. Later in the paper we provide an analytical expression for the maximum possible size of this bias.

2.4 Bias in 2SLS estimates with no control variables

The implication from the above discussion is that even if an instrument is strongly correlated with the key endogenous variable of interest and even if the instrument itself is not correlated with the error term, the inclusion of other endogenous control variables in the system can cause the 2SLS estimate for the main variable of interest to be biased. Given the widespread inclusion of potentially endogenous control variables in 2SLS specifications even in recent applied work in top journals, this point has not been fully appreciated in the empirical literature.

One natural reaction to the prior discussion is to drop the potentially endogenous controls from the model. This can lead to other problems. The tradeoff is that dropping the controls can lead to omitted variable bias but including them leads to contaminated control bias that is exacerbated by weak instruments. This issue is contended in practice with some researchers actively advocating the inclusion of as many controls as possible whereas others implicitly disagreeing with this logic by showing their results without controls. So the question we consider in this section is if the control variables are possibly endogenous, then is it better to drop the controls from the 2SLS system?

We derive the analytical expression for the bias when the control variable is dropped, beginning with the case of a single control variable. Omitted variable bias is driven by the correlation between

the fitted values ($\hat{x}_1$ – now estimated without controls) used in the second stage model and the error term which now includes the effects of the omitted controls. Because the second stage model is now estimated without control variables, the $\hat{\beta}_1$ expression from Equation 9 would include $\hat{x}_1$ rather than $\hat{x}_1^*$. Unlike $\hat{x}_1^*$, which is orthogonal to x_2, $\hat{x}_1$ can be correlated with x_2, which is now part of the second stage error term. The expression for the bias is given by:

$$\begin{aligned}
\text{plim}\,\hat{\beta}_{1,2SLS\,without\,x_2} &= \frac{cov(\hat{x}_1, y)}{var(\hat{x}_1)} \\
&= \frac{\beta_1 cov(\hat{x}_1, x_1) + \beta_2 cov(\hat{x}_1, x_2) + \beta_m cov(\hat{x}_1, x_m) + cov(\hat{x}_1, u)}{var(\hat{x}_1)} \\
&= \beta_1 + \underbrace{\beta_2 \frac{cov(\hat{x}_1, x_2)}{var(\hat{x}_1)} + \beta_m \frac{cov(\hat{x}_1, x_m)}{var(\hat{x}_1)}}_{\text{bias}}
\end{aligned} \tag{14}$$

If the narrow exclusion restriction is satisfied, the last term is zero. The second term is zero only if x_2 is uncorrelated with the instrument z in which case $\hat{x_1}$ would be uncorrelated with x_2. Comparing the bias expressions in Equations 9 and 14, dropping the controls from the 2SLS system of equations does not guarantee in any way that the 2SLS estimate will be less biased without the endogenous controls than it is with the endogenous controls in the model. Indeed, without knowing the signs or sizes of β_m, $cov(\hat{x}_1, x_m)$, $cov(\hat{\gamma}_1 z, x_m)$, and $cov(x_2, x_m)$, all of which are unobservable, it is impossible to know whether dropping the endogenous control variable(s) results in an increase or decrease in the overall bias.

$$\underbrace{\beta_m \frac{cov(\hat{\gamma}_1 z, x_m)}{var(\hat{x}_1^*)} + \beta_m (\hat{\gamma}_2 - \hat{\lambda}_2) \frac{cov(x_2, x_m)}{var(\hat{x}_1^*)}}_{\text{bias including control variable}} \quad versus \quad \underbrace{\beta_2 \frac{cov(\hat{x}_1, x_2)}{var(\hat{x}_1)} + \beta_m \frac{cov(\hat{x}_1, x_m)}{var(\hat{x}_1)}}_{\text{bias not including control variable}} \tag{15}$$

It is worth noting that if both the narrow exclusion condition holds (or is almost satisfied) for the instrument on the key variable of interest and $(\hat{\gamma}_2 - \hat{\lambda}_2)$ is close to zero then the overall bias is likely smaller in the 2SLS estimate with controls than in the estimate without controls. It is also worth noting that one cannot conclude that the 2SLS estimate for β_1 estimated with controls

is biased based simply on whether the 2SLS estimate changes after dropping the controls from the system because the resulting change could be entirely attributable to omitted variable bias associated with the dropped variable(s) which were accounted for when the controls were included as part of the model but are not accounted for when estimating the model without controls. In the case of multiple control variables and instruments, Equations 14 and 15 generalize to:

$$\text{plim}\,\hat{\beta}_{1,2SLS\,without\,\boldsymbol{x_2}} = \beta_1 + \underbrace{\boldsymbol{\beta_2'}\frac{cov(\hat{x}_1,\boldsymbol{x_2})}{var(\hat{x}_1)} + \beta_m\frac{cov(\hat{x}_1,x_m)}{var(\hat{x}_1)}}_{\text{bias}}$$

$$\underbrace{\beta_m\frac{cov(\hat{\boldsymbol{\gamma}}_1'\boldsymbol{z},x_m)}{var(\hat{x}_1^*)} + \beta_m(\hat{\boldsymbol{\gamma}}_2 - \hat{\boldsymbol{\lambda}}_2)'\frac{cov(\boldsymbol{x_2},x_m)}{var(\hat{x}_1^*)}}_{\text{bias including control variables}} \quad versus \quad \underbrace{\boldsymbol{\beta_2'}\frac{cov(\hat{x}_1,\boldsymbol{x_2})}{var(\hat{x}_1)} + \beta_m\frac{cov(\hat{x}_1,x_m)}{var(\hat{x}_1)}}_{\text{bias not including control variables}} \tag{16}$$

It is possible for individual control variables to have opposite effects on the bias for estimates with and without control variables. As in the single control variable case, if the narrow exclusion condition holds for the key variable of interest, and $(\hat{\boldsymbol{\gamma}}_2 - \hat{\boldsymbol{\lambda}}_2)$ is close to zero, the bias is likely smaller in the 2SLS estimate with controls than in the estimate without controls

2.5 Bias for the general case of endogeneity

In this section, we derive the expressions for the general case of endogeneity, rather than the specific case of omitted variable bias. In the case of a single control variable, suppose the data generating process is given by Equation 2, where the error term *w* is potentially correlated with the

x_1 and x_2. Generalizing Equation 9, the bias is given by:

$$\begin{aligned}\text{plim}\,\hat{\beta}_{1,2SLS} &= \beta_1 + \underbrace{\frac{cov(\hat{x}_1^*, w)}{var(\hat{x}_1^*)}}_{\text{bias}} \\ &= \beta_1 + \underbrace{\frac{cov(\hat{\gamma}_1 z, w)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related} \\ \text{to the} \\ \text{exclusion condition}}} + \underbrace{(\hat{\gamma}_2 - \hat{\lambda}_2)\frac{cov(x_2, w)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related to} \\ \text{endogenous controls}}}\end{aligned} \tag{17}$$

With multiple control variables and instruments, suppose the data generating process is given by Equation 10, where the error term w is potentially correlated with the x_1 and $\boldsymbol{x_2}$. Generalizing Equation 13, the bias is given by:

$$\begin{aligned}\text{plim}\,\hat{\beta}_{1,2SLS} &= \beta_1 + \underbrace{\frac{cov(\hat{x}_1^*, w)}{var(\hat{x}_1^*)}}_{\text{bias}} \\ &= \beta_1 + \underbrace{\frac{cov(\hat{\boldsymbol{\gamma}}_1' \boldsymbol{z}, w)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related} \\ \text{to the} \\ \text{exclusion condition}}} + \underbrace{(\hat{\boldsymbol{\gamma}}_2 - \hat{\boldsymbol{\lambda}}_2)' \frac{cov(\boldsymbol{x_2}, w)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related to} \\ \text{endogenous controls}}}\end{aligned} \tag{18}$$

The relevant statistic for testing the impact of the endogeneous control variables is the same as when we derived the expressions for the specific case of omitted variable bias. Last, we derive the expression for the bias with and without control variables for the general case of endogeneity. Generalizing Equations 15 and 16:

$$\underbrace{\frac{cov(\hat{\gamma}_1 z, w)}{var(\hat{x}_1^*)} + (\hat{\gamma}_2 - \hat{\lambda}_2)\frac{cov(x_2, w)}{var(\hat{x}_1^*)}}_{\text{bias including control variable}} \quad versus \quad \underbrace{\beta_2 \frac{cov(\hat{x}_1, x_2)}{var(\hat{x}_1)} + \frac{cov(\hat{x}_1, w)}{var(\hat{x}_1)}}_{\text{bias not including control variable}} \tag{19}$$

$$\underbrace{\frac{cov(\hat{\boldsymbol{\gamma}}_1' \boldsymbol{z}, w)}{var(\hat{x}_1^*)} + (\hat{\boldsymbol{\gamma}}_2 - \hat{\boldsymbol{\lambda}}_2)' \frac{cov(\boldsymbol{x_2}, w)}{var(\hat{x}_1^*)}}_{\text{bias including control variables}} \quad versus \quad \underbrace{\boldsymbol{\beta}_2' \frac{cov(\hat{x}_1, \boldsymbol{x_2})}{var(\hat{x}_1)} + \frac{cov(\hat{x}_1, w)}{var(\hat{x}_1)}}_{\text{bias not including control variables}} \tag{20}$$

3 Testing for contaminated controls

In this section, we present our proposed test of contaminated controls, and derive the relevant test statistic. Our test is motivated by expressions for the bias derived in the previous section. The test is related to the concept of coefficient stability, i.e. the effect of the inclusion of instruments on first stage control variable coefficients in the first stage regression. In related work, Altonji et al. (2005) and Oster (2019) propose methods to estimate omitted variable bias based on coefficient movements after the inclusion of control variables in OLS regressions. Our test differs, in part, in that it examines coefficient stability in the first stage regression of a 2SLS system.

We begin with the case of one control variable and one instrument. We then proceed to the general case of multiple control controls. The test statistic focuses on the quantity $(\hat{\gamma}_2 - \hat{\lambda}_2)$ from Equation 9, or $(\hat{\boldsymbol{\gamma}}_{\mathbf{2}} - \hat{\boldsymbol{\lambda}}_{\mathbf{2}})$ from Equation 13 if there are multiple control variables, and tests whether this term is significantly different from zero. Under the null hypothesis, the expression is equal to zero and there is no contamination control bias in the estimate for the key coefficient of interest. If the null hypothesis is rejected, the term is different from zero, suggesting contamination, and the possibility of bias coming from endogenous control variables.

3.1 One control and one instrument

We need the distribution of $(\hat{\gamma}_2 - \hat{\lambda}_2)$ to be able to determine whether the difference is statistically different from zero. To find the distribution of $(\hat{\gamma}_2 - \hat{\lambda}_2)$ we need an estimate of the covariance between $\hat{\gamma}_2$ and $\hat{\lambda}_2$. This can be derived by estimating the coefficients jointly, using a modified version of the technique of seemingly unrelated regressions (SUR) of Zellner (1962). The SUR setup consists of a set of independent regression equations with correlated error terms. While the equations can be estimated independently using OLS, the original SUR method proposes estimating the regression equations jointly using feasible GLS to get more efficient parameter estimates. For our purposes, both $\hat{\gamma}_2$ and $\hat{\lambda}_2$ can be viewed as OLS parameter estimates of two regression

equations with correlated error terms. The correlation structure between the error terms can be derived analytically, allowing us to obtain the joint distribution of $\hat{\gamma}_2$ and $\hat{\lambda}_2$. Thus, we employ the SUR setup, but estimate the regressions individually by OLS.

First, in the case of one instrument and one control variable, we note that the OLS estimate $\hat{\lambda}_2$ can be computed by regressing x_1 on x_2 and a constant, rather than by regressing $\hat{x}_1$ on x_2 and a constant. Let $\hat{e}$ denote the residual from the first stage regression, as per Equation 4, so that $x_1 = \hat{x}_1 + \hat{e}$. The two approaches are numerically identical since $\hat{e}$ is orthogonal to x_2. To derive the correct distribution of $\hat{\lambda}_2$, we use x_1 rather than $\hat{x}_1$ as the dependent variable. Using $\hat{x}_1$ removes variation from the residual and results in standard errors that are too small.

$$\operatorname{plim} \hat{\lambda}_2 = \frac{Cov(x_1, x_2)}{Var(x_2)} = \frac{Cov(\hat{x}_1 + \hat{e}, x_2)}{Var(x_2)} = \frac{Cov(\hat{x}_1, x_2)}{Var(x_2)} \tag{21}$$

The SUR model stacks the observations of two regressions, and is set up as follows:

$$\begin{bmatrix} x_1 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 & z & x_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_2 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \lambda_1 \\ \lambda_2 \end{bmatrix} + \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \tag{22}$$

Suppose N is the sample size. Let $\boldsymbol{X}$ denote the matrix consisting of the observed data as per the model above, where the first N rows of $\boldsymbol{X}$ correspond to the first regression, and the second N rows to the second regression. The covariance matrix of the OLS estimates Σ is given by:

$$\Sigma = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{23}$$

Assuming that the errors in each regression are homoskedastic, letting I_N denote the identity

matrix of order N, and noting that $Cov(e,\varepsilon) = Var(e)$[8]:

$$\Omega = \begin{bmatrix} Var(e)\, I_N & Cov(e,\varepsilon)\, I_N \\ Cov(e,\varepsilon)\, I_N & Var(\varepsilon)\, I_N \end{bmatrix} = \begin{bmatrix} Var(e)\, I_N & Var(e)\, I_N \\ Var(e)\, I_n & Var(\varepsilon)\, I_N \end{bmatrix} \tag{24}$$

Let $\hat{\Sigma}$ denote the estimated covariance matrix, computed as the sample analog of Σ. Using a Wald test, under the null hypothesis that $\gamma_2 = \lambda_2$, the test statistic follows a chi-square distribution with one degree of freedom.

$$\begin{aligned} (\boldsymbol{R}\hat{\theta})'(\boldsymbol{R}\hat{\Sigma}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\theta}) &\sim \chi^2(1) \\ \hat{\theta} &\equiv (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\lambda}_1, \hat{\lambda}_2)' \\ \boldsymbol{R} &\equiv \begin{bmatrix} 0 & 0 & 1 & 0 & -1 \end{bmatrix} \end{aligned} \tag{25}$$

3.2 Multiple controls and instruments

We now derive and present the test statistic for multiple controls and instruments. Generalizing Equation 22, the two stacked regressions are as follows:

$$\begin{bmatrix} x_1 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{z}' & \boldsymbol{x_2}' & 0 & 0 \\ 0 & 0 & 0 & 1 & \boldsymbol{x_2}' \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \boldsymbol{\gamma_1} \\ \boldsymbol{\gamma_2} \\ \lambda_1 \\ \boldsymbol{\lambda_2} \end{bmatrix} + \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \tag{26}$$

The structure for the covariance matrix is identical to the single control variable case. The joint distribution of $(\hat{\boldsymbol{\gamma}}_{\boldsymbol{2}} - \hat{\boldsymbol{\lambda}}_{\boldsymbol{2}})$ asymptotically follows a chi-square distribution with J degrees of freedom. This quantity is a joint test that each $\gamma_{2j} = \lambda_{2j}$, where $j = 1, \ldots, J$. Let 0_J denote a

[8] $Cov(e,\varepsilon) = Cov(e, x_1 - \lambda_1 - \lambda_2 x_2) = Cov(e, x_1) = Cov(e, \gamma_0 + \gamma_1 z_1 + \gamma_2 x_2 + e) = Var(e)$

column vector of zeros of length J, 0_{JK} denote a (J by K) matrix of zeros (K is the number of instruments), and I_J the identity matrix of order J. Generalizing Equation 25, the test statistic for all of the control variables considered together is given by:

$$
\begin{aligned}
(\boldsymbol{R}\hat{\theta})'(\boldsymbol{R}\hat{\Sigma}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\theta}) &\sim \chi^2(J) \\
\hat{\theta} &\equiv (\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_{\mathbf{1}}', \hat{\boldsymbol{\gamma}}_{\mathbf{2}}', \hat{\lambda}_1, \hat{\boldsymbol{\lambda}}_{\mathbf{2}}')' \\
\boldsymbol{R} &\equiv \begin{bmatrix} 0_J & 0_{JK} & I_J & 0_J & -I_J \end{bmatrix}
\end{aligned} \tag{27}
$$

Rather than performing one joint test across all control variables, researchers can also test individual control variables within a multivariate setting or subsets of control variables using the following test statistics. Suppose we wish to test an individual control variable j. Let $\boldsymbol{R}_j$ denote the j^{th} row of the matrix $\boldsymbol{R}$. The test statistic is given by:

$$
(\boldsymbol{R}_j\hat{\theta})'(\boldsymbol{R}_j\hat{\Sigma}\boldsymbol{R}_j')^{-1}(\boldsymbol{R}_j\hat{\theta}) \sim \chi^2(1) \tag{28}
$$

To test subsets of control variables, let k denote the number of control variables to be tested, and $\boldsymbol{k}$ denote the row indices corresponding to the k control variables to be jointly tested. Let $\boldsymbol{R}_{\boldsymbol{k}}$ denote the $\boldsymbol{R}$ matrix with the relevant k rows. The test statistic is given by:[9]

$$
(\boldsymbol{R}_{\boldsymbol{k}}\hat{\theta})'(\boldsymbol{R}_{\boldsymbol{k}}\hat{\Sigma}\boldsymbol{R}_{\boldsymbol{k}}')^{-1}(\boldsymbol{R}_{\boldsymbol{k}}\hat{\theta}) \sim \chi^2(k) \tag{29}
$$

[9] A note on computing $\hat{\Sigma}$: The dimension of $\boldsymbol{X}$ is [$2N$ by $(2J+K+2)$]. $Var(e)$ and $Var(\varepsilon)$ can be estimated by computing the variance of the residuals using the first half and the second half of the observations from the SUR regression, respectively. Due to the high dimension of $\boldsymbol{\Omega}$ ($2N$ by $2N$), it is computationally preferable to directly compute the matrix $\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}$, which is [$(2J+K+2)$ by $(2J+K+2)$]. Let $\boldsymbol{X_1} \equiv (1, \boldsymbol{z}', \boldsymbol{x}_{\mathbf{2}}')$ and $\boldsymbol{X_2} \equiv (1, \boldsymbol{x}_{\mathbf{2}}')$, where $\boldsymbol{X_1}$ is [N by $(J+K+1)$] and $\boldsymbol{X_2}$ is [N by $(J+1)$]. Then $\boldsymbol{X}'\Omega\boldsymbol{X} = \begin{bmatrix} Var(e)\,\boldsymbol{X_1}'\boldsymbol{X_1} & Var(e)\,\boldsymbol{X_1}'\boldsymbol{X_2} \\ Var(e)\,\boldsymbol{X_2}'\boldsymbol{X_1} & Var(\varepsilon)\,\boldsymbol{X_2}'\boldsymbol{X_2} \end{bmatrix}$.

4 Maximum Possible Bias

4.1 Single Instrumented Variable

The diagnostic test derived in Section 3 allows researchers to test whether the correlations between the instruments and the non-instrumented control variables are large enough to potentially cause bias in the 2SLS estimate on the key variable of interest. This test is valid when a single endogenous variable is being instrumented. In this section we derive a formula to show how large the potential bias could be. Subtracting β_1 from both sides of Equation 9 provides an expression for bias in the 2SLS coefficient related to both the violation of the narrow exclusion condition and the presence of endogenous control variables. If we assume that the narrow exclusion condition holds, then the bias can be written a function of the endogenous controls as shown below.[10] Let σ_j, σ_w, σ_v, and σ_y denote the standard deviation of x_{2j}, w, v, and y, respectively, and let ρ_j denote the correlation between x_{2j} and v.

$$\begin{aligned}
\hat{\beta}_{1,2SLS} - \beta_1 &= (\hat{\gamma}_{\mathbf{2}} - \hat{\boldsymbol{\lambda}}_{\mathbf{2}})' \frac{cov(\boldsymbol{x_2}, w)}{var(\hat{x}_1^*)} \\
&= (\hat{\gamma}_{\mathbf{2}} - \hat{\boldsymbol{\lambda}}_{\mathbf{2}})' \frac{cov(\boldsymbol{x_2}, v)}{var(\hat{x}_1^*)} \\
&= \sum_{j=1}^{J} \frac{(\hat{\gamma}_{2j} - \hat{\lambda}_{2j})\sigma_j \sigma_v \rho_j}{var(\hat{x}_1^*)} \\
\hat{\beta}_{1,2SLS} - \beta_1 &\leq \sum_{j=1}^{J} \frac{(\hat{\gamma}_{2j} - \hat{\lambda}_{2j})\sigma_j \sigma_y \rho_j}{var(\hat{x}_1^*)}
\end{aligned} \tag{30}$$

To calculate the maximum possible bias coming from each control variable j (i.e., MPB_j), we make the simplifying assumption that the other control variables are not correlated with the error term. Let $\hat{\sigma}_v^2$ denote the least squares estimate of σ_v^2. To derive an analytical expression for MPB_j

[10] The bias can be written equivalently using the error term v from the second stage regression.

we start by first deriving the following expression relating σ_v^2 and $\hat{\sigma}_v^2$:[11]

$$\hat{\sigma}_v^2 = \sigma_v^2 - \text{Cov}(\boldsymbol{x}, v)' \text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, v) \tag{31}$$

Under the assumption that control variable j is correlated with the error term and the other controls are not, $\text{Cov}(\hat{x}_1, v) = \hat{\gamma}_{2j} \text{Cov}(x_{2j}, v)$. Let $\tilde{\gamma}_{2j}$ be a [$(J+1)$ by 1] vector with $\hat{\gamma}_{2j}$ the first element, 1 the $(j+1)^{th}$ element, and 0 the remaining elements. This implies that the second part of Equation 31 can be written as follows:

$$\begin{aligned} \text{Cov}(\boldsymbol{x}, v)' \text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, v) &= (\rho_j \sigma_j \sigma_v \tilde{x}_j)^2 \\ \tilde{x}_j &\equiv \sqrt{\tilde{\gamma}_{2j}' \text{Var}(\boldsymbol{x})^{-1} \tilde{\gamma}_{2j}} \end{aligned} \tag{32}$$

Next, if we assume that the control variables in the model explain some of the variation in the dependent variable we can assume that $\sigma_y \geq \sigma_v$.[12] This assumption allows the correlation to be

[11] Let $\boldsymbol{x} \equiv (\hat{x}_1, \boldsymbol{x}_2')'$ and $\boldsymbol{\beta} \equiv (\beta_1, \beta_2')'$. Let $\hat{v}$, $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ denote the least squares estimates of v, β_0 and $\boldsymbol{\beta}$, respectively.

$$\begin{aligned} y &= \beta_0 + \boldsymbol{\beta}' \boldsymbol{x} + v \\ y &= \hat{\beta}_0 + \hat{\boldsymbol{\beta}}' \boldsymbol{x} + \hat{v} \\ \sigma_y^2 &= \boldsymbol{\beta}' \text{Var}(\boldsymbol{x}) \boldsymbol{\beta} + 2\boldsymbol{\beta}' \text{Cov}(\boldsymbol{x}, v) + \sigma_v^2 \\ \sigma_y^2 &= \hat{\boldsymbol{\beta}}' \text{Var}(\boldsymbol{x}) \hat{\boldsymbol{\beta}} + + \hat{\sigma}_v^2 \\ \sigma_y^2 &= [\text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, y)]' \text{Var}(\boldsymbol{x}) [\text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, y)] + \hat{\sigma}_v^2 \\ \sigma_y^2 &= [\boldsymbol{\beta} + \text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, v)]' \text{Var}(\boldsymbol{x}) [\boldsymbol{\beta} + \text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, v)] + \hat{\sigma}_v^2 \\ \sigma_y^2 &= \boldsymbol{\beta}' \text{Var}(\boldsymbol{x}) \boldsymbol{\beta} + 2\boldsymbol{\beta}' \text{Cov}(\boldsymbol{x}, v) + \text{Cov}(\boldsymbol{x}, v)' \text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, v) + \hat{\sigma}_v^2 \\ \sigma_v^2 &= \hat{\sigma}_v^2 + \text{Cov}(\boldsymbol{x}, v)' \text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, v) \\ \hat{\sigma}_v^2 &= \sigma_v^2 - \text{Cov}(\boldsymbol{x}, v)' \text{Var}(\boldsymbol{x})^{-1} \text{Cov}(\boldsymbol{x}, v) \end{aligned}$$

[12] This rules out certain cases of severe correlation between the error term and the regressors.

bounded as follows:

$$\begin{aligned}\hat{\sigma}_v^2 &= \sigma_v^2[1-(\rho_j\sigma_j\tilde{x}_j)^2] \\ \hat{\sigma}_v^2 &\leq \sigma_y^2[1-(\rho_j\sigma_j\tilde{x}_j)^2] \\ |\rho_j| &\leq (\sigma_y\sigma_j\tilde{x}_j)^{-1}\sqrt{\sigma_y^2-\hat{\sigma}_v^2}\end{aligned} \tag{33}$$

If we then substitute Equation 33 into Equation 30 we can bound the total bias and thereby define the maximum possible bias MPB_j as follows:

$$|\hat{\beta}_{1,2SLS}-\beta_1| \leq \left|\frac{(\hat{\gamma}_{2j}-\hat{\lambda}_{2j})\sqrt{\sigma_y^2-\hat{\sigma}_v^2}}{var(\hat{x}_1^*)\,\tilde{x}_j}\right| \equiv MPB_j \tag{34}$$

4.2 Multiple Instrumented Variables

Thus far in the paper, we have assumed 2SLS system includes a single instrumented endogenous variable. In this section, we allow for multiple instrumented endogenous variables and derive expressions for both the bias coming from contaminated controls and the maximum possible bias. Future research is needed to be able to define the appropriate test statistic in a setting with multiple first stages. The appropriate test statistic in this type of setting would be nonlinear and different from the test proposed in this paper for a single endogenous variable.

Let $\boldsymbol{x_3}$ denote an endogenous (G by 1) vector of control variables that are instrumented with a vector of instruments. $\boldsymbol{x_3} \equiv (x_{31},\ldots,x_{3G})'$.[13] Note that $K \geq G+1$. Suppose the true DGP and estimable model are given by:

$$\begin{aligned}y &= \beta_0+\beta_1x_1+\boldsymbol{\beta_2'x_2}+\boldsymbol{\beta_3'x_3}+\beta_m x_m+u \\ y &= \beta_0+\beta_1x_1+\boldsymbol{\beta_2'x_2}+\boldsymbol{\beta_3'x_3}+w\end{aligned} \tag{35}$$

[13]The remaining variables have identical definitions as in previous sections.

The first stage regressions and estimates for x_1 are given by:

$$\begin{aligned} x_1 &= \gamma_0 + \boldsymbol{\gamma_1'} \boldsymbol{z} + \boldsymbol{\gamma_2'} \boldsymbol{x_2} + e_1 \\ \hat{x}_1 &= \hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}_\mathbf{1}' \boldsymbol{z} + \hat{\boldsymbol{\gamma}}_\mathbf{2}' \boldsymbol{x_2} \end{aligned} \tag{36}$$

Let $\boldsymbol{\delta_0}$ be a (G by 1) vector, $\boldsymbol{\delta_1}$ be a (G by K) matrix, and $\boldsymbol{\delta_2}$ be a (G by J) matrix, of parameters. $\boldsymbol{e_3}$ is a (G by 1) vector of errors. The first stage regressions and estimates for $\boldsymbol{x_3}$ are given by:

$$\begin{aligned} \boldsymbol{x_3} &= \boldsymbol{\delta_0} + \boldsymbol{\delta_1} \boldsymbol{z} + \boldsymbol{\delta_2} \boldsymbol{x_2} + \boldsymbol{e_3} \\ \hat{\boldsymbol{x}}_\mathbf{3} &= \hat{\boldsymbol{\delta}}_\mathbf{0} + \hat{\boldsymbol{\delta}}_\mathbf{1} \boldsymbol{z} + \hat{\boldsymbol{\delta}}_\mathbf{2} \boldsymbol{x_2} \end{aligned} \tag{37}$$

Let $\boldsymbol{\lambda_3}$ be a (G by 1) vector of parameters. $\hat{x}_1^*$ is the portion of $\hat{x}_1$ uncorrelated with $\boldsymbol{x_2}$ and $\hat{\boldsymbol{x}}_\mathbf{3}$.

$$\hat{x}_1 = \lambda_1 + \boldsymbol{\lambda_2'} \boldsymbol{x_2} + \boldsymbol{\lambda_3'} \hat{\boldsymbol{x}}_\mathbf{3} + \xi = \hat{\lambda}_1 + \hat{\boldsymbol{\lambda}}_\mathbf{2}' \boldsymbol{x_2} + \hat{\boldsymbol{\lambda}}_\mathbf{3}' \hat{\boldsymbol{x}}_\mathbf{3} + \hat{x}_1^* \tag{38}$$

Equating Equations 36 and 38, we solve for $\hat{x}_1^*$.

$$\begin{aligned} \hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}_\mathbf{1}' \boldsymbol{z} + \hat{\boldsymbol{\gamma}}_\mathbf{2}' \boldsymbol{x_2} &= \hat{\lambda}_1 + \hat{\boldsymbol{\lambda}}_\mathbf{2}' \boldsymbol{x_2} + \hat{\boldsymbol{\lambda}}_\mathbf{3}' \hat{\boldsymbol{x}}_\mathbf{3} + \hat{x}_1^* \\ \hat{x}_1^* &= (\hat{\gamma}_0 - \hat{\lambda}_1) + \hat{\boldsymbol{\gamma}}_\mathbf{1}' \boldsymbol{z} + (\hat{\boldsymbol{\gamma}}_\mathbf{2} - \hat{\boldsymbol{\lambda}}_\mathbf{2})' \boldsymbol{x_2} - \hat{\boldsymbol{\lambda}}_\mathbf{3}' \hat{\boldsymbol{x}}_\mathbf{3} \end{aligned} \tag{39}$$

Substituting using Equation 37, we solve for $\hat{x}_1^*$ as a function of $\boldsymbol{z}$, $\boldsymbol{x_2}$, and the parameters.

$$\begin{aligned} \hat{x}_1^* &= (\hat{\gamma}_0 - \hat{\lambda}_1) + \hat{\boldsymbol{\gamma}}_\mathbf{1}' \boldsymbol{z} + (\hat{\boldsymbol{\gamma}}_\mathbf{2} - \hat{\boldsymbol{\lambda}}_\mathbf{2})' \boldsymbol{x_2} - \hat{\boldsymbol{\lambda}}_\mathbf{3}' (\hat{\boldsymbol{\delta}}_\mathbf{0} + \hat{\boldsymbol{\delta}}_\mathbf{1} \boldsymbol{z} + \hat{\boldsymbol{\delta}}_\mathbf{2} \boldsymbol{x_2}) \\ &= (\hat{\gamma}_0 - \hat{\lambda}_1 - \hat{\boldsymbol{\delta}}_\mathbf{0}' \hat{\boldsymbol{\lambda}}_\mathbf{3}) + (\hat{\boldsymbol{\gamma}}_\mathbf{1} - \hat{\boldsymbol{\delta}}_\mathbf{1}' \hat{\boldsymbol{\lambda}}_\mathbf{3})' \boldsymbol{z} + (\hat{\boldsymbol{\gamma}}_\mathbf{2} - \hat{\boldsymbol{\lambda}}_\mathbf{2} - \hat{\boldsymbol{\delta}}_\mathbf{2}' \hat{\boldsymbol{\lambda}}_\mathbf{3})' \boldsymbol{x_2} \end{aligned} \tag{40}$$

The second stage estimates and expression for the bias are given by:

$$\begin{aligned}\hat{\beta}_{1,2SLS} &= \beta_1 + \underbrace{\beta_m \frac{cov(\hat{x}_1^*, x_m)}{var(\hat{x}_1^*)}}_{\text{bias}} \\ &= \beta_1 + \underbrace{\beta_m \frac{cov[(\hat{\gamma}_1 - \hat{\delta}_1' \hat{\lambda}_3)' z, x_3]}{var(\hat{x}_1^*)}}_{\substack{\text{bias related} \\ \text{to the narrow} \\ \text{exclusion condition}}} + \underbrace{\beta_m (\hat{\gamma}_2 - \hat{\lambda}_2 - \hat{\delta}_2' \hat{\lambda}_3)' \frac{cov(x_2, x_m)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related to} \\ \text{endogenous controls}}}\end{aligned} \tag{41}$$

For the general case of endogeneity, the bias is given by:

$$\hat{\beta}_{1,2SLS} = \beta_1 + \underbrace{\frac{cov[(\hat{\gamma}_1 - \hat{\delta}_1' \hat{\lambda}_3)' z, w]}{var(\hat{x}_1^*)}}_{\substack{\text{bias related} \\ \text{to the narrow} \\ \text{exclusion condition}}} + \underbrace{(\hat{\gamma}_2 - \hat{\lambda}_2 - \hat{\delta}_2' \hat{\lambda}_3)' \frac{cov(x_2, w)}{var(\hat{x}_1^*)}}_{\substack{\text{bias related to} \\ \text{endogenous controls}}} \tag{42}$$

A test for contaminated controls would focus on the quantity $(\hat{\gamma}_2 - \hat{\lambda}_2 - \hat{\delta}_2' \hat{\lambda}_3)$. Unlike the case of a single instrumented variable, this quantity is nonlinear in the parameters, and would require a different and more complex test. We leave this as an area of future research.

Last, we derive the maximum possible bias, defined in the same way as the previous section. Let $\hat{\delta}_{2j}$ denote the j^{th} column of $\hat{\delta}_2$. $\tilde{\gamma}_{2j}$ is now a [$(J+G+1)$ by 1] vector with $\hat{\gamma}_{2j}$ the first element, 1 the $(j+1)^{th}$ element, $\hat{\delta}_{2j}$ the last G elements, and 0 the remaining elements. Generalizing Equations 30 and 34, the maximum possible bias is given by:

$$\begin{aligned}\hat{\beta}_{1,2SLS} - \beta_1 &= (\hat{\gamma}_2 - \hat{\lambda}_2 - \hat{\delta}_2' \hat{\lambda}_3)' \frac{cov(x_2, w)}{var(\hat{x}_1^*)} \\ \hat{\beta}_{1,2SLS} - \beta_1 &\leq \sum_{j=1}^{J} \frac{(\hat{\gamma}_{2j} - \hat{\lambda}_{2j} - \hat{\delta}_{2j}' \hat{\lambda}_3) \sigma_j \sigma_y \rho_j}{var(\hat{x}_1^*)} \\ MPB_j &\equiv \left| \frac{(\hat{\gamma}_{2j} - \hat{\lambda}_{2j} - \hat{\delta}_{2j}' \hat{\lambda}_3) \sqrt{\sigma_y^2 - \hat{\sigma}_v^2}}{var(\hat{x}_1^*) \tilde{x}_j} \right|\end{aligned} \tag{43}$$

5 Simulation

We use simulated data to assess the performance of our proposed test, and examine how the inclusion of contaminated controls affects the bias of the 2SLS of the key variable of interest. We start with a simple case and then consider a variety of extensions and robustness checks. We vary the magnitude of the correlations and the instrument strength in the simulations to provide better understanding for what matters in practice.

5.1 Baseline

We start the simulation exercise considering a setting similar to that observed in empirical work, as described in Equation 1, with one key endogenous variable of interest x_1, one control variable x_2, one omitted variable x_m, and an exogenous instrument z for x_1. In generating the data, the instrument z is constructed to be correlated with x_1 but not correlated with the omitted factor(s); the instrument thus satisfies the narrow exclusion condition. We set β_1, β_2, and β_m equal to 1, and x_1, x_2, x_m, z, and u are all created as standard normal distributions with the correlation structure shown in Table 1. Consistent with real-world data, we allow low-level correlations to exist between x_1 and x_2 as well as between z and x_2. We consider multiple correlation values in some tests, and hence some cells shown below contain multiple values. The sample size is 10,000 and we run 100,000 simulations for each specification.

Table 1: Correlations: one instrument one control

	x_1	x_2	x_m	z	u
x_1	1	0.1 / 0.3	0.3	0.1 / 0.3	0
x_2	0.1 / 0.3	1	0.1 / 0.3	0.0 / 0.2	0
x_m	0.3	0.1 / 0.3	1	0	0
z	0.1 / 0.3	0.0 / 0.2	0	1	0
u	0	0	0	0	1

This table reports the correlation matrix for the baseline simulation with one instrument and and one control.

Table 2 reports the simulation results for one control variable. Figures 1 and 2 plot the distri-

bution of the test statistic, and Figures 3 and 4 plot the distribution of the 2SLS estimate of β_1. For each simulation, we report the fraction of times the null hypothesis is rejected, for tests at the 10%, 5%, and 1% levels. When there is no correlation between x_2 and z, this number should be approximately equal to the significance level. When there is a nonzero correlation between x_2 and z, this is equal to the power of the test, with higher fractions indicating better performance. We also report the bias in the estimate of β_1, when the control variable is included, and when it is excluded. When there is no correlation between x_2 and z, the bias should be approximately equal to zero. Finally, we report the maximum possible bias, computed as the average across all simulations.

Table 2: Simulation results: one instrument one control

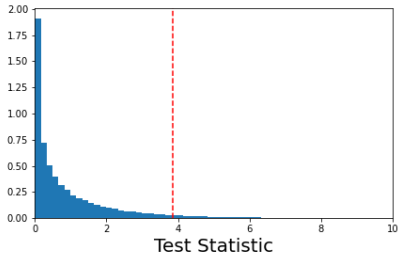
	Model Specification				(1)	(2)	(3)	(4)	(5)	(6)
	$\rho_{x_2,z}$	$\rho_{x_1,z}$	ρ_{x_1,x_2}	ρ_{x_2,x_m}	$R_{0.10}$	$R_{0.05}$	$R_{0.01}$	bias	bias$_{nc}$	mpb
(1)	0.0	0.1	0.1	0.1	0.097	0.046	0.008	-0.004	-0.005	0.097
(2)	0.0	0.1	0.1	0.3	0.096	0.046	0.007	-0.003	-0.004	0.113
(3)	0.0	0.1	0.3	0.1	0.096	0.047	0.008	-0.003	-0.007	0.113
(4)	0.0	0.1	0.3	0.3	0.096	0.046	0.008	-0.002	-0.006	0.129
(5)	0.0	0.3	0.1	0.1	0.100	0.050	0.010	-0.000	-0.000	0.033
(6)	0.0	0.3	0.1	0.3	0.102	0.050	0.010	-0.000	-0.000	0.038
(7)	0.0	0.3	0.3	0.1	0.099	0.049	0.010	-0.000	-0.001	0.038
(8)	0.0	0.3	0.3	0.3	0.099	0.049	0.009	-0.000	-0.001	0.043
(9)	0.2	0.1	0.1	0.1	1.000	1.000	1.000	-0.258	2.016	2.989
(10)	0.2	0.1	0.1	0.3	1.000	1.000	1.000	-0.766	2.016	3.483
(11)	0.2	0.1	0.3	0.1	0.996	0.990	0.954	-0.552	2.014	7.328
(12)	0.2	0.1	0.3	0.3	0.996	0.990	0.954	-1.617	2.014	8.370
(13)	0.2	0.3	0.1	0.1	1.000	1.000	1.000	-0.072	0.667	0.861
(14)	0.2	0.3	0.1	0.3	1.000	1.000	1.000	-0.215	0.667	0.994
(15)	0.2	0.3	0.3	0.1	1.000	1.000	1.000	-0.085	0.666	1.160
(16)	0.2	0.3	0.3	0.3	1.000	1.000	1.000	-0.251	0.667	1.317

This table reports basline simulation results for one instrument and one control variable. The first four columns determine the model specification. The next three columns report the rejection rate at the 10%, 5%, and 1% levels, respectively. The next two columns report the bias in the 2SLS estimate of β_1 with and without (nc) the inclusion of the control variable. The last column reports the maximum possible bias.

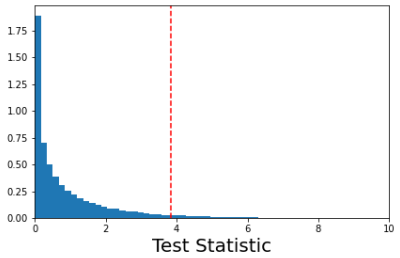
In the models reported in the first half of Table 2, there is no correlation between the control variable x_2 and the instrument in the simulated data. Hence rejection rates should be equal the significance level of the test, and the bias should be zero with and without controls. Our results

Figure 1: Test statistic: $\rho(x_2, z) = 0$ (no contamination)

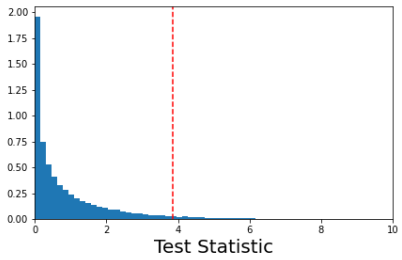
(a) $\rho(x_1, z) = 0.1$, $\rho(x_1, x_2) = 0.1$, $\rho(x_2, x_m) = 0.1$
weak instrument, less relevant control, low endogeneity



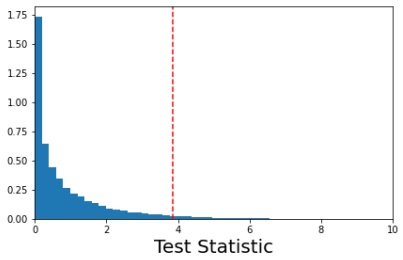
(b) $\rho(x_1, z) = 0.1$, $\rho(x_1, x_2) = 0.1$, $\rho(x_2, x_m) = 0.3$
weak instrument, less relevant control, high endogeneity



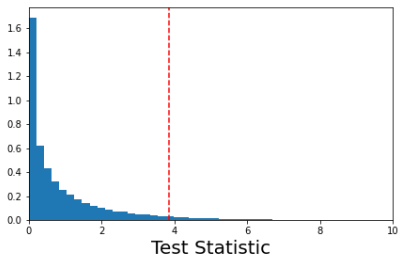
(c) $\rho(x_1, z) = 0.1$, $\rho(x_1, x_2) = 0.3$, $\rho(x_2, x_m) = 0.1$
weak instrument, more relevant control, low endogeneity



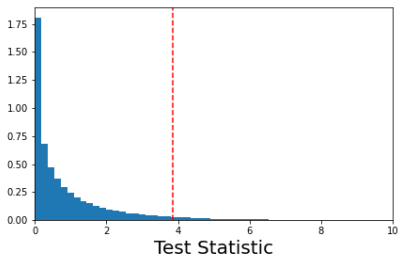
(d) $\rho(x_1, z) = 0.1$, $\rho(x_1, x_2) = 0.3$, $\rho(x_2, x_m) = 0.3$
weak instrument, more relevant control, high endogeneity



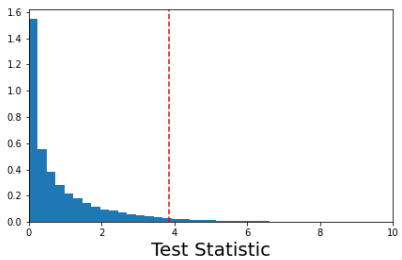
(e) $\rho(x_1, z) = 0.3$, $\rho(x_1, x_2) = 0.1$, $\rho(x_2, x_m) = 0.1$
strong instrument, less relevant control, low endogeneity



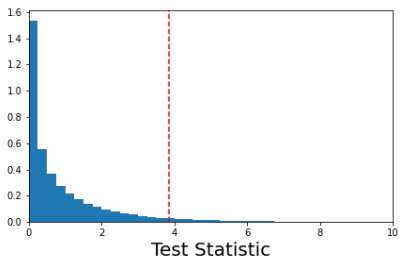
(f) $\rho(x_1, z) = 0.3$, $\rho(x_1, x_2) = 0.1$, $\rho(x_2, x_m) = 0.3$
strong instrument, less relevant control, high endogeneity



(g) $\rho(x_1, z) = 0.3$, $\rho(x_1, x_2) = 0.3$, $\rho(x_2, x_m) = 0.1$
strong instrument, more relevant control, low endogeneity



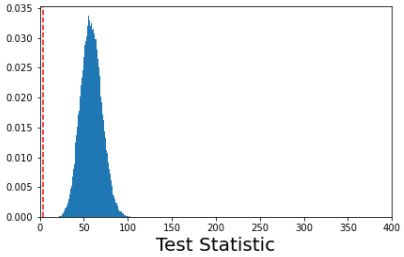
(h) $\rho(x_1, z) = 0.3$, $\rho(x_1, x_2) = 0.3$, $\rho(x_2, x_m) = 0.3$
strong instrument, more relevant control, high endogeneity



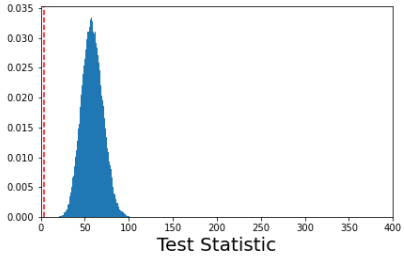
The figures plot the distribution of the test statistic when the instrument is uncorrelated with the control variable. The vertical dashed red line is the critical value for a 5% significance level hypothesis test.

Figure 2: Test statistic: $\rho(x_2, z) = 0.2$ (contaminated control)

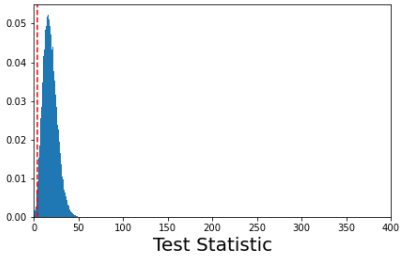
(a) $\rho(x_1,z) = 0.1$, $\rho(x_1,x_2) = 0.1$, $\rho(x_2,x_m) = 0.1$
weak instrument, less relevant control, low endogeneity



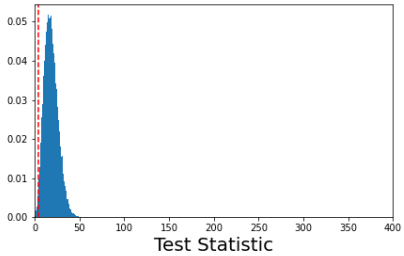
(b) $\rho(x_1,z) = 0.1$, $\rho(x_1,x_2) = 0.1$, $\rho(x_2,x_m) = 0.3$
weak instrument, less relevant control, high endogeneity



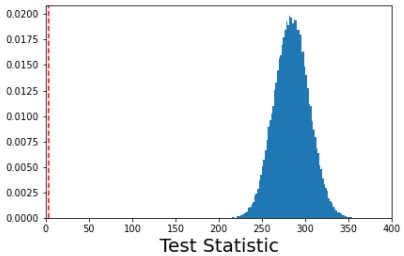
(c) $\rho(x_1,z) = 0.1$, $\rho(x_1,x_2) = 0.3$, $\rho(x_2,x_m) = 0.1$
weak instrument, more relevant control, low endogeneity



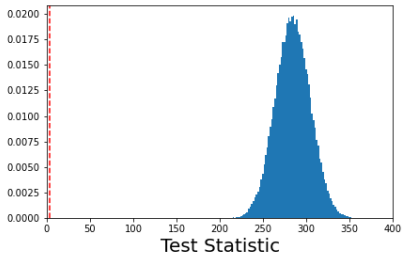
(d) $\rho(x_1,z) = 0.1$, $\rho(x_1,x_2) = 0.3$, $\rho(x_2,x_m) = 0.3$
weak instrument, more relevant control, high endogeneity



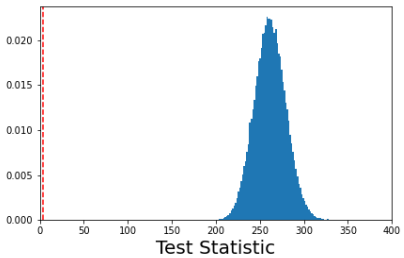
(e) $\rho(x_1,z) = 0.3$, $\rho(x_1,x_2) = 0.1$, $\rho(x_2,x_m) = 0.1$
strong instrument, less relevant control, low endogeneity



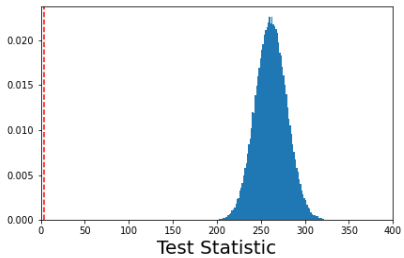
(f) $\rho(x_1,z) = 0.3$, $\rho(x_1,x_2) = 0.1$, $\rho(x_2,x_m) = 0.3$
strong instrument, less relevant control, high endogeneity



(g) $\rho(x_1,z) = 0.3$, $\rho(x_1,x_2) = 0.3$, $\rho(x_2,x_m) = 0.1$
strong instrument, more relevant control, low endogeneity



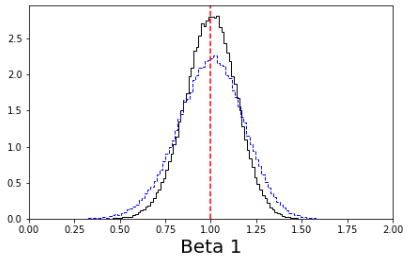
(h) $\rho(x_1,z) = 0.3$, $\rho(x_1,x_2) = 0.3$, $\rho(x_2,x_m) = 0.3$
strong instrument, more relevant control, high endogeneity



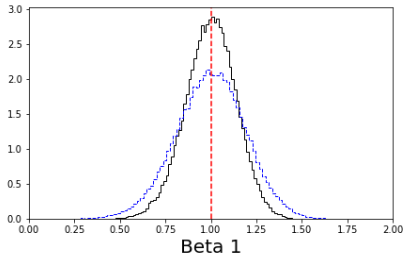
The figures plot the distribution of the test statistic when the instrument is correlated with the control variable. The vertical dashed red line is the critical value for a 5% significance level hypothesis test.

Figure 3: 2SLS estimate: $\rho(x_2, z) = 0$ (no contamination)

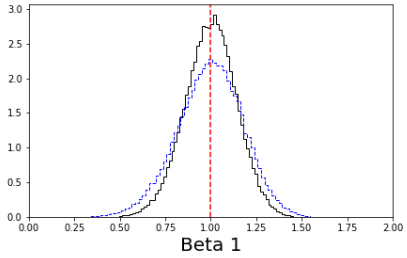
(a) $\rho(x_1,z) = 0.1, \rho(x_1,x_2) = 0.1, \rho(x_2,x_m) = 0.1$
weak instrument, less relevant control, low endogeneity



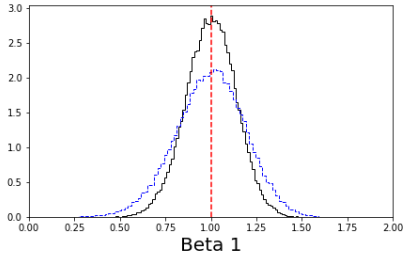
(b) $\rho(x_1,z) = 0.1, \rho(x_1,x_2) = 0.1, \rho(x_2,x_m) = 0.3$
weak instrument, less relevant control, high endogeneity



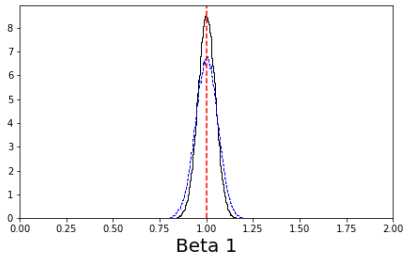
(c) $\rho(x_1,z) = 0.1, \rho(x_1,x_2) = 0.3, \rho(x_2,x_m) = 0.1$
weak instrument, more relevant control, low endogeneity



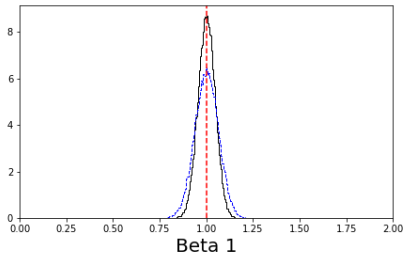
(d) $\rho(x_1,z) = 0.1, \rho(x_1,x_2) = 0.3, \rho(x_2,x_m) = 0.3$
weak instrument, more relevant control, high endogeneity



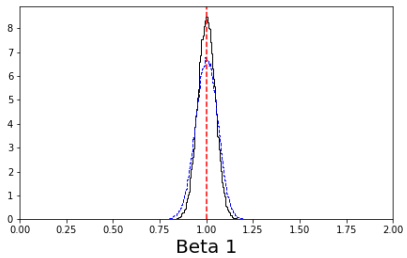
(e) $\rho(x_1,z) = 0.3, \rho(x_1,x_2) = 0.1, \rho(x_2,x_m) = 0.1$
strong instrument, less relevant control, low endogeneity



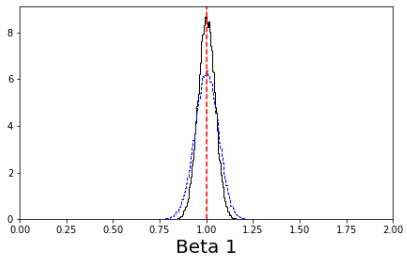
(f) $\rho(x_1,z) = 0.3, \rho(x_1,x_2) = 0.1, \rho(x_2,x_m) = 0.3$
strong instrument, less relevant control, high endogeneity



(g) $\rho(x_1,z) = 0.3, \rho(x_1,x_2) = 0.3, \rho(x_2,x_m) = 0.1$
strong instrument, more relevant control, low endogeneity



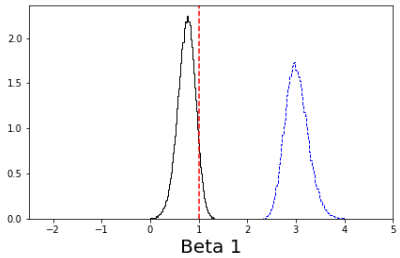
(h) $\rho(x_1,z) = 0.3, \rho(x_1,x_2) = 0.3, \rho(x_2,x_m) = 0.3$
strong instrument, more relevant control, high endogeneity



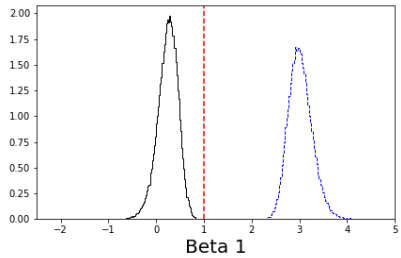
The figures plot the distribution of the 2SLS estimate of β_1 when the instrument is uncorrelated with the control variable, with (solid black) and without (dashed blue) the inclusion of the control variable in the regression. The vertical dashed red line is the true value of β_1.

Figure 4: 2SLS estimate: $\rho(x_2, z) = 0.2$ (contaminated control)

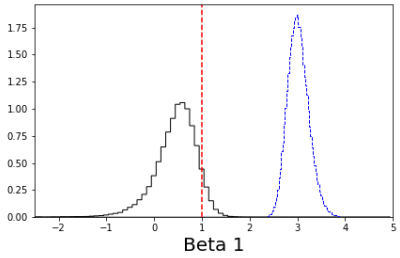
(a) $\rho(x_1, z) = 0.1$, $\rho(x_1, x_2) = 0.1$, $\rho(x_2, x_m) = 0.1$
weak instrument, less relevant control, low endogeneity



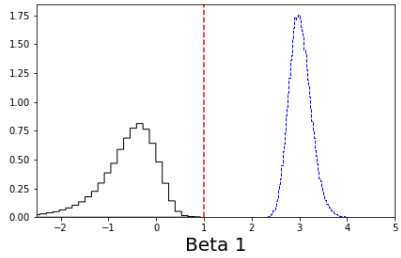
(b) $\rho(x_1, z) = 0.1$, $\rho(x_1, x_2) = 0.1$, $\rho(x_2, x_m) = 0.3$
weak instrument, less relevant control, high endogeneity



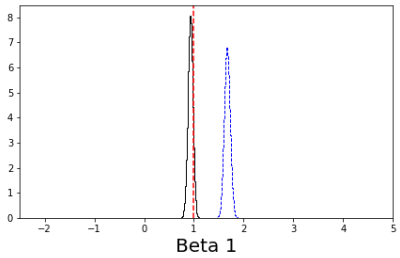
(c) $\rho(x_1, z) = 0.1$, $\rho(x_1, x_2) = 0.3$, $\rho(x_2, x_m) = 0.1$
weak instrument, more relevant control, low endogeneity



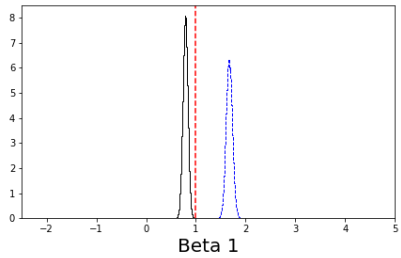
(d) $\rho(x_1, z) = 0.1$, $\rho(x_1, x_2) = 0.3$, $\rho(x_2, x_m) = 0.3$
weak instrument, more relevant control, high endogeneity



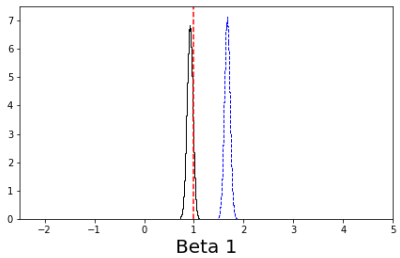
(e) $\rho(x_1, z) = 0.3$, $\rho(x_1, x_2) = 0.1$, $\rho(x_2, x_m) = 0.1$
strong instrument, less relevant control, low endogeneity



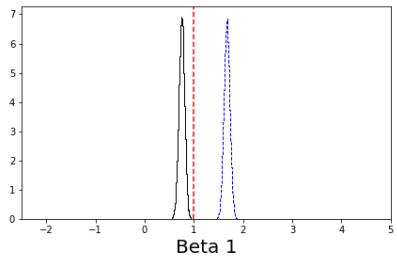
(f) $\rho(x_1, z) = 0.3$, $\rho(x_1, x_2) = 0.1$, $\rho(x_2, x_m) = 0.3$
strong instrument, less relevant control, high endogeneity



(g) $\rho(x_1, z) = 0.3$, $\rho(x_1, x_2) = 0.3$, $\rho(x_2, x_m) = 0.1$
strong instrument, more relevant control, low endogeneity



(h) $\rho(x_1, z) = 0.3$, $\rho(x_1, x_2) = 0.3$, $\rho(x_2, x_m) = 0.3$
strong instrument, more relevant control, high endogeneity



The figures plot the distribution of the 2SLS estimate of β_1 when the instrument is correlated with the control variable, with (solid black) and without (dashed blue) the inclusion of the control variable in the regression. The vertical dashed red line is the true value of β_1.

indicate rejection rates close to significance levels. When the instrument is weak, the rejection rates are modestly lower than significance levels, but always within half a percentage point. The bias is small and negative when the instrument is weak, though very close to zero. When the instrument is strong, rejection rates are equal to their significance levels and the bias is zero regardless of whether the control variable is included. There is no variation in results when the correlation between the control variable and the key variable of interest changes, or when the correlation between the control and the omitted variables changes. Figure 1 plots the distribution of the test statistic. The distribution of the test statistic appears close to a chi-square distribution with one degree of freedom for all specifications, in line with the theory. Figure 3 plots the distribution of β_1 for all specifications, with and without the inclusion of the control variable. In all cases, the distribution of β_1 is narrower when the control variable is included, indicating a more efficient estimate.

In the models reported in the second half of Table 2, there is a nonzero correlation between the control variable and the instrument and hence the bias is expected to be nonzero. Given the known bias in the simulated data, our test should reject the null hypothesis. Higher rejection rates indicate greater power and better performance of the test. The results indicate rejection rates equal to or very close to one. A small fraction of non rejections occur when the instrument is weak and the control variable is strongly correlated with the variable of interest. The bias varies depending on the specification, but is substantial in all cases. The bias is worse in magnitude when the control variable is not included. The bias is also worse in magnitude when the instrument is weaker, and when the correlation between the control variable and the key variable of interest is stronger.[14] In Figure 2, the test statistics have a bell-shaped distribution that varies in location and scale depending on the specification.[15] A weaker instrument (top four plots) reduces the mean of the test statistic. Increased correlation between the control variable and the key variable of interest

[14]The patterns regarding the bias are specific to this particular example, and should not be assumed to hold more generally.

[15]The theoretical distribution of the test statistic is unknown under the alternative hypothesis.

(second and fourth row) modestly reduces both the mean and the variance of the test statistic. Figure 4 plots the distribution of β_1 with and without the control variable. The distribution is generally further away from the true value when the control is not included, indicating generally better performance when the control variable is included.

As expected, the maximum possible bias is generally much larger than the actual bias, with the actual bias being no more than a quarter of the theoretical maximum in the simulated data. This result is expected because the MPB formula is based on a series of assumptions that are intended to calculate the maximum possible bias and hence provides intuition about the worst-case scenario rather than an estimate of the actual bias.

Overall, the results indicate that the test statistic has an accurate rejection rate when the null hypothesis is true, and almost always rejects when the null hypothesis is false. The results also indicate, even when using relatively small real-world-level correlations, that in the presence of contaminated controls, the bias in the variable of interest is substantial, regardless of whether the control variable is included in the regression.

5.2 Extensions

We consider a series of extensions to the baseline simulation results. To save space, we omit the figures for all extensions considered, and only report the tables.

5.2.1 Small sample

First, we consider the impact of a smaller sample size. We repeat the earlier baseline simulation but this time using a sample size of 1000. The results are reported in Table 3. The small sample impacts the performance of the test primarily when the instrument is weak. In the first four rows, the rejection rates are well below their significance levels. In rows nine through twelve the rejection rates are well below one, indicating lower power of the test. When the instrument is strong (rows five through eight and thirteen through sixteen) there is little difference in the results compared

with the baseline larger sample.

Table 3: Simulation results: one instrument one control, small sample

	Model Specification				(1)	(2)	(3)	(4)	(5)	(6)
	$\rho_{x_2,z}$	$\rho_{x_1,z}$	ρ_{x_1,x_2}	ρ_{x_2,x_m}	$R_{0.10}$	$R_{0.05}$	$R_{0.01}$	bias	bias$_{nc}$	mpb
(1)	0.0	0.1	0.1	0.1	0.048	0.014	0.001	-0.036	-0.089	0.379
(2)	0.0	0.1	0.1	0.3	0.049	0.015	0.001	-0.041	-0.067	0.443
(3)	0.0	0.1	0.3	0.1	0.052	0.016	0.001	-0.046	-0.097	0.427
(4)	0.0	0.1	0.3	0.3	0.051	0.015	0.000	0.012	-0.117	0.520
(5)	0.0	0.3	0.1	0.1	0.095	0.047	0.007	-0.003	-0.004	0.105
(6)	0.0	0.3	0.1	0.3	0.095	0.045	0.007	-0.003	-0.005	0.122
(7)	0.0	0.3	0.3	0.1	0.098	0.047	0.008	-0.003	-0.007	0.122
(8)	0.0	0.3	0.3	0.3	0.096	0.047	0.008	-0.002	-0.006	0.139
(9)	0.2	0.1	0.1	0.1	0.813	0.702	0.408	-0.843	2.271	5.801
(10)	0.2	0.1	0.1	0.3	0.813	0.702	0.411	-1.002	2.393	5.678
(11)	0.2	0.1	0.3	0.1	0.364	0.240	0.071	-1.000	1.660	35.032
(12)	0.2	0.1	0.3	0.3	0.364	0.240	0.071	-2.502	2.185	44.310
(13)	0.2	0.3	0.1	0.1	1.000	1.000	1.000	-0.075	0.670	0.872
(14)	0.2	0.3	0.1	0.3	1.000	1.000	1.000	-0.219	0.671	1.007
(15)	0.2	0.3	0.3	0.1	1.000	1.000	1.000	-0.090	0.667	1.179
(16)	0.2	0.3	0.3	0.3	1.000	1.000	1.000	-0.257	0.667	1.338

This table reports simulation results for one instrument and one control variable, with a small sample size The first four columns determine the model specification. The next three columns report the rejection rate at the 10%, 5%, and 1% levels, respectively. The next two columns report the bias in the 2SLS estimate of β_1 with and without (nc) the inclusion of the control variable. The last column reports the maximum possible bias.

5.2.2 High impact of omitted variable

In this section, we set $\beta_m = 3$ in the simulation. This change increases the impact of the omitted variable on y compared to the previous simulation, such that a greater fraction of y is now explained by the omitted variable. The results are reported in Table 4.

The test statistic performs about as well as the baseline case. As expected, the bias is larger across all specifications. In some cases, the bias is larger when the control variable is included relative to when the control variable is not included. The ratio of the actual bias to the maximum possible bias is larger, and in some cases it gets close to one-half.

Table 4: Simulation results: one instrument one control, high impact of omitted variable

	Model Specification				(1)	(2)	(3)	(4)	(5)	(6)
	$\rho_{x_2,z}$	$\rho_{x_1,z}$	ρ_{x_1,x_2}	ρ_{x_2,x_m}	$R_{0.10}$	$R_{0.05}$	$R_{0.01}$	bias	bias$_{nc}$	mpb
(1)	0.0	0.1	0.1	0.1	0.095	0.045	0.008	-0.009	-0.011	0.113
(2)	0.0	0.1	0.1	0.3	0.095	0.046	0.008	-0.007	-0.008	0.162
(3)	0.0	0.1	0.3	0.1	0.097	0.047	0.008	-0.008	-0.012	0.129
(4)	0.0	0.1	0.3	0.3	0.095	0.045	0.008	-0.007	-0.013	0.177
(5)	0.0	0.3	0.1	0.1	0.099	0.050	0.010	-0.001	-0.001	0.038
(6)	0.0	0.3	0.1	0.3	0.099	0.049	0.010	-0.001	-0.001	0.054
(7)	0.0	0.3	0.3	0.1	0.101	0.050	0.010	-0.000	-0.001	0.043
(8)	0.0	0.3	0.3	0.3	0.103	0.050	0.010	-0.000	-0.001	0.059
(9)	0.2	0.1	0.1	0.1	1.000	1.000	1.000	-0.775	2.009	3.482
(10)	0.2	0.1	0.1	0.3	1.000	1.000	1.000	-2.298	2.012	4.986
(11)	0.2	0.3	0.1	0.1	1.000	1.000	1.000	-0.215	0.667	0.994
(12)	0.2	0.3	0.1	0.3	1.000	1.000	1.000	-0.645	0.665	1.403
(13)	0.2	0.1	0.3	0.1	0.996	0.990	0.954	-1.661	2.008	8.371
(14)	0.2	0.1	0.3	0.3	0.995	0.989	0.954	-4.854	2.008	11.547
(15)	0.2	0.3	0.3	0.1	1.000	1.000	1.000	-0.252	0.666	1.317
(16)	0.2	0.3	0.3	0.3	1.000	1.000	1.000	-0.752	0.666	1.800

This table reports simulation results for one instrument and one control variable, with a high impact of the omitted variable. The first four columns determine the model specification. The next three columns report the rejection rate at the 10%, 5%, and 1% levels, respectively. The next two columns report the bias in the 2SLS estimate of β_1 with and without (nc) the inclusion of the control variable. The last column reports the maximum possible bias.

5.2.3 Additional control variables

In this section, we add an additional control variable but otherwise retain the same structure as the first simulation. For this section, let $\boldsymbol{x_2} \equiv (x_{21}, x_{22})'$ be control variables and let $\boldsymbol{\beta_2} \equiv (\beta_{21}, \beta_{22})'$. In the data generating process we set β_1, β_{21}, β_{22}, and β_m equal to 1, and x_1, x_{21}, x_{22}, x_m, z, and u are all created as standard normal distributions with the correlation structure shown in Table 8. For certain cells, we consider multiple values for the correlation. In all simulations, the sample sizes are 10,000.

For each simulation, we run three tests. The first is a joint test that the controls are contaminated. The second and third are tests of whether each of the two controls are individually contaminated. As before, we report the fraction of times the null hypothesis is rejected, for tests at the

Table 5: Correlations: one instrument two controls

	x_1	x_{21}	x_{22}	x_m	z	u
x_1	1	0.2	0.2	0.3	0.1 / 0.3	0
x_{21}	0.2	1	0.0 / 0.2	0.2	0.0 / 0.2	0
x_{22}	0.2	0.0 / 0.2	1	0.2	0.0 / 0.1	0
x_m	0.3	0.2	0.2	1	0	0
z	0.1 / 0.3	0.0 / 0.2	0.0 / 0.1	0	1	0
u	0	0	0	0	0	1

This table reports the correlation matrix for the simulation with two control variables.

10%, 5%, and 1% levels, and we report the bias in the estimate of β_1, when the control variables are included, and when they are excluded. Finally, we report the maximum possible bias coming from each control variable.

In the simulation, both controls are positively correlated with the omitted variable. When both controls are uncorrelated with the instrument, there will be no contamination and no bias. If either or both controls are correlated with the instrument, then there is contamination and nonzero bias. In the simulation, when the controls are correlated with the instrument, the first control variable is calibrated to have a high correlation with the instrument (high correlation control), and the second control is calibrated to have a low correlation with the instrument (low correlation control).

When testing each control individually, the test will show contamination if that control is correlated with the instrument, or if that control is correlated with the other control that is correlated with the instrument. Thus, for the individual control variable tests, there are two channels by which a control can be contaminated. The first is via a direct correlation with the instrument, and the second is via correlation with another control variable that is itself correlated with the instrument.

Table 6 reports the results from the joint test. Table 7 reports the results for each control variable individually. In both tables, in rows one through four, given the assumptions used to create the data there should be no contamination and the bias should be close to zero. Rejection rates should be close to or equal to their nominal levels. The results reported in the tables show the contaminated control test has higher power when the instrument is stronger.

Table 6: Simulation results: one instrument two controls, joint test

	Model Specification				(1)	(2)	(3)	(4)	(5)	(6)	(7)
	$\rho_{x_{21},z}$	$\rho_{x_{22},z}$	$\rho_{x_1,z}$	$\rho_{x_{21},x_{22}}$	$R_{0.10}$	$R_{0.05}$	$R_{0.01}$	bias	bias_{nc}	mpb_1	mpb_2
(1)	0.0	0.0	0.1	0.0	0.091	0.042	0.007	-0.002	-0.007	0.160	0.160
(2)	0.0	0.0	0.1	0.2	0.089	0.042	0.006	-0.002	-0.006	0.166	0.167
(3)	0.0	0.0	0.3	0.0	0.100	0.050	0.010	-0.000	-0.000	0.053	0.053
(4)	0.0	0.0	0.3	0.2	0.098	0.049	0.010	-0.000	-0.001	0.056	0.056
(5)	0.0	0.1	0.1	0.0	1.000	1.000	1.000	-0.257	1.003	0.200	2.501
(6)	0.0	0.1	0.1	0.2	1.000	1.000	1.000	-0.207	1.003	0.516	2.553
(7)	0.0	0.1	0.3	0.0	1.000	1.000	1.000	-0.072	0.333	0.057	0.711
(8)	0.0	0.1	0.3	0.2	1.000	1.000	1.000	-0.059	0.333	0.151	0.748
(9)	0.2	0.0	0.1	0.0	1.000	1.000	1.000	-0.690	2.013	6.636	0.265
(10)	0.2	0.0	0.1	0.2	1.000	1.000	1.000	-0.516	2.012	6.322	1.291
(11)	0.2	0.0	0.3	0.0	1.000	1.000	1.000	-0.154	0.667	1.504	0.060
(12)	0.2	0.0	0.3	0.2	1.000	1.000	1.000	-0.125	0.667	1.560	0.319
(13)	0.2	0.1	0.1	0.0	0.983	0.965	0.887	-1.621	3.025	10.378	5.271
(14)	0.2	0.1	0.1	0.2	0.999	0.998	0.987	-1.049	3.024	7.760	2.626
(15)	0.2	0.1	0.3	0.0	1.000	1.000	1.000	-0.251	1.000	1.626	0.825
(16)	0.2	0.1	0.3	0.2	1.000	1.000	1.000	-0.201	1.000	1.502	0.508

This table reports simulation results for two control variables. The first four columns determine the model specification. The next three columns report the rejection rate of the joint test at the 10%, 5%, and 1% levels, respectively. The next two columns report the bias in the 2SLS estimate of β_1 with and without (nc) the inclusion of the control variable. The last two columns report the maximum possible bias for the first and second control variable.

In rows five through eight, the low correlation (second) control is contaminated and we would expect the contaminated control test to reject the null. As reported, the rejection rates for both the joint and individual tests are equal to one for these rows. For the uncontaminated control, as expected the rejection rates are close to their nominal levels when it is not correlated with the contaminated control, and closer to one when it is correlated. In rows nine through twelve, the high-correlation (first) control is contaminated. The pattern of results in these rows is similar to the previous four rows, with improved power on individual tests of the uncontaminated control. In rows thirteen through sixteen, both controls are contaminated. Rejection rates are close to or equal to one for all tests.

Across all specifications, the bias is worse when control variables are not included, regardless

Table 7: Simulation results: one instrument two controls, individual tests

	Model Specification				(1)	(2)	(3)	(4)	(5)	(6)
	$\rho_{x_{21},z}$	$\rho_{x_{21},z}$	$\rho_{x_1,z}$	$\rho_{x_{21},x_{22}}$	$R_{1,0.10}$	$R_{1,0.05}$	$R_{1,0.01}$	$R_{2,0.10}$	$R_{2,0.05}$	$R_{2,0.01}$
(1)	0.0	0.0	0.1	0.0	0.097	0.046	0.008	0.096	0.046	0.008
(2)	0.0	0.0	0.1	0.2	0.094	0.045	0.008	0.096	0.046	0.008
(3)	0.0	0.0	0.3	0.0	0.098	0.049	0.010	0.100	0.050	0.010
(4)	0.0	0.0	0.3	0.2	0.099	0.049	0.010	0.099	0.049	0.010
(5)	0.0	0.1	0.1	0.0	0.094	0.044	0.007	1.000	1.000	1.000
(6)	0.0	0.1	0.1	0.2	0.646	0.513	0.257	1.000	1.000	1.000
(7)	0.0	0.1	0.3	0.0	0.099	0.050	0.010	1.000	1.000	1.000
(8)	0.0	0.1	0.3	0.2	0.659	0.536	0.296	1.000	1.000	1.000
(9)	0.2	0.0	0.1	0.0	1.000	1.000	1.000	0.088	0.039	0.005
(10)	0.2	0.0	0.1	0.2	1.000	1.000	1.000	0.993	0.983	0.915
(11)	0.2	0.0	0.3	0.0	1.000	1.000	1.000	0.098	0.049	0.010
(12)	0.2	0.0	0.3	0.2	1.000	1.000	1.000	0.994	0.986	0.943
(13)	0.2	0.1	0.1	0.0	0.995	0.989	0.954	0.995	0.989	0.947
(14)	0.2	0.1	0.1	0.2	1.000	1.000	0.997	1.000	0.999	0.992
(15)	0.2	0.1	0.3	0.0	1.000	1.000	1.000	1.000	1.000	1.000
(16)	0.2	0.1	0.3	0.2	1.000	1.000	1.000	1.000	1.000	1.000

This table reports simulation results for two control variables. The first four columns determine the model specification. The next three columns report the rejection rate of the individual test for the first control variable at the 10%, 5%, and 1% levels, respectively. The last three columns report the rejection rate of the individual test for the second control variable at the 10%, 5%, and 1% levels, respectively.

of whether they are contaminated. The bias is worst when both controls are contaminated, followed by when the high correlation control is contaminated, followed by when the low correlation control is contaminated. The bias is zero whether neither control is contaminated. As expected, a weaker instrument always results in greater bias. When the control variables are correlated, the bias is marginally smaller.

Overall, the simulation results indicate that the contaminated control test has the correct size when there is no contamination, and adequate power to detect contamination when contaminated control bias exists. The results hold for a wide variety of specifications.

5.2.4 Additional instrument

We now consider a simulation with an additional instrument and a single control variable, so that $\boldsymbol{z} \equiv (z_1, z_2)'$. In the data generating process we set β_1, β_2, and β_m equal to 1, and x_1, x_2, x_m, z_1, z_2 and u are all created as standard normal distributions with the correlation structure shown in Table 8. For certain cells, we consider multiple values for the correlation, which are reported below separated by a forward slash. In all simulations, the sample sizes are 10,000.

Table 8: Correlations: two instruments one controls

	x_1	x_2	x_m	z_1	z_2	u
x_1	1	0.1 / 0.3	0.3	0.1 / 0.3	0.1 / 0.3	0
x_2	0.1 / 0.3	1	0.1 / 0.3	0.2	0	0
x_m	0.3	0.1 / 0.3	1	0	0	0
z_1	0.1 / 0.3	0.2	0	1	0.2	0
z_2	0.1 / 0.3	0	0	0.2	1	0
u	0	0	0	0	0	1

This table reports the correlation matrix for the simulation with two instrumental variables.

In this setting, the first instrument z_1 is correlated with the control variable x_2, while the second instrument z_2 is uncorrelated. The first set of results are reported in Table 9. In addition to reporting the rejection rates, bias, and maximum possible bias, we also report the 5% rejection rate of the Sargan J-test, as the model is overidentified.

Since the first instrument is correlated with the control in the simulated data, we have contamination in all the model specifications. As reported in the table, the rejection rates for the contaminated control test are equal to one whenever the first instrument is strong (rows nine through sixteen). When the first instrument is weak and the second is strong (rows five through eight), the power of the test is substantially reduced. This is however not a cause for concern, as the actual bias (and the mpb) are both very close to zero in these cases. When both instruments are weak (rows one through four), the test has good power when the control is less relevant (rows one and two) and less power when the control is more relevant (rows three and four). Again, our test has low power generally when the bias itself is economically small and hence unlikely, with or with-

Table 9: Simulation results: two instruments one control, overidentified model

	Model Specification				(1)	(2)	(3)	(4)	(5)	(6)	(7)
	ρ_{x_1,z_1}	ρ_{x_1,z_2}	ρ_{x_1,x_2}	ρ_{x_2,x_m}	$R_{0.10}$	$R_{0.05}$	$R_{0.01}$	bias	bias$_{nc}$	mpb	$R_{s,0.05}$
(1)	0.1	0.1	0.1	0.1	1.000	1.000	0.999	-0.094	1.000	1.123	0.341
(2)	0.1	0.1	0.1	0.3	1.000	1.000	0.999	-0.281	1.000	1.307	0.980
(3)	0.1	0.1	0.3	0.1	0.622	0.498	0.264	-0.041	1.001	0.590	0.316
(4)	0.1	0.1	0.3	0.3	0.618	0.494	0.260	-0.125	1.000	0.672	0.993
(5)	0.1	0.3	0.1	0.1	0.335	0.228	0.086	-0.005	0.091	0.065	0.316
(6)	0.1	0.3	0.1	0.3	0.333	0.227	0.085	-0.015	0.091	0.075	0.995
(7)	0.1	0.3	0.3	0.1	0.338	0.231	0.087	0.005	0.091	0.075	0.316
(8)	0.1	0.3	0.3	0.3	0.338	0.230	0.087	0.014	0.091	0.085	0.995
(9)	0.3	0.1	0.1	0.1	1.000	1.000	1.000	-0.068	0.636	0.818	0.554
(10)	0.3	0.1	0.1	0.3	1.000	1.000	1.000	-0.203	0.636	0.943	0.821
(11)	0.3	0.1	0.3	0.1	1.000	1.000	1.000	-0.076	0.637	1.065	0.513
(12)	0.3	0.1	0.3	0.3	1.000	1.000	1.000	-0.229	0.636	1.209	0.844
(13)	0.3	0.3	0.1	0.1	1.000	1.000	1.000	-0.034	0.333	0.417	0.354
(14)	0.3	0.3	0.1	0.3	1.000	1.000	1.000	-0.101	0.333	0.479	0.978
(15)	0.3	0.3	0.3	0.1	1.000	1.000	1.000	-0.031	0.333	0.448	0.343
(16)	0.3	0.3	0.3	0.3	1.000	1.000	1.000	-0.094	0.333	0.507	0.984

This table reports simulation results for two instruments and one control variable. In all models $\rho_{x_2,z_1} = 0.2$ and $\rho_{x_2,z_2} = 0$. The first four columns determine the model specification. The next three columns report the rejection rate at the 10%, 5%, and 1% levels, respectively. The next two columns report the bias in the 2SLS estimate of β_1 with and without (nc) the inclusion of the control variable. The next column reports the maximum possible bias. The last column reports the rejection rate of the Sargan test at the 5% level.

out being able to detect the bias, to change the inference on the key coefficient estimate. In most specifications, the Sargan test tends to have lower power when the correlation between the control variable and the omitted variable is low, and vice versa. The contaminated control test and the Sargan test capture different features of the data and are complementary in their usefulness. The Sargan test is only possible in overidentified models.

The second set of results are reported in Table 10. Here, we estimate two just identified models using each instrument one at a time. Rejection rates at the 5% levels, bias, and maximum possible bias are reported for each model.

Our test performs well across all specifications with the just identified models. Rejection rates are close to the significance level when we use the second instrument that is not correlated with the

Table 10: Simulation results: two instruments one control, just identified models

	Model Specification				(1)	(2)	(3)	(4)	(5)	(6)
	ρ_{x_1,z_1}	ρ_{x_1,z_2}	ρ_{x_1,x_2}	ρ_{x_2,x_m}	$R_{1,0.05}$	$R_{2,0.05}$	bias_1	bias_2	mpb_1	mpb_2
(1)	0.1	0.1	0.1	0.1	1.000	0.046	-0.258	-0.003	2.990	0.097
(2)	0.1	0.1	0.1	0.3	1.000	0.046	-0.766	-0.003	3.485	0.113
(3)	0.1	0.1	0.3	0.1	0.990	0.046	-0.552	-0.003	7.349	0.113
(4)	0.1	0.1	0.3	0.3	0.989	0.046	-1.612	-0.002	8.403	0.129
(5)	0.1	0.3	0.1	0.1	1.000	0.049	-0.258	-0.000	2.987	0.033
(6)	0.1	0.3	0.1	0.3	1.000	0.049	-0.766	-0.001	3.483	0.038
(7)	0.1	0.3	0.3	0.1	0.990	0.049	-0.554	-0.000	7.323	0.038
(8)	0.1	0.3	0.3	0.3	0.989	0.050	-1.619	-0.000	8.384	0.043
(9)	0.3	0.1	0.1	0.1	1.000	0.046	-0.072	-0.003	0.861	0.097
(10)	0.3	0.1	0.1	0.3	1.000	0.046	-0.215	-0.003	0.994	0.113
(11)	0.3	0.1	0.3	0.1	1.000	0.047	-0.084	-0.002	1.160	0.114
(12)	0.3	0.1	0.3	0.3	1.000	0.046	-0.250	-0.003	1.317	0.129
(13)	0.3	0.3	0.1	0.1	1.000	0.051	-0.072	-0.000	0.861	0.033
(14)	0.3	0.3	0.1	0.3	1.000	0.050	-0.215	-0.000	0.994	0.038
(15)	0.3	0.3	0.3	0.1	1.000	0.050	-0.084	-0.000	1.160	0.038
(16)	0.3	0.3	0.3	0.3	1.000	0.050	-0.251	-0.000	1.317	0.043

This table reports simulation results for two instruments and one control variable. In all models $\rho_{x_2,z_1} = 0.2$ and $\rho_{x_2,z_2} = 0$. The first four columns determine the model specification. The next two columns report the rejection rate at the 5% levels for the two just identified models. The next two columns report the bias in the 2SLS estimate of β_1 for the two just identified models. The last two columns report the bias in the 2SLS estimate of β_1 for the two just identified models.

control. Rejection rates are close to one when we use the first instrument that is correlated with the control.

6 Empirical Example

The results in Section 5 were based on simulated data. That discussion was important to show how the contaminated control test and MPB calculations performed as expected when the differences between the observed model and true data generating process were known. That exercise was also important to show empirically how the power of the test and the usefulness of the MPB calculation can be affected by weak instruments. In this section, we leave the simulated data aside and illustrate

the use of the new contaminated control test together with the MPB formula with a well-known empirical example based on a paper by Campa and Kedia published in the Journal of Finance in 2002.

6.1 Illustration of how to use the new contaminated control test and MPB calculations with an empirical example

The Campa and Kedia (2002) paper is part of a large literature that explores the diversification discount of multi-division firms and includes publications in both economics and finance journals across several decades. Across these years different studies have utilized different samples and econometric approaches to explore the diversification discount, and depending on the specific sample and approach used, have reported varying levels of a discount with many papers in this literature finding at least some evidence of a discount consistent with a multi-divisional firm's market value being less than the sum of the imputed values of its individual segments if they had each existed outside the conglomerate. Various explanations for the discount have been suggested including the idea that corporate diversification could be associated with inefficient investment and/or internal capital market policies (e.g., Shin and Stulz (1998); Rajan et al. (2000), Ozbas and Scharfstein (2009)), lower acquisition market reactions and/or lower valued target firms (e.g.,Morck et al. (1990); Graham and Wolf (2002)), and agency and governance issues (e.g., Denis et al. (1997); Hoechle et al. (2012); Ellis et al. (2018); Andreou et al. (2019)).

For our purposes we are interested in a result reported by Campa and Kedia (2002) suggesting a diversification premium rather than a discount. This paper called attention to the fact that the decision to diversify is endogenous and suggested several instrumental variables to account for the endogeneity. In their empirical approach they include the various instruments in a pre-first-stage probit model and then use the predicted probability from this model as a single generated instrument in the first stage equation of a 2SLS system. The dependent variable in the second

stage is a measure of the excess value at the firm compared to the sum of the imputed values of the firm's segments. The dependent variable in the first stage in the 2SLS system is an indicator variable for whether the firm is diversified in that year (D=1). We use this setting to illustrate how the diagnostic test and MPB formula suggested in this paper can help researchers explore 2SLS results in important ways.

To facilitate the discussion and to streamline the example, we make some simplifying assumptions for the empirical approach. The first change we make is to drop the probit model that was used in advance of the first stage equation in Campa and Kedia (2002). Rather than using the instruments in a pre-first-stage model to generate a single instrument, we include the instruments directly in the first stage to instrument D and then estimate a traditional 2SLS system of equations. This change allows us to model the effect of different instruments individually rather than altogether as part of a single generated instrument and recasts the three-equation approach using a simpler two-equation approach.[16] The second change we make is to include each control variable once in the model rather than including the controls along with their respective lagged values. This change makes the example more parsimonious and eliminates control variables that are highly correlated.

The simplified 2SLS system of equations we use is shown below. Following Campa and Kedia (2002) and Berger and Ofek (1995) we estimate a firm's excess value each year as the log of the ratio of the firm's total capital to the firm's imputed value.[17] D is the key variable of interest and

[16]Campa and Kedia (2002) suggest multiple firm-level, year-level, and industry-year level instruments that are all included in their pre-first-stage probit model. In our approach, we include year and industry controls in the first stage equation and hence do not include the various year and industry-level instruments used in their paper in our first stage model. Of the remaining instruments suggested by Campa and Kedia we include (exclude) only the subset of strong instruments that have F-statistics from just-identified first stage models above (below) 10. This approach leads to 3 strong instruments (PNDIV, PSDIV, and MAJOREX). In our model we include indicator variables for which country the firm is incorporated in rather rather than using whether the firm is incorporated outside the US as an instrument.

[17]The firm's excess value each year is measured as the log of the ratio of the firm's total capital to the firm's imputed value. A positive (negative) ratio suggests the firm is trading at a premium (discount) compared to what it would trade if its various segments existed as separate entities. To calculate the imputed value, a sales multiplier is calculated for each industry each year as the median total capital-to-sales ratio based only on US single segment firms in that industry. The sales multiplier is then used to find the imputed value of segments that are in the same industry by multiplying the segment sales by the sales multiplier. The firm's overall imputed value in a given year is the sum of the

is an indicator for the firm having more than one business division in a given year. The control variables $\boldsymbol{x_2}$ are similar to the control variables used by Campa and Kedia (2002) and include the capital expenditure-to-sales ratio, the EBIT-to-sales ratio, book leverage, the log of assets, and an indicator for whether the firm is one of the largest 500 US firms by market value each year. Fixed effects are included to control for year, industry, and country of incorporation. The first and second stage equations are shown below.

$$ExcessValue = \beta_0 + \beta_1 D + \boldsymbol{\beta_2'} \boldsymbol{x_2} + w$$

$$D = \gamma_0 + \boldsymbol{\gamma_1'} \boldsymbol{z} + \boldsymbol{\gamma_2'} \boldsymbol{x_2} + e \tag{44}$$

To create our sample we follow the approach described in Campa and Kedia (2002) but using more recent data from 1986 through 2022[18]. Using this sample we obtain 2 empirical results that are similar to the 2002 paper. First, in untabulated results using the second stage equation above as a simple OLS model rather than as a two-stage model, we obtain a $\beta_{1,OLS}$ estimate of -0.133. This result is similar to the OLS result reported in Campa and Kedia (2002) and is also broadly consistent with earlier papers in this literature that used similar methods and report evidence of a diversification discount. Second, and as reported in Table 11, we find a positive and significant coefficient on D using the 2SLS approach described above with several of the instruments from the Campa and Kedia (2002) paper. This second result shows that we obtain a similar outcome

imputed segment values. Campa and Kedia (2002) use both a sales multiplier and an asset multiplier in their analysis. For the purpose of demonstrating the effect that contaminated controls may have on the 2SLS results we focus only on the sales-based calculation.

[18]To create our sample we start with the full Compustat segment database and then generally follow the sample creation criteria described in Campa and Kedia (2002) and Berger and Ofek (1995) using data starting in 1986 and extending forward through 2022. This means we eliminate segments that do not report sales information, are missing a SIC code, or that are not identified as business segments. Following these earlier papers we also eliminate any firm-year if the overall sales are less than $20 million, if the firm reports segments in the financial sector (SIC 6000-6999), if the sum of the segment sales is more than 1% different than the overall sales reported for the firm, or if the inputs to calculate total capital are missing. Total capital is calculated as the sum of Compustat's market value of equity, long-term debt, current portion of long-term debt, and preferred stock. The preferred stock is assumed to be 0 if missing. Following these earlier papers we eliminate any firm-year where the estimated excess firm value is above 1.386 or below -1.386. Control variables are winsorized at the 1% level. If a firm appears in the main Compustat file but not in the segments database, we assume it is a single segment firm.

as reported in Campa and Kedia suggesting a diversification premium using our simplified 2SLS model and updated sample.

Given the various other papers in this literature that document a discount rather than a premium, it is worth exploring whether contaminated control bias could explain this unexpected result. It is important to note that Campa and Kedia's identification strategy, and indeed the empirical approach embraced by most finance papers that utilize 2SLS models, critically requires that the other control variables included in the 2SLS system either be exogenous variables or that they at least not be correlated with the specific instrument(s) being used in conjunction with the key endogenous variable of interest. We use the contaminated control test and MPB formula proposed in this paper (1) to show that the instruments used in the Campa and Kedia (2002) model are correlated with the control variables and hence the potential for contaminated control bias in the key coefficient of interest exists in this model, (2) to estimate the potential size of the bias coming from the contaminated controls using the MPB formula described above, and (3) to illustrate how the above tools can help a researcher explore the robustness of their 2SLS results.

In column 1 of Table 11 we report the second stage results from an overidentified 2SLS model where D (the dependent variable of the fist stage) has been instrumented with the PNDIV, PSDIV, and MAJOREX instruments described in Campa and Kedia (2002)[19] Similar to Campa and Kedia, we find that the coefficient on D is positive after using a 2SLS approach. In untabulated tests, the p-values associated with the contaminated control test statistics for each of the control variables for the model reported in column 1 of Table 11 were each less than 1% indicating that in each case we reject the null hypothesis of no contaminated control bias for each of these variables. Given the strength of the instruments in our example, and the simulation results in this paper, the

[19]Following Campa and Kedia (2002), PNDIV is intended to capture the attractiveness of a firm being diversified and is defined as the "fraction of all firms in the industry which are conglomerates" that year. PSDIV provides similar information but is defined as the "fraction of sales by other firms in the industry accounted for by diversified firms" that year. Campa and Kedia (2002) argue that firms are more likely to diversify if they are more visible to investors due to a reduction in information asymmetries and that being listed on the NYSE, Nasdaq, or AMEX exchanges would lead to this visibility. MAJOREX is a indicator variable for whether a firm is on one of these exchanges in a given year.

Table 11: Diversification Discount Empirical Example

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	3 IVs	MPB in	Just Identified Models			3 IVs	2 IVs
	1 End	β_D	PNDIV	PSDIV	MAJOREX	2 End	1 End
D	0.062**		-0.099***	-0.003	2.062***	-0.177***	-0.101***
	(0.029)		(0.001)	(0.963)	(0.000)	(0.000)	(0.000)
Log(Assets)	0.043***	0.229	0.050***	0.046***	-0.047***	0.198***	0.050***
	(0.000)		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
CAPX/Sales	0.770***	0.098	0.741***	0.758***	1.121***	0.621***	0.741***
	(0.000)		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
EBIT/Sales	0.073***	0.053	0.080***	0.076***	-0.008	-0.062***	0.080***
	(0.000)		(0.000)	(0.000)	(0.628)	(0.000)	(0.000)
Leverage	0.029***	0.031	0.034***	0.031***	-0.029	-0.211***	0.034***
	(0.002)		(0.000)	(0.001)	(0.130)	(0.000)	(0.000)
SP500	0.198***	0.099	0.196***	0.197***	0.232***	-0.247***	0.195***
	(0.000)		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	-0.780**		-0.775***	-0.778**	-0.847	-1.458***	-0.775***
	(0.013)		(0.008)	(0.011)	(0.190)	(0.000)	(0.008)
Year FE	Yes		Yes	Yes	Yes	Yes	Yes
Industry FE	Yes		Yes	Yes	Yes	Yes	Yes
Country FE	Yes		Yes	Yes	Yes	Yes	Yes
Observations	108,782		108,782	108,782	108,782	108,782	108,782
1st Stage F	631.982		1,715.891	300.179	149.281	740.933	859.826
Sargan χ^2	341.053					20.341	2.641
Sargan p-value	$<.001$					$<.001$	0.104

The second stage dependent variable is a measure of excess firm value calculated as the log of the ratio of the firm's total capital to the firm's imputed value. Column 1 reports the second stage results from an overidentified 2SLS model that instruments the endogenous variable D using all 3 instruments (PNDIV, PSDIV, and MAJOREX). Column 2 reports the maximum possible bias (MPB) that could exist in the coefficient on D in column 1 due to the contaminated control bias coming from each control variable. Columns 3 - 5 report the 2SLS results from just identified models with the instruments listed in the column headers. Column 6 reports the 2SLS results from an overidentified model that uses the 3 instruments to instrument both D and Log(Assets). Column 7 reports the 2SLS results from an overidentified model that uses 2 instruments (PNDIV, PSDIV) to instrument D. P-values are shown below the coefficients in parenthesis. Significance is shown at the 1%, 5% and 10% levels using ***, **, and * superscripts, respectively. Industry controls are defined using 2-digit SICs. Sargan (1958) and Basmann (1960) chi-squared overidentification test results are reported for the overidentified models.

contaminated control tests should have sufficient power in this setting. Column 2 of Table 11 reports the maximum possible bias that could be affecting the $\beta_{1,2SLS}$ estimate in column 1 from

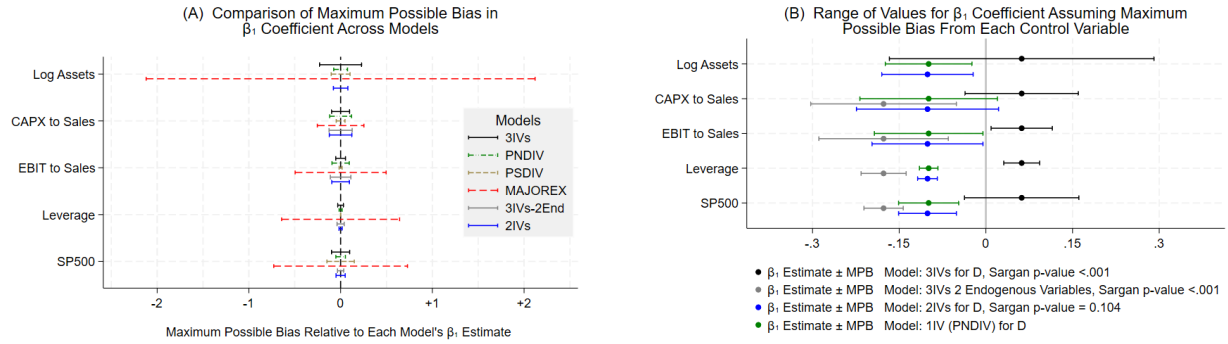
each control variable. Comparing the size of the MPB for each control variable with the size of the $\beta_{1,2SLS}$ estimate helps clarify whether the bias could be large enough to possibly change the sign on the key coefficient. This is shown visually in Panel B of Figure 5 where the $\beta_{1,2SLS}$ estimate $\pm$ the MPB crosses zero in the first model for 3 of the control variables.

So what should a researcher do in a situation like this when one or more of the control variables fails the contaminated control test and the MPB is large enough to change the sign on the key 2SLS estimate?

First, the contaminated control test derived in Section 3 identifies control variables that *may* be creating bias in the 2SLS coefficient of interest if the control variables are indeed also correlated with an omitted variable. In current practice, authors that use 2SLS already spend time discussing the (narrow) exclusion condition as it applies to the instrument(s) on the key variable of interest. If one or more of the control variables fails the contaminated control test then the researcher may still be able to use the 2SLS result if they can argue that the flagged control variable is likely to have little or no correlation with the error. In this sense, the diagnostic test developed in this paper helps researchers know which control variables need to be discussed in terms of their possible endogeneity. In our example, it would not be plausible to argue that the affected controls are exogenous given that assets, CAPX, earnings, leverage, and firm size are all plausibly related to firm valuation and are also likely co-determined with other factors not included in the model that are in the error term.

In exploring the 2SLS system, if the control variables are not plausibly exogenous then it is important to examine whether the MPB is large enough to change the sign on the main 2SLS coefficient. As noted above, the MPB is the maximum possible bias rather than actual bias and hence represents the worst possible scenario. Based on the simulation results, it may be reasonable to assume in practice that the actual bias is likely less than half the MPB. This would suggest that a researcher may still be able to use the inference from the 2SLS result if the sign on the key variable of interest remains unchanged after adjusting the coefficient $\pm$ 0.5*MPB. On the other hand, there

Figure 5: Maximum Possible Bias



Panel A plots the maximum possible bias coming from each of the control variables in the different models listed in the legend. In this figure if the horizontal bands are close to 0 (e.g., the small $\pm$ MPB bands shown for Leverage in the 2IVs model) this means the maximum possible bias in the $\beta_{1,2SLS}$ estimate for that model coming from that specific control variable is close to 0. In contrast, if the horizontal bands are large, for example the $\pm$ 2.12 MPB bands shown in the just identified MAJOREX model, then the possible bias in the $\beta_{1,2SLS}$ estimate for that model could be on the order of $\pm$ 2.12 which is much larger than the size of the $\beta_{1,2SLS}$ estimate for that model. The 3IVs, PNDIV, PSDIV, MAJOREX, 3IVs-2End, and 2IVs models in the figure refer to the same models also described in columns 1, 3, 4, 5, 6, and 7 in Table 11, respectively. The circles plotted in the center of the bands in Panel B represent the $\beta_{1,2SLS}$ estimates for the models listed at the bottom of the figure. The bands show how different the true $\beta_{1,2SLS}$ estimate could be based on the MPB calculation for each control variable. The 3IVs, 3IVs 2 Endogenous, 2IVs, and 1IV models refer to the same models reported in columns 1, 6, 7, and 3 in Table 11, respectively.

may be contaminated control bias coming from more than one control variable and together the bias may still be large enough to potentially create problems when drawing strong inference around the key variable of interest if the size of the MPB is relatively large for multiple control variables.[20]

Another possible way forward would be to find a different instrument that is less correlated with the control variables. Or, if the 2SLS system is already overidentified, then researchers can either explore which of their instruments creates the least bias and use that instrument in a just identified model, or choose to use the various instruments to instrument not only the key endogenous variable of interest but also the control variable with the largest possible bias using two first stages in a 2SLS system. In the empirical example reported in column 1 of Table 11, the 2SLS system was overidentified with 3 strong instruments so we use both of these approaches and re-estimate the

[20]The MPB calculation is done variable-by-variable assuming in each case that the other variables are not contaminated. This means that the MPB is not additive across variables.

2SLS results using 3 separate just identified models (columns 3 - 5 of Table 11) and estimate an overidentified model with 2 endogenous variables (column 6 of Table 11). Given that the MPB was largest for the Log(Assets) control variable in the model reported in column 1, we use the same 3 instruments in column 6 to instrument both Log(Assets) and D in a new 2SLS model. As reported in column 6 of Table 11 after addressing the variable with the largest possible bias, the $\beta_{1,2SLS}$ estimate turns negative and, as shown in Figure 5, the MPB values for this model are considerably smaller than the possible bias values calculated for column 1 suggesting that the reason for the original positive coefficient on D was in fact due to bias.

To better understand our system we also estimate just identified models. As reported in columns 3 - 5 each of the 3 instruments is strong based on the F-statistic from the first stage but the estimated marginal effect of D on a firm's excess value ranges from a -0.099 in column 3 to a positive 2.06 in column 5 suggesting that at least one of these instruments is invalid. The conclusion that at least one of the instruments is invalid is also supported by the Sargan overidentification test in column 1. To find out which of the 3 instruments is likely invalid we compute the MPB for each of the control variables in each of the just identified models reported in columns 3 - 5 of Table 11. The MPB values for these models are shown side-by-side in Panel A of Figure 5. As shown in the figure the MPB values associated with the control variables in the MAJOREX just identified model are much larger than the possible bias associated with the other instruments suggesting that MAJOREX is the problematic instrument. This conclusion is also supported by 3 other observations: (1) the marginal effect of D reported in column 5 of Table 11 is too large given the range of values in the dependent variable, (2) MAJOREX is the only instrument that suggests a diversification premium instead of a discount and this result contradicts many other studies, and (3) in column 7 of Table 11 we report the 2SLS results from a model that uses only PNDIV and PSDIV to instrument D and find that the Sargan overidentification test does not fail suggesting again that MAJOREX was the instrument creating validity issues in column 1. The logic and discussion above suggests the 2SLS models in columns 3, 6, and 7 of Table 11 would be better for finding the marginal effect

of diversification on excess firm value compared to the other 2SLS models in that these models exhibit less contaminated control bias. Panel B of Figure 5 presents this same information visually showing that (1) the $\beta_{1,2SLS}$ estimate in each of these models is negative, and (2) the MPB values in these models are too small to flip the sign positive. From this analysis we conclude that the $\beta_{1,2SLS}$ estimate is likely negative, or possibly zero, but not positive suggesting a diversification discount and not a premium.

7 Conclusion

Identifying a causal relationship between variables is often difficult given the many unobservable factors that relate to most financial topics. In recent years, many researchers have used instrumental variables in a 2SLS setting to deal with the endogeneity. A survey of the use of 2SLS in papers at the Journal of Finance, Journal of Financial Economics, and the Review of Financial Studies over recent years indicates that hundreds of papers have used 2SLS as part of their analysis and that almost all of them provide minimal or no discussion of the potential endogeneity of the control variables included in the model. Indeed, standard practice appears to be to discuss the relevancy and exclusion conditions for a given instrument as far as these conditions relate specifically to the key variable of interest and then to include, as though exogenous, an assortment of other control variables that may themselves also be endogenous. Many of these papers simply assert or assume that the control variables are exogenous.

Yet, despite these assertions and the general lack of discussion around the potential endogeneity of the control variables, it is likely that most of these empirical settings have at least weakly endogenous control variables. Our paper shows analytically, and via simulation, that ignoring the low-level correlations that can exist between the control variables and the error term can have a direct and strong effect on the researcher's ability to draw inference from the 2SLS results if the control variable(s) are also correlated with the instrument for the key variable of interest.

Along these lines, our paper provides guidance related to the following 4 questions: First, what effect does the inclusion of potentially endogenous control variables have on the 2SLS estimate for the key variable of interest given a strong instrument for the key variable that itself is not correlated with the error term? Answer: Including endogenous control variables can generate large bias in the 2SLS estimate of interest even if the instrument for that key variable is strong and is itself not correlated with the error term. The bias that comes from the inclusion of endogenous control variables only affects the 2SLS estimate on the key variable interest if the control variables are both endogenous and correlated with the instrument for the key variable of interest. Contaminated control variable bias is exacerbated by weak instrument(s) for the key variable of interest. The contaminated control test and the MPB formula introduced in this paper can help researchers assess whether the size of the bias is likely large enough to affect the inference on the main variable of interest.

Second, is the 2SLS bias for the key variable of interest made larger or smaller with or without including the other potentially endogenous control variables in the system of equations? Answer: It is not possible to say whether the 2SLS bias will increase or decrease when dropping the endogenous control variables from the system. In some settings the bias increases whereas in others it decreases. However, based on the analytical form of the bias discussed in Section 2, if both (1) the narrow exclusion condition holds (or is almost satisfied) for the instrument on the key variable of interest and (2) the difference described in this paper, ($\hat{\gamma}_2$ - $\hat{\lambda}_2$), is close to zero then the overall bias is likely smaller in the 2SLS estimate for the key variable of interest with the controls included than in the estimate without controls.

Third, what information can be inferred from estimating the 2SLS estimate both with and without the control variables and then comparing the estimates? Answer: Dropping an important variable (endogenous or not) from the system creates the potential for omitted variable bias in the key estimate if the dropped variable is correlated with the other control variables and the instrument(s). Thus estimating the 2SLS estimate both with and without control variables is trading

off potential bias from the inclusion of endogenous control variables that are correlated with the instrument against overall omitted variable bias. Hence, observing a large change in the 2SLS estimate when comparing the key result with and without the control variables need not indicate that the 2SLS estimate with controls is biased given that the change could be attributable to omitted variable bias created when dropping the control variables. However, observing little or no change in the 2SLS estimate both with and without the control variables provides corroborating evidence that the 2SLS estimate is not largely affected by bias from the control variables.

And, fourth, is there a test that would reveal whether the 2SLS estimate for the key variable of interest could be affected by contaminated controls? Answer: Yes. In order for the inclusion of other control variables to affect the key 2SLS estimate, two conditions have to occur: (1) the control variable(s) must be endogenous, and (2) the control variable(s) must be correlated with the instrument for the key variable of interest. It is not possible to ascertain the first condition but the second is testable. We propose testing whether $(\hat{\boldsymbol{\gamma}}_{\mathbf{2}} - \hat{\boldsymbol{\lambda}}_{\mathbf{2}})$ is statistically different from zero when investigating this bias. Failing to reject the null hypothesis of this test statistic being equal to 0 supports the conclusion that the control variables are not creating material bias in the key variable of interest. This test has the advantage of being able to rule out the presence of contaminated control bias but is limited in that it rules out a necessary but not a sufficient condition for this type of bias.

As highlighted in Section 6 with the diversification discount example, this paper suggests an approach for researchers using 2SLS to first test for the possibility of contaminated control bias in specific control variables and then to calculate the maximum possible bias in the key coefficient coming from the flagged variables. Using the new contaminated control test together with the proposed MPB calculations will allow researchers in the future to examine whether their 2SLS results are robust or whether contaminated control bias may be affecting the results in a material way. As noted above, hundreds of recent papers in top finance journals have previously simply assumed that the control variables in their 2SLS systems are not biasing their key 2SLS result, but

this assumption is unlikely true in many if not most cases. The tools and approaches proposed in this paper will allow researchers to examine this assumption in detail.

The discussion in Section 6 also provides practical advice for how to explore 2SLS results in the event that one or more control variables are flagged by the contaminated control test. If specific control variables fail the proposed test, and the MPB on those variables is relatively large compared to the key coefficient of interest, the researcher can either instrument those variables or explain why they are unlikely to be endogenous using arguments similar to the literature's current approach when motivating the narrow exclusion condition for the instrument on the key variable of interest. The proposed test and MPB calculations will help researchers understand which control variables may need additional discussion. One of the advantages of using this approach in an overidentified model is that the researcher can check whether specific instruments worsen the potential bias. Researchers often use the Sargan overidentification test to identify validity issues in 2SLS models, but this test is only possible in overidentified models and does not identify which of the instruments is invalid. The MPB calculations proposed in this paper provide an alternative way to identify potential problems with validity and have the added benefit of allowing researchers to assess which of several instruments is likely causing the problem and being available in both just identified and overidentified models.

In summary, the diagnostic test derived in this paper and the related MPB calculations will help researchers know when they need to explore their 2SLS system of equations in more detail, know which control variables need specific consideration, be able to choose between specific instruments to minimize bias if the different instruments are suggesting different inferences, and be able to assess the robustness of their main 2SLS results and whether the inference is likely affected by contaminated control bias.

References

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *Journal of Human resources*, 40(4):791–821.

Andreou, P. C., Doukas, J. A., Koursaros, D., and Louca, C. (2019). Valuation effects of overconfident ceos on corporate diversification and refocusing decisions. *Journal of Banking amp; Finance*, 100:182–204.

Angrist, J. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, pages 69–85.

Angrist, J. and Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press.

Basmann, R. L. (1960). On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association*, 55(292):650–659.

Berger, P. G. and Ofek, E. (1995). Diversification's effect on firm value. *Journal of Financial Economics*, 37:39–65.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443.

Campa, J. M. and Kedia, S. (2002). Explaining the diversification discount. *The Journal of Finance*, 57:1731–1762.

Davidson, R. and MacKinnon, J. G. (2004). *Econometric theory and methods*. Oxford University Press.

Denis, D. J., Denis, D. K., and Sarin, A. (1997). Agency problems, equity ownership, and corporate diversification. *The Journal of Finance*, 52:135.

Ellis, J. A., Fee, C. E., and Thomas, S. (2018). Playing favorites? industry expert directors in diversified firms. *Journal of Financial and Quantitative Analysis*, 53:1679–1714.

Goldberger, A. (1991). *A Course in Econometrics*. Harvard University Press.

Graham, M. L. and Wolf, J. (2002). Does corporate diversification destroy value. *The Journal of Finance*.

Greene, W. H. (2003). *Econometric analysis*. Prentice Hall.

Hoechle, D., Schmid, M., Walter, I., and Yermack, D. (2012). How much of the diversification discount can be explained by poor corporate governance? *Journal of Financial Economics*, 103:41–60.

Lovell, M. H. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010.

Morck, R., Shleifer, A., and Vishny, R. W. (1990). Do managerial objectives drive bad acquisitions? *The Journal of Finance*, 45:31.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives*, 20(4):111–132.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Ozbas, O. and Scharfstein, D. S. (2009). Evidence on the dark side of internal capital markets. *Review of Financial Studies*, 23:581–599.

Rajan, R., Servaes, H., and Zingales, L. (2000). The cost of diversity: The diversification discount and inefficient investment. *The Journal of Finance*, 55:35–80.

Roberts, M. R. and Whited, T. M. (2012). Endogeneity in empirical corporate finance. *SSRN Electronic Journal*.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the econometric society*, pages 393–415.

Shin, H.-H. and Stulz, R. M. (1998). Are internal capital markets efficient? *The Quarterly Journal of Economics*, 113:531–552.

Wooldridge, J. (2003). *Introductory Econometrics: A Modern Approach, 2Ed*. South-Western/Thomson Learning.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, pages 348–368.